First measurement of the jet mass in events with highly boosted top quarks and studies with top tagging at CMS

Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Physik
der Universität Hamburg

vorgelegt von
Torben Dreyer

Hamburg
2019

Gutachter/innen der Dissertation:	Prof. Dr. Johannes Haller
	Prof. Dr. Elisabetta Gallo
Zusammensetzung der Prüfungskommission:	Prof. Dr. Johannes Haller
	Prof. Dr. Elisabetta Gallo
	Prof. Dr. Günter Sigl
	Dr. Roman Kogler
	Dr. Christian Schwanenberger
Vorsitzende/r der Prüfungskommission:	Prof. Dr. Günter Sigl
Datum der Disputation:	22.10.2019
Vorsitzender Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Michael Potthoff
Leiterdes Fachbereichs PHYSIK:	Prof. Dr. Wolfgang Hansen
Dekander Fakultät MIN:	Prof. Dr. Heinrich Graener

List of publications

CMS Collaboration, “Measurement of the jet mass in highly boosted $\mathrm{t\bar{t}}$ events from pp collisions at $\sqrt{s} = 8\,\mathrm{TeV}$”, Eur. Phys. J. C77 no. 7, (2017) 467, doi:10.1140/epjc/s10052-017-5030-3, arXiv:1703.06330.

CMS Collaboration, "W and top tagging scale factors", CMS-DP-2017-026 (2017), https://cds.cern.ch/record/2275225.

Abstract

This thesis includes two analyses with high-momentum top quarks in data collected by the CMS detector in proton-proton collisions at the Large Hadron Collider (LHC). The high center-of-mass energies at the LHC up to 13 TeV lead to a large production of high-momentum top quarks from standard model processes and allows for a production of very heavy new particles decaying into high-momentum top quarks. The large Lorentz boost poses a challenge to the reconstruction of high-momentum top quarks. Their decay products are collimated and a reconstruction in three separate jets is not efficient. A boosted top quark is therefore reconstructed in one large jet and jet substructure information is important to identify boosted top quark jets. This leads to a large interest on boosted top quarks and jet substructure from the experimental and theory communities.

The first analysis is the first measurement of the jet-mass distribution in fully-merged top quark decays in $\mathrm{t\bar{t}}$ production. The measurement is performed with data collected at a center-of-mass energy of $\sqrt{s} = 8\,\mathrm{TeV}$ in 2012. The data is corrected for detector effects and compared to Monte Carlo (MC) simulated distributions. The sensitivity to the top quark mass is used to extract a value of the top quark mass of $m_\mathrm{t} = 170.8 \pm 9.0$ GeV. The uncertainties are dominated by statistical uncertainties. The ultimate goal is an extraction of a well-defined top quark mass from a comparison to analytic calculations once they are available for the measured phase space.

The second analysis includes studies on the performance of the CMSTopTagger v2 and the HOTVR algorithm in data and simulation. Top tagging algorithms are important to identify boosted top quarks. A new method is presented to measure the official CMS top tagging data-to-simulation scale factors. The scale factors are measured in data collected in the years 2016 and 2017 at $\sqrt{s} = 13\,\mathrm{TeV}$ for 'fully-merged', 'semi-merged', and 'not-merged' jets using a template fit method. This leads to less dependence on the measurement phase space compared to previous methods. The scale factors are used in several CMS publications to correct for differences in the top tagging efficiency between data and simulation. Mistag rates are studied in 2016 data using a cut-and-count method.

Zusammenfassung

Diese Arbeit beinhaltet zwei Analysen mit Top-Quarks mit hohem Impuls in hochenergetischen Proton-Proton-Kollisionen am Large-Hadron-Collider (LHC). Die hohen Schwerpunktsenergien bis zu 13 TeV am LHC führen zu einer hohen Produktion von Top-Quarks mit hohem Impuls durch Standard-Modell-Prozesse und ermöglichen die Produktion schwerer neuer Teilchen, welche vornehmlich in Top-Quarks mit entsprechend hohem Impuls zerfallen. Der resultierende hohe Lorentz-Faktor der Top-Quarks stellt eine Herausforderung für die Rekonstruktion dar, da die Zerfallsprodukte in Flugrichtung des Top-Quarks gebündelt sind und die Rekonstruktion in drei einzelnen Jets nicht effizient ist. Ein überlagerter Top-Quark-Zerfall kann stattdessen in einem großen Jet rekonstruiert werden und die spezifische Substruktur des Jets kann verwendet werden, um den überlagerte Top-Quark-Jets zu identifizieren. Das alles führt zu einem zunehmenden Interesse an Jet-Substruktur von experimenteller und theoretischer Seite.

Die erste Analyse ist die erste Messung der Jetmasse in vollständig überlagerten Top-Quark-Zerfällen in $\mathrm{t\bar{t}}$-Produktion. Die Daten wurden im Jahr 2012 bei einer Schwerpunktsenergie von 8 TeV aufgezeichnet. Die Daten wurden um Detektoreffekte korrigiert und mit Monte Carlo (MC) simulierten Verteilungen verglichen. Die Sensitivität der Verteilung auf die Top-Quark-Masse wird genutzt um einen Wert für die Masse von $m_\mathrm{t} = 170.8 \pm 9.0$ GeV zu extrahieren. Die Unsicherheiten auf diese Messung werden dominiert von statistischen Unsicherheiten. Das zukünftige Ziel dieser Messung ist eine Extraktion einer wohldefinierten Masse aus einem Vergleich mit analytischen Rechnungen, sobald diese für den gemessenen Phasenraum verfügbar sind.

Die zweite Analyse beinhaltet Studien zur Effizienz des CMSTopTagger v2 und des HOTVR-Algorithmus in Daten und Simulation. Top-Tagging-Algorithmen werden benötigt um, vollständig überlagerte Top-Quark-Zerfälle zu identifizieren. Eine neue Methode zur Messung der offiziellen CMS-Top-Tagging-Skalenfaktoren wird präsentiert. Sie werden für "vollständig-überlagerte", "teilweise-überlagerte", und "nicht-überlagerte" Top-Quark-Jets gemessen, um gegenüber früheren Messungen weniger abhängig vom Messungsphasenraum zu sein. Die Top-Tagging-Skalenfaktoren wurden für Daten bei einer Schwerpunktsenergie von 13 TeV aus den Jahren 2016 und 2017 gemessen und werden in mehreren CMS-Analysen verwendet um Effizienzunterschiede zwischen Daten und Simulation auszugleichen. Mitag-Raten wurden in 2016 Daten mit einer ”cut-and-count”-Methode studiert.

Contents

1 Introduction

The standard model of particle physics [1, 2] is the most successful theory to describe all known fundamental particles and their interactions. Despite the good agreement with experimental data, some shortcomings indicate that there might be new physics beyond the standard model.

The top quark plays a special role in the standard model and is of central importance for this thesis. It is the heaviest known fundamental particle and might therefore play an important role in the electroweak sector of the standard model. The mass of the top quark is one of the fundamental standard model parameters. Many measurements of the top quark mass and other properties have been preformed in proton-anti-proton collisions at the Tevatron collider and in proton-proton collisions at the Large Hadron Collider (LHC).

The top quark is also important for many new physics models predicting heavy new particles decaying into top quarks. A good reconstruction and identification of top quarks is therefore important for standard model measurements and searches for new physics with top quarks in the final state.

The reconstruction becomes especially challenging for top quarks with high momentum and resulting high Lorentz boost. The decay products of high-momentum top quarks are boosted in the direction of flight of the top quark and the jets from the decay products start to overlap. This leads to a decrease of the reconstruction efficiency of a hadronic top quark decay in three separate jets. In this case a top quark can be reconstructed in one large jet and the specific substructure of the jets can be used to identify boosted top quarks.

The high center-of-mass energy in proton-proton collisions at the LHC leads to a significant production of high-momentum top quarks and allows studies in kinematic regions with boosted top quarks in the final state. This leads to a large interest on jet substructure from both the theoretical and experimental communities. Special variables and methods have been developed to classify jets and to identify jets containing heavy-particle decays like top quark decays or decays of boosted W bosons. Experimental measurements of jet-substructure variables are important to validate calculations and simulations and to gain knowledge about the underlying physics of jet substructure.

This thesis includes two analyses using data collected by the CMS experiment in proton-proton collisions at the LHC. The first analysis is a measurement of the jet-mass distribution in highly boosted top quark decays. It uses data collected in the year 2012 in proton-proton collisions with a center-of-mass energy of 8 TeV. The measured data is corrected for detector effects and used to validate the description of the jet mass in Monte Carlo (MC) event generators. The data is provided for future studies with new MC event generators and analytic calculations. The jet mass for highly boosted top quarks is sensitive to the mass of the top quark and a mass value is extracted with the use of MC generated distributions to study the sensitivity of the measurement. This variable provides the possibility to extract a well-defined top quark mass from a comparison of the data to analytic calculations once they are available for the measured phase space. The measurement is published by CMS in reference [3]. Pioneering studies for this measurement have been performed in the master thesis in reference [4]. The final measurement was completed within the scope of this thesis including a careful treatment of systematic uncertainties and the extraction of the top quark mass.

The second analysis includes studies with top tagging algorithms in proton-proton data with a center-of-mass energy of 13 TeV. Top tagging algorithms are used to identify large jets that include a fully-merged hadronic top quark decay. Within this analysis the performance of the standard top tagger in CMS, the CMSTopTagger v2 [5], and a newer approach, the HOTVR algorithm [6], are studied in simulation. A new method is used to measure the top tagging efficiency and the official CMS top tagging data-to-simulation scale factors in data collected in the years 2016 and 2017. The validation in data is important for analyses using these algorithms and the data-to-simulation scale factors can be used to correct for differences in the efficiency between data and simulation. The scale factors that are measured within the scope of this thesis are used in several CMS publications on searches for new physics with boosted top quarks in the final state using the 2016 and 2017 data sets.

The thesis is structured in the following way. It starts with a short description of the theoretical background on the standard model and possible extensions, on jet-substructure calculations, and on MC event generators in chapter 2. Chapter 3 includes a brief description of the CMS experiment and the LHC followed by an overview on the event reconstruction in chapter 4. Chapter 5 describes two experimental methods that are used in the analysis chapters on the measurement of the jet-mass distribution in chapter 6 and on studies of top tagging algorithms in data and simulation in chapter 7. The thesis closes with a conclusion briefly summarizing the most important results of the two analysis chapters.

2 Theory

This chapter provides a short overview of the theoretical background needed for this thesis. It begins with a brief overview of the standard model of particle physics and highlights the importance of the top quark for the standard model and for searches for new physics beyond the standard model. It includes a short description of the status of theoretical calculations of jet-substructure variables and closes with a short overview of the general structure of high-energy physics simulations.

2.1 The standard model of particle physics

The standard model of particle physics is a renormalizable quantum field theory describing all known fundamental particles and their interactions except for gravity. It is invariant under the local gauge symmetry group $SU(3)_c \times SU(2)_L \times U(1)_Y$, where the $SU(3)_c$ group describes the strong interaction coupling to color charge c and the $SU(2)_L \times U(1)_Y$ groups describe the electroweak interaction coupling to the weak isospin and the hypercharge Y. The interactions and the gauge bosons that serve as mediators are a consequence of the requirement of local gauge invariance. Additional information on the different interactions can be found in the following subsections and a comprehensive overview of the standard model can be found in references [1, 2].

Figure 2.1 shows the particle content of the standard model. The fundamental particles are divided into fermions carrying a spin of $\frac{1}{2}$ and bosons with integer spin. All particles have a set of quantum numbers that allow different interactions between them. The interactions are described in the following sections. All fermions carry a weak isospin with its third component I being $\pm\frac{1}{2}$. The fermions can be further divided into quarks and leptons. The electrically neutral leptons, the neutrinos, carry only the weak isospin while the electrically charged leptons carry an electrical charge of $-\mathrm{e}$. Quarks carry electrical charge and color charge in addition to the weak isospin. Up-type quarks carry an electrical charge of $\frac{2}{3}\mathrm{e}$ and down-type quarks a charge of $-\frac{1}{3}\mathrm{e}$. The color charge exists

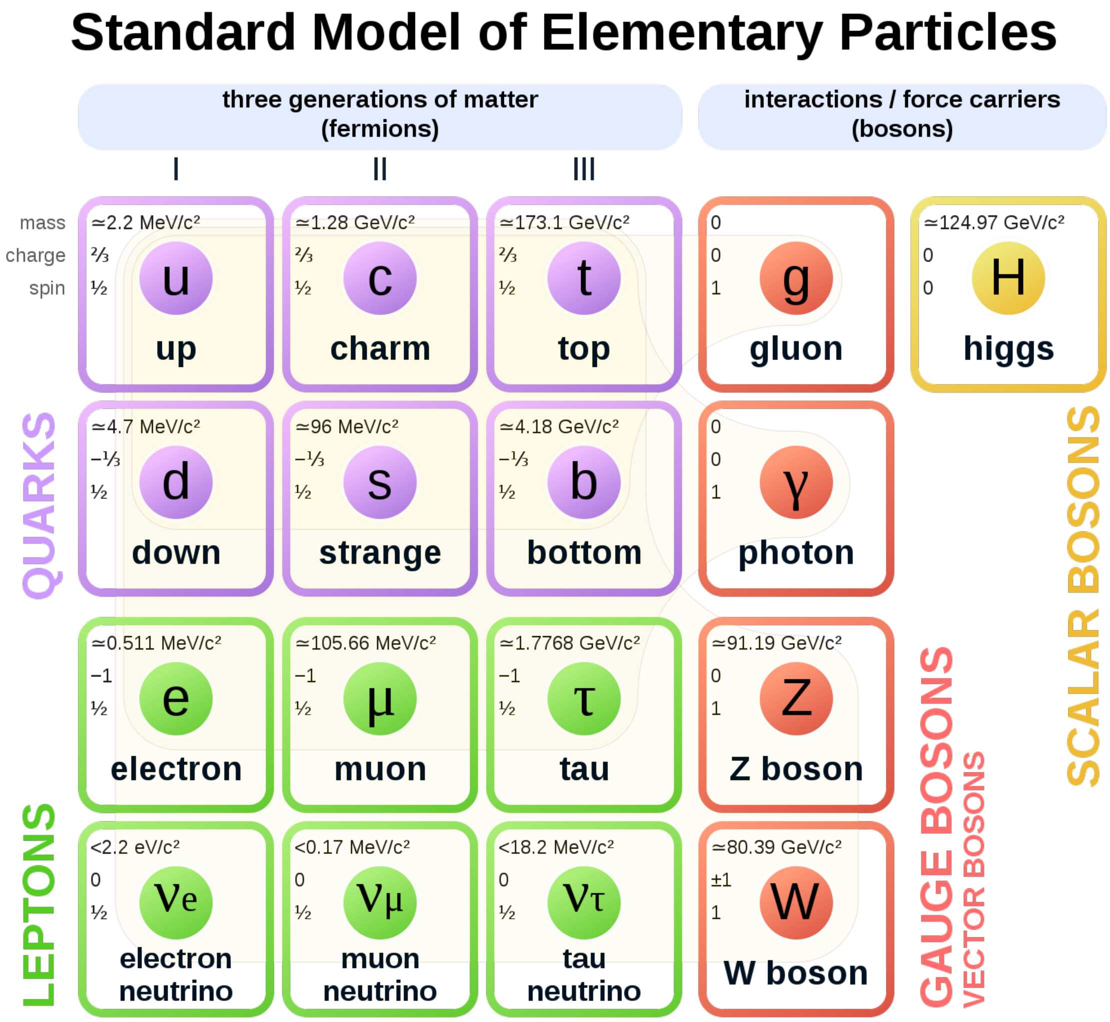

Figure 2.1: Particle content of the standard model of particle physics taken from reference [7].

in three colors "red", "blue", and "green" with respective anti-colors. Quarks and leptons each come in three generations differing in the mass of the particles. Each fermion has an antiparticle with inverted electrical charge. The gauge bosons with a spin of one serve as mediators of the fundamental forces. The photon is a massless boson with a weak isospin of zero and no electrical or color charge. The W bosons are massive bosons with an electrical charge of ±e and a weak isospin of one. They do not carry color charge. The Z boson is a massive particle with a weak isospin of zero and no electrical or color charge. Gluons are massless and electrically neural bosons that carry color and anti-color charge.

In addition to the gauge bosons with a spin of one the standard model includes a scalar boson, the Higgs boson, resulting from a spontaneous symmetry breaking in the electroweak sector needed to explain the measured masses of the W and Z bosons. The Higgs mechanism provides further a framework to include fermion masses into the standard model. More details are given in section 2.1.5 and in reference [2].

2.1.1 Quantum Electro Dynamics (QED)

The QED describes the electromagnetic interactions coupling to particles carrying electrical charge. It is invariant under the $U(1)_\gamma$ symmetry group which results from the spontaneous symmetry breaking of the $SU(2)_L \times U(1)_Y$ group described later. The requirement of local gauge invariance leads to one massless electrically neutral mediator, the photon. Due to the massless mediator the range of the electromagnetic interaction is infinite.

2.1.2 Quantum Chromo Dynamics (QCD)

The field theory of Quantum Chromo Dynamics describes strong interactions between color-charged particles. The color charge exists in three different colors, “red”, “blue”, and “green” with respective anti-colors. The strong interaction is mediated by massless gluons carrying color and anti-color charge themselves. The QCD is invariant under the local gauge group $SU(3)_c$ leading to eight gluon fields. Combinations of the three color charges allow nine gluon fields, a color octet and a color singlet. The color singlet would lead to long-range strong interactions between color-neutral particles, like hadrons, which is not observed in nature.

The strong interaction has a different behavior compared to the electromagnetic interaction resulting from the self interaction of gluons which carry color charge themselves. Similar to the electromagnetic interaction, vacuum polarization in the form of quark loops shield the color charge at high distances leading to a decreasing strength with increasing distance. The self interactions between the colored gluons, however, lead to an opposite effect increasing the strength of the interaction with increasing distance. This leads to the so-called confinement of colored objects. Colored objects like quarks cannot exits as free particles but only in color-neutral bound states of two or more quarks called hadrons. With larger distance between two color-charged particles the field energy between these particles increases until it is large enough to produce quark-anti-quark pairs from the vacuum. In this way new hadrons are produced to obtain a color-neutral state. This process is called hadronization.

Quarks produced in particle collisions can not be observed as free quarks because they hadronize before they reach the detector[1]. The result is a jet consisting of many hadrons flying in the direction of flight of the quark. The energy and momentum of the particles in a jet can be measured in a detector and jets can be reconstructed using dedicated algorithms described in section 4.4. Therefore a precise understanding of jets is crucial for the interpretation of data collected in high-energy particle collisions.

2.1.3 Weak interaction

The weak interaction is mediated by massive mediators leading to a short-range interaction. It is mediated by the charged W^+ and W^- bosons and by the neutral Z boson. The W and Z bosons have masses of $m_\mathrm{W} = 80.379 \pm 0.012\,\mathrm{GeV}$ and $m_\mathrm{Z} = 91.1876 \pm 0.0021\,\mathrm{GeV}$ [2]. The weak interaction is the only interaction that violates the parity symmetry and also the only interaction in the standard model that violates the charge-parity (CP) symmetry. It couples to the left-handed component of fermions and the right-handed component of anti-fermions. The states that take part in the interaction can be grouped in weak isospin doublets. Right-handed fermion states are isospin singlets. A coupling between different flavors is possible in the quark sector but not for leptons. These couplings can be realized by a rotation of the down-type quarks in the weak isospin space. The weak isospin states can then be written as:

$$\begin{pmatrix} u \\ d' \end{pmatrix}_L , \quad \begin{pmatrix} c \\ s' \end{pmatrix}_L , \quad \begin{pmatrix} t \\ b' \end{pmatrix}_L , \quad \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L , \quad \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}_L , \quad \begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}_L . \tag{2.1}$$

[1] One exception is the top quark that decays before hadronization and can be reconstructed from its decay products (more in section 2.2)

The down-type quark states are related to the mass eigenstates by a rotation with the Cabbibo-Kobayashi-Maskawa (CKM) matrix,

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} U_{ud} & U_{us} & U_{ub} \\ U_{cd} & U_{cs} & U_{cb} \\ U_{td} & U_{ts} & U_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}. \tag{2.2}$$

The entries of this matrix are not predicted by the standard model and need to be measured in experiments. They are, however, not independent and the CKM matrix can be described by three angles and an additional phase factor.

2.1.4 Electroweak unification

A theoretical description of the weak and the electromagnetic interaction within one theoretical framework can be realized with a $SU(2)_L \times U(1)_Y$ symmetry group. A hyper charge Y is introduced and defined as

$$Y = 2(Q - I), \tag{2.3}$$

where Q is the electrical charge and I is the third component of the weak isospin. The hyper charge is the charge introduced by the $U(1)_Y$ symmetry group. The requirement of local gauge invariance leads to three W^μ fields from the $SU(2)_L$ group, two electrically charged and a neutral one, and a neutral field B^μ from the $U(1)_Y$ group. The physical W boson fields are obtained by a linear combination of the W^1 and W^2 fields. The physical fields for the Z boson and the photon A are obtained by a mixing of the fields B^μ and W^3 of the form:

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_\mathrm{w} & \sin\theta_\mathrm{w} \\ -\sin\theta_\mathrm{w} & \cos\theta_\mathrm{w} \end{pmatrix} \begin{pmatrix} B_\mu \\ W^3_\mu \end{pmatrix}, \tag{2.4}$$

with the electroweak mixing or Weinberg-angle θ_w.

The electroweak unification gives a consistent description of electroweak interactions but it does not include mass terms for the W and Z bosons. Mass terms for the electroweak gauge bosons can be introduced by a spontaneous symmetry breaking in the electroweak sector described in the following.

2.1.5 Spontaneous symmetry breaking

The electroweak unification describes only massless gauge bosons but the W and Z bosons are observed to be massive particles. The solution to this problem is a spontaneous electroweak symmetry breaking by introducing a symmetric scalar potential

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda \left(\Phi^\dagger \Phi\right)^2 \tag{2.5}$$

with a complex self-interacting scalar field $\Phi \equiv \begin{pmatrix} \Phi^+ \\ \Phi^0 \end{pmatrix}$, the Higgs field. In the case of $\mu^2 < 0$ the Higgs field has a non-zero vacuum expectation value of

$$\langle \Phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \tag{2.6}$$

with $\frac{v}{\sqrt{2}} = \sqrt{-\frac{\mu^2}{2\lambda}}$ and $v \approx 246\,\mathrm{GeV}$. The symmetry is spontaneously broken by the choice of the non-zero ground state. The requirement of local gauge invariance under $SU(2)_L \times U(1)_Y$ leads to mass terms for the W and Z bosons of the form

$$m_\mathrm{W}^2 = \frac{g^2 v^2}{4}, \quad m_\mathrm{Z}^2 = \frac{(g'^2 + g^2)}{4}, \tag{2.7}$$

where g and and g' are the coupling constants from the $SU(2)_L$ and $U(1)_Y$ symmetry groups from the electroweak unification. The remaining degree of freedom leads to a scalar particle, the Higgs boson, with a mass of $m_\mathrm{H} = \lambda v$. The generators corresponding to the photon and gluon fields remain unbroken and the respective gauge bosons stay massless.

The fermion masses m_{f_i} can be integrated into the standard model by a Yukawa coupling h_{f_i} between the Higgs field and the fermions. The fermions acquire mass corresponding to the strength of the Yukawa coupling

$$m_{f_i} = \frac{h_{f_i} v}{\sqrt{2}}. \tag{2.8}$$

The Yukawa couplings are free parameters of the standard model and need to be constrained with experimental data. Measurements are performed by the ATLAS and CMS collaborations at the LHC sensitive to the Yukawa couplings in processes like the production of a Higgs boson in association with a $\mathrm{t\bar{t}}$ pair ($\mathrm{pp} \rightarrow \mathrm{Ht\bar{t}}$) and in decays of the Higgs boson to fermions like $\mathrm{H} \rightarrow \mathrm{bb}$ and $\mathrm{H} \rightarrow \tau\tau$ [8–13].

The search for the Higgs boson was one of the main motivations for the Large Hadron

Collider (LHC) at CERN. In the year 2012 the CMS and the ATLAS collaborations finally reported the observation of a Higgs-like boson with a mass of $\sim 125\,\mathrm{GeV}$ [14, 15] which could constitute the observation of the last missing piece of the standard model, if the measured Higgs properties stay compatible with a standard model Higgs boson.

2.1.6 Shortcomings of the standard model

The standard model shows an excellent agreement with experimental data. It has, however, still some shortcomings that implicate the existence of physics beyond the standard model. A few important shortcomings are listed in the following.

- The gravitational force is not part of the standard model. At the Planck scale at $10^{19}\,\mathrm{GeV}$ gravity should play a significant role but it is not yet clear how it can be incorporated in a quantum field theory.
- Only about 5 % of the energy in our universe consists of baryonic matter associated to the standard model. Recent cosmological observations [16, 17] suggest that about 26 % of the matter in the universe consists of electrically neutral and probably only weakly interacting Dark Matter. The standard model does not provide a candidate for cold Dark Matter, needed to explain the observations. Additionally it has no explanation for the remaining $\sim 69\,\%$ consisting of Dark Energy.
- With the current knowledge of the standard model an equal distribution of matter and anti-matter is expected in our universe. Observations, however, show that the universe mostly consists of matter and there is a large asymmetry between matter and anti-matter. The CP violation introduced by weak interaction in the CKM matrix is not sufficient to explain the observed asymmetry in nature. A review of this matter can be found in reference [18].
- Neutrinos in the standard model are massless but measurements of neutrino oscillations indicate that neutrinos have masses.
- The standard model does not provide a fundamental mechanism that explains the large hierarchy between the electroweak scale and the Planck scale at which gravity becomes significant. A more technical argument is associated to the Higgs mass. In the standard model the mass of the Higgs boson gets large corrections from fermionic and bosonic loops that exceed the value of the bare mass. This leads to the expectation that the Higgs mass should rather be close to the Planck scale than at the measured value of $\sim 125\,\mathrm{GeV}$. This difference in the scales can in the standard model only be handled by a severe fine tuning of the parameters which is seen as un-natural by many physicists.

- The electromagnetic and weak forces are unified into one underlying theory at the electroweak scale. A unification of the electroweak and the strong interaction at a higher scale does not exist yet.

2.2 The top quark

The top quark plays a key role in this thesis. It was first observed by the CDF [19] and D0 [20] collaborations in proton-anti-proton collisions with a center-of-mass energy of 1.8 TeV at the Tevatron collider. It is the heaviest known fundamental particle with a mass of 173.34 ± 0.76 GeV [21]. The most precise single measurement was performed by the CMS collaboration and resulted in a mass of 172.35 ± 0.51 GeV [22]. The high mass of the top quark makes it special in comparison to other quarks. Its lifetime is shorter than the characteristic time of the hadronization which means that the top quark is the only quark that decays before it hadronizes. Therefore it is possible to reconstruct the top quark directly from its reconstructed decay products.

The large mass of the top quark leads to a strong coupling to the Higgs field and the top quark might therefore play an important role in the electroweak sector. The mass of the top quark is an important input for self-consistency tests of the standard model [23].

2.2.1 Production

At the LHC top quarks are mainly produced in pairs of a top quark and an anti-top quark via the strong interaction. Single production via the weak interaction is also possible but sub-dominant because of the colored initial state in proton-proton collisions. The leading-order production processes of top quark pair ($\mathrm{t\bar{t}}$) production are shown in figure 2.2. The production can either be realized by a quark-anti-quark annihilation or by gluon-gluon fusion.

Calculations of inclusive cross sections at the LHC require a folding of the partonic cross section with the Parton Density Functions (PDFs) of the proton using the factorization

$$\sigma_{\mathrm{t\bar{t}}} = \sum_{i,j\in[\mathrm{q},\bar{\mathrm{q}},\mathrm{g}]} \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; f_i(x_1,\mu_f) \; f_j(x_2,\mu_f) \; \hat{\sigma}_{ij\to\mathrm{t\bar{t}}}(x_1,x_2,\mu_r,\mu_f), \tag{2.9}$$

where x_1 and x_2 are the momentum fractions of the interacting partons with respect to

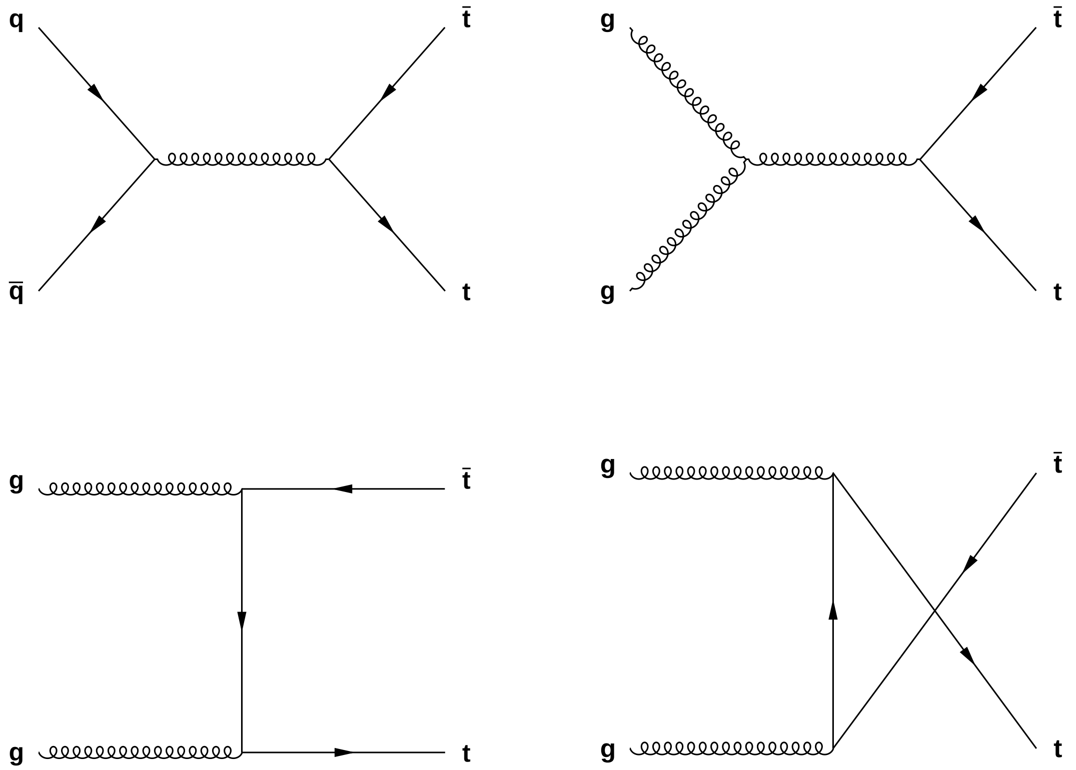

Figure 2.2: Leading-order Feynman diagrams for $t\bar{t}$ production in proton-proton collisions. The diagram on the top left shows the production via quark-anti-quark annihilation and all the others show the production via gluon-gluon fusion.

the respective proton, $f_i(x, \mu_f)$ and $f_j(x, \mu_f)$ are the PDFs for the parton types i and j, μ_r and μ_f are the renormalization and factorization scales, and $\hat{\sigma}_{ij\rightarrow t\bar{t}}(x_1, x_2, \mu_r, \mu_f)$ is the partonic cross section. The PDFs have been measured explicitly in Deep Inelastic Scattering (DIS) experiments at the Hera Collider. Current PDF sets like NNPDF 3.0 [24] contain additional constraints from other experiments like recent constraints from LHC measurements.

The production via gluon-gluon fusion is the dominant production mode at the LHC because of a high number of gluons in the proton with small fractional momentum. The inclusive $t\bar{t}$ production cross sections at 8 and 13 TeV are predicted at next-to-next-to-leading-order (NNLO) to be $252.9^{+6.4}_{-8.6}(\text{scale}) \pm 11.7(\text{PDF} + \alpha_S)$ pb and $831.8^{+19.8}_{-29.2}(\text{scale}) \pm 35.6(\text{PDF} + \alpha_S)$ pb using the TOP++2.0 program [25] and assuming a top quark mass of 172.5 GeV.

The production of single top quarks is divided into three production channels. The leading-order Feynman diagrams are shown in figure 2.3. Single top quarks can be pro-

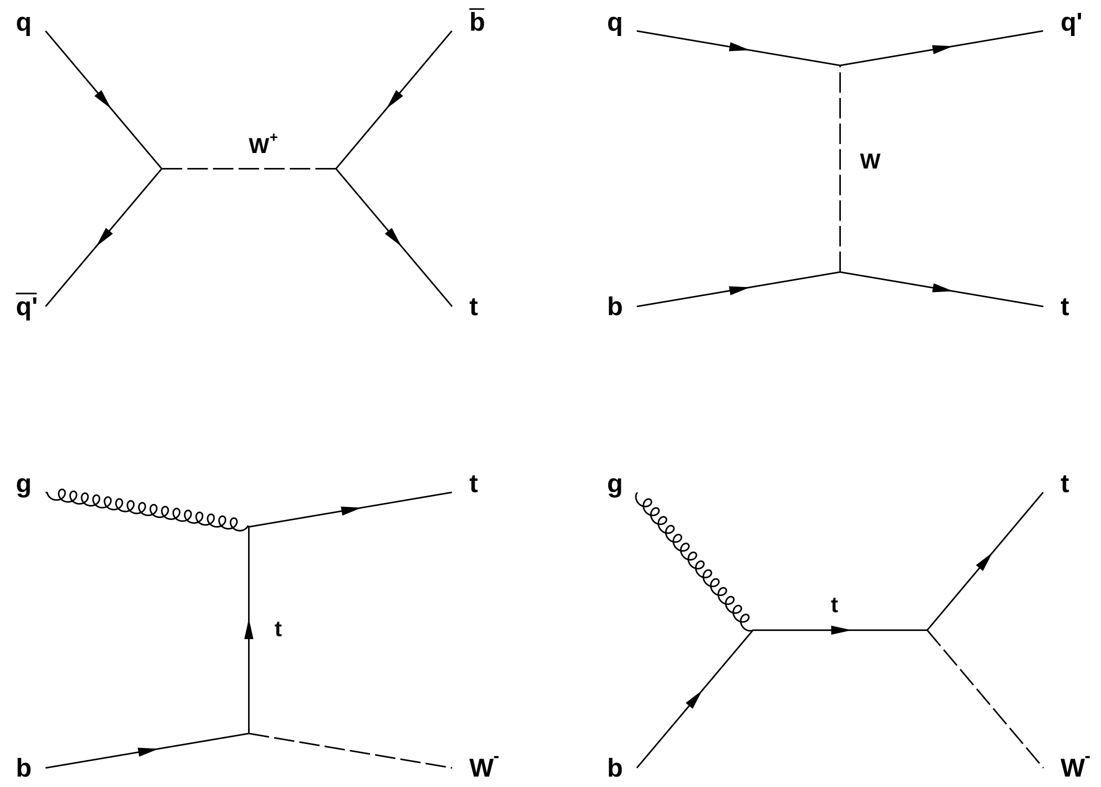

Figure 2.3: Leading-order Feynman diagrams for single top quark production in proton-proton collisions. The diagrams on the top show the s-channel (left) and the t-channel (right) production by an exchange of a W boson. The diagrams on the bottom show the production in association with a W boson.

duced in the s-channel and t-channel by an exchange of a W boson, or in association with a W boson (tW-channel).

2.2.2 Decay

The top quark decays via the weak interaction with a probability of more than 99 % into a W boson and a b quark. Decays including other quarks are also possible but strongly suppressed by the CKM matrix. The W boson can further decay hadronically into two quarks ($\mathrm{W} \rightarrow \mathrm{q}\bar{\mathrm{q}}'$) or leptonically into a lepton and the respective neutrino ($\mathrm{W} \rightarrow \mathrm{l}\nu$). The branching ratio for $\mathrm{W} \rightarrow \mathrm{l}\nu$ is about 33 % and to quarks about 67 % [2]. This leads to the branching ratios of the different $\mathrm{t\bar{t}}$ decay channels of $\sim$ 45 % for fully-hadronic decays where both top quarks decay hadronically, $\sim$ 44 % for lepton+jets decays where one top decays hadronically and the other one leptoincally, and $\sim$ 11 % for dileptonic decays in

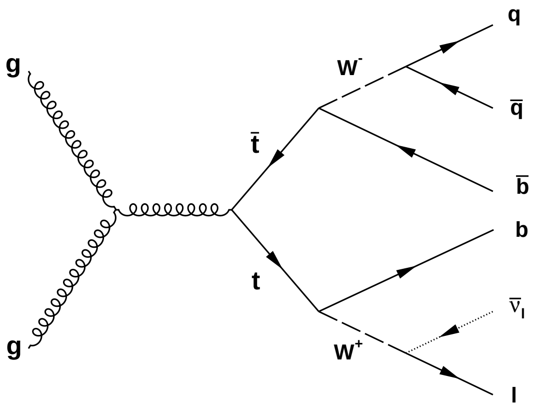

Figure 2.4: Leading-order Feynman diagram for a $\mathrm{t\bar{t}}$ pair produced by gluon-gluon fusion decaying in the lepton+jets decay channel.

which both top quarks decay leptonically.

In the first analysis in this thesis in chapter 6 the lepton+jets decay channel refers just to decays including an electron or a muon without the τ lepton leading to a branching ratio of $\sim 29\,\%$, excluding also $\tau \rightarrow \mathrm{e}$ and $\tau \rightarrow \mu$ decays. The reason for this definition is the more challenging reconstruction of the τ lepton especially for hadronic decays of the τ and the more difficult separation against the fully-hadronic decay channel. A Feynman diagram for a lepton+jets $\mathrm{t\bar{t}}$ decay is shown in figure 2.4.

2.2.3 Mass definition

The mass of the top quark is a fundamental parameter of the standard model. Together with the masses of the W boson and the Higgs boson it is important for consistency tests of the standard model [23]. The mass is further connected to considerations of the vacuum stability suggesting that the current measured value leads to a meta-stable universe[2] [26]. This makes the mass of the top quark an interesting parameter to study and to measure.

The most precise measurements are performed at the LHC by the CMS and ATLAS collaborations in so-called direct measurements. The most precise single measurement is performed by CMS and resulted in a top quark mass of $172.35 \pm 0.51\,\mathrm{GeV}$ [22]. The most precise measurement from ATLAS results in a mass of $172.08 \pm 0.91\,\mathrm{GeV}$ [27]. These direct measurements usually rely on a kinematic reconstruction of the full $\mathrm{t\bar{t}}$ decay from jets, leptons, and missing transverse momentum $p_{\mathrm{T}}^{\mathrm{miss}}$. A discussion is ongoing on how

[2] Considering no contributions from physics beyond the standard model.

exactly the measured mass is related to a well-defined mass in a proper renormalization scheme. Doubts often arise because the reconstruction of the $\mathrm{t\bar{t}}$ system relies on non-perturbative models of soft QCD effects in the parton-shower and hadronization models, which introduce an effective cutoff at the order of 1 GeV [28]. Commonly used theoretical mass defintions for the top quark are the $\overline{\mathrm{MS}}$ mass $\overline{m}_{\mathrm{t}}(\mu_r)$ depending on the renormalization scale μ_r and the pole mass $m_{\mathrm{t}}^{\mathrm{pole}}$ where all self energy corrections are included in the mass definition. The pole mass suffers from a renormalon ambiguity at the order of Λ_{QCD} [29, 30]. Another interesting top quark mass defintion is the MSR mass $m^{\mathrm{MSR}}(R)$ [31, 32] that includes corrections up to a scale R (typically $R = \Lambda_{\mathrm{QCD}}$) leading to a mass definition close to the pole mass withount the renormalon ambiguity.

Extractions of the top quark mass from kinematic distributions in dileptonic $\mathrm{t\bar{t}}$ prodction have also been studied. Examples are the invariant mass of the lepton-b-jet system m_{lb}, the transverse mass from the $\mathrm{b\bar{b}}$ sytem m_{T2}, and the invariant mass of the $\mathrm{lb}\nu$ system $m_{\mathrm{lb}\nu}$ [33–35]. These variables are expected to lead to a lower dependence on systematic uncertainties compared to the full reconstruction of the $\mathrm{t\bar{t}}$ system. Measurements of this kind have been performed by CMS [36] and lead to uncertainties reduced by $\sim$ 25 % compared to other measurements in the dilepton channel.

Systematic studies of the determination of the top quark mass comparing new generators with increased parton-shower precision can be found in reference [37]. Shifts of the top quark mass of the order of 200 MeV are observed in the absence of detector effects, increasing up to more than 1 GeV when a smearing is applied to account for the detector resolution.

Indirect measurements from the inclusive $\mathrm{t\bar{t}}$ cross section can be found in references [38, 39]. These measurements are expected to measure the pole mass of the top quark directly because they do not rely directly on the parton shower. The resulting uncertainties on the top quark mass are significantly larger than 1 GeV and exceed the uncertainties of the direct measurements.

In references [40–42] it is suggested that the measured value of the top quark mass in the direct measurements is shifted with respect to the pole mass by non-perturbative effects. A measurement of the top quark mass in the highly boosted regime is suggested for future $\mathrm{e^+e^-}$ collisions where the top quark is reconstructed in one single jet allowing a systematic treatment of the soft effects in the soft-collinear effective theory (SCET) [43–46]. A numerical description of the shift between the mass in a Monte Carlo generator and different well-defined mass definitions was studied in reference [47] in $\mathrm{e^+e^-}$ calculations using the 2-Jettiness [48] distribution in the boosted top quark regime. Analytic calculations from first principles are compared to PYTHIA 8.2 [49] leading to the conclusion that the measured top quark mass is closer to the MSR mass than to the pole mass. The shift to the pole mass was estimated to be $0.57 \pm 0.28\,\text{GeV}$. First calculations for proton-proton collisions under LHC conditions can be found in reference [50]. A first measurement of the jet mass in boosted top quark decays at the LHC using 8 TeV data was performed within the scope of this thesis (chapter 6) and is published by CMS in reference [3]. A preliminary result of a measurement by CMS in 13 TeV data can be found in reference [51].

2.3 Jet substructure

The study of jet substructure plays and important role at the LHC. It can be used to identify decays of heavy particles, like W bosons or top quarks, that are reconstructed in one large jet. Therefore, the use of jet substructure improves the sensitivity in searches for new physics where the sensitivity would decrease for the reconstruction of a heavy-particle decay in separate jets. All this leads to an increasing theoretical interest in jet substructure and to the development of new jet-substructure variables that are calculable from first principles. A comprehensive review can be found in reference [52].

A challenge in the calculations is the presence of soft QCD processes in the jets that lead to divergences. These divergences cancel out in the case of infrared and collinear-safe (IRC safe) variables when summing over the full soft and collinear phase space. IRC safe observables are insensitive to collinear splittings and infinitesimal soft emissions.

2.3.1 Jet mass

An important jet-substructure variable is the invariant mass of a jet called jet mass. The jet mass is the invariant mass of all particles associated to the jet. In case of groomed mass definitions only particles that pass a jet grooming procedure, described in section 2.3.3, are considered. The jet mass provides a good way to separate large jets that contain a full decay of a heavy particle, like a top quark or a W boson, from jets induced by light quarks or gluons. The jet mass of a heavy-particle jet is related to the mass of the respective particle. The mass of a jet that contains a hadronic top quark decay is expected to be close to the mass of the top quark. Jets from light-quark decays are expected to have lower masses. The separation between heavy-particle jets and light-quark or gluon jets is reduced by the parton shower and additional soft radiation which can significantly increase the jet mass and lead to very heavy jets from QCD multijet production.

Calculations of the jet mass are performed for $m_{\text{jet}} \ll p_{\text{T, jet}}$. Large logarithms of the form $\log^n(p_{\text{T, jet}}/m_{\text{jet}})$ contribute at each order of the perturbation theory. The influence of the logarithms is reduced by a resummation up to certain order called leading-logarithm (LL), next-to-leading-logarithm (NLL) and so on. First calculations of the jet mass have been performed for $\mathrm{e^+e^-}$collisions in references [53, 54] with NLL precision. More recent calculations reach an accuracy on NNNLL+NLO [55].

A calculation of the jet mass for boosted top quarks in $\mathrm{e^+e^-}$ collisions is performed in references [40–42]. A double-differential cross section of the right and left-hemisphere masses m_{R} and m_{L} is calculated at NLL precision using a factorization into different energy scales with the soft-collinear effective theory (SCET).

A simplified cross section in SCET, differential in the left and right-hemisphere masses m_{L} and m_{R}, can be expressed for $m_{\mathrm{L}}, m_{\mathrm{R}} \ll Q$ as

$$\frac{\mathrm{d}\sigma}{\mathrm{d}m_{\mathrm{R}}\mathrm{d}m_{\mathrm{L}}} = \sigma_0 H(Q^2;\mu) \cdot J(m_{\mathrm{L}};\mu) \otimes J(m_{\mathrm{R}};\mu) \otimes S(m_{\mathrm{L}}, m_{\mathrm{R}};\mu)[52], \tag{2.10}$$

where σ_0 in the cross section for $\mathrm{e^+e^-} \to \mathrm{q\overline{q}}$, the hard function H describes virtual corrections at the center-of-mass energy Q^2, the jet functions J describe collinear radiation within the respective jets at the scale of the respective hemisphere mass, and the soft function S includes effects of perturbative and non-perturbative soft QCD radiation. Figure 2.5 shows in a sketch for a single jet at which scales the different functions contribute. The hard function contributes at the scale of the jet p_{T}, the jet function at the scale of the jet mass m_{jet}, and the soft function at the scale of $m_{\text{jet}}^2/p_{\mathrm{T}}$. Non-perturbative effects within the soft function become important the scale of $\Lambda_{\mathrm{QCD}} \sim 1\,\mathrm{GeV}$.

Non-perturbative effects like hadronization have not yet been calculated successfully from first principles. They can only be introduced by a shape function that can be convoluted with the perturbative calculation. A common approach is the removal of the soft effects by a jet grooming explained later. Jet grooming becomes important for proton-proton collisions because of additional soft effects from the interaction of the proton remnants called underlying event (UE) and from additional interactions during the same bunch crossing called pileup (PU). Pileup is not correlated to the physics process and is therefore often handled by the experiment.

The influence of jet grooming on jet-substructure calculations was fist studied in references

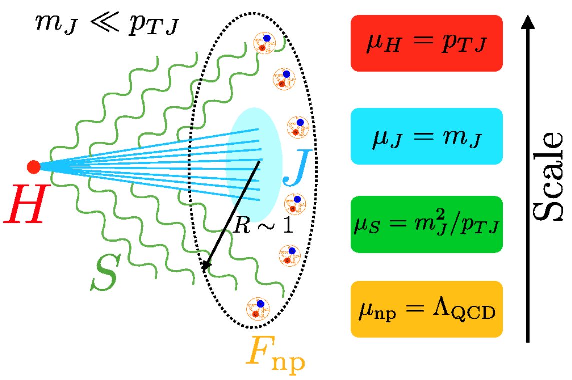

Figure 2.5: Sketch to visualize the relevant energy scales in a jet important for a factorization in SCET taken from reference [52]. The hard function H is responsible for virtual corrections in the production, the jet function J for collinear radiation within the jet, and the soft function S for perturbative and non-perturbative soft radiation. Non-perturbative effects become important at the scale of $\Lambda_{\text{QCD}} \sim 1\,\text{GeV}$.

[56, 57]. Calculations of the jet mass for light jets with Soft Drop [58] grooming are performed in references [59, 60]. A measurement of the groomed jet mass have been performed by CMS [61] and compared to the analytic calculations showing an agreement within the uncertainties.

A first calculation of the jet mass for boosted top quarks with light grooming in pp collisions is performed in reference [50] with NLL precision. The differential cross section as a function of the jet mass is shown in figure 2.6 compared to a MC generated distribution with PYTHIA 8. The distribution peaks close to the top quark mass and shows a good agreement between calculations and MC simulation in the peak region from 173 to 180 GeV. Calculation and MC simulation do not agree within uncertainties for low values of the mass.

A first measurement of the jet-mass distribution for boosted top quarks in data collected at a center-of-mass energy of 8 TeV was performed within the scope of this thesis and is described in chapter 6. The data could not be compared to analytic calculations yet because the measurement was published before the first proton-proton calculations became available and the calculations are not available for the measurement phase space described in chapter 6. A preliminary result of a jet-mass measurement in boosted $\text{t}\bar{\text{t}}$ production in CMS at 13 TeV can be found in reference [51].

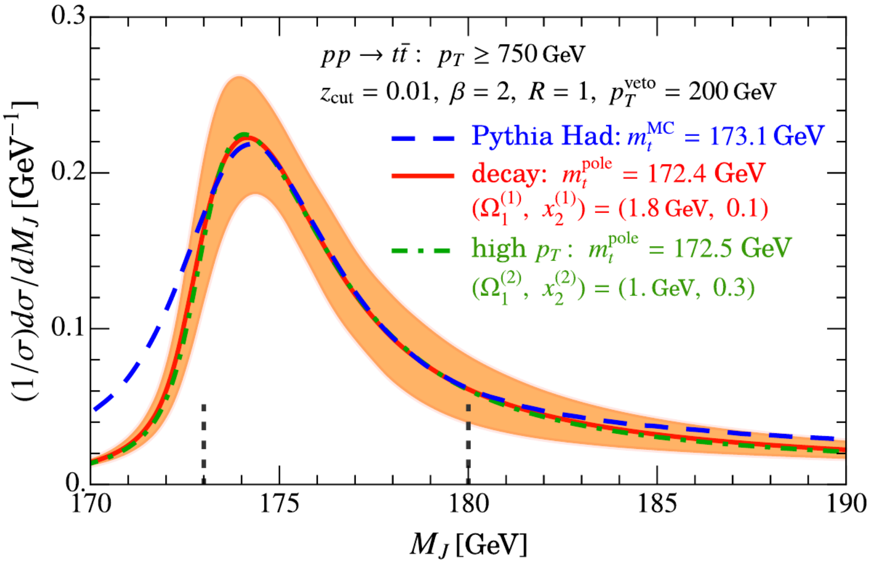

Figure 2.6: Calculated jet-mass distribution with light Soft Drop grooming in highly boosted top quark decays taken from reference [50].

2.3.2 N-subjettiness

Another important jet-substructure variable is called N-subjettiness τ_N [62, 63] which serves as a measure on how well a jet is compatible with an N-subjet hypothesis. It defines N subjet axes in the jet and the value τ_N is defined as

$$\tau_N = \frac{1}{d_0} \sum_k p_{\mathrm{T},k} \min\{\Delta R_{1,k}, \Delta R_{2,k}, ..., \Delta R_{N,k}\}, \tag{2.11}$$

where $p_{\mathrm{T},k}$ is the transverse momentum of the respective particle k, $\Delta R_{x,k}$ is the distance to the defined axis x in the η-ϕ plane, and d_0 is a normalization factor defined as:

$$d_0 = \sum_k p_{\mathrm{T},k} R_0, \tag{2.12}$$

with the jet distance parameter R_0 used in the jet clustering. The subjet axes are found by an iterative procedure starting with axes found by an exclusive k_T algorithm.

The energy distribution in a jet is expected to be different between jets that contain a fully-merged top quark decay and light-quark or gluon-induced jets. In the case of a hadronic top quark jet the energy is expected to be distributed in three subjets resulting from the three quarks from the top quark decay. A more uniform energy distribution is expected for a light-quark jet. In case of a top quark jet the value τ_3 is expected to be small with a higher value expected for a light-quark jet. The value τ_N gets smaller the better the energy flow is aligned with the N subjet axes. Especially sensitive to the different energy distributions are ratios of the form τ_N/τ_{N-1}. The ratio τ_3/τ_2 is often used for top tagging since it is expected to be small for jets with a three-prong substructure and higher for two or one-prong jets. Another example is the ratio τ_2/τ_1 that is interesting for hadronic W boson tagging where a two-prong structure is expected.

The N-subjettiness values τ_n are IRC safe but their ratios are not. The ratios are, however, Sudakov safe [64] which makes them still calculable. Calculations of the N-subjettiness ratio τ_2/τ_1 are performed [65] for jets with two-prong structure from W boson decays and for QCD jets. Figure 2.7 shows the calculated N-subjettiness distributions in WW production and in QCD dijet production, with and without grooming, and for different selections with the modified Mass-Drop Tagger (mMDT) [57]. The distributions for W jets peak at lower values compared to the light jets from QCD dijet production. Grooming helps to improve the separation between W jets and light jets.

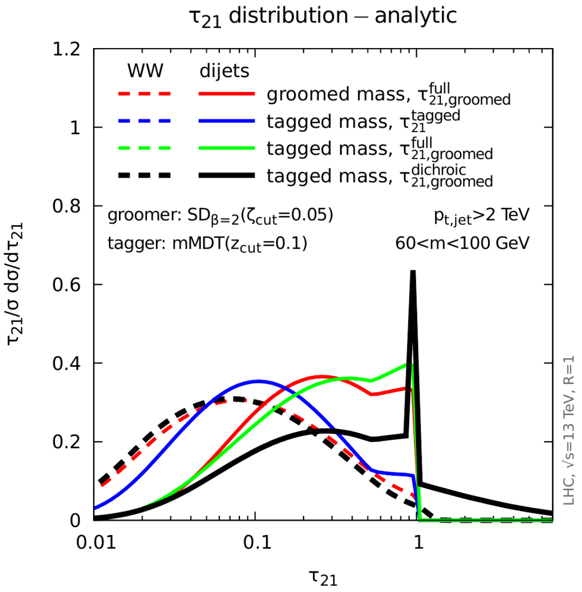

Figure 2.7: Analytic calculations of the N-subjettiness ratio τ_2/τ_1 taken from [65]. The N-subjettiness ration is shown for QCD dijet production and for WW production.

2.3.3 Jet grooming

The purpose of jet grooming is a removal of soft and wide-angle radiation coming from processes like initial-state radiation (ISR), the underlying event, and pileup. It helps to remove non-perturbative effects in the theoretical calculations and to improve the resolution of substructure variables. It should further remove non-global logarithms (NGLs) which arise in higher-order jet-substructure calculations as a result of the correlation between in-jet and out-of-jet radiation.

Several grooming algorithms have been studied in the past, like filtering [66], trimming [67], pruning [68] and others. These grooming algorithms have been shown not to remove NGLs efficiently in the calculations [56, 57]. An algorithm that efficiently removes the NGLs was developed within the scope of the modified Mass-Drop Tagger (mMDT) [57] and was generalized in the Soft Drop algorithm [58]. This thesis uses mainly the Soft Drop algorithm to define a groomed jet mass for top tagging purposes. The algorithm is briefly discussed below.

Soft Drop

The Soft Drop [58] grooming algorithm is an extension of the grooming in the modified Mass-Drop Tagger (mMDT) in reference [57]. It sequentially reverts steps of the clustering history of a jet[3]. At each step the softer of the two resulting pseudojets is removed and the algorithm continues with the remaining one until the following Soft Drop condition is fulfilled:

$$\frac{\min(p_{\mathrm{T}1}, p_{\mathrm{T}2})}{p_{\mathrm{T}1} + p_{\mathrm{T}2}} > z_{\mathrm{cut}} \left(\frac{\Delta R_{12}}{R_0} \right)^{\beta}, \tag{2.13}$$

where $p_{\mathrm{T}1}$ and $p_{\mathrm{T}2}$ are the transverse momenta of the two pseudojets, ΔR_{12} the angular distance between the pseudojets, and R_0 the original jet distance parameter used in the jet clustering. The parameters z_{cut} and β define the strength of the Soft Drop grooming.

Once the Soft Drop condition is passed the two pseudojets are called subjets and assigned to the original jet. The groomed jet four-vector is defined as the sum of the four-vectors of the Soft Drop subjets.

The Soft Drop grooming in CMS uses a β value of 0, a z_{cut} value of 0.1, and a Cambridge/Aachen (CA) [69, 70] jet clustering history. The CA algorithm is a sequential jet clustering algorithm that is explained in more detail in section 4.4. The values for the Soft Drop grooming used in CMS have been found and tested in reference [5].

2.4 New physics decaying into boosted top quarks

The short comings of the standard model are a motivation for many theories extending the standard model. Many of these extensions predict heavy hypothetical new particles, some of which decay dominantly into top quarks. Top quarks from the decays of such

[3] More information jet clustering algorithms can be found in section 4.4.

new particles can have a large momentum and a large Lorenz boost if the mass of the hypothetical particle is much larger than the mass of the top quark. The reconstruction and the identification of such highly boosted top quarks is challenging and special reconstruction techniques are needed. A good understanding of boosted top quarks is needed to gain sensitivity in the respective searches. More information on these techniques can be found in section 5.2. Three examples of such new particles are given below.

Vector-like quarks Vector-like quarks are colored fermions that are still allowed by the experimental data in contrast to a fourth generation of chiral quarks which is excluded by the Higgs measurements. Vector-like quarks have the same quantum numbers for left and right-handed components. Their left-handed and right-handed components transform in the same way under the $SU(3)_c \times SU(2)_L \times U(1)_Y$ symmetry of the standard model. Vector-like quarks do not obtain their mass by a Yukawa coupling to the Higgs boson but they can mix with standard model particles and in this way modify the coupling of other fermions to the Higgs boson. In this way they can be a part of a solution for the hierarchy problem. They further introduce additional sources of CP violation needed in many theories. A detailed review can be found in reference [71]. Recent searches for vector-like quarks at the LHC can be found in references [72–80].

Leptoquarks Leptoquarks are hypothetical new bosons that carry electrical charge and color charge. They mediate a new interaction between quarks and leptons and introduce a new symmetry between the quark and the lepton sector. Leptoquarks decay into a quark and a lepton. They are predicted by several extensions of the standard model like grand unified theories (GUTs) [81–83] that try to unify all interactions in one fundamental theory, compositeness models in which quarks and leptons are not fundamental particles but consist of so-called preons themselves [84, 85], and technicolor models [86]. Recent searches for leptoquarks decaying into top quarks can be found in references [87–89].

Resonances decaying into $\mathrm{t\bar{t}}$ pairs Several extensions predict particles decaying into $\mathrm{t\bar{t}}$ pairs like for example leptophobic $\mathrm{Z'}$ bosons predicted by extended gauge theories [90–92] or Kaluza-Klein excitations of gluons in theories with warped extra dimensions [93]. Searches for resonances decaying into $\mathrm{t\bar{t}}$ pairs at the LHC have been performed by CMS and ATLAS in references [94–96].

2.5 Simulation of proton-proton collisions

The complexity of a high-energy physics event, from the hard interaction to the detector response, poses a challenge to the interpretation of the measured data. Monte Carlo (MC) techniques are usually used in several stages to simulate the full event evolution and to obtain predictions that can be compared to the reconstructed data. The different stages of the event generation are briefly discussed below.

Hard matrix element calculation

The hard matrix element of an interaction is usually calculated by specialized MC generators. Three different generators for this purpose are used in this thesis. POWHEG [97–101] and MC@NLO [102] are able to produce events with up to next-to-leading-order (NLO) precision while MadGraph [103] provides leading-order (LO) plus additional jets. The cross section from the hard matrix element is convoluted with the parton density functions (PDFs) of the proton (see equation (2.9)). The PDF sets used in this thesis are given in the respective analysis chapters.

Parton shower and hadronization

The hard matrix element generators are interfaced with the multi-purpose generators PYTHIA [49, 104] or HERWIG [105] which are responsible for the simulation of the parton shower and hadronization. The evolution of the parton shower relies on perturbative calculations down to an energy of $\sim 1\,\mathrm{GeV}$. Below this scale non-perturbative soft QCD effects in the parton shower and hadronization become significant which can not be calculated and rely on tuned models instead. PYTHIA uses a string fragmentation model while HERWIG uses a cluster fragmentation model. Both models are often compared to estimate an uncertainty on the choice of the parton-shower model. Interactions of the proton remnants are handled by tuned models, so called underlying-event tunes, in PYTHIA or HERWIG.

Detector simulation

On top of the previous steps a full simulation of the CMS detector is applied using the GEANT4 framework [106]. This includes a simulation of the interactions of the stable particles with the material in the different parts of the CMS detector. Pileup is added to the detector-level simulation by the production of additional interactions in the same event.

Parton, particle, and detector level

The definition of different levels is important to compare calculations from first principles, MC simulation, and data at the same level of included processes and corrections. The parton level for heavy particles is often defined before their decay and might include radiation effects. It does not include any effects of the hadronization or the underlying event. The particle level or hadron level includes effects of the parton shower as well as non-perturbative hadronization effects. It includes a description of the underlying event but within this thesis no effects of pileup. The detector level includes all effects of the particle level. It further includes the description of pileup and a full simulation of CMS detector response. The detector level is the level at which the data is accessible. At this level only comparisons between data and MC simulation are possible because analytic calculations can only be performed at the parton or at the particle level. For comparisons of the data to analytic calculations the data needs to be corrected for detector effects first.

3 Experiment

The data analyzed in this thesis was recorded by the Compact Muon Solenoid (CMS) detector in proton-proton collisions at the LHC at the European Organization for nuclear research (CERN) near Geneva in Switzerland. This section will give a short introduction to the LHC and the CMS experiment and will briefly summarize the most important principles.

3.1 The Large Hadron Collider

The Large Hadron Collider is a circular hadron collider designed for proton-proton and heavy-ion collisions at CERN. Detailed information on the LHC accelerator can be found in reference [107]. It is built in the tunnel of the Large Electron Positron collider LEP with a circumference of 26.7 km. It is designed to reach a center-of-mass energy of 14 TeV for pp collisions with an instantaneous design luminosity of $10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. Particles are pre-accelerated by a chain of pre-accelerators and injected into the LHC ring where they are ramped up to the final collision energy. A sketch of the full accelerator complex can be found in figure 3.1. The particles in the LHC are accelerated using superconducting cavities. Superconducting dipole magnets are used to keep the particles on a circular trajectory and higher magnetic moments are used to focus and stabilize the beams. The LHC has four interaction points in which the proton beams are focused and crossed to allow interactions between the particles. Each of the interaction points holds one of the four main experiments at the LHC. ATLAS [109] and CMS [110] are large multi-purpose detectors mainly designed to the search for the Higgs boson, to study its properties, and to search for heavy new particles predicted by new physics models in proton-proton collisions. They are further capable to analyze heavy-ion collisions. The ALICE experiment [111] is designed to study heavy-ion collisions and the LHCb experiment [112] is designed to perform precise measurements of the decays of b-mesons.

After an incident in the year 2008 and the following repairs, the LHC operated with a

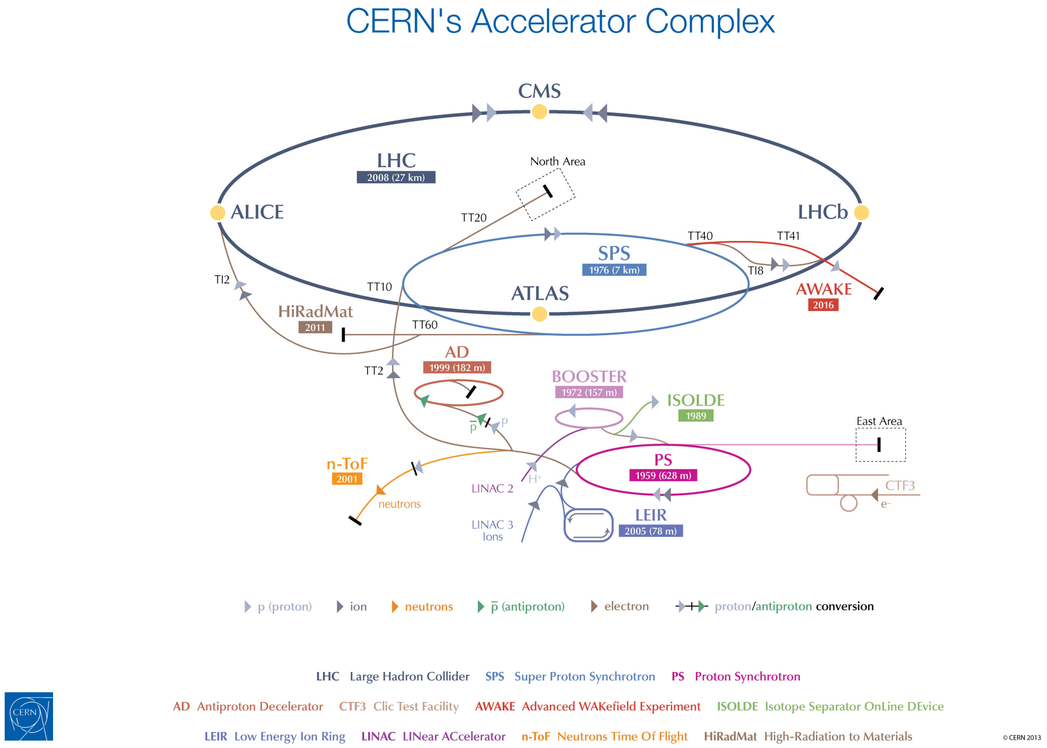

Figure 3.1: Sketch of the full accelerator complex at CERN taken from reference [108].

reduced center-of-mass energy of 7 TeV in 2010 and 2011 and with 8 TeV in 2012. After the first long shutdown the center-of-mass energy was increased to 13 TeV in the years 2015 to 2018. The integrated luminosity delivered to the CMS experiment is shown in figure 3.2 showing the cumulative luminosity as a function of time for the different runs of the LHC. The total amount of delivered luminosity for a specific center-of-mass energy increases significantly with increasing center-of-mass energy.

3.2 The Compact Muon Solenoid

The CMS detector is a multi-purpose detector stationed in one of the interaction points of the LHC. A detailed description of the detector can be found in the technical design report [110]. The general structure and some information on the various detector components are summarized below. An important motivation for the design of the CMS detector was the reconstruction of Higgs boson decays but it is also designed for several other standard model measurements, for searches for new heavy particles, and to reconstruct heavy-ion collisions to gain new insights in QCD. The detector is designed for an excellent reconstruction of muons, a good energy resolution for electromagnetic showers, a good

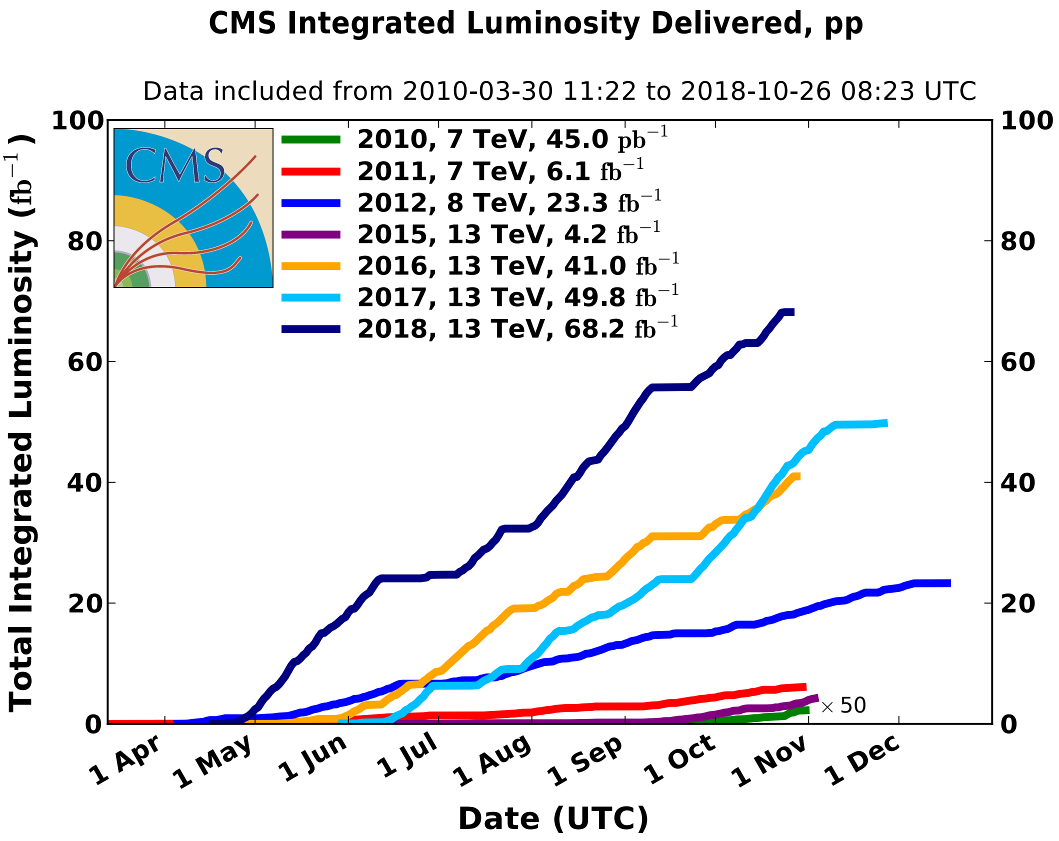

Figure 3.2: Cumulative luminosity of proton-proton collisions delivered by the LHC to the CMS experiment for different runs in the years 2010 to 2018. The figure is taken from reference [113].

charged-particle resolution and reconstruction efficiency close to the interaction point, and a good missing transverse momentum and dijet-mass resolution. The general concept is an onion-like structure of several different detector layers placed around the interaction point. An important component of the detector is a large superconducting solenoid with an inner radius of 3 m and a magnetic field up to 4 T. The trajectories of charged particles are bend by the magnetic field allowing a measurement of their momenta from the bending radius of the reconstructed tracks. A silicon-based tracking detector is placed inside of the magnet close to the interaction point to reconstruct the trajectories of charged particles from hits in the different layers of the tracking system. The tracker is followed by an electromagnetic calorimeter (ECAL) to absorb and measure the energy of electromagnetically interacting particles, mainly electrons and photons. The ECAL is extended by a hadronic calorimeter to measure the energy of charged and neutral hadrons. Outside of the magnet, an iron return yoke is used to return the magnetic field and to provide a high magnetic field for the muon system that is integrated into the return yoke. Several layers of muon chambers are used to improve the measurement of the trajectories of muons which pass the inner detector components. A sketch of the CMS detector can be found in figure 3.3. More information on the different detector components is given below or in reference [110].

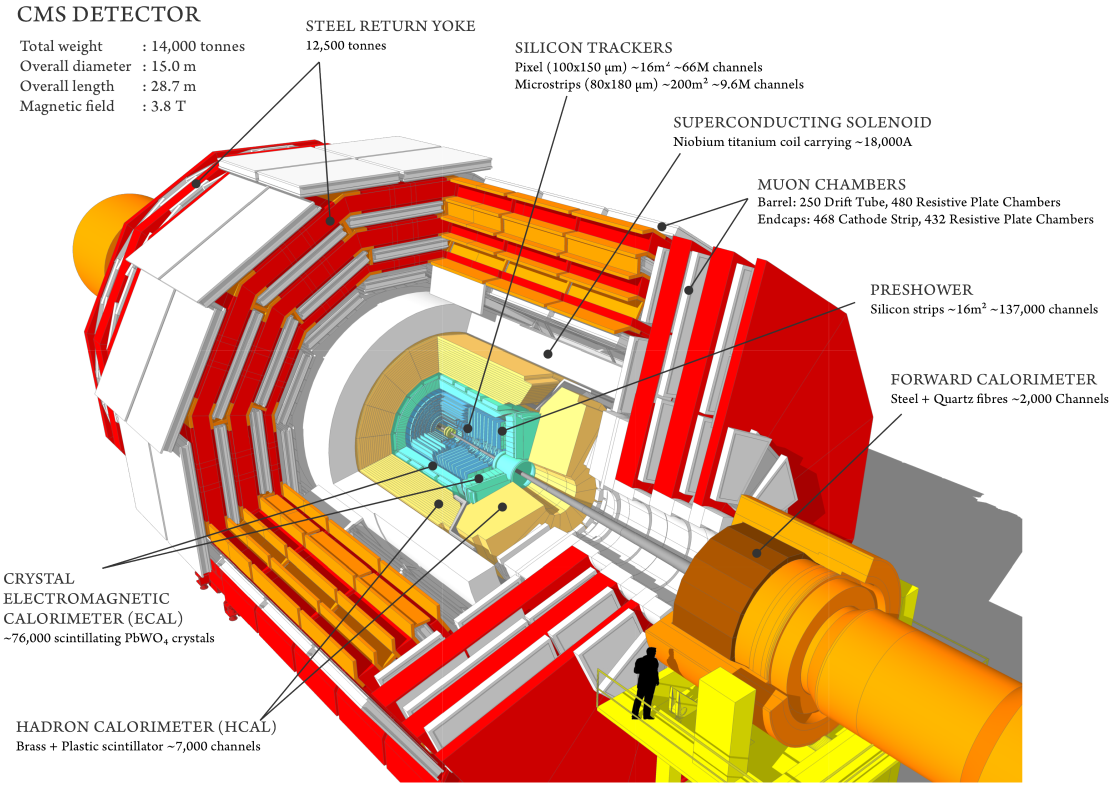
CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO$_4$ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 3.3: Sketch of the CMS detector, taken from reference [114].

3.2.1 Coordinate system

The coordinate system used throughout the thesis is a right-handed coordinate system with the x axis pointing horizontally towards the center of the LHC ring and the y axis pointing vertically upwards. The azimuthal angle Φ is defined in the x-y plane and the polar angle Θ with respect to the z axis. Usually the Lorentz-invariant pseudo-rapidity $\eta = -\ln\tan(\Theta/2)$ is used instead of Θ because the initial momentum of the interacting particles in the z direction is unknown.

3.2.2 Tracking system

The tracking system is the first component of the detector closest to the beam pipe. It is used to reconstruct tracks of charged particles. The tracker is placed in a magnetic field of 4 T, provided by the solenoid, which bends the tracks of charged particles. The transverse momentum of a charged particle can be measured from the bending radius of its track. The tracker is also used to reconstruct secondary vertices close to the interaction point, for example from decays of b mesons within a jet. The whole tracker is based on several layers of silicon detectors. Tracks of charged particles are reconstructed from several hits in the different layers of the tracking system. More details on the reconstruction can be found in chapter 4.

Closest to the beam pipe are three layers of pixel detectors in the barrel part[1] and two discs for the forward region covering up to $|\eta| < 2.5$. A sketch visualizing the η-coverage of the different pixel layers is given in figure 3.4. The pixels in each barrel layer have a size of $100 \times 150\,\mu\text{m}^2$. The small size of the pixels helps to improve the track resolution close to the interaction point and therefore improves the reconstruction of secondary vertices.

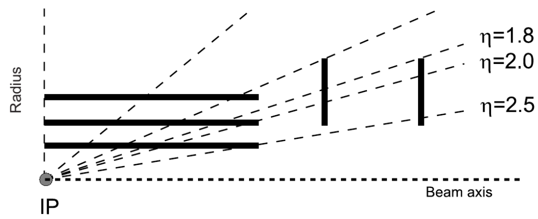

Figure 3.4: Sketch of the η-coverage of the CMS pixel layers up to the 2016 data taking. The sketch is taken from reference [114].

[1] A fourth pixel layer was installed in the shutdown between the 2016 and the 2017 data-taking periods.

Following the pixel layers are 10 layers of silicon strip detectors in the barrel part and 13 discs at each end of the barrel region extending the tracker up to $|\eta| < 2.5$. The silicon strip detectors have a worse spacial resolution compared to the pixel layers.

3.2.3 Calorimeters

Calorimeters are used to absorb particles and to measure the deposited energy. In CMS, a homogeneous electromagnetic calorimeter is used following the tracking system to measure the energy of electromagnetically interacting particles like electrons and photons. The electromagnetic calorimeter consists of lead tungstate ($PbWO_4$) crystals. Electromagnetically interacting particles passing the material induce electromagnetic showers. Particles in the electromagnetic shower excite electrons in the scintillator material to a higher energetic state and scintillation light is produced by returning to the ground state. The amount of scintillation light is proportional to the deposited energy. The scintillation light is read out by photo multipliers. The homogeneous crystals lead to a good energy resolution for electrons and photons.

The electromagnetic calorimeter is followed by a hadronic calorimeter HCAL used to measure the energy of charged and neutral hadrons that are not fully absorbed in the ECAL. The main part of the HCAL is a sampling calorimeter that consists of several layers of brass plates to absorb energy and plastic scintillator as active material. The sampling structure of brass and plastic scintillator is used in the barrel part for $|\eta| < 1.3$ and for the endcaps with $1.3 < |\eta| < 3$. In the forward region for $3 < |\eta| < 5$ the sampling calorimeter is supplemented by a radiation-hard Cherenkov-based detector using quartz fibers as active material. An additional hadronic calorimeter is placed outside of the solenoid in the barrel part consisting of one or two layers of scintillator material depending on the η range. These additional layers are necessary because the depth of the calorimeters inside the magnet is not enough to entirely absorb some of the hadronic showers. The additional layers are used to catch the tails of those showers and thus to improve their energy resolution.

The resolution of the calorimeters can be parametrized as in reference [110]:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2, \tag{3.1}$$

where S is the stochastic term associated to statistical fluctuations in the shower and photostatistics, N is the noise term caused by electrical noise, digitization noise, and pileup noise, and C is a constant term caused by non-uniform longitudinal light collection, calibration errors, and energy leakage.

3.2.4 Magnet

The magnet is a superconducting solenoid with a length of 12.5 m and a free bore with a radius of 3 m. It is designed to reach an inner magnetic field of 4 T. The magnetic field is crucial for the detector because it bends the tracks of charged particles and allows a measurement of the charged-particle momentum from the curvature of the track. The magnet is surrounded by an iron return yoke to return the magnetic field and ensure a high magnetic field close to the magnet for the muon chambers.

3.2.5 Muon system

Muon champers are placed outside of the magnet and are embedded into the iron return yoke. The muon system is used for three different purposes, for the identification of muons, muon reconstruction, and triggering. Three different kinds of muon detectors are used in the muon system matching the requirements for the different purposes at the different positions.

Drift tubes (DTs) are used in the barrel region for $|\eta| < 1.2$ where the muon rate and the background from neutrons is low. A drift tube chamber consists of several drift cells. Drift cells are gas-filled cells with an anode wire in the middle and cathode strips at the two ends of the cell. Muons passing a drift cell ionize the gas and the resulting electrons and ions drift to the anode wire and to the cathode strips where the signal can be read out.

Cathode strip chambers (CSCs) are used in the endcap discs for $0.9 < |\eta| < 2.4$ where the muon rate and backgrounds are high and the magnetic field is not uniform. The CSCs are gas-filled multiwire chambers with layers of wires interchanged with layers of cathode strips. The wires are orientated azimuthal and the strips radially. The chambers are arranged circular around the beam. Particles passing the chambers ionize the gas, electrons and ions drift to the anodes and cathodes at which the electrical signals can be read out.

Resistive plate chambers (RPCs) are used both in the barrel and endcap regions. They provide a very fast read out for an ionizing signal, which makes them suitable for triggering purposes. The muon trigger system is based on six layers of RPCs in the barrel and three in the endcaps.

3.2.6 Trigger

Bunch crossings take place every 25 ns within the CMS detector which leads to an interaction rate of 40 MHz. This poses a challenge to the handling of the data because a full reconstruction of such a large amount of events is not possible. CMS uses a two-step trigger system to reduce the rate to an amount that can be handled. The first step is the Level-1 (L1) trigger. The L1 trigger is based on electronics and uses only rough data from calorimeters and the muon system accessible at this time scale while the high-resolution data is kept in memory pipelines. The L1 trigger is able to reduce the rate to a maximum of 100 kHz. The second step of the trigger is the software-based High-Level Trigger (HLT). The HLT algorithms have access to the full detector output and can use more complex variables based on reconstructed objects similar to the ones used in later analyses. The HLT is used to reduce the rate further by a factor of 10^3 to be able to handle the data.

4 Event reconstruction

The event reconstruction in CMS is based on the Particle Flow PF algorithm [115] using information from all sub-detector components to identify different types of particles and to reconstruct their four-momenta. The CMS detector is well suited for the PF algorithm because of the fine granularity of the tracker and the ECAL, the high magnetic field, and the good reconstruction of muons. The calorimetry inside the magnet also helps the PF algorithm by avoiding dead material between the tracker and the calorimeters and therefore allowing a good matching efficiency between tracks and calorimeter clusters.

Different particles can leave signals in various detector components. Electrons for example leave a track in the tracking system and deposit energy in the ECAL. Photons are expected to deposit energy in the ECAL but do not leave a track in the tracking system. Similar is the case for charged and neutral hadrons which both deposit energy in the calorimeters but only the charged hadrons lead to a track in the tracking system. The PF algorithm helps to improve the reconstruction of jets by linking tracking information for charged hadrons to the calorimeter clusters. Muons can be reconstructed from hits in the tracker and in the muon system.

More details on the PF event reconstruction and the reconstruction of specific objects are given in the following sections and in reference [115]. Exact definitions of the objects used in the analyses presented in this thesis can be found in the respective chapters.

4.1 Basic reconstruction steps

The PF algorithm is based on a reconstruction of basic objects in the different detector parts and on a linking of these objects. The combination of information from the various sub-detectors leads to an improved energy and angular resolution.

4.1.1 Track reconstruction

Charged-particle tracks are an important input to the PF algorithm and important for the jet reconstruction because about two thirds of the energy of a jet comes from charged particles. For low p_{T} the energy resolution of the tracker is superior to the one of the calorimeters.

Tracks are reconstructed using an iterative-tracking algorithm [116]. For each iteration, hits associated to tracks found in the previous iteration are removed from the input list. The quality criteria on tracking seeds and the track fit are reduced for different iterations to achieve a high tracking efficiency with a moderate misidentification rate. In the last few iterations the requirements on the primary vertex are relaxed to reconstruct tracks from secondary vertices from photon conversions or decays of long-lived particles.

4.1.2 Calorimeter clustering

Energy deposits in the calorimeters are clustered for several purposes. The calorimeter clusters are used to identify and measure neutral particles and to separate energy deposits of neutral particles from the deposits of charged particles. They are further used to measure the energy of charged hadrons in cases where the calorimeter measurement is superior to the tracker information and the ECAL is used to identify clusters from electrons. More information on the clustering can be found in reference [115].

4.1.3 Linking of detector signals

Most particles are expected to leave signals in several sub-detectors. A linking of the respective detector signals is therefore important. Tracks are extrapolated from the last layer of the tracker to the calorimeters to link the tracks to calorimeter clusters. Calorimeter clusters from photons produced by bremsstrahlung of an electron are linked to electron tracks by a tangential extrapolation. ECAL and HCAL clusters are linked similar to the linking of tracks to calorimeter clusters. For the muon reconstruction tracks from the tracker are linked to tracks from the muon system. They are called ’global muon’ if the global fit returns an acceptable χ^2.

4.2 Particle-flow event reconstruction

The actual particle-flow reconstruction follows several steps. Global muons are marked as PF muon candidates and are removed from the input list if the combined momentum is comparable with the one of the tracker-only muon within three standard deviations. Electron candidates are identified by a combination of tracker and calorimeter information. Tracks and calorimeter clusters associated to electron candidates are removed. This includes calorimeter clusters of identified photon candidates from bremsstrahlung. Tighter quality criteria on tracks are applied in the following steps. PF charged hadron candidates are identified from tracks matched to calorimeter clusters. PF photon candidates and neutral hadron candidates are identified from calorimeter clusters that can not be matched to tracks.

4.3 Primary vertices

Primary vertex candidates are found using a deterministic annealing algorithm [116] on selected tracks consistent with coming from a prompt decay in the interaction region. Primary-vertex candidates with more than two tracks are then fitted with an adaptive vertex fitter [117] to obtain the vertex parameters. The primary event vertex is defined as the primary-vertex candidate with the highest sum of the quadratic transverse momenta p_{T}^2 of the associated tracks.

4.4 Jet reconstruction

Jets in CMS are reconstructed from a list of all PF candidates in the event using the FastJet software package [118]. All jet algorithms used in this thesis are sequential clustering algorithms. Very commonly used algorithms of this kind are the k_{T} [119], the anti-k_{T} [120], and the Cambridge/Aachen (CA) [69, 70] algorithms.

They are based on the same principle. They start with a list of all PF candidates in the event and define two distance measures:

$$d_{ij} = \min(p_{\mathrm{T},i}^{2p}, p_{\mathrm{T},j}^{2p}) \frac{\Delta R_{ij}^2}{R^2} \tag{4.1}$$

$$d_{iB} = p_{\mathrm{T},i}^{2p}, \tag{4.2}$$

where d_{ij} is the distance measure between the particles or pseudojets i and j and d_{iB} is the distance measure to the beam. The jet distance parameter R is a constant parameter that defines the size of the resulting jets. The parameter p is different for the three algorithms and will be discussed later. The algorithms proceed in the following way.

- The smallest distance measure is determined from the list of all possible d_{ij} and d_{iB} values.
- Pseudojets i and j are combined if d_{ij} is the smallest measure.
- If d_{iB} is the smallest measure, i is called a jet and is removed from the list.
- This is repeated until the input list is empty.

The k_{T} algorithm is now obtained by setting the parameter p to 1. This means that d_{ij} gets larger for higher p_{T} of the pseudojets. Therefore the k_{T} algorithm tends to cluster soft particles first.

The Cambridge/Aachen algorithm is obtained for $p = 0$. The CA clustering does not depend on the p_{T} of the particles leading to purely geometrical distance measures. It is often used in studies with jet substructure because the clustering history can be used to define geometrically separated energy clusters within large jets called subjets.

The standard jet algorithm used in CMS is the anti-k_{T} algorithm using $p = -1$. This leads to the opposite behavior compared to the k_{T} algorithm. It tends to cluster the harder particles first leading to very conical jets.

Another jet algorithm used in this thesis is the algorithm from the Heavy Object Tagger with Variable R (HOTVR) [6] based on the variable-R algorithm [121]. It defines the same distance measures as defined in equations (4.1) and (4.2), only the fixed distance parameter R is replaced with a p_{T}-dependent parameter $R_{\mathrm{eff}}(p_{\mathrm{T}}) = \rho/p_{\mathrm{T}}$, where the parameter ρ defines how strong the jet radius scales with p_{T}. Jets with higher momentum will be reconstructed with a smaller radius compared to jets with a lower momentum. A cutoff on R_{eff} is used to avoid very small or very large jet radii leading to a definition of R_{eff} as:

$$R_{\mathrm{eff}} = \begin{cases} R_{\mathrm{min}} & \text{for } \rho/p_{\mathrm{T}} < R_{\mathrm{min}} \\ R_{\mathrm{max}} & \text{for } \rho/p_{\mathrm{T}} > R_{\mathrm{max}} \\ \rho/p_{\mathrm{T}} & \text{else.} \end{cases} \tag{4.3}$$

The value p is chosen to be $p = 0$ leading to a CA-like clustering.

The HOTVR jet algorithm includes a jet grooming within the clustering to remove soft radiation from effects like pileup or underlying event. This is needed because jets with a large radius are sensitive to collecting additional soft radiation because of their large jet area. The grooming is included by a mass-jump criterion

$$\theta \cdot m_{ij} < \max(m_i, m_j), \tag{4.4}$$

that is checked at each combination step if the mass of the combined pseudojets m_{ij} is larger than a fixed value μ. If the mass-jump criterion fails the lighter pseudojet will be identified as soft radiation and removed from the input list. The value of θ can be used to set the strength of the mass-jump grooming. The full algorithm works now in the following way:

- At each step it looks for the smallest distance measure.
- Pseudojet i is called a jet and removed from the list if d_{iB} is the smallest distance measure.
- If d_{ij} is the smallest measure, the mass of the combination m_{ij} will be checked.
    - i and j are combined if $m_{ij} < \mu$.
    - Else, the mass-jump criterion is checked.
        - If the mass jump fails the pseudojet with the lower mass is removed from the input list.
        - If i and j pass the mass jump the transverse momenta of i and j need to fulfill $p_{\mathrm{T}\,i,j} > p_{\mathrm{T,sub}}$.
            - The respective pseudojets are removed if their p_{T} is too soft.
            - If both fulfill $p_{\mathrm{T}\,i,j} > p_{\mathrm{T,sub}}$ i and j are combined and the pseudojets i and j are assigned as subjets to the resulting pseudojet. If i and j already have subjets, their subjets are assigned to the resulting pseudojet instead of i and j.
- These steps are repeated until the input list is empty.

The HOTVR algorithm is available as a FastJet plug-in the FastJet contrib package [122]. The default values for the jet clustering recommended in the paper are used for the jet clustering in this thesis and summarized in the following list: $\rho = 600\,\mathrm{GeV}$, $R_{\mathrm{min}} = 0.1$, $R_{\mathrm{max}} = 1.5$, $\theta = 0.7$, $\mu = 30\,\mathrm{GeV}$, $p_{\mathrm{T,sub}} = 30\,\mathrm{GeV}$.

4.5 Pileup-removal techniques

Due to the high luminosity of the LHC it is very likely that many proton-proton interactions take place within the same bunch crossing. These additional interactions to the hard physics of the primary event vertex are referred to as pileup interactions. The pileup interactions lead to additional particles in the event that affect physical objects like jets or missing transverse momentum. In the case of jets the additional particles are clustered into the jet leading to an increase of the jet momentum or mass and to worse resolutions. Different methods to reduce the influence of pileup particles are available and two methods that are used in this thesis are described below.

4.5.1 Charged Hadron Subtraction (CHS)

Primary vertices are reconstructed as described above. Vertices other than the primary event vertex are considered as pileup vertices. Charged hadrons associated to one of the pileup vertices are removed from the list of particles that is used to reconstructs physical object like jets. This pileup removal is called Charged Hadron Subtraction (CHS) and is described in reference [115].

4.5.2 Pileup Per Particle ID (PUPPI)

The PUPPI algorithm is a per-particle approach that assigns a weight to each PF candidate in the event. It uses charged particles to identify the pileup contributions and then locally applies weights to other candidates like neutral hadrons. The weights range between zero and one, where a candidate from pileup should get a weight of zero and a candidate from the hard interaction a weight of one. The weights are used to rescale the four-momentum of each PF candidate. Physical objects like jets are reconstructed from the rescaled PF candidates. More details on the PUPPI algorithm can be found in reference [123].

4.6 B tagging

B tagging is used to identify jets that include decays of b mesons. In this thesis b tagging is used to identify and reconstruct b quarks from top quark decays. It is important to enrich the selected data with processes from $\mathrm{t\bar{t}}$ production. The long lifetime of b mesons leads to secondary decays close to the primary vertex. The resulting secondary decay vertices from b mesons within the jet can be reconstructed. Most b tagging algorithms make use of the information of reconstructed secondary vertices within jets but also soft leptons from semileptonic b quark decays can be used to identify b meson decays inside a jet. A description of the three b tagging algorithms that are used in this thesis is given below.

4.6.1 Combined Secondary Vertex (CSV) algorithm

The CSV algorithm combines information from reconstructed secondary vertices with information on tracks associated with the jet. Secondary vertices are reconstructed with an adaptive vertex fit [117] using high-quality input tracks. A multivariate analysis (MVA) approach is used to combine several variables based on the reconstructed secondary vertices and track information. The additional track information allows discrimination even without a reconstructed secondary vertex. The final CSV discriminator is a combination of two likelihood ratios built from the input variables to separate b from c quark jets and b jets from light-flavor jets. More details on the algorithm and the exact input variables can be found in reference [124]. The CSV algorithm was commonly used in CMS for the analysis of data taken at a center-of-mass energy of 8 TeV.

4.6.2 CSVv2

The CSVv2 algorithm is based on the principle of the CSV algorithm and was developed for the analysis of data taken at a center-of-mass energy of 13 TeV. It uses a new inclusive-vertex-finding (IVF) algorithm using all tracks in the event instead of just the tracks associated to the jet. The discriminator is built by a neural network, a multi-layer perceptron with one hidden layer, instead of the previous likelihood ratios. More input variables have been added to the MVA. Details on the tagging algorithm and on the input variables can be found in reference [125].

4.6.3 DeepCSV

The DeepCSV algorithm was also developed for the analysis of 13 TeV data and is based on the CSVv2 algorithm. In contrast to CVSv2 it uses a deep neural network with several hidden layers. The same input variables are used with more tracks for each of the track-based variables leading to a better performance compared to the CSVv2 algorithm. More details can be found in reference [125].

5 Experimental methods

This chapter includes a brief description of two experimental methods that are of fundamental importance for this thesis and that go beyond the event reconstruction described in the previous chapter.

5.1 Unfolding

The experimental data in high-energy physics is usually obtained by a relatively complex reconstruction including signals from various detector components. In measurements with real data the fundamental physics quantities are therefore folded with detector and reconstruction effects. This makes it difficult to compare the data to theory predictions because the measured quantity might be smeared and shifted due to limited detector resolution and reconstruction effects. The different experiments use detector simulations to simulate the detector response for MC generated events and to compare MC generated distributions to real data at the detector level. However, dedicated studies for differences in the detector response between data and simulation are needed and are usually done within the collaborations. That makes it difficult to reinterpret the results with new simulations outside the experimental collaborations. It is further not possible to compare the data to analytic calculations because the detector simulation is only applicable to generated particles and not to analytic functions describing specific distributions.

Specific differential cross section distributions are often corrected for detector effects to the particle level or to the parton level to allow for an easier comparison to different MC generators and calculations. This is often done by unfolding algorithms. The fist analysis presented in this thesis uses the TUnfold framework [126] which is briefly described in the following section.

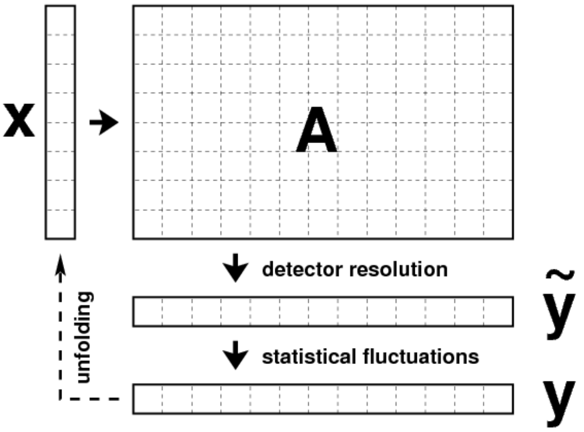

Figure 5.1: Sketch of the unfolding problem where the particle or parton-level distribution $\boldsymbol{x}$ if folded with detector effects described by a response matrix $\mathbf{A}$. This results in an average detector-level distribution $\tilde{\boldsymbol{y}}$ where the real measured distribution $\boldsymbol{y}$ can differ because of statistical fluctuations. The unfolding aims on a determination of the distribution $\boldsymbol{x}$ from the measured detector-level distribution $\boldsymbol{y}$. The sketch is taken from reference [126].

5.1.1 TUnfold

The unfolding problem can be posed in the following way. A true distribution $\boldsymbol{x}$ is folded with detector effects leading to migrations between the bins in $\boldsymbol{x}$, resulting in a measured distribution $\boldsymbol{y}$. These migrations can be described by a migration or response matrix $\mathbf{A}$. The response matrix $\mathbf{A}$ holds the probabilities A_{ij} that an event that is generated in bin j is measured in bin i. The unfolding problem can be written as

$$\tilde{y}_i = \sum_j A_{ij} x_j, \tag{5.1}$$

where $\tilde{y}_i$ is the expected mean of the detector-level distributions. The truly measured value y_i can differ because of statistical fluctuations. The problem is visualized in a sketch in figure 5.1 taken from reference [126].

TUnfold uses a regularized unfolding approach to solve the unfolding problem and to obtain the true distribution $\boldsymbol{x}$ from the detector-level distribution $\boldsymbol{y}$ with the help of a simulated response matrix. The TUnfold method is looking for the stationary point of a Lagrangian with three contributions:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3. \tag{5.2}$$

It is based on the minimization of the likelihood

$$\mathcal{L}_1 = (\boldsymbol{y} - \mathbf{A}\boldsymbol{x})^{\mathrm{T}} \mathbf{V}_{\boldsymbol{yy}}{}^{-1} (\boldsymbol{y} - \mathbf{A}\boldsymbol{x}), \tag{5.3}$$

where $\mathbf{V}_{\boldsymbol{yy}}$ is the covariance matrix of the input bins in $\boldsymbol{y}$. This minimization amplifies statistical fluctuations on the detector-level input distribution leading to non-physical fluctuations on the result $\boldsymbol{x}$.

An additional regularization term is used to damp large differences between the result and a bias distribution $\boldsymbol{x}_0$ often taken from simulation. The regularization term can be written as:

$$\mathcal{L}_2 = \tau^2(\boldsymbol{x} - f_b\boldsymbol{x}_\mathbf{0})^\mathrm{T}(\mathbf{L}^\mathrm{T}\mathbf{L})(\boldsymbol{x} - f_b\boldsymbol{x}_\mathbf{0}), \tag{5.4}$$

where the parameter τ determines the strength of the regularization, f_b is an optional bias scale factor used to scale the bias distribution $\boldsymbol{x}_0$, and the matrix $\mathbf{L}$ holds the regularization conditions.

The third component of the Lagrangian is an optional area constraint

$$\mathcal{L}_3 = \lambda(Y - \boldsymbol{e}^\mathrm{T}\boldsymbol{x}), \tag{5.5}$$

with the total number of events on detector level $Y = \sum_i y_i$ and the efficiency vector $e_j = \sum_i A_{ij}$. It is used to make sure that the resulting values in $\boldsymbol{x}$ corrected with the efficiencies in $\boldsymbol{e}$ match the total number of events on detector level Y.

The stationary point of the Lagrangian is found using the partial derivatives of $\mathcal{L}(\boldsymbol{x}, \lambda)$. More details can be found in reference [126].

Determination of the regularization strength

A careful choice of the regularization strength is very important. A regularization term always introduces a small bias towards the bias distribution that is often taken from simulation. If the regularization strength is too low it does not damp the non-physical statistical fluctuations and if the regularization strength is too high the result tends to follow the bias distribution and not the true distribution. Two methods to estimate the optimal regularization strength are implemented in the TUnfold framework and described below.

L-curve scan The L-curve scan defines two terms that should probe the influence of the first two terms of the Lagrangian:

$$L_x^\mathrm{curve} = \log \mathcal{L}_1 \quad \text{and} \quad L_y^\mathrm{curve} = \log \frac{\mathcal{L}_2}{\tau^2}. \tag{5.6}$$

The value L_x^{curve} will be small for low values of τ because in these cases the first term of the Lagrangian will dominate. For high values of τ the second term $\mathcal{L}_2$ will become dominant leading to low values for L_y^{curve} and high values for L_x^{curve}. The unfolding is repeated with different τ values and the two terms are scanned as a function of τ. They form an L-shaped curve in the L_x^{curve}-L_y^{curve} plane and the optimal τ value is found to be the point of largest curvature in the L_x^{curve}-L_y^{curve} plane.

Scan over global correlation coefficients A second approach is a minimization of global correlation coefficients defined as:

$$\rho_i = \sqrt{1 - \frac{1}{(\mathbf{V_{\boldsymbol{xx}}}^{-1})_{ii}(\mathbf{V}_{\boldsymbol{xx}})_{ii}}}, \tag{5.7}$$

where $\mathbf{V}_{\boldsymbol{xx}}$ is the covariance matrix of the output $\boldsymbol{x}$. The statistical fluctuations are expected to introduce negative correlations while the regularization is expected to introduce positive correlations. The point of minimal global correlation is therefore a good compromise.

Different ways to perform the scan are implemented minimizing the average or maximum global correlations and using statistical uncertainties only in the covariance matrix or using additionally systematic uncertainties.

Regularization conditions

The form of the $\mathbf{L}$ matrix determines how the regularization is applied. If a unity matrix is used for the $\mathbf{L}$ matrix the regularization reduces large differences between the bin contents of the unfolding output and the bias distribution. This is often called size regularization. TUnfold also supports forms of the $\mathbf{L}$ matrix to apply the regularization to the first or second derivative of the distribution. Large differences in the first or second derivative between the unfolding output and the bias distribution are reduced in these cases respectively.

5.2 Top tagging

Top quarks with high momentum have a large Lorentz boost leading to their decay products being collimated in the direction of flight of the top quark. At very high momentum the jets formed by the top quark decay products overlap and the reconstruction of a hadronic top quark decay in three separate jets becomes inefficient. In these cases it becomes more efficient to reconstruct the top quark decay in one large jet. Top tagging algorithms typically use substructure information of large jets to decide if the jet contains a fully-merged hadronic top quark decay or if the jet is induced by a light quark or gluon.

Top tagging is important for searches for heavy new particles decaying into top quarks with high momentum, where sensitivity would get lost by a reconstruction of the top quark in three separate jets. It is also important for some standard model measurements to enrich a selected phase space with boosted top quarks. The second analysis presented in this thesis consists of studies of the efficiency of top tagging algorithms in data and simulation, while the first analysis presents a measurement of an important substructure variables used in many top tagging algorithms, the jet mass for boosted top quarks.

5.2.1 Top tagging algorithms

Different kinds of top tagging algorithms have been developed during the last years. They have in common that they use information on the substructure of large jets in one or other way. A very common approach is to define a set of selection criteria on specially designed jet-substructure variables (see also section 2.3). Those taggers will be called cut-based taggers in the following. Examples of cut-based taggers used in CMS are among others the CMSTopTagger [127] based on the John Hopkins tagger [128], the CMSTopTagger v2 [5] that is described below, the HepTopTagger [129–131], and the Heavy Object Tagger with Variable R (HOTVR) [6], also described below. To make better use of correlations between different variables some taggers use multivariate analysis (MVA) methods to combine several input variables into one discriminator. Example for such MVA based taggers studied in CMS are the Boosted Event Shape Tagger (BEST) [132] and a tagger based on energy correlation functions described in reference [133]. Some recent approaches like DeepAK8 [134] or Lola [135] make extensive use of machine learning, using deep neural networks with low-level input information like particle four-vectors or jet images.

The second analysis presented in chapter 7 contains studies of the performance of two cut-based approaches in real data and in simulation. The respective taggers are briefly discussed below.

CMSTopTagger v2

The CMSTopTagger v2 is a cut-based approach that was first presented in reference [5]. It uses large anti-k_T jets with a radius of $R = 0.8$ and is studied for two different pileup-removal techniques, for CHS and PUPPI jets. The tagger is based the Soft Drop mass, the N-subjettiness ratio τ_3/τ_2 calculated without the Soft Drop grooming, and subjet b-tagging. A jet is tagged as a top jet if its Soft Drop mass lies within a mass window around the top quark mass and the τ_3/τ_2 value lies below a threshold that defines the working point of the tagger. A list of working points for CHS and PUPPI jets is given in table 5.1. All listed working points are studied with and without a requirement on at least one subjet b tag.

Table 5.1: Working points of the CMSTopTagger v2.

Jet collection	Soft Drop mass	τ_3/τ_2
PUPPI	$105 < m_{SD} < 210\,\text{GeV}$	0.4
		0.46
		0.54
		0.65
		0.8
CHS	$105 < m_{SD} < 220\,\text{GeV}$	0.5
		0.57
		0.67
		0.81

Heavy Object Tagger with Variable R (HOTVR)

HOTVR is a relatively new cut-based top tagging algorithm [6]. It is based on the specially developed HOTVR jet algorithm described in section 4.4. The variable distance parameter in the jet clustering allows the jets to become larger for lower momenta of the jets. This allows the HOTVR algorithm to reconstruct hadronic top quark decays with lower momentum compared to the CMSTopTagger v2, that uses a fixed jet radius of 0.8, while having a similar performance for high-momentum jets. The jet grooming of the HOTVR jet clustering should avoid a strong dependence on additional soft radiation, especially for the larger low-momentum jets which are otherwise strongly affected. The variable jet radius together with the integrated grooming leads to a tagger stable over a large p_T range down to top quark momenta of $p_T > 200\,\text{GeV}$.

HOTVR jets are tagged as top jets if they fulfill the set of requirements listed below.

- The invariant mass of the jet has to be within a window around the top quark mass of $140 < m_{\mathrm{jet}} < 220\,\mathrm{GeV}$.
- The leading subjet should carry less than 80% of the full jet p_{T} ($f_{p_{\mathrm{T}}} = p_{\mathrm{T},1}/p_{\mathrm{T,jet}} < 0.8$). For light-quark or gluon jets the leading subjet is expected to carry most of the jet momentum.
- The number of subjets is expected to be greater or equal to three ($N_{\mathrm{jets}} \geq 3$) because of the three quarks from the top quark decay.
- The minimum pairwise mass of the three leading subjets m_{min} is in the case of hadronic top quark jets expected to be close to the W boson mass in most of the cases and can be much lower for light-quark jets. This leads to the requirement of $m_{\mathrm{min}} < 50\,\mathrm{GeV}$.

6 Measurement of the jet-mass distribution

6.1 Introduction

This chapter describes the measurement of the differential $\mathrm{t\bar{t}}$ production cross section as a function of jet mass of the jet with the highest transverse momentum in the event (leading jet). The measurement presented in this chapter provides the first measurement of a jet-mass distribution for highly boosted top quarks at the particle level. It is important to test the modeling of different MC event generators and the underlying physics. The peak region of the jet-mass distribution is further sensitive to the mass of the top quark which is a fundamental parameter of the standard model (see also section 2.2.3).

The measured jet-mass distribution in this chapter is used to verify different MC event generators and two different parton-shower models in PYTHIA and HERWIG in the boosted top quark regime. A top quark mass is extracted from data and MC-generated distributions and the compatibility with the traditional direct measurements in resolved top quark decays is tested. An extraction of a well-defined top quark mass using analytic calculations, as proposed in [40–42, 50] (see also sections 2.2 and 2.3), is not yet possible because no analytic calculations exist yet for the measured phase space. The data at the particle level is however published for future studies with analytic calculations or MC generators.

The analysis in this chapter starts with a definition of the measurement phase space at the particle level that is enriched with events in which the leading jet contains a fully-merged hadronic top quark decay. The jets are chosen large enough to contain top quark decays down to a top quark p_{T} of $\sim 400\,\mathrm{GeV}$. A similar phase space is defined at the detector level and the data is corrected for detector effects to the particle level using a regularized unfolding approach. The result is a differential and a normalized differential $\mathrm{t\bar{t}}$ production cross section as a function of the mass of the leading jet. The ultimate goal of this measurement would be to compare the mass distribution at the particle level to the analytic calculations mentioned above and to extract a well-defined top quark mass from the data.

Because no analytic calculations for the analyzed phase space exist yet only a comparison to different MC generators is possible at the moment. A value of the top quark mass is extracted using different MC-generated templates to test the sensitivity that can be achieved with the 8 TeV data.

Pioneering work for the analysis presented in this chapter was performed in the master thesis in reference [4]. There, a first simple unfolding of the jet-mass distribution was studied. Within the scope of this thesis a proper measurement of the differential cross section was performed with a more complex unfolding setup and a careful study of all relevant systematic uncertainties. A value of the top quark mass is estimated from a comparison of data to predictions from MC event generators. The results of this chapter have been published by the CMS collaboration in reference [3].

6.2 Data and simulation

Data

The data used for this measurement was collected by the CMS detector in the year 2012 in proton-proton collisions at a center-of-mass energy of 8 TeV. Only certified runs are used for this measurement corresponding to an integrated luminosity of 19.7 fb^{-1}.

Simulation

The default $\mathrm{t\bar{t}}$ sample used for this measurement is obtained using POWHEG 1.380 [97–101] for the calculation of the hard matrix element interfaced with PYTHIA v6.424 [104] for the simulation of the parton shower and hadronization. A value of the top quark mass of 172.5 GeV was used in the production of this sample. Two exclusive samples for high invariant masses of the $\mathrm{t\bar{t}}$ system $m_{\mathrm{t\bar{t}}}$ are used in addition to the inclusive sample to increase the number of simulated events for high values of $m_{\mathrm{t\bar{t}}}$ and hence for high top quark p_{T}. These samples are produced for $700 < m_{\mathrm{t\bar{t}}} < 1000$ GeV and $m_{\mathrm{t\bar{t}}} > 1000$ GeV.

Additional $\mathrm{t\bar{t}}$ samples are produced with MadGraph 5.1.5.11 [103] interfaced with pythia. The Madspin [136] package is used for the decay of the heavy resonances. The MadGraph +pythia samples are produced for seven different values of the top quark mass of 166.5 GeV, 169.5 GeV, 171.5 GeV, 172.5 GeV, 173.5 GeV, 175.5 GeV, and 178.5 GeV. Systematic effects on the parton-shower model are studied with a $\mathrm{t\bar{t}}$ sample obtained with mc@nlo v3.41 [102] interfaced with herwig 6.520 [105]. Samples of $\mathrm{t\bar{t}}$ production with renormalization a factorization scales μ_r and μ_f varied fully correlated by factors of 0.5 and 2 are produced with powheg +pythia for $m_{\mathrm{t\bar{t}}} > 700\,\mathrm{GeV}$.

The production of W bosons in association with jets is simulated with MadGraph +pythia. Single top quark production is obtained with powheg +pythia. Electron and muon-enriched samples of QCD multijet production are simulated with pythia.

The MLM algorithm [137] is used for all MadGraph +pythia $\mathrm{t\bar{t}}$ samples to match the hard matrix element objects to the parton shower. All MadGraph samples are produced with the CTEQ6L PDF set [138], $\mathrm{t\bar{t}}$ samples simulated with powheg use the CT10 [139] PDF set, and the powheg samples for single top quark production use the CTEQ6M [140] PDF set. The Z2* [141, 142] underlying-event tune is used for all samples simulated with pythia.

A full detector simulation is applied for all MC samples within the CMS software with releases CMSSW_5_3_X using the Geant4 v9.2 [106] framework for interactions of particles with the detector material.

6.3 Object definitions

This section includes a detailed description of all object definitions which are used for this measurement. The objects are defined similarly but not identically at the particle and the detector level. The particle-level objects will be discussed first, followed by the detector-level objects.

6.3.1 Particle-level objects

The objects at the particle level are defined in a way that a comparison to analytic calculations and MC event generators at the particle level is possible. Soft effects like hadronization effects and color reconnections are consistently included in the analytic calculations and need to be taken into account when extracting the top quark mass. A strong dependence of the measurement phase space on specific parton-shower models should be avoided. More complex concepts like missing transverse momentum $p_{\mathrm{T}}^{\mathrm{miss}}$ or the identification of jets originating from the hadronization of b quarks could introduce a dependence on the parton shower and are therefore not considered at the particle level.

Leptons

The leptons used in this measurement are electrons and muons from prompt W boson decays. Electrons and muons from τ decays are not considered. Leptons are only used for $p_{\mathrm{T}} > 45\,\mathrm{GeV}$ and $|\eta| < 2.1$ to match the trigger acceptance at the detector level.

Jets

The jets at the particle level are clustered with the Cambridge/Aachen (CA) jet algorithm with a distance parameter of $R = 1.2$. Additional studies have been performed with CA jets with distance parameters of 0.8 and 1.5. All jets are clustered from all stable particles except for neutrinos.

Four-momenta of leptons are subtracted from the jets to avoid a double counting of the lepton energy and to match the definitions at the detector level. The four-momentum of a lepton is subtracted from the four-momentum of a jet if the distance in the $\eta - \phi$

plane between the lepton and the jet is smaller than the distance parameter of the jet $\Delta R(\text{lepton, jet}) < R$.

Jets are only considered for an η range below $|\eta| < 2.5$ to match the acceptance at the detector level.

6.3.2 Detector-level objects

The detector-level object definitions are more complex compared to the particle-level objects because they are based on reconstructed detector information and have to be accessible in real data. Several corrections have to be applied to the simulation at detector level to correct for differences between data and simulation concerning identification efficiencies and energy measurements. The following subsections include a brief description of the definitions used for the different physical objects used in this analysis. These definitions are based on the event reconstruction described in chapter 4.

Muons

The muon candidates used in this measurement have to fulfill the following quality criteria connected to the TightID working point [143].

- The muon candidate has to be reconstructed as a global muon and as a tracker muon,
- the χ^2 of the fitted track normalized to the number of degrees of freedom has to be smaller than 10,
- there must be at least one hit in at least two muon stations,
- the transverse impact parameter $|d_{xy}|$ with respect to the primary event vertex has to be smaller than 2 mm,
- the longitudinal distance to the primary vertex d_z has to be smaller than 5 mm,
- and at least one hit in the pixel detector and at least five hits in the tracker are required.

Muon candidates in this measurement are only considered with $p_{\mathrm{T}} > 45\,\mathrm{GeV}$ and $|\eta| < 2.1$. No isolation criteria are applied to muon candidates.

Scale factors to correct for differences in the identification efficiency between data and simulation are derived centrally in CMS [143] and applied in this analysis.

Electrons

The electron candidates used in this measurement are identified with a multivariate analysis (MVA) approach using several variables from the tracks and calorimeter clusters [144]. They have to fulfill the following identification criteria.

- The candidates have to pass a conversion veto.
- A hit in each layer of the tracker is required.
- The value of the MVA discriminator has to be above a certain threshold depending on the η of the super cluster η_{SC}:
  - MVAdiscr. > 0.94 for $|\eta_{SC}| < 0.8$,
  - MVAdiscr. > 0.85 for $0.8 < |\eta_{SC}| < 1.479$,
  - MVAdiscr. > 0.92 for $1.479 < |\eta_{SC}| < 2.5$.

Electron candidates for this measurement are only considered with $p_{\mathrm{T}} > 45\,\mathrm{GeV}$ and $|\eta| < 2.1$. No isolation criterion is used to select electron candidates.

The efficiency of the electron identification in data and simulation is studied centrally in CMS [144] and data-to-simulation scale factors are applied in this analysis to correct for differences between data and simulation.

Jets

Two different jet collections are used for this measurement. Anti-k_{T} jets with a distance parameter of $R = 0.5$ (called AK5 jets in the following) are used for the selection of potentially boosted $\mathrm{t\bar{t}}$ decays and background suppression. Cambridge/Aachen jets with a distance parameter of $R = 1.2$ (called CA12 jets in the following) are used for definition of the measurement phase space and the definition of the jet mass.

All jets are clustered from all PF candidates after charged hadron subtraction (CHS). Isolated leptons are not considered in the clustering. Loose identification criteria are applied to all jets. Jet energy corrections (JECs) [145] are applied as p_{T} and η-dependent factors to the four-momenta of the jets to correct for differences in the jet energy scale between particle and detector level and to correct for residual differences between data and simulation at the detector level. The JECs that are applied to the AK5 jets have been derived within the CMS collaboration for AK5 jets with CHS applied. Corrections derived for anti-k_T jets with $R = 0.7$ (AK7) are used to correct the CA12 jets since no JECs have been evaluated for CA12 jets in CMS and the AK7 jets are the closest to the CA12 jets with available corrections. The uncertainty on the corrections applied to the CA12 jets is studied in appendix A.4 and increased to cover the differences between AK7 and CA12 jets. A jet energy resolution (JER) smearing [145] is applied to all jets to account for a worse JER in data compared to simulation.

The four-momenta of leptons are subtracted from the four-momenta of jets in both jet collections to avoid a double counting of lepton energies and to make sure that the JECs are applied correctly. The JECs are undone before the cleaning and reapplied on the corrected four-momentum afterwards. The four-momenta of non-isolated electron or muon candidates are subtracted from the AK5 jet four-momentum if the distance between the jet and the lepton is smaller $\Delta R(\text{jet,lepton}) < 0.5$. For the CA12 jets a list with all clustered PF candidates is stored and the lepton four-vector is subtracted if the lepton can be found in the list of clustered PF candidates.

The combined secondary vertex (CSV) algorithm is used to identify AK5 jets originating from decays of b mesons. The tight working point (CSVT) is used to identify b jets in this measurement.

Missing transverse momentum

Missing transverse momentum $\vec{p}_\mathrm{T}^{\,\mathrm{miss}}$ is defined as the negative vectorial sum of all PF particles in the event. This measurement uses a Type-1 correction [146] on $\vec{p}_\mathrm{T}^{\,\mathrm{miss}}$ using fully-corrected anti-k_T jets with $R = 0.5$. The corrected value is obtained by

$$\vec{p}_\mathrm{T}^{\,\text{miss, corr}} = \vec{p}_\mathrm{T}^{\,\mathrm{miss}} - \sum_{\text{jets}} \left(\vec{p}_{\text{T, jet}}^{\,\text{corr}} - \vec{p}_{\text{T, jet}} \right), \tag{6.1}$$

where the values with the superscript “corr” are corrected values.

6.4 Measurement phase space definition

The definition of the measurement phase space at the particle level is a crucial step for this analysis. It should be a good compromise between the ability to perform analytic calculations and to perform the experimental measurement at the detector level. The measurement is performed in the lepton+jets $\mathrm{t\bar{t}}$ decay channel containing one electron or muon from the leptonic top quark decay to ensure a good suppression of background processes at the detector level. The leptons are considered for $p_{\mathrm{T}} > 45\,\mathrm{GeV}$ and $|\eta| < 2.1$ to match the trigger acceptance at the detector level.

At the time this measurement was performed, calculations of the jet-mass distribution for boosted top quarks have only been performed in $\mathrm{e^+e^-}$ collisions as a double-differential cross section $\mathrm{d}^2\sigma/(\mathrm{d}m_{\mathrm{jet},1}\mathrm{d}m_{\mathrm{jet},2})$ [40–42, 47, 147] using a hemisphere mass. Recently also first calculations of the differential cross section as a function of the boosted top-jet mass in proton-proton collisions have been published in reference [50]. The calculations are performed in the framework of the soft-collinear effective theory (SCET) [43–46] leading to an expansion in lambda $m/Q \sim m/(2p_{\mathrm{T}}) \sim 0.2$, where Q is the momentum transfer. Therefore large values of Q and hence of the p_{T} of the produced top quarks are needed for the analytic calculations in SCET. Another requirement on a high p_{T} of the top quarks is set by the requirement of the analytic calculations that all decay products of the top quark have to be contained within the jet. The spatial separation of the top quark decay products depends on the top quark momentum. The higher the top quark momentum the higher is its Lorentz boost and the more collimated are its decay products in the direction of flight on the top quark. A high top quark p_{T} much larger than the top quark mass is therefore preferred by the analytic calculations. The experimental measurement, on the other side, needs enough events in data after the full event selection at the detector level for a stable unfolding leading to a preference for lower values of the top quark p_{T}.

A good compromise on the jet p_{T} between the needs of analytic calculations and experimental measurement is evaluated by studying the mass of the jet with the highest p_{T} in the event ('leading jet') for different p_{T} thresholds in events containing a lepton. The leading jet is supposed to contain the hadronic top quark decay. Figure 6.1 shows the distribution of the invariant mass of the leading jet for $p_{\mathrm{T}} > 300\,\mathrm{GeV}$, $p_{\mathrm{T}} > 400\,\mathrm{GeV}$, and $p_{\mathrm{T}} > 500\,\mathrm{GeV}$. A second jet is required with $p_{\mathrm{T}} > 100\,\mathrm{GeV}$ because of the presence of a b quark from the leptonic top quark decay. The jets are clustered with the CA algorithm with a distance parameter of 1.2. The figure shows only the default $\mathrm{t\bar{t}}$ simulation with POWHEG +PYTHIA normalized to an integrated luminosity of $19.7\,\mathrm{fb}^{-1}$. All distributions in figure 6.1 show a peak around 190 GeV connected to events in which the full top quark

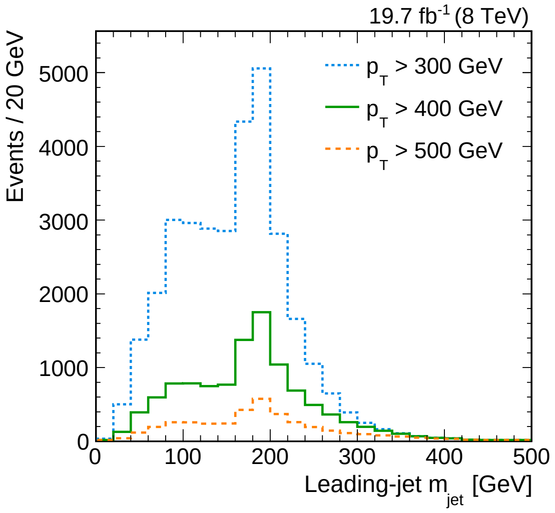

Figure 6.1: Jet-mass distributions of the leading jet in lepton+jets $t\bar{t}$ decays simulated with POWHEG +PYTHIA and normalized to an integrated luminosity of 19.7 fb^{-1}. The distribution of the leading-jet mass is shown for different p_T thresholds. A second jet with $p_T > 100$ GeV is required. The jets are clustered with the Cambridge/Aachen algorithm with a distance parameter of $R = 1.2$

decay is clustered into the leading jet. The peak position is shifted to values higher than the top quark mass by additional soft radiation from effects like initial-state radiation, final-state radiation, or the underlying event. Jets showing lower masses do not contain a full top quark decay. It is possible that only the two quarks from the hadronic W boson decay are clustered into the leading jet leading to jet masses connected to the W boson mass of 80.38 GeV [2]. It is also possible that the leading jet is formed by just one light-quark jet from the W decay, from one of the b quarks or even from additional radiation leading to low masses. Of high importance for this measurement is the amount of events expected within the top quark mass peak in data after the full selection at the detector level. The expected number of events in data for a p_T of the leading jet larger than 500 GeV is not enough assuming a reconstruction efficiency of about 10-20% (shown at a later stage of this measurement). A leading-jet p_T larger than 400 GeV was found to be a good compromise between enough events expected in data for the measurement and high top quark p_T for the analytic calculations.

The next important step is a study of the exact jet definition to be used for this measurement. The Cambridge/Aachen algorithm is used to cluster the jets because it is commonly used for applications of jet substructure at 8 TeV like top tagging. Its distance measure used in the jet clustering is purely geometrical. The distance parameter of the jet algorithm has to be chosen large enough to cover the full top quark decay down to the p_T threshold of 400 GeV and at the same time not too large to avoid large dependencies on additional soft radiation. Additional radiation from soft effects like pileup, underly-

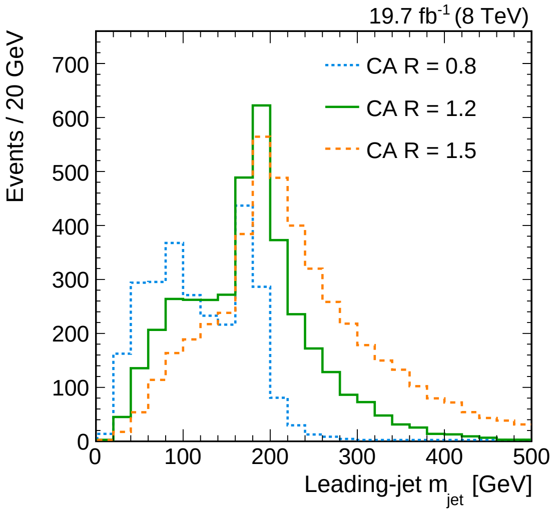

Figure 6.2: Jet-mass distributions of the leading jet in electron+jets $\mathrm{t\bar{t}}$ decays simulated with POWHEG +PYTHIA and normalized to an integrated luminosity of 19.7 fb^{-1}. The distributions of the leading-jet mass are shown for jets reconstructed with the Cambridge/Aachen algorithm and different distance parameters. Events shown in this figure are required to contain a leading jet with $p_{\mathrm{T}} > 400$ GeV, a second jet with $p_{\mathrm{T}} > 150$ GeV, and a veto on additional jets with $p_{\mathrm{T}} > 150$ GeV is applied.

ing event, or initial-state radiation shift the jet mass to higher values and broaden the jet-mass distribution. For the study of different jet distance parameters one electron or muon is required together with a leading jet with $p_{\mathrm{T}} > 400$ GeV and a second jet with $p_{\mathrm{T}} > 150$ GeV. A veto is set on additional jets with $p_{\mathrm{T}} > 150$ GeV. The veto is needed by the analytic calculations and should be chosen as hard as possible because it leads to non-global logarithms of the order of $\log(p_{\mathrm{T,veto}}/p_{\mathrm{T,jet}})$ [148] and should therefore fulfill the criterion $p_{\mathrm{T,veto}}/p_{\mathrm{T,jet}} < 1$. At the same time a hard jet veto again reduces the amount of events in data available for the measurement. A veto on additional jets with $p_{\mathrm{T}} > 150$ GeV was chosen as a compromise leading to $p_{\mathrm{T,veto}}/p_{\mathrm{T,jet}} = 150\,\mathrm{GeV}/400\,\mathrm{GeV} = 0.375$. The veto has no significant effect on the shape of the jet-mass distribution. Figure 6.2 shows the jet-mass distribution of the leading jet for different distance parameters used in the jet clustering of $R = 0.8$, 1.2, and 1.5. The distribution for jets with $R = 0.8$ shows beside the top quark mass peak a second peak around 90 GeV containing semi-merged events with just the hadronically decaying W boson reconstructed in the leading jet and not the full top quark decay. It has in general the largest fraction of events reconstructed at low masses originating from semi-merged, light-quark, or gluon jets. These events are not covered by the analytic calculations and a distance parameter of 0.8 is shown to be too small to reconstruct the full top quark decay for the chosen p_{T} threshold. The distribution for jets with $R = 1.5$ shows the highest contribution of fully-merged top quarks but at the same time a worse mass resolution compared to the other distributions leading to a large tail to higher masses. A jet distance parameter of $R = 1.2$ was found to be a good compromise between a high number of events in the peak and a relatively good jet-mass resolution.

The basic selection after the studies above is based on the following criteria:

- The event is required to contain one e/μ from the leptonic top quark decay,
- a leading jet with $p_{\mathrm{T}} > 400\,\mathrm{GeV}$,
- a second jet with $p_{\mathrm{T}} > 150\,\mathrm{GeV}$,
- and a veto is set on additional jets with $p_{\mathrm{T}} > 150\,\mathrm{GeV}$.

The distribution of the leading-jet mass after this selection is shown in figure 6.3 on the top left. The full selection is shown in black. For illustration reasons the $\mathrm{t\bar{t}}$ sample is divided into a fully-merged and a not-merged contribution by matching particles from the MC generator to the jet. The leading jet is called fully-merged if all three quarks from the hadronic top quark decay have a distance to the jet smaller than the jet distance parameter $\Delta R(\text{leading jet}, \mathrm{q_i}) < 1.2$, otherwise the jet is called not-merged. Both contributions are shown together with the inclusive distribution. After this selection a significant amount of events are not matched originating from jets that do not include the full top quark decay.

Two additional selection criteria are introduced to further enrich the phase space with fully-merged hadronic top quark decays. First the distance in ΔR between the lepton and the second jet is required to be smaller than the jet distance parameter, $\Delta R(\text{jet 2, lepton}) < 1.2$, to make sure that the leptonically decaying top quark is boosted. Secondly the mass of the leading jet is required to be larger than the invariant mass of the combined four-vectors of the second jet and the lepton ($m_{\text{leading jet}} > m_{\text{jet 2 + lepton}}$). In correctly reconstructed events the leading jet contains a full top quark decay while the combination of the second jet and the lepton should only contain the b quark and the lepton from the leptonic top quark decay and not the neutrino. The combination of the second jet and the lepton should therefore have a lower invariant mass than the leading jet. The distributions of the leading-jet mass before and after these two selection steps are shown in figure 6.3. It can be seen that the two additional selection criteria help to enrich the measurement phase space with jets that contain a hadronic top quark decay within the leading jet. The full selection shown at the bottom of figure 6.3 is used for the measurement performed in this thesis and was published in reference [3].

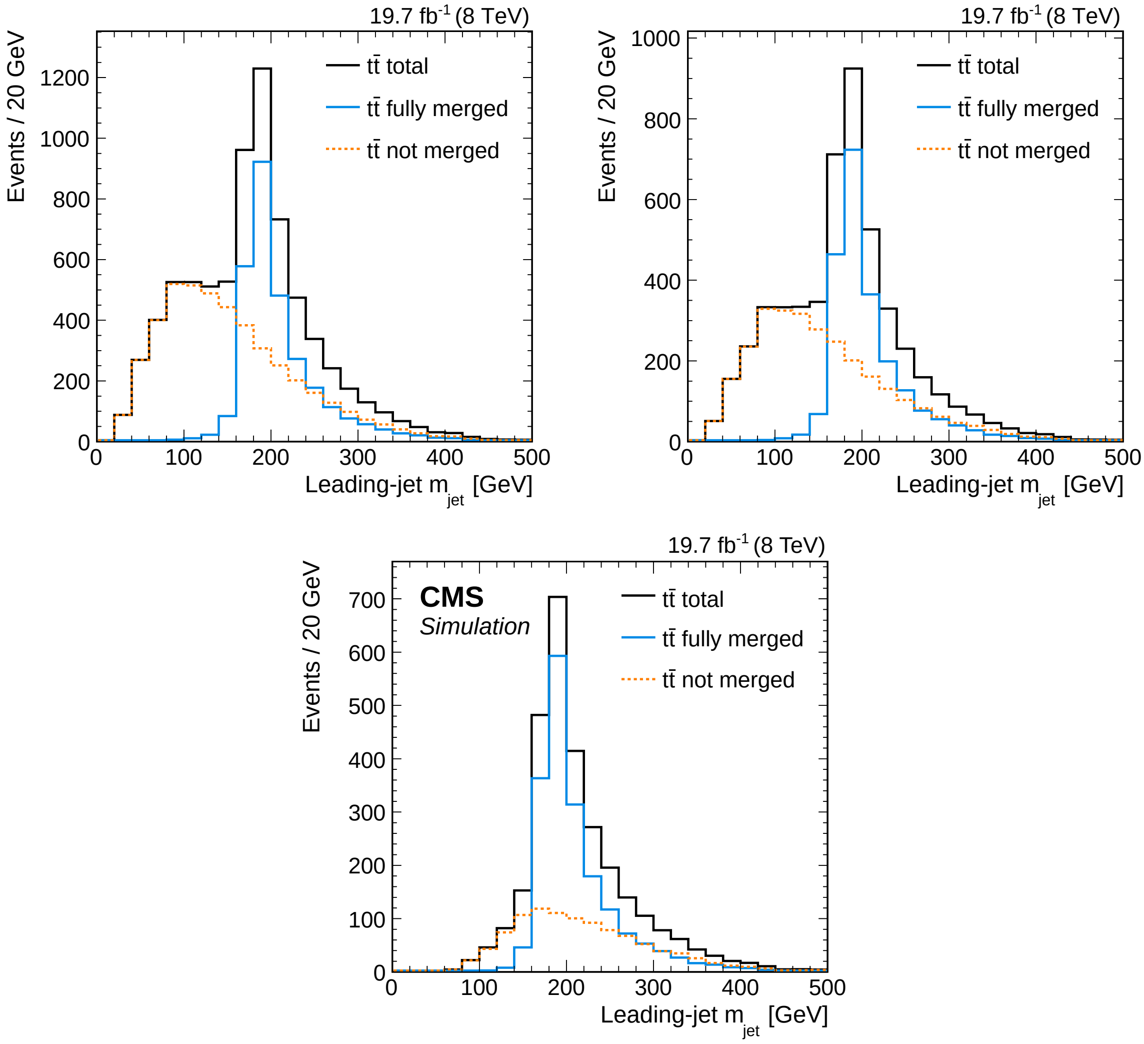

Figure 6.3: Jet-mass distributions of the leading jet in lepton+jets $t\bar{t}$ decays simulated with POWHEG +PYTHIA and normalized to an integrated luminosity of $19.7\,\mathrm{fb}^{-1}$. The distributions of the leading-jet mass are shown at different steps of the selection. Each figure shows the inclusive distribution together with a fully-merged and a not-merged contribution obtained by matching the decay products of the hadronic top quark decay on generator level to the leading jet. The figure on the top left shows the distribution for the baseline selection. The figure on the top right further includes a selection on $\Delta R(\text{jet 2, lepton}) < 1.2$ and the figure on the bottom shows the full selection including also the selection on $m_{\text{leading jet}} > m_{\text{jet 2 + lepton}}$. The figure for the full selection is published in reference [3].

6.5 Monte Carlo generators

The measurement should be as independent as possible of the simulation model that is used for the unfolding of the data to the particle level. The jet-mass distribution is therefore studied in this section at the particle level for different simulation setups. Figure 6.4 shows the distribution of the leading-jet mass after the selection developed in section 6.4 for events simulated with POWHEG +PYTHIA, with MadGraph +PYTHIA, and MC@NLO +HERWIG. All distributions are normalized to an integrated luminosity of $19.7\,\mathrm{fb}^{-1}$. About 20% more events are observed with MadGraph +PYTHIA compared to the other simulations. This can be explained by two effects shown in figure 6.5. On the left the p_T distribution of the hadronic top quark from the MC generator is shown after selecting events with a leading jet with $p_\mathrm{T} > 300\,\mathrm{GeV}$ and a second jet with $p_\mathrm{T} > 100\,\mathrm{GeV}$. The top quark p_T spectrum is harder in MadGraph +PYTHIA compared to POWHEG +PYTHIA and MC@NLO +HERWIG, leading to more events with high top quark p_T and therefore high jet p_T. The figure on the right shows the p_T distribution of the third jet in the event after a selection of events with a leading jet with $p_\mathrm{T} > 400\,\mathrm{GeV}$ and a second jet with $p_\mathrm{T} > 150\,\mathrm{GeV}$. The third-jet p_T is softer in MadGraph +PYTHIA leading to less events rejected by the veto and therefore more events in the measurement phase space. MC@NLO +HERWIG shows a similar cross section after the full selection but also a softer p_T spectrum of the third jet.

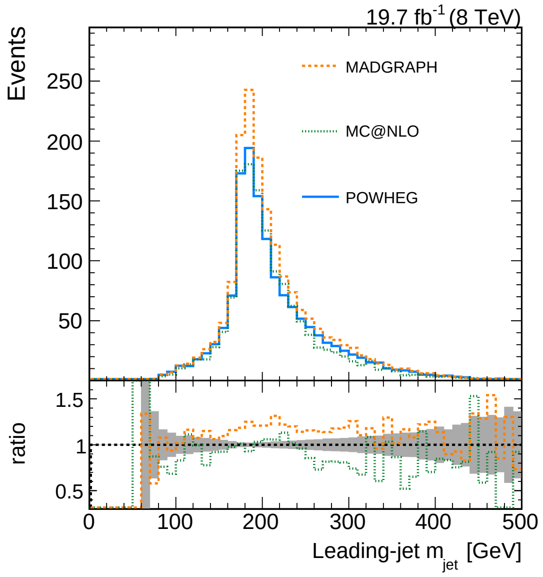

Figure 6.4: Distribution of the leading-jet mass for the full measurement phase space at the particle level obtained with different MC generators. The distributions are simulated with POWHEG +PYTHIA, MadGraph +PYTHIA, and MC@NLO +HERWIG. All distributions are normalized to an integrated luminosity of 19.7 fb^{-1}.

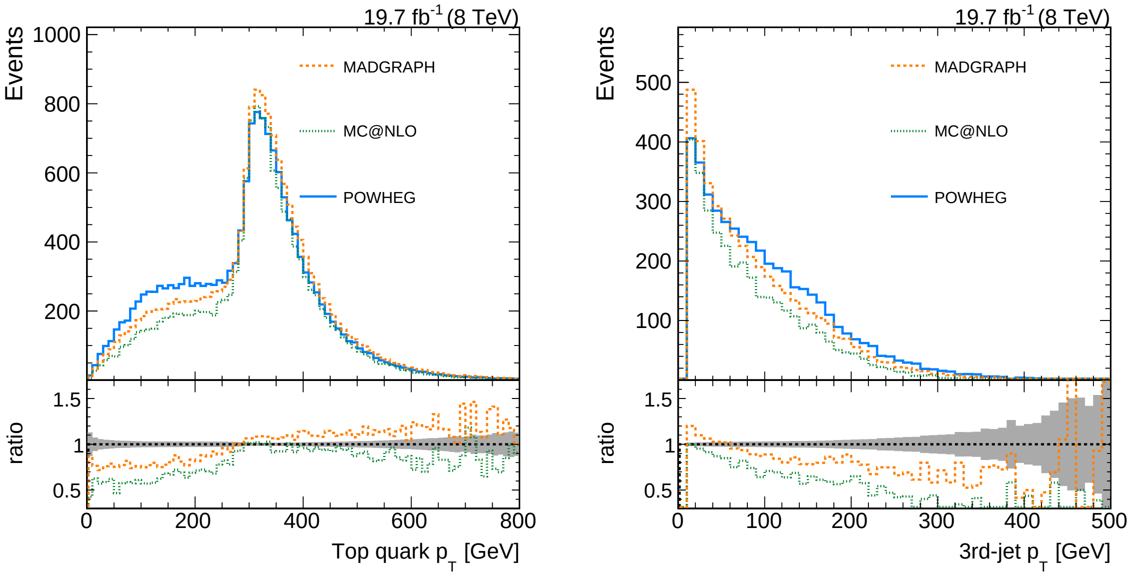

Figure 6.5: The p_{T} distribution of the hadronically decaying top quark at the generator level is shown on the left in lepton+jets $\mathrm{t\bar{t}}$ decays with one e/μ, a leading jet with $p_{\mathrm{T}} > 300\,\mathrm{GeV}$ and a second jet with $p_{\mathrm{T}} > 100\,\mathrm{GeV}$. The figure on the right shows the p_{T} distribution of the third jet for a events with a leading jet with $p_{\mathrm{T}} > 400\,\mathrm{GeV}$ and a second jet with $p_{\mathrm{T}} > 150\,\mathrm{GeV}$. Both figures show distributions obtained with POWHEG +PYTHIA, MadGraph +PYTHIA, and MC@NLO +HERWIG. All distributions are normalized to an integrated luminosity of 19.7 fb^{-1}.

6.6 Event selection at the detector level

The event selection at the detector level has two purposes. On the one hand it should suppress the dominant background processes like W boson production in association with jets and QCD multijet production. On the other hand it should select a similar phase space as at the particle level. It is therefore divided into two steps.

The first selection step aims on a selection of a pure sample of potentially boosted $\mathrm{t\bar{t}}$ events in the e/μ+jets decay channel. This $\mathrm{t\bar{t}}$ selection was developed within the scope of a $\mathrm{t\bar{t}}$-resonance search in reference [149] and was slightly adjusted for this analysis. It poses only soft selection criteria on the jet p_{T} and should have a minor effect on the shape of the CA12 jet-mass distribution.

In case of the muon channel events are triggered by a single-muon trigger requiring one muon with $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ and $|\eta| < 2.1$. In the case of the electron channel a logical "or" of two triggers is used. One trigger requires an electron with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$ together with two AK5 jets with $p_{\mathrm{T}} > 100\,\mathrm{GeV}$ for the leading jet and $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ for the sub-leading jet. This trigger is used together with a single-jet trigger requiring an AK5 jet with $p_{\mathrm{T}} > 320\,\mathrm{GeV}$.

The efficiency for the single-muon trigger was measured centrally in CMS in a $\mathrm{Z} \to \mu^+\mu^-$ sample for muons with $p_{\mathrm{T}} > 45\,\mathrm{GeV}$. The efficiency is 95% for $|\eta| < 0.9$, 85% for $0.9 < |\eta| < 1.2$, and 83% for $1.2 < |\eta| < 2.1$. Differences in the efficiency are corrected for by applying corresponding scale factors. The efficiency of the combined electron trigger was studied within the scope of the $\mathrm{t\bar{t}}$-resonance search in reference [149] in a $\mathrm{Z}/\gamma^* \to ll + \mathrm{jets}$ sample to be 90% for a leading-jet $p_{\mathrm{T}} < 320\,\mathrm{GeV}$ and fully efficient above a leading-jet $p_{\mathrm{T}} > 350\,\mathrm{GeV}$. Scale factors are applied to simulation to correct for differences between data and simulation.

Events are selected if they contain exactly one electron or muon candidate with $p_{\mathrm{T}} > 45\,\mathrm{GeV}$ and $|\eta| < 2.1$. A veto on additional muon or electron candidates is set to avoid overlap between the two channels and to suppress dileptonic $\mathrm{t\bar{t}}$ decays. One AK5 jet with $p_{\mathrm{T}} > 150\,\mathrm{GeV}$ and $|\eta| < 2.4$ and another one with $p_{\mathrm{T}} > 50\,\mathrm{GeV}$ and $|\eta| < 2.4$ are required to select potentially boosted $\mathrm{t\bar{t}}$ decays. At least one AK5 jet in the event has to be b-tagged because two b quarks are expected from the top quark decays. The b tag reduces the contributions from W+jets and QCD multijet production. A missing transverse momentum of $p_{\mathrm{T}}^{\mathrm{miss}} > 20\,\mathrm{GeV}$ is required because of the presence of a neutrino from the leptonic top quark decay. Furthermore, a high value of $H_{\mathrm{T}}^{\mathrm{lep}} > 150\,\mathrm{GeV}$ is required, where

$H_{\mathrm{T}}^{\mathrm{lep}}$ is the scalar sum of the lepton p_{T} and $p_{\mathrm{T}}^{\mathrm{miss}}$. Both requirements on $p_{\mathrm{T}}^{\mathrm{miss}}$ and $H_{\mathrm{T}}^{\mathrm{lep}}$ are used to suppress events from QCD multijet production. A two-dimensional lepton isolation is used to further suppress QCD multijet events by the requirement of $\Delta R_{\mathrm{min}} > 0.5$ in a logical "or" with $p_{\mathrm{T,rel}} > 25\,\mathrm{GeV}$, where ΔR_{min} is the distance between the lepton and the closest AK5 jet and $p_{\mathrm{T,rel}}$ is the perpendicular component of the lepton momentum with respect to the closest AK5 jet. All AK5 jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$ are considered for this two-dimensional isolation. The efficiency for this 2D isolation has been studied in data and simulation within the scope of the $\mathrm{t\bar{t}}$-resonance search [149] in a $\mathrm{Z} \to ll$ sample. A good agreement between data and simulation was observed and no corrections on simulation are needed. Only in the electron channel an additional triangular selection criterion is used to make sure that $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ points along the transverse direction with respect to the lepton or the leading jet. It is used to further reduce the QCD multijet background and requires $-\frac{1.5}{75\,\mathrm{GeV}} p_{\mathrm{T}}^{\mathrm{miss}} + 1.5 < \Delta\phi(\mathrm{lepton}, p_{\mathrm{T}}^{\mathrm{miss}}) < \frac{1.5}{75\,\mathrm{GeV}} p_{\mathrm{T}}^{\mathrm{miss}} + 1.5$ and $-\frac{1.5}{75\,\mathrm{GeV}} p_{\mathrm{T}}^{\mathrm{miss}} + 1.5 < \Delta\phi(\mathrm{leading\ jet}, p_{\mathrm{T}}^{\mathrm{miss}}) < \frac{1.5}{75\,\mathrm{GeV}} p_{\mathrm{T}}^{\mathrm{miss}} + 1.5$.

On top of the $\mathrm{t\bar{t}}$ selection a measurement phase space selection is applied similar to the definition at the particle level developed in section 6.4. A leading CA12 jet with $p_{\mathrm{T}} > 400\,\mathrm{GeV}$ and $|\eta| < 2.5$ is required together with a second CA12 jet with $p_{\mathrm{T}} > 150\,\mathrm{GeV}$ and $|\eta| < 2.5$. A veto is set on additional CA12 jets with $p_{\mathrm{T}} > 150\,\mathrm{GeV}$ and $|\eta| < 2.5$. The distance between the sub-leading CA12 jet and the lepton should be smaller than 1.2 and the mass of the leading jet is required to be higher than the mass of the sub-leading jet. This selection is similar to the particle-level selection, except for a slightly softer jet-mass selection criterion using only the second-jet mass instead of the combination of second jet and lepton.

Figure 6.6 shows the p_{T} and η distributions of the leading CA12 jet after the full selection in the combination of both channels with full systematic uncertainties. All simulations except for the $\mathrm{t\bar{t}}$ simulation are normalized to an integrated luminosity of $19.7\,\mathrm{fb}^{-1}$. The $\mathrm{t\bar{t}}$ simulation is further scaled to match the number of events in data to allow for a better comparison. This is needed because the top quark p_{T} spectrum is expected to be softer in data compared to simulation leading to less events in data for a high top quark p_{T}. This behavior was measured in references [150–154] for different ranges in the top quark p_{T}. The additional scaling for the $\mathrm{t\bar{t}}$ simulation is only used for the figures. The following unfolding does not depend on the normalization of the $\mathrm{t\bar{t}}$ simulation. The jet-mass distribution is shown in figure 6.7 for a leading jet with $400 < p_{\mathrm{T}} < 500\,\mathrm{GeV}$ on the left and with $p_{\mathrm{T}} > 500\,\mathrm{GeV}$ on the right. Data and simulation agree within uncertainties.

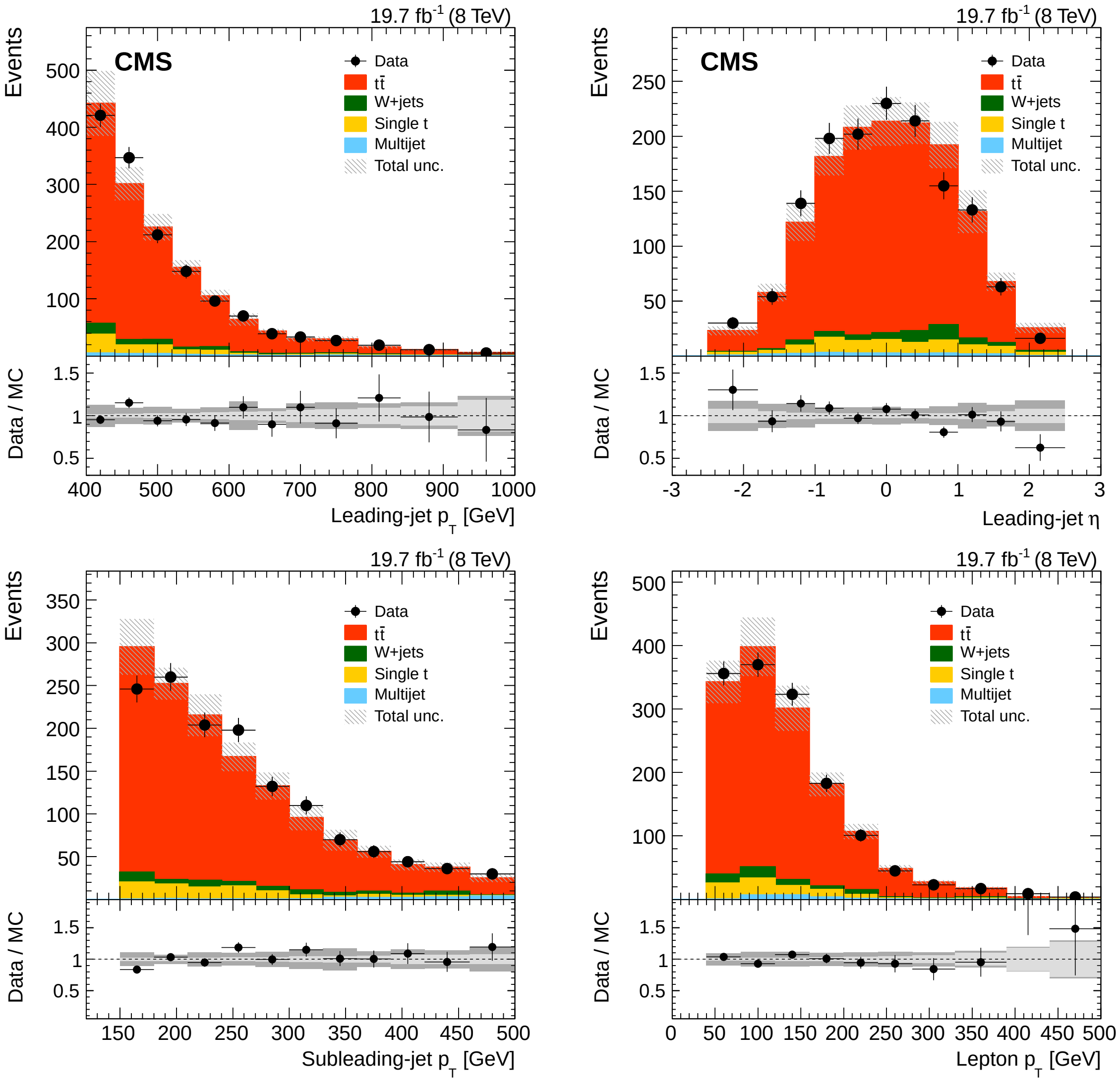

Figure 6.6: Distributions of the leading-jet p_T (top left) and η (top right), of the subleading jet p_T (bottom left), and on the lepton p_T (bottom right). All distributions show the combination of the electron and muon channels at the detector level. The full event selection is applied. The distributions are shown in data (black points) and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution, the light gray area shows the statistical uncertainty on the simulation, and the dark gray area shows the total uncertainty including systematic uncertainties. These distributions can be found in the publication in reference [3].

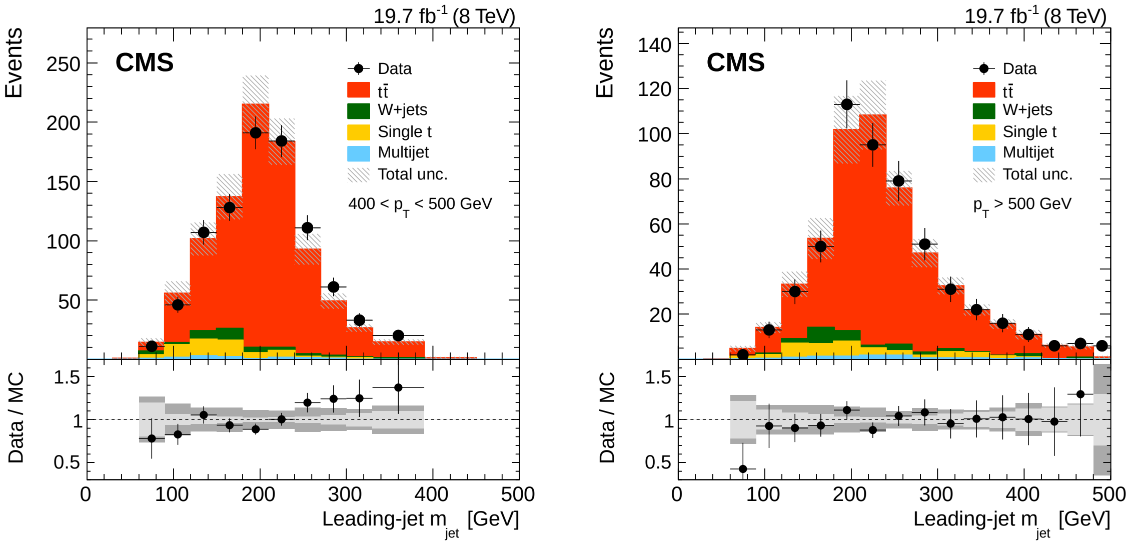

Figure 6.7: Distributions of the leading-jet mass for $400 < p_T < 500\,\text{GeV}$ (left) and $p_T > 500\,\text{GeV}$ (right). The distributions are shown for the combination of the electron and muon channels with the full event selection applied. The distributions are shown in data (black points) and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution, the light gray area shows the statistical uncertainty on the simulation, and the dark gray area shows the total uncertainty including systematic uncertainties. These jet-mass distributions are published in reference [3].

Background processes from other $\mathrm{t\bar{t}}$ decay channels

After the event selection discussed above some $\mathrm{t\bar{t}}$ events are still left from $\mathrm{t\bar{t}}$ decay channels other than the e/μ+jets channel. These events are considered as background for this analysis. Figure 6.8 shows the distribution of the leading-jet mass in simulation at the detector level for the combination of the electron and the muon channels and for different contributions from different $\mathrm{t\bar{t}}$ decay channels. The dominant fraction after the selection comes from e/μ+jets $\mathrm{t\bar{t}}$ decays with 80.7% but also small fractions of τ+jets and dileptonic decays are observed. Fully-hadronic $\mathrm{t\bar{t}}$ decays are well suppressed and contribute with only 0.4% to the full selection. The τ+jets background contributes with about 7.3% to the full selection, has a similar shape as the signal, and peaks also at the top quark mass because of the presence of a hadronic top quark decay in the event. The dileptonic $\mathrm{t\bar{t}}$ events show a background-like shape and contribute with 11.6% to the full selection. The $\mathrm{t\bar{t}}$ background processes are treated differently to other background processes which are subtracted before the unfolding. Subtracting the τ+jets background evaluated in simulation from the data would lead to a dependence of the measurement on position of the top quark mass peak in simulation and therefore on the top quark mass used in the simulation. For all $\mathrm{t\bar{t}}$ backgrounds a subtraction prior to the unfolding would lead to a dependence of the measurement on the normalization of the $\mathrm{t\bar{t}}$ simulation which is not well known because the top quark p_{T} spectrum is not well modeled by the simulation. All $\mathrm{t\bar{t}}$ backgrounds are therefore included in the response matrix of the unfolding to be treated as relative corrections in the unfolding instead of an absolute subtraction.

6.7 Unfolding

After a careful definition of the measurement phase space at the particle level and a similar selection at the detector level the next step is an unfolding of the data to the particle level. The unfolding is performed with the TUnfold framework [126] described in section 5.1. The response matrix for the unfolding of the data is evaluated with the default $\mathrm{t\bar{t}}$ sample simulated with POWHEG +PYTHIA. This section includes studies for the choice of the fundamental parameters of the unfolding setup, the setup of the response matrix, and studies on the dependence of the unfolding on the simulation model used to evaluate the response matrix.

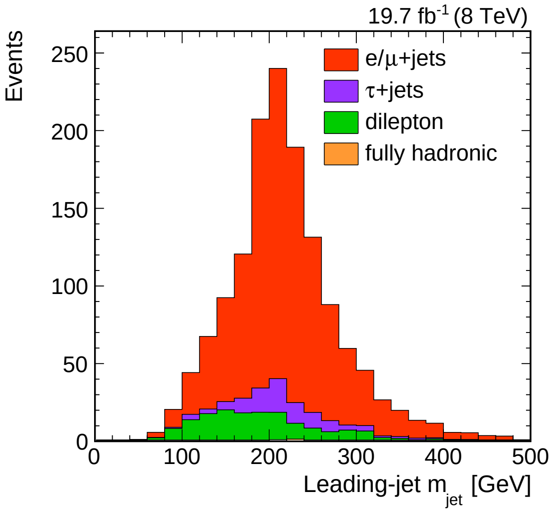

Figure 6.8: Distribution of the leading-jet mass at the detector level for the combination of the electron and muon channels. The figure shows $t\bar{t}$ events simulated with POWHEG +PYTHIA and normalized to an integrated luminosity of 19.7 fb^{-1}. The simulation is divided into different decay channels of the $t\bar{t}$ system. Contributions from e/μ+jets, τ+jets, dileptonic, and fully-hadronic $t\bar{t}$ decays are stacked.

6.7.1 Binning

The binning at the particle level should not be too small to make sure that events are not smeared over too many bins by the limited detector resolution. However, the bin width at the particle level should not be chosen smaller than one standard deviation of the reconstruction resolution to allow a stable unfolding. Furthermore, the amount of events in the measurement phase space that also pass the detector-level selection should be reasonably high to have enough statistical precision. The binning at the detector level should be chosen finer as the one at the particle level. There should be about twice as many bins at the detector level than at the particle level to give enough freedom to the minimization in the unfolding process.

The binning at the particle level in this measurement is chosen with the help of the standard deviation (σ) of the distribution of the relative difference between m_{jet} at the detector and at the particle level defined as $(m_{\mathrm{jet}}^{\mathrm{reco}} - m_{\mathrm{jet}}^{\mathrm{gen}})/m_{\mathrm{jet}}^{\mathrm{gen}}$. Figure 6.9 shows the mean and the σ of $(m_{\mathrm{jet}}^{\mathrm{reco}} - m_{\mathrm{jet}}^{\mathrm{gen}})/m_{\mathrm{jet}}^{\mathrm{gen}}$ as a function of $m_{\mathrm{jet}}^{\mathrm{gen}}$ in the electron channel. It shows the inclusive distributions (top) together with high-pileup and low-pileup contributions (bottom) obtained by requiring a number of primary vertices (NPV) higher or lower than 15. The figures show that jets with a low mass at the particle level are often reconstructed at higher masses. With higher particle-level masses the reconstructed mass at the detector level gets closer to the value at the particle level. A similar effect is observed for the resolution which is low at low masses and improves with higher masses until it reaches a minimum of about 15%. Both effects can be explained by a relatively high dependence of the large CA12 jets on soft effects, especially on pileup which is included in the simulation at the detector level but not at the particle level. The additional radiation by pileup shifts the jet mass to higher values and leads to a worse jet-mass resolution. This effect can be observed by looking at the quantities for high and low-pileup scenarios. For the high-pileup scenario the mean jet-mass difference gets larger and the resolution gets worse while both quantities improve in the low-pileup case. The pileup effects are included in the simulation and should be handled by the unfolding process. Figure 6.10 shows the distribution of the leading-jet mass in data and simulation for the high and low-pileup region. Both regions show a good data to simulation agreement and show that the differences due to pileup are reasonably well modeled by the simulation. Hence, no corrections are applied before the unfolding. The bin width at the particle level is now chosen to be roughly one σ of the reconstruction resolution and at the same time large enough that each bin is expected to contain at least 50 events in data for the combination of the electron and muon channels leading to seven bins at the particle level. Distributions of the purity and stablility, which are often used to define the binning, are shown in appendix A.2 together with the reconstruction effciency. Purity and stability are not used in this unfolding because the definition is difficult in the case of a shift between the particle and the detector level. Further discussion is given in appendix A.2.

The binning at the detector level is chosen such that the distribution of the leading-jet mass is flat for the default simulation with POWHEG +PYTHIA by requiring the same number of events for each bin. This binning is chosen to reduce effects from the choice of the simulation model that might be amplified by a sharp peak in the detector-level distribution. Thirteen bins are used at the detector level making sure that the bin borders do not overlap with the bin borders at the particle level to avoid border effects.

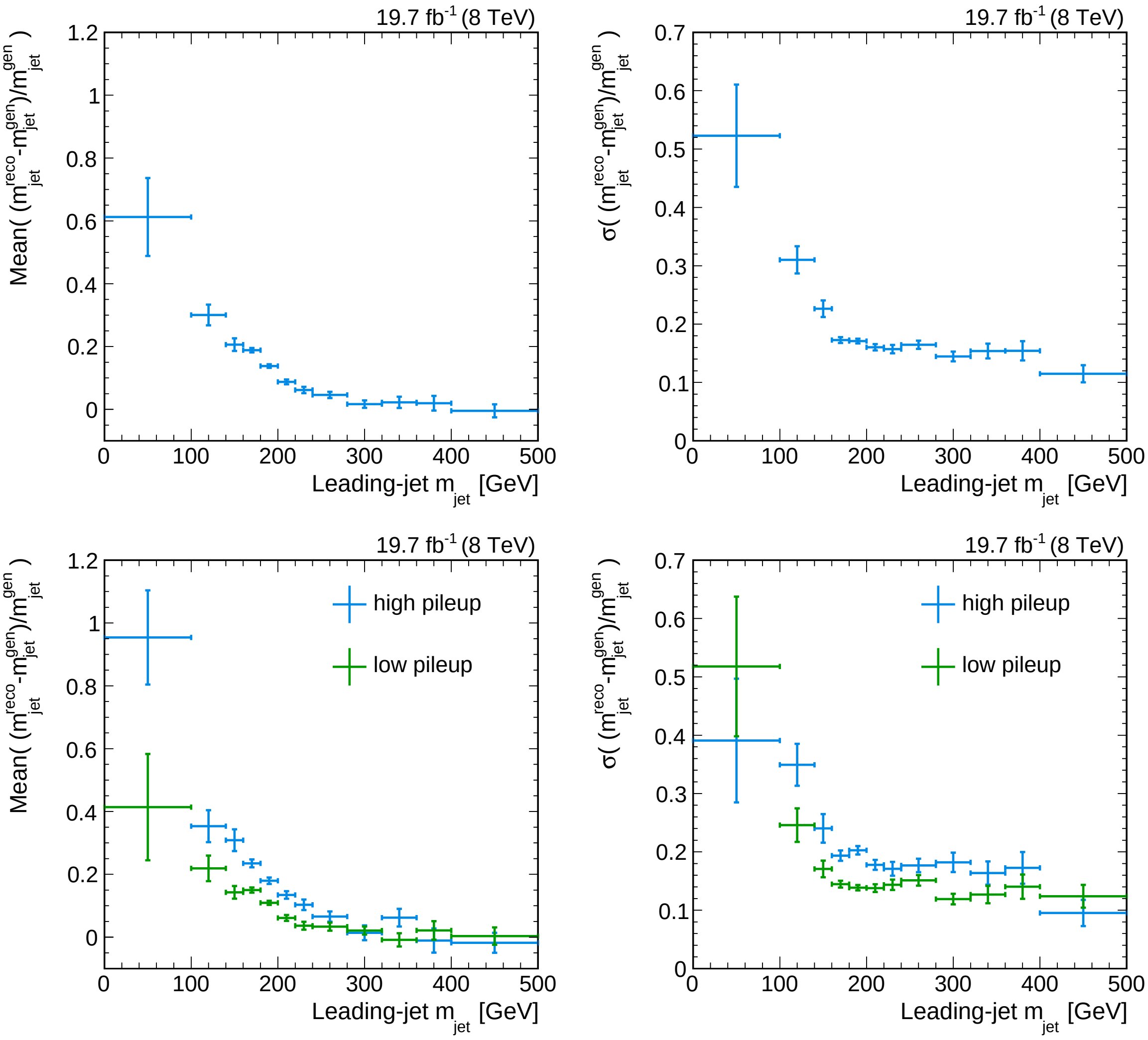

Figure 6.9: Mean (left) and σ (right) of the relative jet-mass difference between detector and particle level as a function of m_{jet} at the particle level. The plots are obtained in $\mathrm{t\bar{t}}$ simulation in the electron+jets channel. Distributions obtained from the inclusive selection can be found at the top. Distributions for a high-pileup and a low-pileup scenario requiring more or less than 15 primary vertices are shown at the bottom.

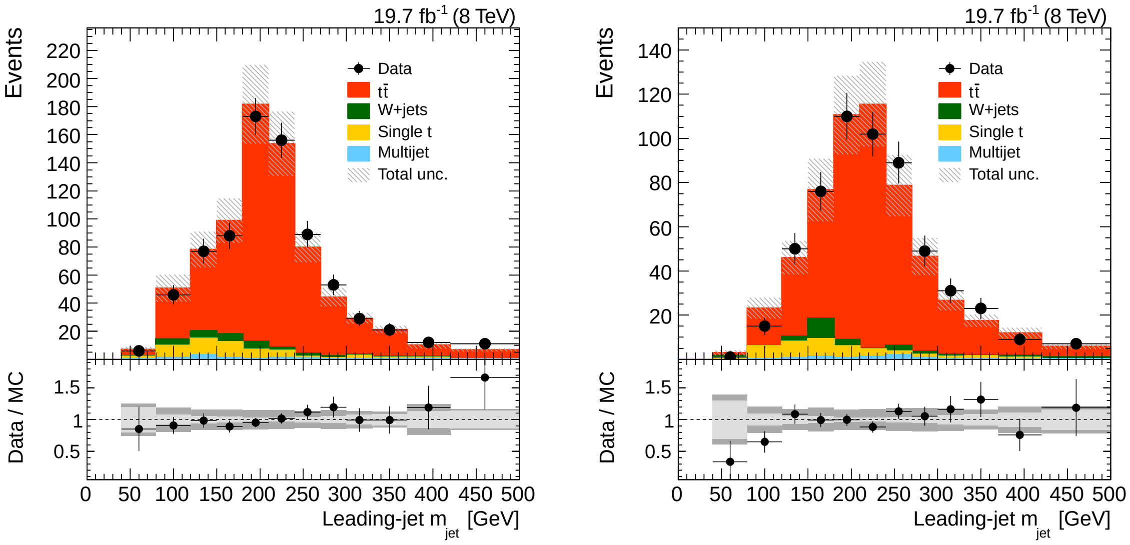

Figure 6.10: Distributions of the leading-jet mass for the high (left) and low-pileup (right) scenario selecting more or less than 15 primary vertices in the event. The distributions are shown for the combination of the electron and muon channels with the full event selection applied. The distributions are shown in data (black points) and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution, the light gray area shows the statistical uncertainty on the simulation, and the dark gray area shows the total uncertainty including systematic uncertainties.

6.7.2 Response matrix

The response matrix is used to handle migrations between bins from the particle level to the detector level. It is a two-dimensional matrix holding the particle-level information in lines (y-axis) and the reconstruction-level information in rows (x-axis). It is estimated from simulation. The main part of the response matrix holds probabilities that an event generated in one bin at the particle level is reconstructed in another bin at the detector level. Each row is therefore normalized to the total number of events in the corresponding bin at the particle level. Events that are part of the particle-level phase space but do not pass the detector-level selection are filled for each particle-level bin in the respective underflow bin of the detector level distribution. Events that are selected at the detector level but not at the particle level are included in respective underflow bins of the particle-level distribution. Such events can come from $\mathrm{t\bar{t}}$ background processes like τ+jets or from events migrating from outside the measurement phase space into the detector-level phase space. Scale factors accounting for differences between data and simulation at the detector level are included in the main part of the response matrix and are compensated in the underflow bins of the detector-level distribution to make sure they do not change the number of events in the particle-level phase space. A sketch of the general structure of the response matrix is shown in figure 6.11 for visualization purposes.

Division of the phase space into different p_T bins

The shape of the jet-mass distribution of boosted top quarks is expected to depend on the momentum of the top quarks and in this way on the p_T of the jets. The influence of additional radiation from initial-state radiation, final-state radiation, and underlying event becomes more prominent with increasing p_T and is expected to shift the jet mass to higher values and to broaden the distribution. Figure 6.12 shows the distribution of the leading-jet mass at the particle level for the full phase space and both channels combined for different ranges in jet p_T. All distributions are normalized to an integral of one to allow a pure shape comparison. A dependence of the shape of the jet-mass distribution on the p_T of the jet is clearly visible. The peak position is shifted to higher values for increasing jet p_T and the tail of the distribution to higher masses gets more prominent as expected. Because of the clear dependence of the shape of the jet-mass distribution on the jet p_T and the knowledge that the top quark p_T spectrum is not well modeled by the simulation, the measurement phase space is divided into two p_T bins at both the particle and the detector level. A two-dimensional unfolding is set up with two p_T bins for $400 < p_\mathrm{T} < 500\,\mathrm{GeV}$ and $p_\mathrm{T} > 500\,\mathrm{GeV}$. Migrations between the two p_T regions are

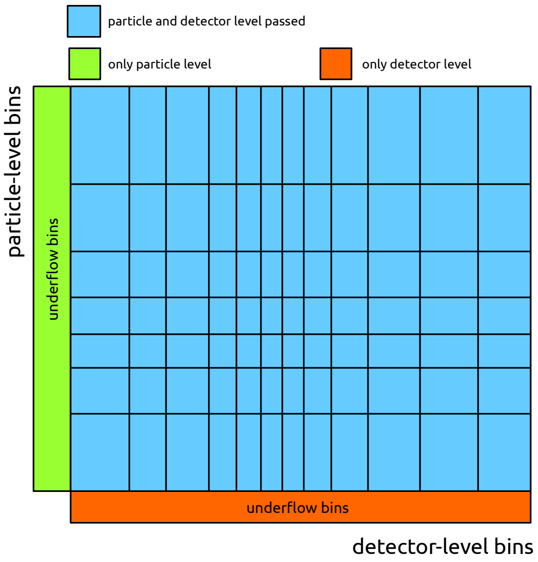

Figure 6.11: Sketch to illustrate the general structure of a response matrix.

accounted for by respective areas in the response matrix. The binning in p_{T} further helps to reduce dependence of the unfolding on simulation models that affect the p_{T} distribution of the jets.

An additional side-band region is introduced on both levels for a leading-jet p_{T} between 300 and 400 GeV to account for migrations from a lower jet-p_{T} region into the measurement phase space and to reduce model dependencies in the unfolding. This side-band region is further divided into two bins at the detector level for $300 < p_{\mathrm{T}} < 360\,\mathrm{GeV}$ and $360 < p_{\mathrm{T}} < 400\,\mathrm{GeV}$. The bins in the jet-mass distribution are reduced for the side-band region from seven to five bins at the particle level and from thirteen to nine bins at the detector level.

Additional side-band regions

Additional side-band regions are added to the unfolding by relaxing individual requirements in the phase-space definition at both the particle level and the detector level. The side-band regions help to constrain selection efficiencies from data and to reduce the dependence on the simulation model.

A side-band region for a lower p_{T} of the second jet is introduced for $100 < p_{\mathrm{T},2} < 150\,\mathrm{GeV}$ with one bin at the particle and two bins at the detector level. Another side-band region is defined by inverting the veto on additional jets and explicitly requiring a third CA12 jet with $150 < p_{\mathrm{T},3} < 200\,\mathrm{GeV}$. This second side-band region adds one bin at both

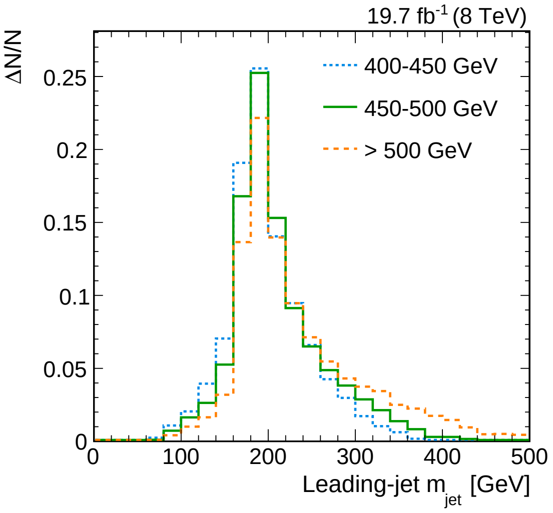

Figure 6.12: Distribution of the leading-jet mass at the particle level for different ranges in the p_{T} of the leading jet. The distributions are obtained for $\mathrm{t\bar{t}}$ simulation with POWHEG +PYTHIA. They are normalized to an integral of one to allow a shape comparison.

the particle and the detector level. Inversions of the requirements on $m_{\text{jet},1} > m_{\text{jet},2}$ and $\Delta R(\text{jet 2, lepton}) < 1.2$ have been studied but found not to be useful because an inversion of these requirements leads to different event kinematics and do not help to improve the unfolding. Particle-level information is added in the last 6 bins of the particle-level axis of the response matrix for events which are selected at the detector level but not part of the particle-level phase space. Figure 6.13 shows the full response matrix with all p_{T} regions and side-band regions. Particle-level bins are shown on the y-axis and detector-level bins on the x-axis. The response matrix is derived with the default $\mathrm{t\bar{t}}$ sample simulated with POWHEG +PYTHIA. A sketch of the response matrix is shown together with the simulation to visualize the structure explained above. The effect of low jet masses being reconstructed too high as discussed in section 6.7.1 is visible in the response matrix.

6.7.3 Regularization

A regularization term is used within the TUnfold framework to damp large statistical fluctuations amplified by the unfolding process as described in section 5.1.1. Technically the regularization term reduces large differences between the unfolding output and the bias distribution used in the unfolding. In this measurement the bias distributions is the particle-level distribution simulated with the same $\mathrm{t\bar{t}}$ simulation used to evaluate the response matrix. A regularization always introduces a small bias of the measurement towards the simulated bias distribution. The exact definition and strength of the regu-

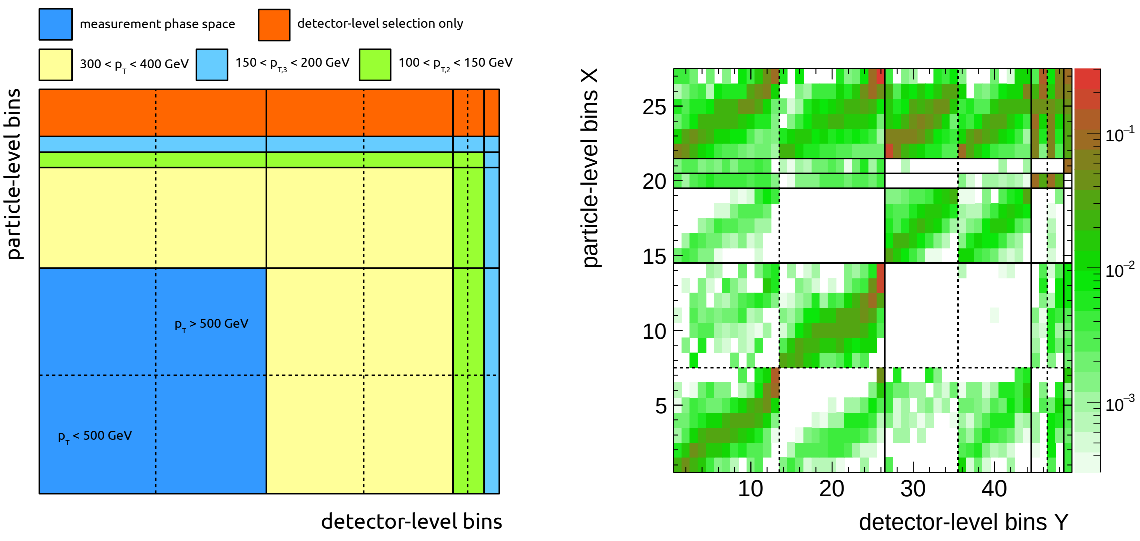

Figure 6.13: A sketch is shown on the left to visualize the structure of the response matrix used in the final measurement. The actual response matrix derived with $t\bar{t}$ events simulated with POWHEG +PYTHIA is shown on the right for the combination of the electron and muon channels.

larization has to be chosen carefully to find a good compromise between the suppression of statistical fluctuations and a small bias towards the simulation.

This section includes a brief overview of the exact definitions for the regularization used for this measurement followed by studies of different regularization parameters. The regularization for the measurement is defined in the following way:

- The regularization is applied to all bins in the measurement phase space and the side bands except for overflow or underflow bins.
- The regularization is applied with respect to the number of events in each bin (size regularization) and no density factors are applied. This leads to a unity matrix as **L** matrix.
- The bias distribution is taken from the particle-level distribution of the simulation used for the response matrix. It is scaled such that the number of events at the detector level matches the number of events in data using the bias scale factor f_b. The scaling reduces the influence on differences in the cross section in data and simulation.
- The optimal regularization strength τ is chosen by a minimization of the average global correlation coefficients as described in section 5.1.1. The covariance matrix that is used for the correlations includes uncertainties from statistical input uncertainties ($\mathbf{V}_{yy}$ in equation (5.3)), statistical uncertainties on the response matrix **A** in equation (5.3), and background uncertainties.

Determination of the optimal τ value

The strength of the regularization is defined by the τ value. If this value is chosen too small the unfolding output shows large non-physical statistical fluctuations. If chosen too large, the unfolding output will be strongly biased towards the bias distribution from simulation. Two approaches to determine the optimal τ value are described in section 5.1.1 and tested in this chapter. Figure 6.14 shows on the top left a scan over the global correlation coefficients together with the point corresponding to the optimal value found by the scan in an unfolding of pseudo-data simulated with MadGraph+pythia and a top quark mass of 166.5 GeV unfolded with a response matrix simulated with MadGraph+pythia and a top quark mass of 172.5 GeV in the electron channel. The corresponding L-curve is shown on the top right together with the points corresponding to the optimal values obtained with the L-curve scan and the global-correlation scan. The L-curve scan finds in this case a much lower value of τ. This leads to large fluctuations on the unfolding output which can be seen on the bottom of figure 6.14, where the respective unfolding output distributions are shown together with the particle-level distributions in MadGraph +pythia with $m_{\mathrm{t}} = 166.5\,\mathrm{GeV}$ and the bias distribution used in the unfolding. The result obtained with the global-correlation scan is shown on the left showing reasonable statistical uncertainties and no clear dependence on the bias distribution. The result obtained with the L-curve scan on the right, however, shows large fluctuations. The L-curve scan clearly failed in this case. In another example unfolding pseudo-data simulated with MadGraph +pythia with $m_{\mathrm{t}} = 172.5$ with a response matrix simulated with powheg +pythia, the results of the L-curve scan and the global-correlation scan are much closer. Figure 6.15 shows as before on the top the two scans with the optimal points obtained by the respective scans. The values derived by the L-curve scan and the global-correlation scan are much closer in this case. This is also visible in the unfolded distributions shown on the bottom which show very similar statistical uncertainties in this case.

In summary, the L-curve scan was observed to fail in some cases especially for low input statistics while the scan over global correlation coefficients showed a much more stable behavior. The difference between the two methods are small in cases in which both methods work. The optimal τ value for the final unfolding is therefore chosen by a scan over global correlation coefficients.

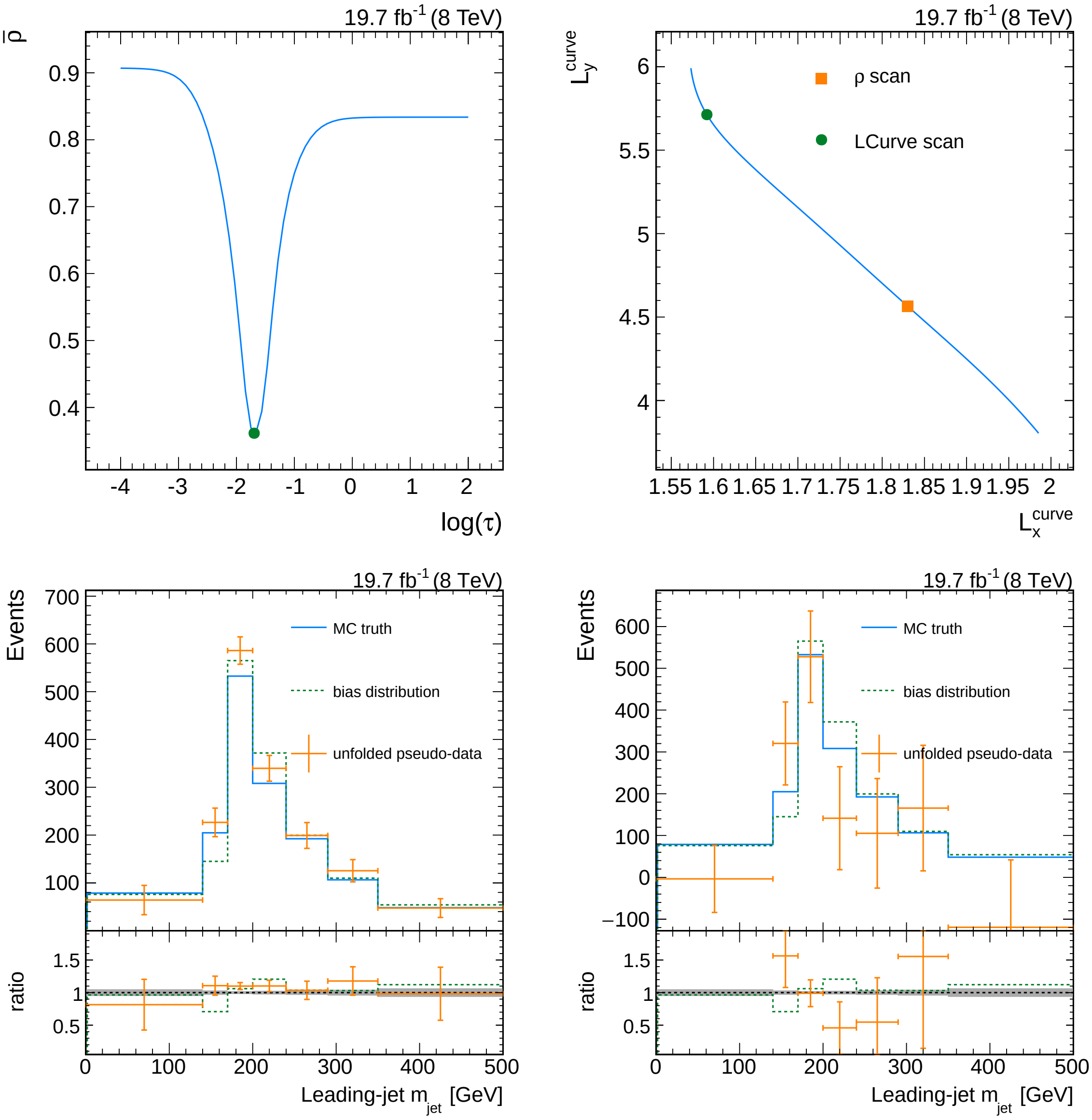

Figure 6.14: Tests of different methods to determine the optimal regularization strength using pseudo-data simulated with MadGraph +pythia and $m_\mathrm{t} = 166.5\,\mathrm{GeV}$ unfolded with a response matrix derived with MadGraph +pythia and $m_\mathrm{t} = 172.5\,\mathrm{GeV}$ in the electron channel. The scan over global correlation coefficients is shown on the top left together with the optimal point. The corresponding L-curve scan is shown on the top right together with the optimal points found by the L-curve scan and by the correlation scan. The figure on the bottom left show the unfolded pseudo-data obtained with a scan over the global correlation coefficients together with the respective particle-level distribution and the bias distribution from the unfolding. The figure on the bottom right shows the same for the L-curve scan.

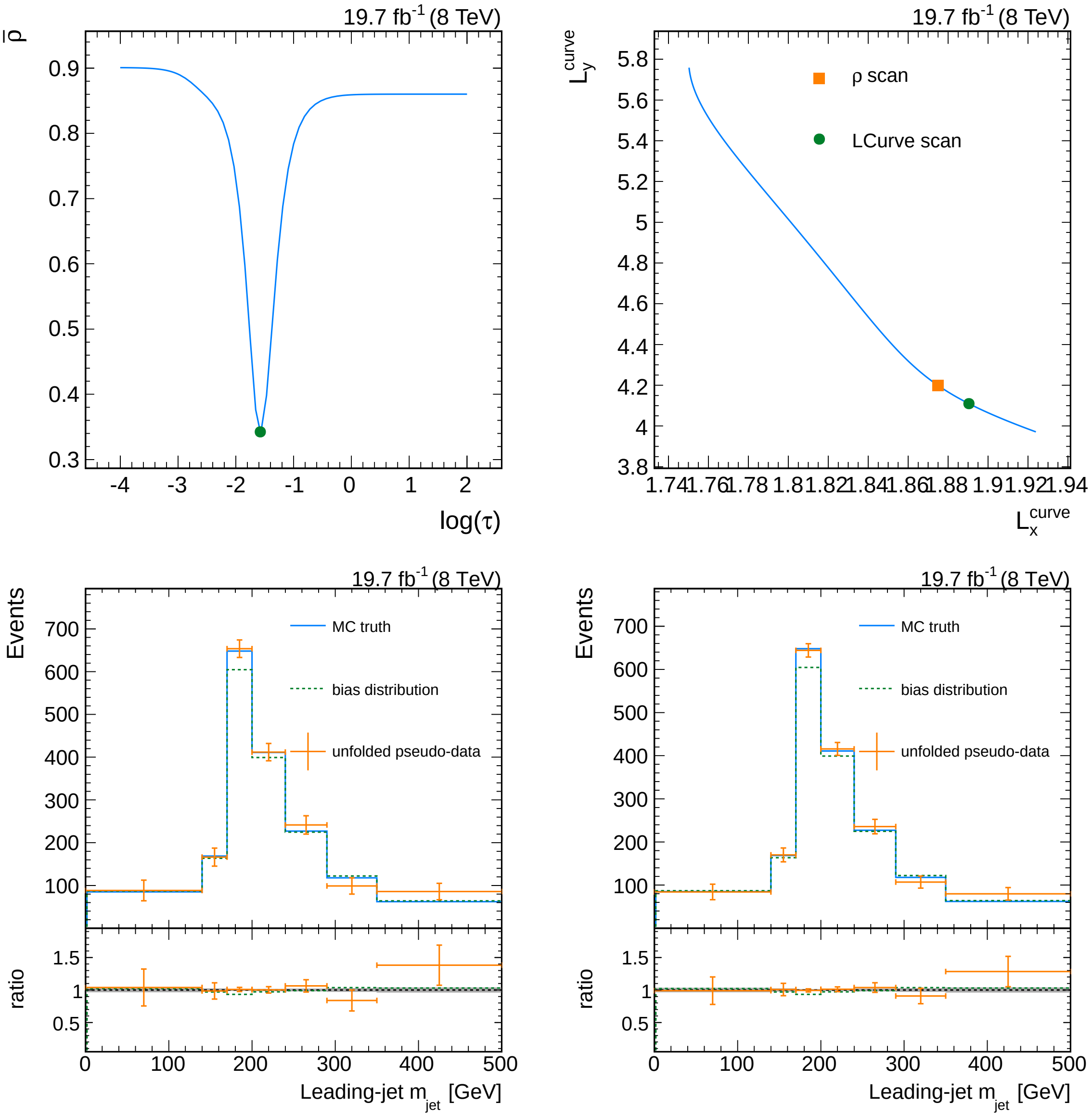

Figure 6.15: Tests of different methods to determine the optimal regularization strength using pseudo-data simulated with MadGraph +pythia and unfolded with a response matrix derived with powheg +pythia and in the muon channel. The scan over global correlation coefficients is shown on the top left together with the optimal point. The corresponding L-curve scan is shown on the top right together with the optimal points found by the L-curve scan and by the correlation scan. The figure on the bottom left show the unfolded pseudo-data obtained with a scan over the global correlation coefficients together with the respective particle-level distribution and the bias distribution from the unfolding. The figure on the bottom right shows the same for the L-curve scan.

Form of the L matrix

The choice of the **L** matrix determines if the regularization is applied to the bin content, to the first derivative, or to the second derivative of the distribution as mentioned in section 5.1.1. Two different choices of the **L** matrix are tested in this analysis, a size regularization and a regularization to the first derivative. A priori it is not clear which regularization should perform better for an asymmetric peaking spectrum like the one measured in this analysis. The size regularization of a peaking spectrum is expected to lead to a small bias in the peak region of the spectrum. A regularization to the first derivative leads to a regularization with respect to a distribution with two peaks which could lead to a bias in the tails of the distribution. Furthermore, the fist derivative is different for high and low masses which could lead to an asymmetric bias. Figure 6.16 shows an unfolding of pseudo-data simulated with MadGraph +pythia with $m_{\mathrm{t}} = 172.5\,\mathrm{GeV}$ unfolded with a response matrix simulated with powheg +pythia with size regularization on the left and with first derivative regularization on the right. Both unfolding outputs are consistent with the particle-level distribution and do not follow the bias distribution. The unfolding with first derivative regularization shows slightly larger statistical uncertainties compared to size regularization and it does not give an improvement on the unfolding. A size regularization is therefore used for the final measurement.

6.7.4 Model dependence

The response matrix has to be evaluated in simulation because particle-level information is only available in simulation and not in data. The unfolding should, in the ideal case, only correct for detector and reconstruction effects and should not depend on the simulation model used to derive the response matrix. A dependence on the simulation model, however, can be introduced if the detector response depends on some event kinematics that are changed by the different simulation models. In order to suppress model dependencies on the response in different p_{T} regions the unfolding is performed in two jet-p_{T} bins. Additional side-band regions are defined to reduce the effect of the simulation model on migrations from outside the measurement phase space. The remaining effects of different choices for the simulation model on the unfolding results are tested by an unfolding of simulated pseudo-data obtained with different simulation models with a response matrix simulated with the default simulation. The unfolded pseudo-data is compared to the particle-level distribution obtained with the same simulation (MC truth) and the bias distribution obtained with the simulation used for the response matrix. The unfolding output should be consistent with the particle-level distribution if the unfolding

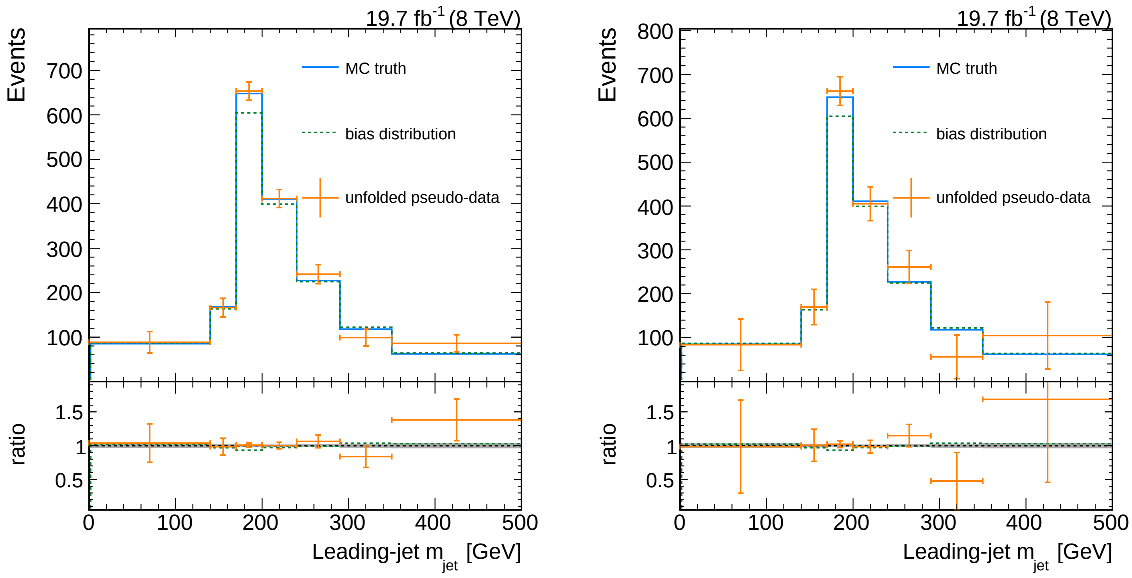

Figure 6.16: Unfolded pseudo-data obtained with size regularization on the left and a regularization on the first derivative on the right. Both figures show unfolded pseudo-data simulated with MadGraph+pythia and unfolded with powheg +pythia in the muon channel together with the corresponding particle-level distribution and the bias distribution from the unfolding.

is independent of the differences between the two simulation models. An unfolding output following rather the bias distribution than the MC truth would be a sign for a too strong regularization. Differences between the unfolded pseudo-data and the respective particle-level distribution are considered as model uncertainties on the results of this measurement. More details on the handling of model uncertainties are given in the following (section 6.8.3).

Figure 6.17 shows an example of a model-dependence test with pseudo-data simulated with MadGraph +pythia and unfolded with migrations simulated with powheg +pythia. All unfolding tests are performed in both the electron and the muon channels separately and in the combination of both channels. The figures show the unfolded distributions together with the respective particle-level distributions simulated with MadGraph +pythia as well as the bias distributions simulated with powheg +pythia and scaled to the number of events in MadGraph +pythia at the detector level. The unfolding outputs are consistent with the particle-level distributions and do not follow the bias distributions. No model uncertainty for the difference between MadGraph and powheg is therefore considered in the measurement. Further studies of model dependencies are shown in appendix A.1. The different model dependencies considered in this measurement are listed in the following:

- The dependence on the choice of the renormalization and factorization scales μ_r and

μ_f are studied with extra samples produced with simultaneous variations of μ_r and μ_f by factors of 0.5 and 2. These samples are simulated with POWHEG +PYTHIA and have only been studied for an invariant mass of the $\mathrm{t\bar{t}}$ system $m_{\mathrm{t\bar{t}}} > 700\,\mathrm{GeV}$ which makes up the largest part of the measurement phase space. These samples are used as pseudo-data and unfolded with the default simulation with $m_{\mathrm{t\bar{t}}} > 700\,\mathrm{GeV}$.

- A possible dependence on the top quark mass used in simulation is studied using different $\mathrm{t\bar{t}}$ samples simulated with MadGraph +PYTHIA and different values of m_{t} as pseudo-data. The different samples are unfolded with a response matrix simulated with MadGraph +PYTHIA with $m_{\mathrm{t}} = 172.5\,\mathrm{GeV}$. The considered mass points are $m_{\mathrm{t}} = 166.5$, 169.5, 171.5, 173.5, 175.5, and 178.5 GeV.
- The dependence on the parton-shower model is studied by unfolding pseudo-data simulated with MC@NLO +HERWIG with a response matrix simulated with POWHEG +PYTHIA.
- A possible influence of the shape of the top quark p_{T} spectrum is studied by reweighting the top quark p_{T} spectrum for the pseudo-data simulated with POWHEG +PYTHIA. The pseudo-data is unfolded with the nominal POWHEG +PYTHIA sample. This study was just performed as a test and is not considered as an extra model uncertainty. Other model effects like the variation of μ_r and μ_f also change the p_{T} spectrum and should cover the effect.

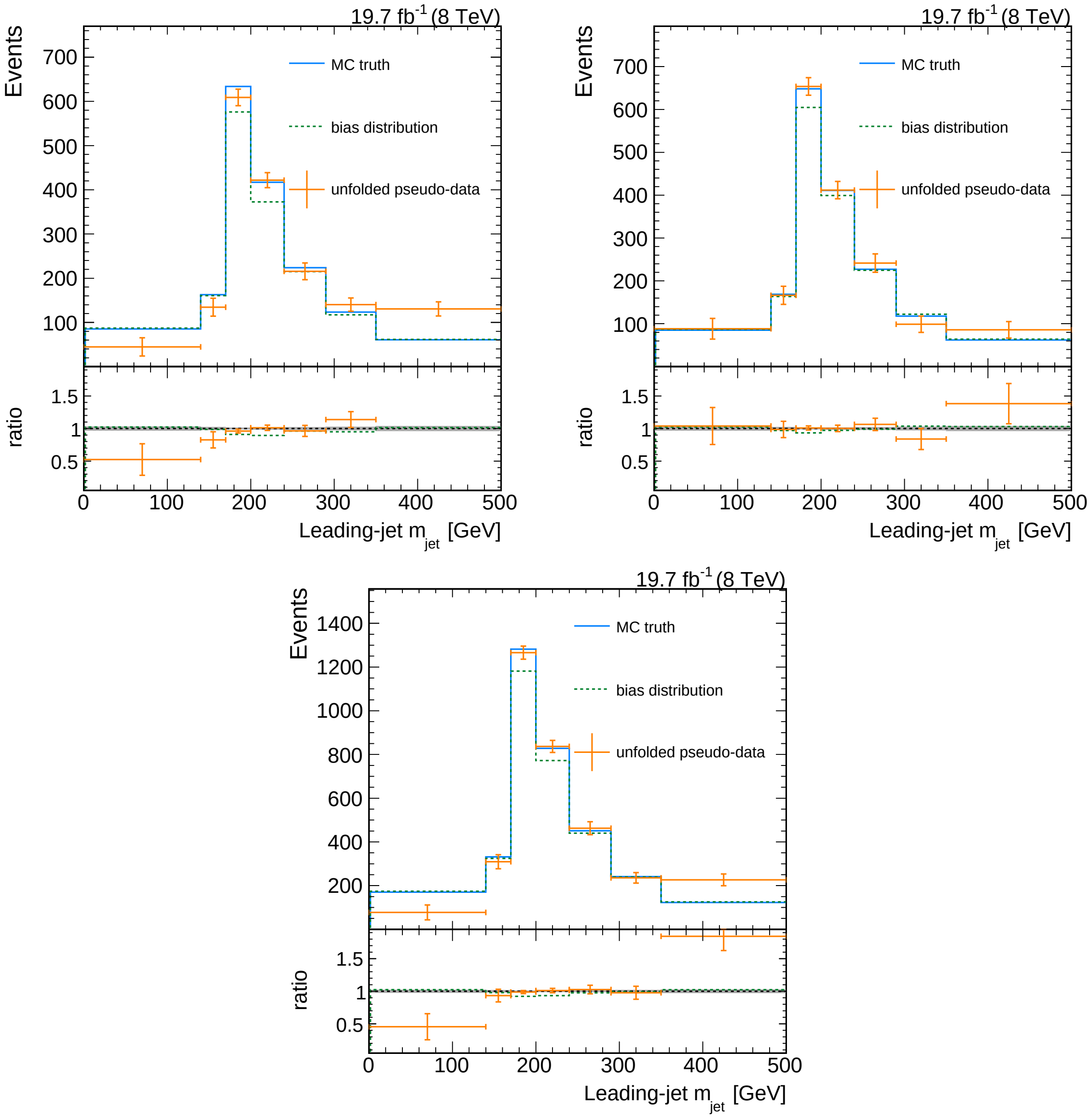

Figure 6.17: Jet-mass distribution of the leading jet for unfolded pseudo-data simulated with MadGraph +pythia and unfolded with a response matrix evaluated with powheg +pythia. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

6.8 Uncertainties of the unfolding output

This section gives a detailed description of all uncertainties considered for the final cross section measurement. The unfolding is affected by three different sources of uncertainties. Uncertainties can arise from a limited number of events in data or simulation, from systematic effects at the detector level, and from the choice of the simulation model used to determine the response matrix.

6.8.1 Statistical uncertainties

Three different sources of statistical uncertainties are considered for the measurement. The first and largest statistical uncertainty on the unfolding output is associated to the statistical uncertainty of the input data ($\mathbf{V}_{\boldsymbol{yy}}$ in equation (5.3)). A second uncertainty is connected to the limited number of events in simulation used to determine the response matrix and leading to statistical uncertainties on each bin of the response matrix ($\mathbf{A}$ in equation (5.3)). The third part of the statistical uncertainty comes from a limited number of events in the simulation of background processes subtracted prior to the unfolding.

All three sources are uncorrelated between different bins before the unfolding process but lead to correlated uncertainties on the unfolding output. A covariance matrix is estimated within the TUnfold framework for each of the three sources. The resulting covariance matrices are summed and the result is called statistical uncertainty in the following.

6.8.2 Systematic uncertainties at the detector level

The uncertainties on the unfolding output connected to systematic uncertainties at the detector level can also be divided into different categories which are treated in different ways. A detailed description on the different sources is given below.

Uncertainties on the response matrix

The unfolding is effected by several corrections applied to the simulation at the detector level to correct for differences in reconstruction efficiencies between data and simulation. These corrections lead to changes in the response matrix. The effects of the uncertainties of these correction factors on the unfolding output are estimated with response matrices produced with varied correction factors by ± 1 σ respectively. The resulting shifts between the nominal and varied response matrices are propagated to the unfolding output using a linear error propagation within the TUnfold framework. This leads to shifts on the unfolding output which are treated as systematic uncertainties and assumed to be fully correlated between individual bins. Either the up or the down variation of each correction is used for the final result depending on which variation gives the larger overall effect to keep the correlations of each uncertainty. A covariance matrix is built for each uncertainty and all covariance matrices are added up to the full systematic uncertainty. The effects from systematic uncertainties on the response matrix are evaluated in an unfolding of pseudo-data simulated with MadGraph +pythia instead of data. The MadGraph +pythia sample provides lower statistical uncertainties compared to real data and the systematic effects from variations in the response matrix should not depend on the input data. The evaluation in an unfolding of simulated pseudo-data is done to avoid a double counting of statistical uncertainties from the input data in the propagation of the systematic effects. The considered systematic uncertainties on the response matrix are listed in the following.

- The jet energy corrections (JECs) are varied within their uncertainties fully correlated for AK5 jets and CA12 jets. The uncertainties on the AK5 jet energy corrections are provided by CMS [145]. Jet energy corrections derived for AK7 jets by CMS [145] are used on the CA12 jets. The uncertainties on these correction have been increased for this analysis to cover the differences between AK7 and CA12 jets. More details on the JECs for the CA12 jets are given in appendix A.4. The jet mass of the CA12 jets is left untouched by the variations of the JECs to avoid a double counting of the uncertainty by the following jet-mass scale variations.
- The jet energy resolution smearing is varied on both AK5 and CA12 jets simultaneously.
- The jet-mass scale was studied in a sample enriched with CA12 jets containing a hadronic W decay as described in appendix A.3. The peak position of the W-jet mass was found to be consistent between data and simulation within uncertainties and no jet-mass correction is applied. The jet-mass scale is still varied by ±1.5% corresponding to the difference in the peak position between data and simulation in appendix A.3.
- Pileup corrections are introduced to match the number of primary interactions in simulation to the instantaneous luminosity profile in data. The uncertainties are obtained by a variation of the total inelastic cross section by ±5% [155, 156].
- Correction factors to correct for differences in the b tagging efficiency between data

and simulation are derived centrally in CMS for different jet flavors [124, 157]. They are varied fully correlated within their uncertainties.

- The uncertainties on the electron and muon ID scale factors are varied within their uncertainties as evaluated in CMS [143, 144].
- The uncertainties on the muon trigger efficiency is derived centrally within CMS while the uncertainties on the combined electron trigger have been measured within the scope of the $\mathrm{t\bar{t}}$-resonance search in reference [149] to be 1%. The respective scale factors are varied within their uncertainties.

Normalization of background processes

Background processes are estimated in simulation and subtracted from the data prior to the unfolding. The cross sections of all background processes are varied within the respective uncertainties. The resulting effect on the unfolding output is handled within TUnfold and treated as a fully-correlated uncertainty. A covariance matrix is built as before and added to the total systematic uncertainty. The uncertainty on W+jets production cross section is chosen conservatively to be 19% as it was measured for W+heavy-flavor production in reference [158]. An uncertainty of 23% is applied to the cross section of single top quark production as measured in single top quark production in association with a W boson in reference [159]. The uncertainty on QCD multijet production is set to 100%.

Normalization

A constant uncertainty on the measurement of the integrated luminosity of 2.6% [160] is added quadratically to the other systematic uncertainties after the unfolding. It is treated fully correlated between the different bins.

6.8.3 Model uncertainties

Model uncertainties are included to cover effects on the unfolding output by the choice of the simulation model used to simulate the response matrix. The estimation of these uncertainties is based on the model-dependence tests described in section 6.7.4. The uncertainty is estimated in an unfolding of an alternative simulation model used as pseudo-data unfolded with a response matrix simulated with the default simulation. The unfolding output is compared to the particle-level distribution in the alternative model and differences between the unfolding output and the particle-level distribution are considered as a model uncertainty for the final measurement. A list of model uncertainties and their treatment is given in the following.

- The uncertainty on the renormalization and factorization scales μ_r and μ_f are obtained in an unfolding of pseudo-data simulated with POWHEG +PYTHIA for an invariant mass of the $\mathrm{t\bar{t}}$ system $m_{\mathrm{t\bar{t}}} > 700\,\mathrm{GeV}$ and μ_r and μ_f varied simultaneously by factors of 0.5 and 2. The pseudo-data is unfolded with the default POWHEG +PYTHIA sample for $m_{\mathrm{t\bar{t}}} > 700\,\mathrm{GeV}$. Only the variation giving the overall larger effect is considered in the total uncertainties. The uncertainty is considered fully correlated or anti-correlated between the different bins depending on the direction of the shifts.

- An uncertainty on the choice of the parton shower is estimated by unfolding pseudo-data simulated with MC@NLO+HERWIG with the default response matrix simulated with POWHEG +PYTHIA. The uncertainty is considered to be fully correlated or anti-correlated between different bins depending on the direction of the shifts. No additional uncertainty on the choice of the MC generator is used because the unfolding of MADGRAPH +PYTHIA with POWHEG +PYTHIA in section 6.7.4 showed no significant model effect on the unfolding output.

- Several samples simulated with MADGRAPH +PYTHIA and different values of m_{t} are used as pseudo-data to estimate an uncertainty on the choice of m_{t} used in the evaluation of the response matrix. The samples are produced with variations of m_{t} by ±1, ±3, and ±6 GeV and unfolded with a response matrix simulated with MADGRAPH +PYTHIA with the central value of $m_{\mathrm{t}} = 172.5\,\mathrm{GeV}$. The envelope of all mass variations is taken as an uncertainty on m_{t}. The uncertainty is therefore treated uncorrelated between different bins of the unfolding output.

- The PDF uncertainties are treated differently compared to the other model uncertainties. The POWHEG +PYTHIA $\mathrm{t\bar{t}}$ sample is reweighted with PDF weights corresponding to the 51 eigenvectors of the CT10 PDF set [139] and a response matrix is derived for each variation. The systematic error propagation in TUnfold is used to obtain the shifts on the unfolding output as it is done for the systematic uncertainties on the response matrix. The uncertainties are scaled to 68% confidence level and added in quadrature to obtain the total PDF uncertainty.

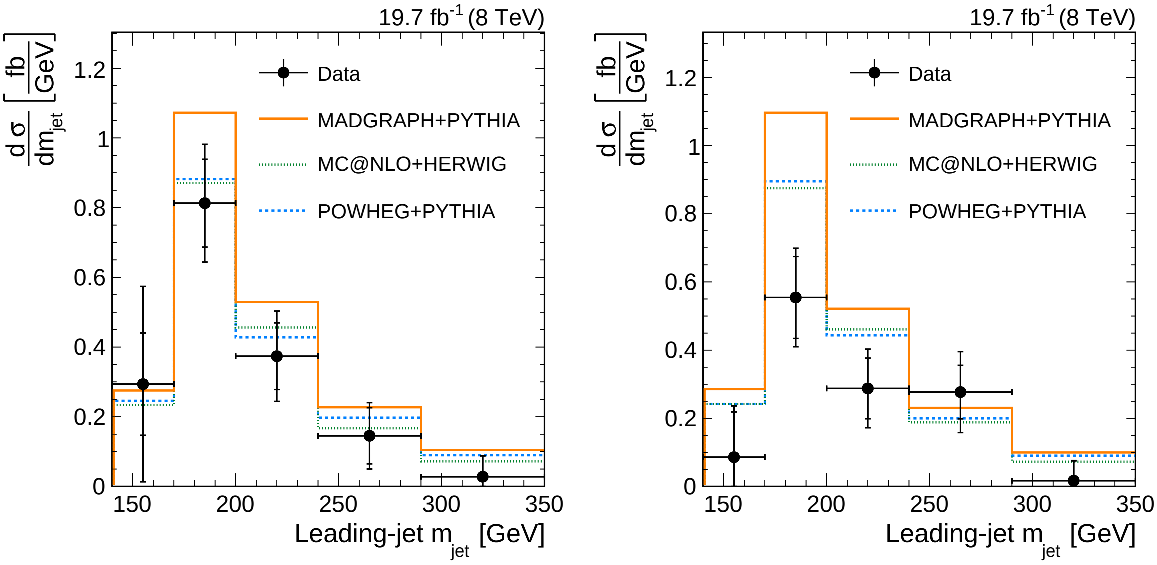

Figure 6.18: Differential $t\bar{t}$ production cross section as a function of the leading-jet mass at the particle level in the electron channel (left) and in the muon channel (right). The measurement is performed in a fiducial phase space defined in section 6.4. The measured data is shown as black points, the inner error bars show the statistical uncertainties form the unfolding and the outer error bars show the total uncertainty. The data is compared to particle-level distributions obtained with different simulation models using POWHEG +PYTHIA, MadGraph +PYTHIA, and MC@NLO +HERWIG.

6.9 Differential cross sections

The differential $t\bar{t}$ production cross section as a function of the leading-jet mass is first measured in the electron and in the muon channels separately before it is measured in the combination of the two channels. Figure 6.18 shows the differential cross section at the particle level in data measured in the electron channel on the left and in the muon channel on the right. The bins below 150 GeV and above 350 GeV are not considered for the final result because these bins showed significant instabilities in the unfolding tests. The data is compared to particle-level distributions obtained with different MC generators. The cross section in data is in general a bit lower compared to the MC generators which is well consistent with observations of a softer top quark p_T spectrum in the resolved case in references [150–152] and with other cross section measurements in the boosted regime in references [153, 154]. The top quark p_T spectrum in MadGraph +PYTHIA is even a bit harder compared to the other MC samples as studied in section 6.5. Figure 6.19 shows a comparison of both channels for data and POWHEG +PYTHIA simulation at the particle level. Both channels are consistent within the uncertainties and can be combined before the unfolding. The combination is performed prior to the unfolding to reduce statistical uncertainties on the input data and on the response matrix.

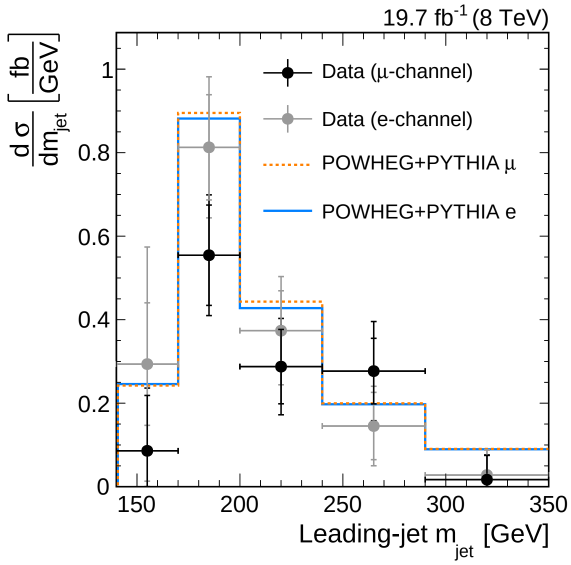

Figure 6.19: Comparison of the differential $t\bar{t}$ production cross section as a function of the leading-jet mass at the particle level measured in data in the electron and muon channels. The inner error bars on the data points show the statistical uncertainty from the unfolding and the outer error bars the total uncertainty. The data is shown together with the respective particle-level distributions simulated with POWHEG +PYTHIA.

6.10 Results

The measurement is finally performed in the combination of the electron and muon channels. The central result is a differential $t\bar{t}$ production cross section as a function of the leading-jet mass in a fiducial phase space defined in section 6.4 and summarized in table 6.1. The measured particle-level distribution in data is shown in figure 6.20 and compared to particle-level distributions simulated with POWHEG +PYTHIA, MADGRAPH +PYTHIA and MC@NLO +HERWIG. The overall cross section in the fiducial phase space is measured to be $\sigma = 101 \pm 11(\text{stat.}) \pm 13(\text{syst.}) \pm 9(\text{model})$ fb. It is a bit lower compared to the cross sections from POWHEG+PYTHIA with 133^{+18}_{-28} fb and MADGRAPH+PYTHIA with 159^{+17}_{-18} fb

Table 6.1: Selection criteria defining the measurement phase space at the particle level.

lepton+jets $t\bar{t}$ decays		
e/μ	$p_T > 45$ GeV	$\lvert\eta\rvert < 2.1$
leading jet	$p_T > 400$ GeV	$\lvert\eta\rvert < 2.5$
second jet	$p_T > 150$ GeV	
veto on additional jets	$p_T > 150$ GeV	
ΔR(jet 2, lepton) < 1.2		
$m_{\text{jet 1}} > m_{\text{jet 2 + lepton}}$		

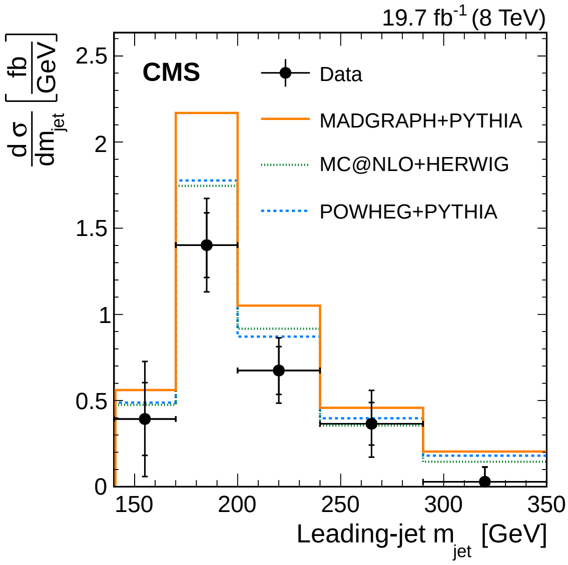

Figure 6.20: Differential $t\bar{t}$ production cross section as a function of the leading-jet mass at the particle level measured in a fiducial measurement phase space described above. The unfolded data is shown as black points. The uncertainties on the measurement are shown as vertical bars, where the inner error bars show the statistical uncertainty and the outer error bars the full uncertainty. The horizontal bars show the bin width. The data is compared to particle-level distributions simulated with POWHEG+PYTHIA, MadGraph+PYTHIA, and MC@NLO+HERWIG. This result was published in reference [3].

assuming an inclusive $t\bar{t}$ cross section of 253 pb [25, 161–166]. The uncertainties come from variations of μ_r and μ_f. This effect is well consistent with other measurements in boosted $t\bar{t}$ production in references [153, 154] and can be explained by the observation of a softer top quark p_T spectrum in data compared to simulation in references [150–152], leading to a lower $t\bar{t}$ cross section in data for high top quark p_T. Comparisons to next-to-next-to-leading order (NNLO) calculations [167] show a better agreement with the shape of the top quark p_T spectrum in data. The top quark p_T spectrum in MadGraph +PYTHIA was observed to be even harder compared to the other simulations leading to an even higher fiducial cross section.

Figure 6.21 shows the relative systematic uncertainties on the unfolding output compared to the statistical uncertainty and to the combination of statistical and systematic or model uncertainties, respectively. The different uncertainty sources and their treatment are discussed in section 6.8. The statistical uncertainty is the dominant uncertainty in each bin. The largest systematic uncertainties come from the jet energy scale and the jet-mass scale. The largest model uncertainties are connected to the parton shower and the choice of m_t in the simulation of the response matrix. The measured cross-section values for each bin together with statistical, systematic, and model uncertainties are given in table 6.2. They can be found in the publication in reference [3]. A more detailed list with

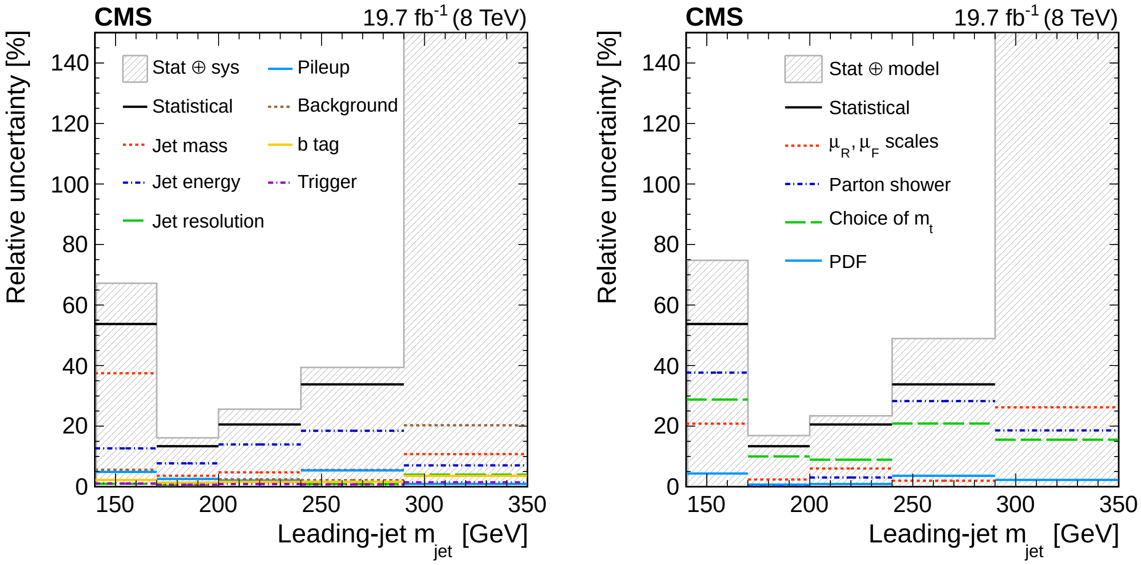

Figure 6.21: Relative uncertainties on the measurement of the differential $t\bar{t}$ production cross section as a function of the leading-jet mass shown in figure 6.20. Relative systematic experimental uncertainties are shown on the left and model uncertainties on the right. Each set of uncertainties is compared to the statistical uncertainties on the measurement in black. The gray hatched region shows the combination of statistical and systematic or model uncertainties respectively. These figures are published in reference [3].

all individual uncertainty contributions is given in table A.1 in appendix A.5 together with correlation coefficients between different output bins in figure A.16. Covariance matrices can be found in reference [3] or in tables A.3 and A.4 in appendix A.5.

Normalized cross section

The normalized differential cross section is measured in addition to the differential cross section. The cross section is normalized to the total cross section measured in the fiducial

Table 6.2: Measured values of the differential $t\bar{t}$ production cross section with uncertainties for the individual bins shown in figure 6.20. These values are published in reference [3].

Range in m_{jet} [GeV]	140–170	170–200	200–240	240–290	290–350
Integrated cross section [fb]	12	42	27	18	1.7
Statistical uncertainty [%]	54	13	21	34	300
Systematic uncertainty [%]	40	9	16	20	25
Model uncertainty [%]	52	10	11	35	36
Total uncertainty [%]	85	19	28	53	300

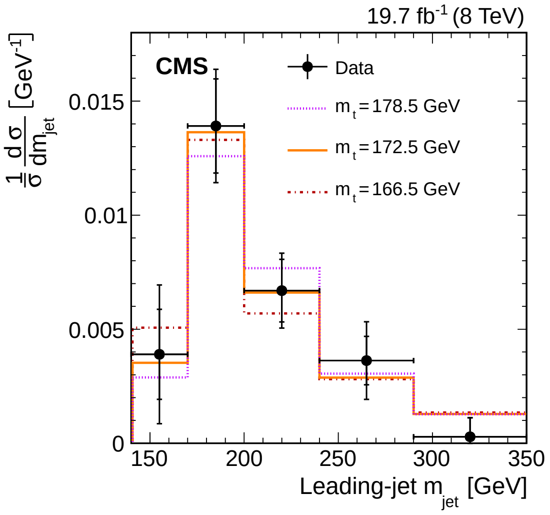

Figure 6.22: Normalized differential $t\bar{t}$ production cross section as a function of the leading-jet mass at the particle level. The cross section is measured in a fiducial measurement phase space described above. The unfolded data is shown as black points. The uncertainties on the measurement are shown as vertical bars, where the inner error bars show the statistical uncertainty and the outer error bars the full uncertainty. The horizontal bars show the bin width. The data is compared to particle-level distributions simulated with POWHEG +PYTHIA, MadGraph +PYTHIA, and MC@NLO +HERWIG. This result was published in reference [3].

phase space. The systematic and model uncertainties are evaluated in the absolute case and the respective covariance matrices are normalized through Gaussian error propagation. Figure 6.22 shows the normalized differential $t\bar{t}$ production cross section as a function of the leading-jet mass measured in the fiducial measurement phase space. The data is compared to distributions at the particle level obtained with MadGraph +PYTHIA and different values of m_t. Once the normalization difference is accounted for, the shape of the leading-jet mass spectrum is well described by the simulation. A sensitivity on the top quark mass is clearly visible from the different MC samples. This sensitivity is tested in the following section 6.11. Table 6.3 includes the measured normalized cross-section values for the individual bins together with the different uncertainty sources. A more detailed list of individual uncertainties, as well as correlation coefficients between individual output bins, and covariance matrices can be found in table A.2, figure A.17, and tables A.5 and A.6 in appendix A.5.

Table 6.3: Measured values of the normalized differential $\mathrm{t\bar{t}}$ production cross section with uncertainties for the individual bins shown in figure 6.22. The values are published in reference [3].

Range in m_{jet} [GeV]	140–170	170–200	200–240	240–290	290–350
Integrated normalized cross section	0.12	0.42	0.27	0.18	0.017
Statistical uncertainty [%]	51	15	21	29	290
Systematic uncertainty [%]	34	5	9	13	27
Model uncertainty [%]	48	9	10	34	36
Total uncertainty [%]	78	18	25	47	300

6.11 Top quark mass extraction

The ultimate goal of this kind of measurement is a comparison of the data to analytic calculations from first principles at the particle level and an extraction of a well-defined top quark mass. Analytic calculations are, however, not yet available for this measurement phase space. An extraction of m_{t} using MC event generators is done instead to test the sensitivity of the method with the data collected at a center-of-mass energy of 8 TeV. This extraction is meant as a proof of principle. The top quark mass is extracted from the normalized differential cross section because only the shape of the jet-mass spectrum can be reliably calculated. A χ^2 value is calculated for the difference between the data and different MC generated templates obtained with MadGraph +pythia and different values of m_{t}. Seven templates are used with a central top quark mass of $m_{\mathrm{t}} = 172.5\,\mathrm{GeV}$ and variations of m_{t} by $\pm 1\,\mathrm{GeV}$, $\pm 3\,\mathrm{GeV}$, and $\pm 6\,\mathrm{GeV}$. For each template the χ^2 value is calculated as $\chi^2 = d^{\mathrm{T}}V^{-1}d$, where d is a vector of differences between data and simulation in the different bins and V^{-1} is the inverted covariance matrix. The covariance matrix is not invertable using all bins because the content of one bin is always constrained by all other bin contents. Therefore only four of the five bins are used for the extraction of m_{t} and it was checked that the result does not depend on the choice of the bins used for the calculation of the χ^2 value. An additional theory uncertainty is included in the fit for variations of the renormalization and factorization scales by factors of 0.5 and 2 on the MC templates. This uncertainty is evaluated on the central MadGraph +pythia sample and the relative uncertainty is applied to all other mass points.

A parabola is now fitted to the distribution of the χ^2 values as a function of the top quark mass used in the different templates. Figure 6.23 shows the distribution of $\chi^2 - \chi^2_{\text{min}}$ as a function of m_{t} for different sources of uncertainties included in the calculation. Not all points lie perfectly on the fitted parabola, some fluctuate slightly. To test if this effect can be related to a limited statistical precision on the simulation, each bin of the MC templates

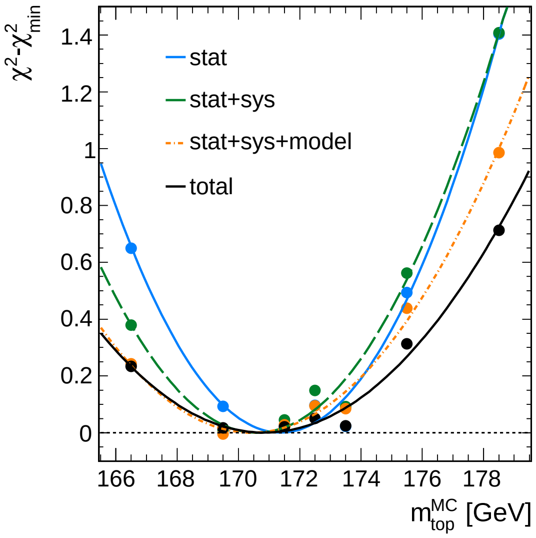

Figure 6.23: Values of $\chi^2 - \chi^2_{\mathrm{min}}$ for a comparison of the unfolded data compared to MC templates generated with different values of m_{t}. The χ^2 values are shown as a function of m_{t} and for different amounts of uncertainties included into the fit. A parabola is fitted to each $\chi^2 - \chi^2_{\mathrm{min}}$ distribution as a function of m_{t}.

is smeared randomly by a Gaussian distribution within the statistical uncertainties and the χ^2 values are recalculated. This procedure is performed 100 times and a mean χ^2 value is estimated for each value with an uncertainty defined by the Gaussian width of the distribution of the 100 χ^2 values. Only statistical uncertainties on the unfolding output are considered for this study. The resulting distribution as a function of m_{t} is shown in figure 6.24 together with the unsmeared distribution. Parabolas are fitted to both distributions. All smeared values are consistent with the fitted parabola within their uncertainties. The top quark mass is extracted from the unsmeared distribution since no significant differences for the position of the minimum of the parabolas can be observed.

The best estimator of the top quark mass is obtained by the position of the minimum of the fitted χ^2 distribution including all uncertainties. The 1 σ uncertainty is obtained by the position at which $\chi^2 - \chi^2_{\mathrm{min}}$ is equal to one. Individual uncertainty contributions are obtained by a quadratic subtraction of uncertainties derived in fits with a different amount of uncertainty sources included. This leads to a top quark mass of:

$$\begin{aligned} m_{\mathrm{t}} =& 170.8 \pm 6.0 \text{ (stat.)} \pm 2.8 \text{ (syst.)} \pm 4.6 \text{ (model)} \pm 4.0 \text{ (th.)} \ \text{GeV} \\ =& 170.8 \pm 9.0 \ \text{GeV}. \end{aligned}$$

The minimum χ^2 is 1.6 for three degrees of freedom. This is the first measurement of the top quark mass in the boosted regime for fully-merged top quark decays. The measured value of the top quark mass is consistent with current direct measurements of the top quark

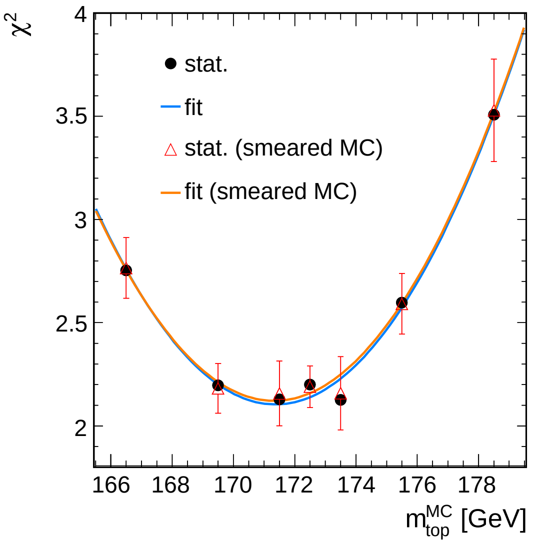

Figure 6.24: Values of χ^2 for a comparison of the unfolded data compared to MC templates generated with different values of m_t. The χ^2 values are shown as a function of m_t with only statistical uncertainties used in the fit as black points. The red points are obtained by smearing the MC templates 100 times within their uncertainties and taking the mean value. The vertical bars show the Gaussian width of the 100 χ^2 values. Parabolas are fitted to both distributions.

mass [22, 168–172]. Because of considereations mentioned in section 2.2.3 this measured value might not be directly comparable to indirect measurements of the pole mass [38, 39]. The uncertaity on the measured top quark mass is much larger compared to the direct measurements and to the indirect measurements of the pole mass. It is, however, still dominated by the statistical uncertainties and large improvements are expected on both the statistical and the systematic uncertainties for a measurement with more data at a center-of-mass energy of 13 TeV. More discussion can be found in the outlook following below.

6.12 Summary and outlook

The first measurement of the differential and the normalized differential $t\bar{t}$ production cross section as a function of the jet mass in fully-merged top quark decyas was presented in this chapter. The measurement was performed in a fiducial phase space enriched with events in which the leading jet contains a fully-merged hadronic top quark decay. Large CA jets are used with a distance parameter of $R = 1.2$, large enough to include a full top quark decay down to a p_T threshold of 400 GeV. The data has been corrected for detector effects to the particle level. The total cross section in the fiducial region was measured to be $\sigma = 101 \pm 11(\text{stat.}) \pm 13(\text{syst.}) \pm 9(\text{model})$ fb, lower than the estimated cross section in simulation with POWHEG +PYTHIA of 133^{+18}_{-28} assuming a total $\mathrm{t\bar{t}}$ production cross section of 253 pb. This is well consistent with other measurements in the boosted regime in references [153, 154] and with the observation of a softer top quark p_T spectrum in data compared to simulation in resolved measurements in references [150–152].

The measured data can be compared to MC simulations at the particle level and can help to improve the understanding of jet substructure in simulation. The ultimate goal of this measurement would be a comparison to analytic calculations from first principles at the particle level as they are performed for $\mathrm{e^+e^-}$ collisions in references [40–42] and for pp collisions in reference [50]. A comparison to analytic calculations could lead to an extraction of a well-defined top quark mass. Since no analytic calculations exist yet for the measured phase space, a top quark mass was extracted from a comparison to MC simulated templates to test the sensitivity with the 8 TeV data and to prove that the method works. This results in a mass of 170.8 ± 9.0 GeV, where the largest uncertainty contribution comes from statistical uncertainties. The measured value is well consistent with the direct measurements of the top quark mass.

The uncertainty on the extracted top quark mass is quite large with the 8 TeV data compared to the direct measurements but large improvements are expected for a similar measurement performed with 13 TeV data. The uncertainties in the 8 TeV measurement are dominated by statistical uncertainties. Much more $\mathrm{t\bar{t}}$ events with a high top quark p_T are expected at 13 TeV, not only because the data sets are larger but, more important, because the $\mathrm{t\bar{t}}$ cross section increases. This should lead to a decrease of the statistical uncertainties in the measurement. Also the systematic uncertainties are expected to decrease since the low number of events at 8 TeV had a large influence on the whole definition of the measurement. With more events it would be possible to increase the p_T threshold and to use smaller jets which should lead to a better mass resolution and a higher sensitivity on the top quark mass. Grooming methods like Soft Drop can be used to further improve the resolution. A finer binning can be used if the reconstruction resolution improves. The jet-mass scale can be studied with higher precision using more data. Modeling uncertainties, like the uncertainties on the renormalization and factorization scales, might decrease in an unfolding with more p_T bins and more side-band regions which is possible with more data.

A preliminary result of a measurement by CMS at 13 TeV can be found in reference [51] with significantly lower uncertainties. It uses a new jet reconstruction with the XCone [173] algorithm leading together with the larger dataset to uncertainties much closer to the direct measurements.

7 Measurement of top tagging efficiencies

7.1 Introduction

Top tagging is an important tool for many new-physics searches looking for heavy new particles decaying into top quarks. Top quarks from the decay of very heavy hypothetical new particles with masses much larger than the top quark mass have a high momentum. With increasing momentum of the top quark and increasing Lorentz boost the decay products of the top quark are more and more collimated in the direction of flight of the top quark. At some point a reconstruction of a hadronic top quark decay in three separate jets becomes inefficient and the decay can rather be reconstructed in one large jet. Top tagging algorithms are needed to identify large jets that contain all decay products of a hadronic top quark decay. They use jet-substructure information to separate jets containing a hadronic top quark decay from light-quark and gluon jets. Top tagging is therefore important to increase the sensitivity in many searches for new physics at high masses of the hypothetical particles. It is also used in standard model measurements in references [153, 154] to enrich the measurement phase space with high-momentum $\mathrm{t\bar{t}}$ production.

This chapter starts with studies on the performance of the CMSTopTagger v2 and HOTVR in MC simulation produced for the 2016 data taking at a center-of-mass energy of 13 TeV. Their performance is compared for two different pileup-removal techniques, for PUPPI and CHS. Following the simulation studies, a new method is used to measure the efficiency in real collision data and to determine data-to-simulation scale factors needed to correct for the difference in the efficiency between data and simulation. Previous measurements of the top tagging efficiency and scale factors for 8 TeV data [5] and for 13 TeV data collected in 2015 [174] measured general scale factors for $\mathrm{t\bar{t}}$ production in a specific measurement phase space. The disadvantage of these measurements is a potential dependence on the measurement phase space because not all selected jets in $\mathrm{t\bar{t}}$ events contain all decay products of a fully-merged hadronic top quark decay. Jets that contain a fully-merged hadronic top quark decay have a different efficiency compared to jets that contain only parts of the top decay. This might lead to different data-to-simulation scale factors for the different contributions. An inclusive scale factor for all $\mathrm{t\bar{t}}$ events can therefore depend on the numbers of events from the different contributions in the full measurement phase space.

Efficiencies and scale factors have been measured within the scope of this thesis in data collected at a center-of-mass energy of 13 TeV in the years 2016 and 2017. They are measured for 'fully-merged', 'semi-merged', and 'not-merged' $\mathrm{t\bar{t}}$ contributions as a function of the jet p_{T}. The different $\mathrm{t\bar{t}}$ contributions are defined by a matching of the top quark decay products from the MC generator to the jets in simulation. The efficiency in data is obtained by fitting the different contributions defined in simulation to the data using a maximum-likelihood fit in a pass and a fail region. The measurement of the scale factors for different $\mathrm{t\bar{t}}$ contributions makes them less dependent on the phase space they are measured in and allows an application to phase spaces different to the measurement phase space. Furthermore, the maximum-likelihood fit provides the possibility to constrain some of the systematic uncertainties on the scale factors. Mistag rates and corresponding scale factors are studied in 2016 data in addition to the signal scale factors. The mistag rate is studied using a simple cut-and-count method in a phase space enriched with QCD multijet production.

7.2 Data and simulation

Data

This chapter includes studies with data collected by the CMS detector in proton-proton collisions at a center-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$ in the years 2016 and 2017. Only certified runs are used corresponding to an integrated luminosity of $35.9\,\mathrm{fb}^{-1}$ in 2016 and $41.4\,\mathrm{fb}^{-1}$ in 2017. The two data sets are studied separately.

Simulation

Two separate productions of simulation samples have been done in CMS for the 2016 and for the 2017 data taking.

Simulations for the 2016 studies The default $\mathrm{t\bar{t}}$ simulation is produced with POWHEG v2 [97–101] for the calculation of the hard matrix element interfaced with PYTHIA 8 [49] for the parton shower and the hadronization. Single top quark production in the s-channel is simulated with MadGraph5_aMC@NLO [103] interfaced with PYTHIA 8 and the t- and tW-channels with POWHEG +PYTHIA. W boson production in association with jets is simulated with MadGraph5_aMC@NLO +PYTHIA 8 where the FXFX algorithm [175] is used to match the additional radiation to the hard matrix element. Drell Yan events in association with jets are simulated with MadGraph +PYTHIA where the MLM algorithm [176] is used for the matching of the additional radiation to the hard matrix element. QCD multijet production is simulated with PYTHIA 8.

All samples use the NNPDF 3.0 [24] PDF set. The simulations of $\mathrm{t\bar{t}}$ and single top quark production in the t-channel use the CUETP8M2T4 [177] underlying-event tune and all other samples use the CUETP8M1 [142, 178] tune.

An additional $\mathrm{t\bar{t}}$ sample for systematic studies is simulated with POWHEG interfaced with HERWIG++ [179] using the EE5C tune [180].

Additional samples are produced for studies of mistag rates. Three additional samples of QCD multijet production are simulated with MadGraph +PYTHIA, PYTHIA, and HERWIG++. The HERWIG++ sample uses the CUETHS1 tune [142]. W and Z production to qq in association with jets are simulated with MadGraph +PYTHIA for $H_\mathrm{T} > 50\,\mathrm{GeV}$ using MLM matching.

Samples of Z' production decaying into $\mathrm{t\bar{t}}$ pairs are simulated with MadGraph +PYTHIA 8 for the study of the top tagging performance in simulation. Several samples are produced for different Z' masses between 500 and 5000 GeV with a resonance width of 1 %.

Simulations for the 2017 studies The main difference to the 2016 simulation is a new underlying-event tune, the CP5 tune [181]. The NNPDF 3.1 [182] PDF set is used for the 2017 production.

The $\mathrm{t\bar{t}}$ production is simulated with POWHEG v2+PYTHIA 8. Single top quark production is simulated in the s-channel with MadGraph5_aMC@NLO interfaced with PYTHIA 8 and in the t- and tW-channels with POWHEG +PYTHIA. Drell Yan and W production in association with jets are simulated with MadGraph +PYTHIA with MLM matching. Muon enriched QCD multijet production is simulated with PYTHIA 8.

One additional $t\bar{t}$ sample with the same setup that was used for the default sample but with the CUETP8M2T4 tune instead of the CP5 tune was produced with a selection on the p_T of the generated W boson from the top quark decay larger than 150 GeV. This sample is used to study the influence of the new PYTHIA tune.

7.3 Performance in simulation

This section contains studies on the performance of the CMSTopTagger v2 and HOTVR in simulated events. It starts with a definition of signal and background processes and continues with comparisons of the tagging efficiency and mistag rate between the two taggers. Both tagging algorithms are studied with two different pileup-removal techniques CHS and PUPPI. Performance studies of the CMSTopTagger v2 have been previously performed in reference [174]. The performance studies in this section have been performed within the scope of studies towards a comparison with other top tagging algorithms in CMS in reference [134].

7.3.1 Simulation and jet definition

The performance in simulation is studied on samples simulated for the 2016 data analysis. Simulated samples of heavy Z' bosons decaying into a pair of top quarks are used to define hadronically decaying top quarks with high momentum. They contain a large number of high-momentum top quarks. Samples of QCD multijet production are used for studies of the mistag rates. These samples are obtained with PYTHIA. More information on the samples is given in section 7.2.

Two different kinds of jets are used for the CMSTopTagger v2 and the HOTVR algorithm. The CMSTopTagger v2 uses anti-k_T jets with a distance parameter of $R = 0.8$ (called AK8 in the following). Two different AK8 jet collections are used for jets with CHS or with PUPPI pileup-removal applied. Jet energy corrections for both collections have been derived within CMS and are applied as p_T and η-dependent correction factors to the four-momentum of the jet. A jet energy resolution (JER) smearing is applied to account for a different energy resolution between data and simulation. The JER smearing is provided by the CMS collaboration. Both jet collections contain subjets found by the Soft Drop algorithm. L2L3 JECs as derived for AK4 jets in CMS are applied to the subjets. The Soft Drop mass is defined as the invariant mass of the combination of the corrected subjet four-vectors. The CSVv2 algorithm is used to apply b tagging on the Soft Drop subjets.

The HOTVR algorithm uses its own jet algorithm. Jet energy corrections derived for AK4 jets in CMS are applied on the HOTVR subjets. The full HOTVR jet is built from the corrected subjets. HOTVR jets are studied with CHS and with PUPPI pileup removal.

7.3.2 Boosted top quark definition

Boosted top quarks are defined as hadronically decaying top quarks right before their decay. The performance is studied for different regions in the top quark p_T. The hadronic top quarks need to fulfill different criteria depending on the p_T range. In the low-p_T region between 300 and 470 GeV the top quarks are selected for $|\eta| < 2.4$ and the distance between the top quark and all its three decay products $\mathrm{q_i}$ has to be smaller than 1.2 ($\max(\Delta R(\mathrm{top},\mathrm{q_i})) < 1.2$). For the high-$p_\mathrm{T}$ region with $1000 < p_\mathrm{T} < 1400\,\mathrm{GeV}$ an η range of $|\eta| < 1.5$ is required and the distance between the top quark and its decay products is chosen to be smaller than 0.6 ($\max(\Delta R(\mathrm{top},\mathrm{q_i})) < 0.6$) because the top quark has a larger Lorentz boost and the decay products are more collimated. The criteria are summarized in table 7.3.2. A large jet is matched to the boosted top quarks if the distance between the jet and the top quark fulfills $\Delta R(\mathrm{jet, top}) < 0.6$, where the jet is either an AK8 jet or a HOTVR jet depending on the studied tagger.

Table 7.1: Definition of boosted top quarks for the top tagging performance studies in simulation.

low-p_T region	high-p_T region
hadronically decaying top quarks before decay	
$300 < p_\mathrm{T} < 470\,\mathrm{GeV}$ $\lvert\eta\rvert < 2.4$ $\max(\Delta R(\mathrm{top},\mathrm{q_i})) < 1.2$	$1000 < p_\mathrm{T} < 1400\,\mathrm{GeV}$ $\lvert\eta\rvert < 1.5$ $\max(\Delta R(\mathrm{top},\mathrm{q_i})) < 0.6$

7.3.3 QCD background definition

The main background for top tagging algorithms are jets from QCD multijet production. The mistag rates of the different taggers are studied on light quarks, b quarks, and gluons from the matrix-element generator. The same p_T and η criteria as for the top quarks are applied on the light particles. A jet is again matched if $\Delta R(\text{particle, jet}) < 0.6$. The jet type depends on the respective tagger.

7.3.4 Tagging variables in simulation

Figure 7.1 shows the distributions of the Soft Drop mass and the N-subjettiness ratio τ_3/τ_2 in simulation for AK8 jets matched to boosted top quarks and matched to quarks from QCD production as defined above. The distributions compare jets clustered with PUPPI with jets clustered with CHS in two p_T regions. The Soft Drop mass distributions for top quarks show two peaks, one peak close to the mass of the W boson at $\sim 80\,\mathrm{GeV}$ corresponding to jets in which just the hadronic W decay is clustered, and one close to the top quark mass at $\sim 175\,\mathrm{GeV}$ for jets in which the full top quark decay is clustered. Jets from QCD production have a falling Soft Drop mass distribution. It can be seen that the mass is shifted to higher values for CHS jets compared to the PUPPI jets especially in the region of low p_T because of a stronger sensitivity to additional radiation from pileup. As expected the τ_3/τ_2 distributions show lower values for jets matched to top quarks compared to the QCD jets. The peak of the N-subjettiness distribution for top quark jets at high values in the low-p_T region corresponds to jets which do not have a three-prong structure because only parts of the top quark decay are clustered. The N-subjettiness distributions also show larger values for CHS compared to PUPPI.

Figure 7.2 shows the jet mass of HOTVR jets in both p_T regions comparing again CHS and PUPPI jets. The jet-mass distributions show a peak close to the top quark mass for jets matched to top quarks and a falling distribution for QCD jets. Larger masses are observed for CHS jets compared to PUPPI similar to the AK8 jets because the CHS jets are more sensitive to additional radiation from pileup. More tagging variables for HOTVR jets are shown in figure 7.3 for the low-p_T region. The figure shows the f_{p_T} distribution, the τ_3/τ_2 distribution, the number of subjets, and the m_min distribution. The m_min distribution is shown only for jets with at least three subjets. All distributions show the expected behavior for jets matched to top quarks and for QCD jets (see also section 5.2.1). Significant differences are visible between the HOTVR jets with CHS and PUPPI in all distributions.

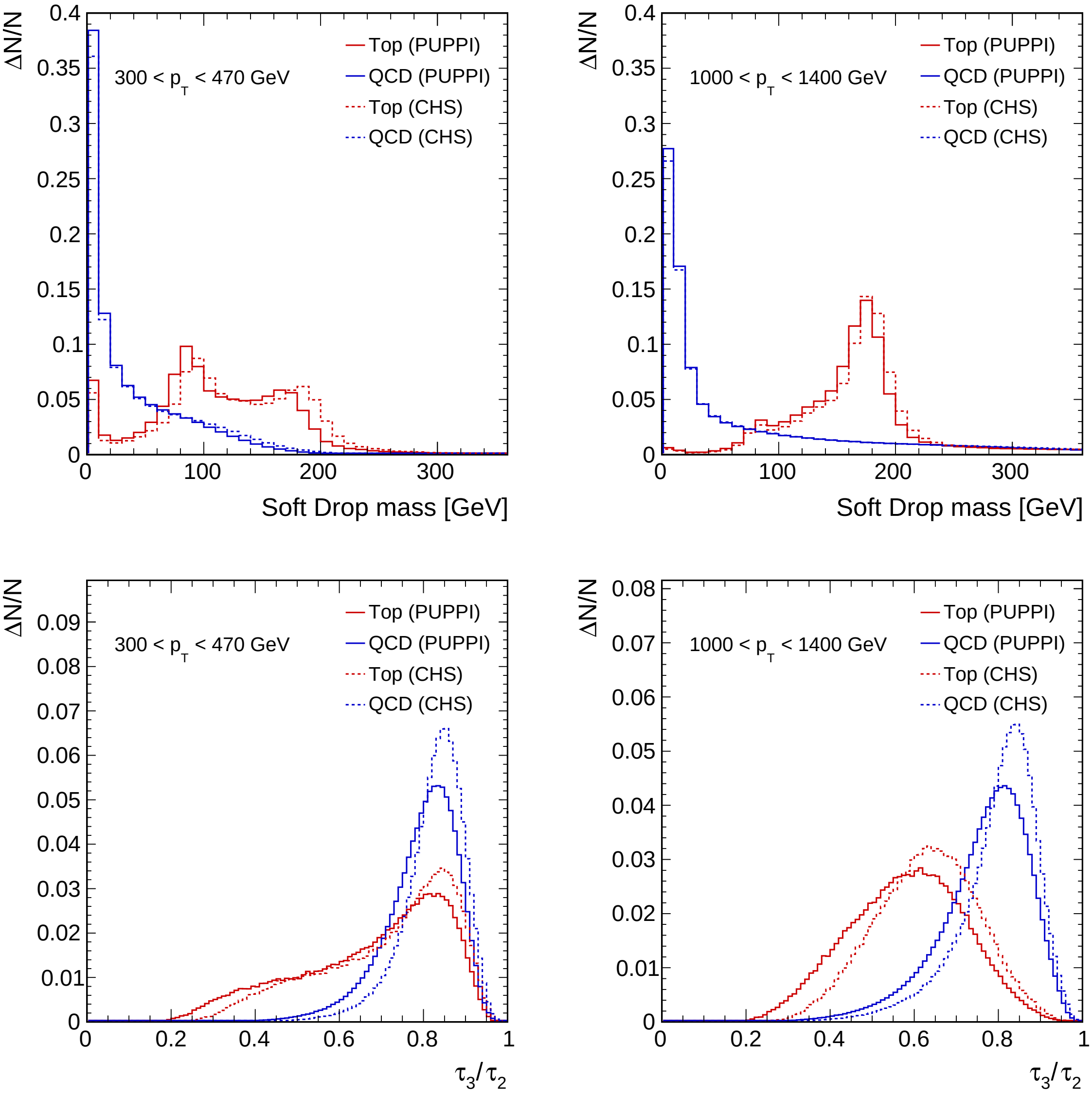

Figure 7.1: Distributions of the Soft Drop mass (top) and the N-subjettiness ratio τ_3/τ_2 (bottom) for AK8 jets in simulation. Jets matched to top quarks are compared to light-quark jets. Both distributions are shown for the low-p_{T} region (left) and for the high-p_{T} region (right). Jets clustered with PUPPI are shown as solid lines and jets with CHS as dashed lines.

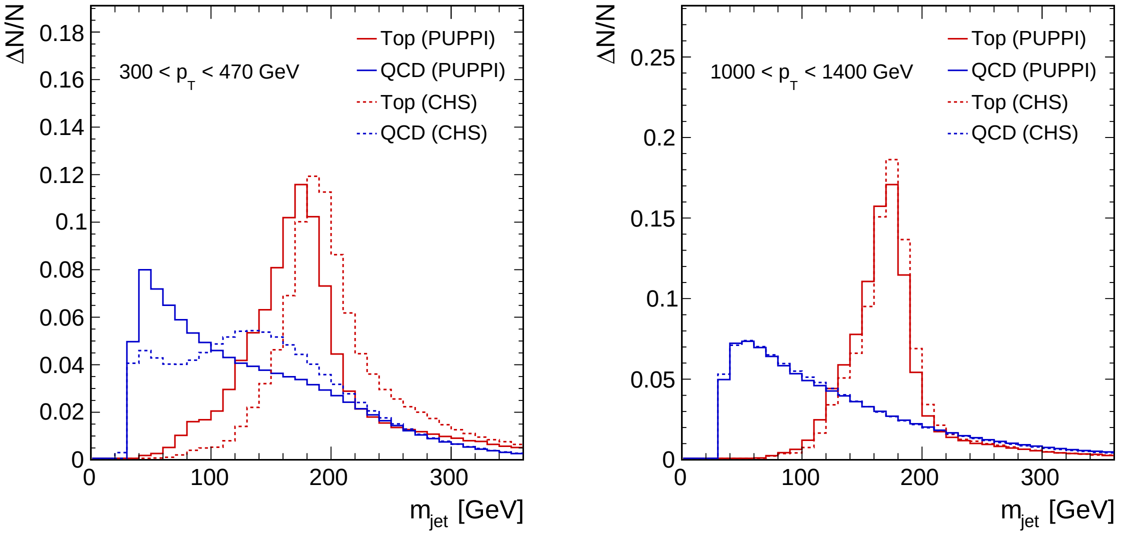

Figure 7.2: Jet-mass distributions for HOTVR jets in simulation. Jets matched to top quarks are compared to light-quark jets. HOTVR jets clustered with PUPPI (solid lines) are compared to jets clustered with CHS (dashed lines). The jet-mass distribution is shown in the low-p_{T} region (left) and in the high-p_{T} region (right).

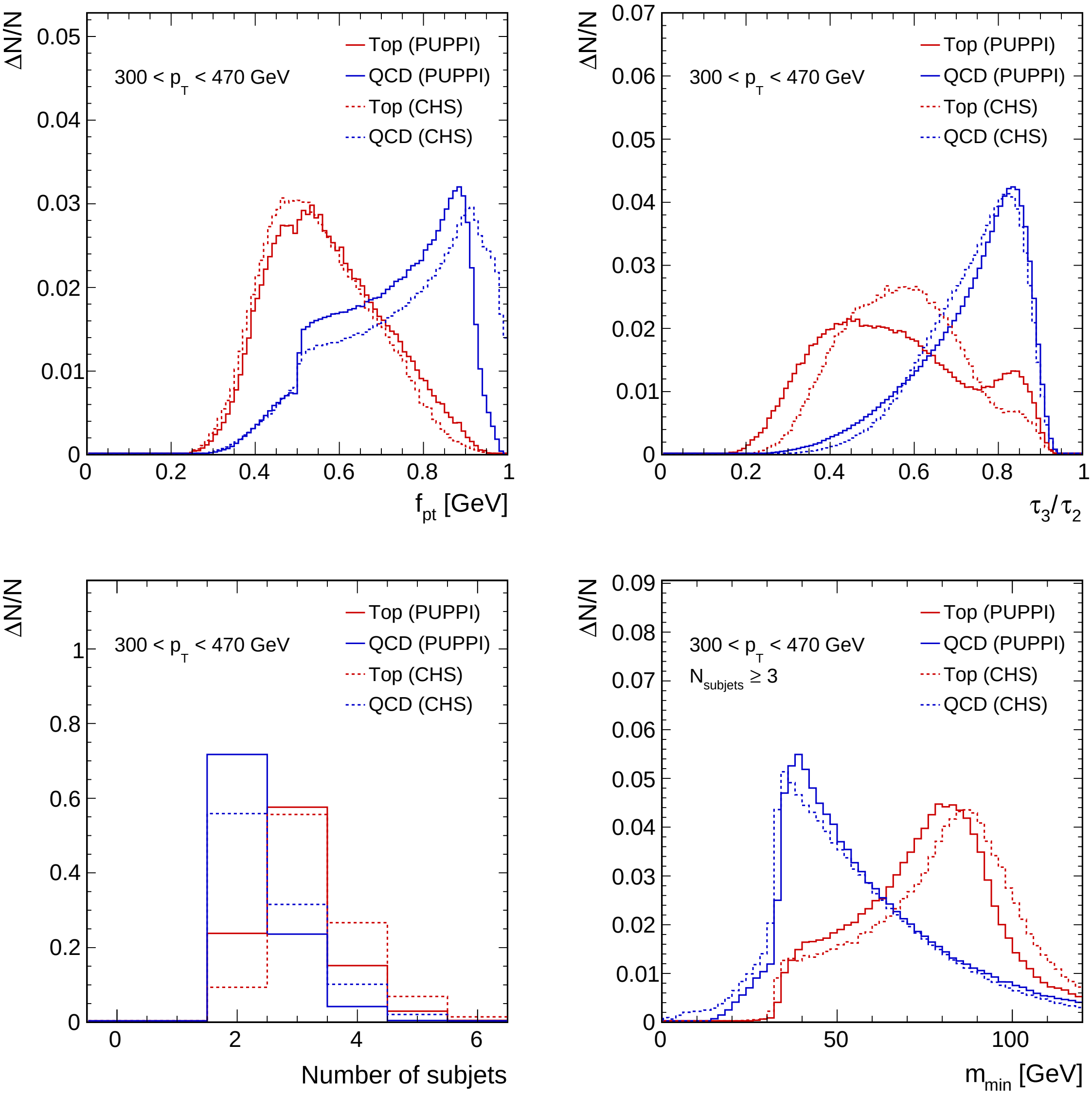

Figure 7.3: Distributions of HOTVR tagging variables in simulation for the low-p_T region. The figure shows the ratio of the leading-subjet p_T to the full jet p_T (f_{p_T}) (top left), the N-subjettiness ratio τ_3/τ_2 (top right), the number of HOTVR subjets (bottom left), and the minimum pairwise mass of the three leading subjets (m_{min}) (bottom right). The m_{min} distribution is show for jets with at least three subjets. In all distributions jets matched to top quarks are compared to light-quark jets. Jets clustered with PUPPI are shown as solid lines and jets with CHS as dashed lines.

7.3.5 Efficiency and mistag rate in simulation

The top tagging efficiency and the mistag rate are defined with respect to all boosted top quarks defined in section 7.3.2 or all light particles defined in section 7.3.3 respectively. The efficiency of matching the different kind of jets for the different taggers to the particles is part of the tagging efficiency. A particle without a matched jet is automatically not tagged. The signal efficiency is therefore defined as the number of tagged boosted top quarks divided by the number of all boosted top quarks. The mistag rate is defined consistently as the number of tagged light particles divided by the number of all light particles. Including the particles without a matched jet in the calculation allows a better comparison between taggers with different jet distance parameters because the matching efficiency can be different for different jet algorithms and should be included in the efficiency definition.

Receiver operating characteristics (ROC curves) are often used to compare the performance of different tagging algorithms. They show on the x axis the signal efficiency versus the corresponding mistag rate on the y axis. ROC curves for all studied taggers are shown in figure 7.4 for both p_T regions. Only the selection requirement on the N-subjettiness ratio τ_3/τ_2 is scanned for all taggers, all other requirements are kept fixed. The CMSTopTagger v2 shows a very similar performance with CHS and PUPPI pileup removal. Subjet b tagging clearly helps to improve the performance. The HOTVR algorithm shows a better performance in the low-p_T region for signal efficiencies larger than 10% compared to the CMSTopTagger v2 with subjet b tagging. It still shows a performance comparable to the CMSTopTagger v2 with subjet b tagging in the high-p_T region and is even slightly better for a signal efficiency above $\sim$ 35%. HOTVR with PUPPI shows a slightly better performance compared to CHS especially in the low-p_T region.

Figure 7.5 shows the signal efficiency for the different taggers as a function of the top quark p_T (top) and as a function of the number of primary vertices (NPV) (bottom). The low-p_T region is shown on the left and the high-p_T region on the right. The N-subjettiness criterion in each tagger is chosen such that the overall signal efficiency is approximately 30% to allow a better comparison between the different taggers. The CMSTopTagger v2 shows a turn-on behavior with increasing p_T in the low-p_T region because it uses a fixed jet distance parameter of $R = 0.8$ which is not sufficient to cluster the full top quark decay at low momenta. The HOTVR tagger uses a variable-R approach and is able to cluster top quarks with lower momenta in larger jets. The efficiency of HOTVR is stable as a function of p_T in both regions. The efficiency as a function of the NPV is rather flat for the jets with PUPPI pileup removal and shows a clear dependence on the NPV for jets with CHS applied. The same behavior is observed for the CMSTopTagger v2. The

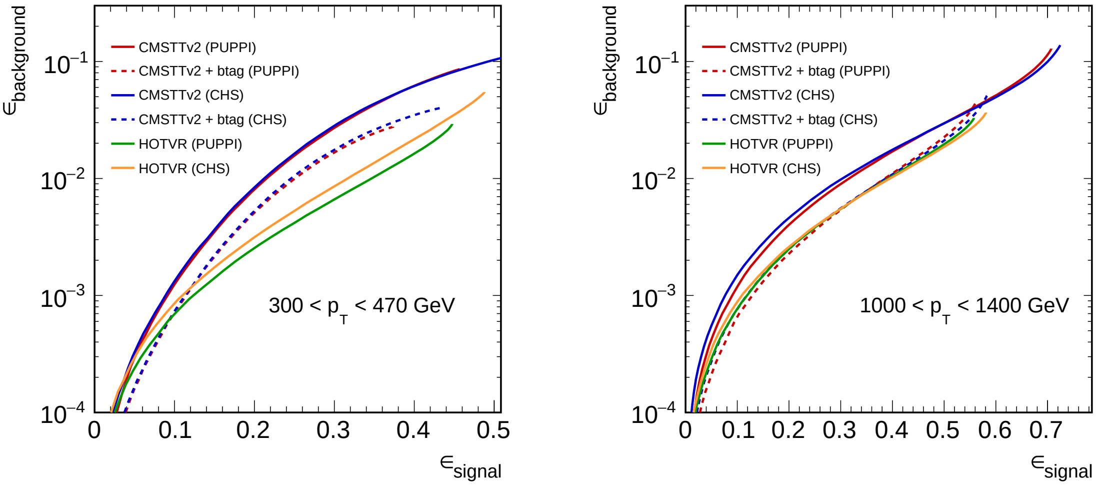

Figure 7.4: Receiver operating characteristics (ROC) in two p_{T} bins for $300 < p_{\mathrm{T}} < 470\,\mathrm{GeV}$ (left) and for $1000 < p_{\mathrm{T}} < 1400\,\mathrm{GeV}$ (right) for different taggers and different pileup-removal techniques. Only the N-subjettiness ratio τ_3/τ_2 is scanned to obtain the curves, all other tagging requirements are fixed.

PUPPI method seems to be better suited to reduce pileup effects on the tagging efficiency. Figure 7.6 shows the corresponding mistag rates as a function of the light-particle p_{T} and as a function of the NPV. A turn-on behavior as a function of p_{T} in the low-p_{T} region is observed for the CMSTopTagger v2 as before for the signal efficiency. The mistag rate for HOTVR jets is flat as a function of p_{T} in both p_{T} regions. The jets clustered with PUPPI show in general a more stable behavior against pileup compared to the jets with CHS applied. They show a lower dependence on the number of primary vertices. The mistag rate for the HOTVR tagger is in general lower compared to the CMSTopTagger v2 especially in the low-p_{T} region.

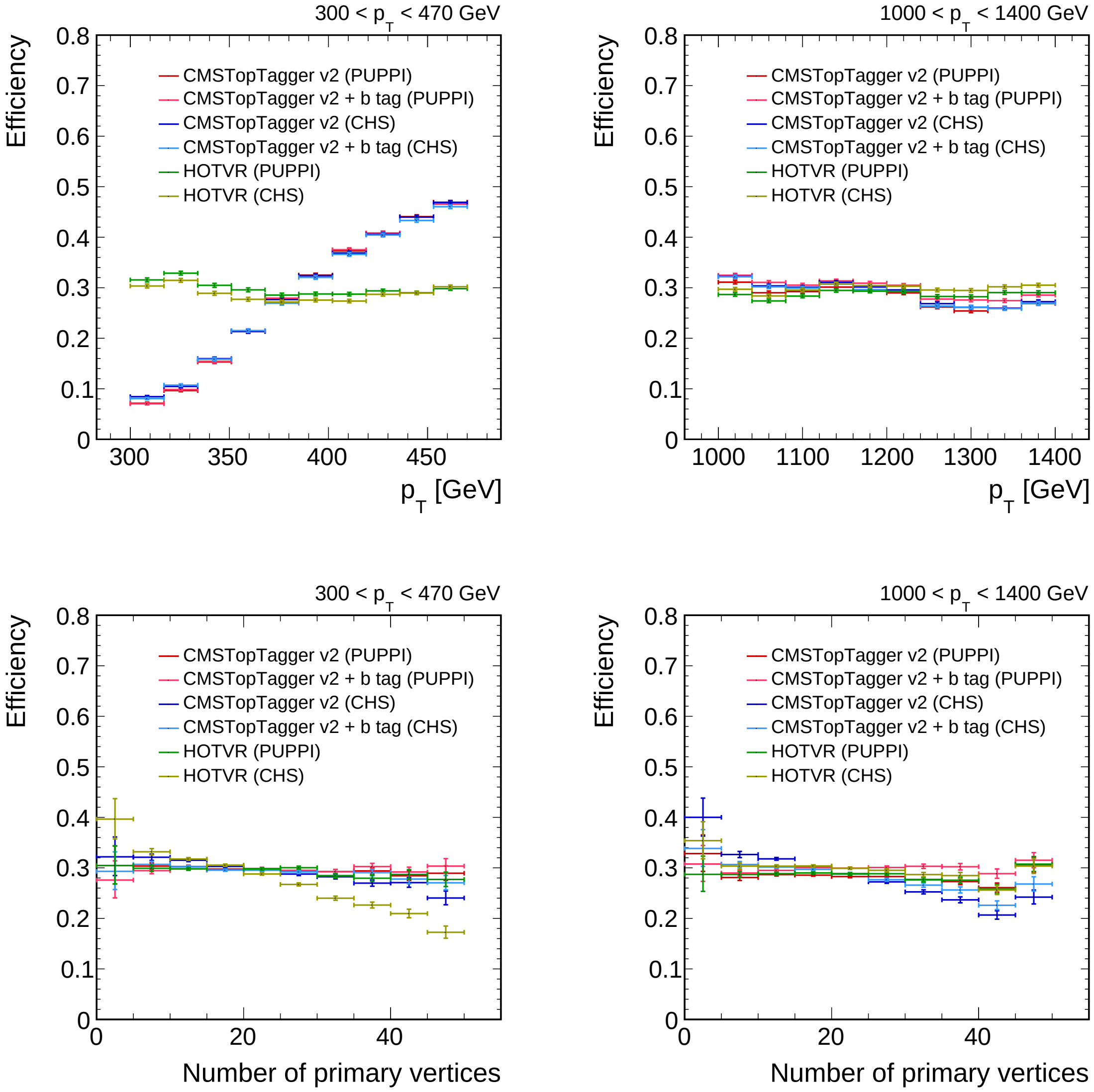

Figure 7.5: Top tagging efficiency for different top tagging algorithms as a function of the top quark p_{T} (top) and as a function the number of primary vertices (bottom). The efficiency is shown in two p_{T} regions for $300 < p_{\mathrm{T}} < 470\,\mathrm{GeV}$ (left) and for $1000 < p_{\mathrm{T}} < 1400\,\mathrm{GeV}$ (right).

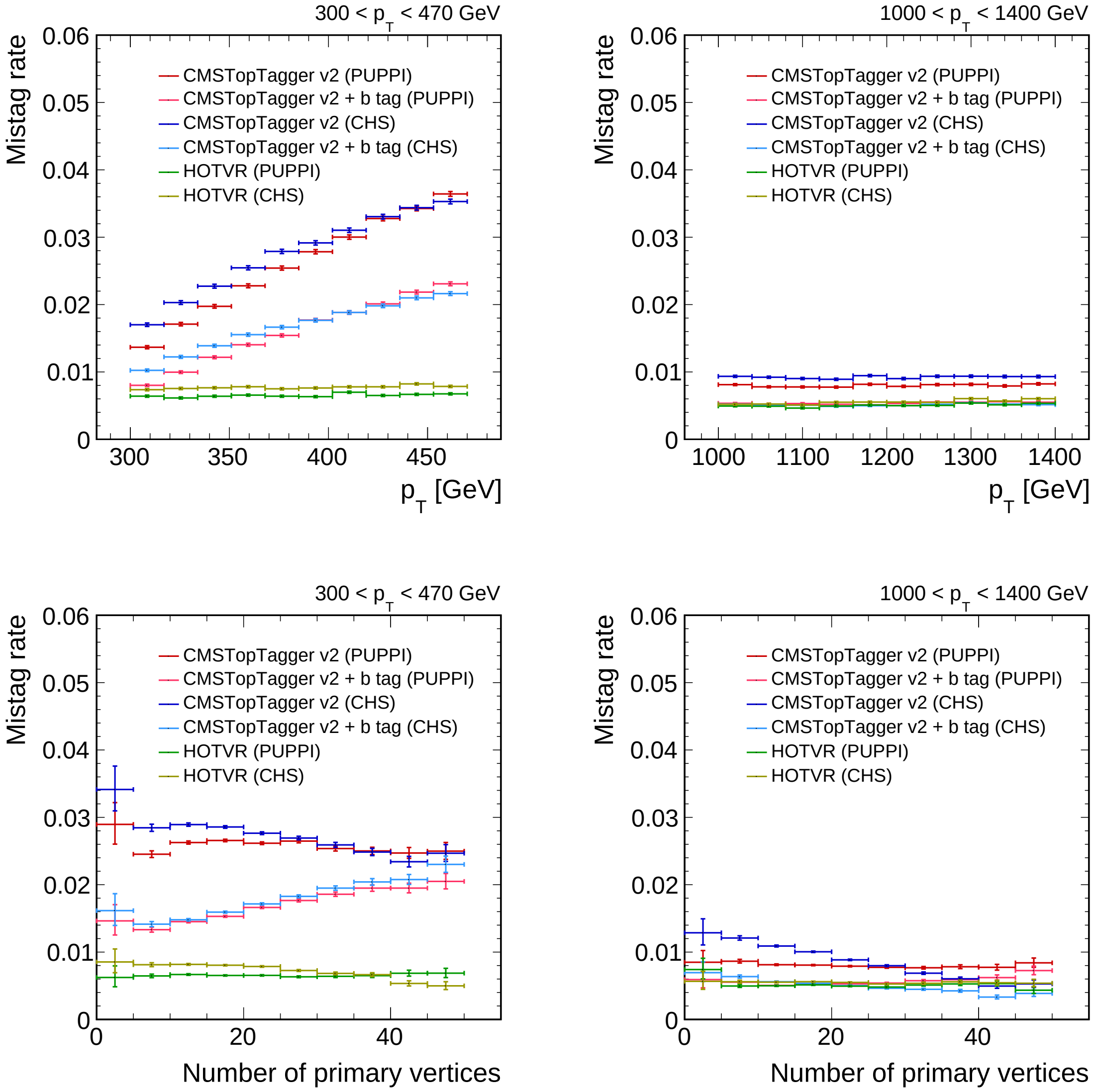

Figure 7.6: Top tagging mistag rate for different top tagging algorithms as a function of the particle p_T (top) and as a function the number of primary vertices (bottom). The efficiency is shown in two p_T regions for $300 < p_\mathrm{T} < 470\,\mathrm{GeV}$ (left) and for $1000 < p_\mathrm{T} < 1400\,\mathrm{GeV}$ (right).

7.4 Validation in 2016 data

This section includes a measurement of the tagging efficiency for the different top tagging algorithms in data collected by the CMS detector in the year 2016. The efficiency in data is compared to the efficiency in simulation and data-to-simulation scale factors are derived. Previous measurements of top tagging scale factors can be found for 8 TeV data in reference [5] and 13 TeV data collected in the year 2015 in reference [174]. The new aspect of this measurement with respect to the previous ones is a measurement of scale factors for different contributions of the $\mathrm{t\bar{t}}$ events, using a template fit method, in order to be less dependent on the measurement phase space.

7.4.1 Object definitions

Muons are selected by quality criteria corresponding to the TightID working point. The criteria are the same as for the 8 TeV analysis and are listed in section 6.3.2. Muons are considered for $p_{\mathrm{T}} > 55\,\mathrm{GeV}$ and $|\eta| < 2.4$. Scale factors to cover differences in the identification efficiency between data and simulation have been derived within CMS and are applied in this analysis [183].

The missing transverse momentum $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ is similar to the 8 TeV analysis build from the negative vectorial sum of all PF particles and corrected by a Type-1 correction [184] using anti-k_{T} jets with $R = 0.4$ and with CHS applied (see also section 6.3.2).

Anti-k_{T} jets with a distance parameter of $R = 0.4$ (called AK4 jets in the following) are used for the selection of a pure $\mathrm{t\bar{t}}$ sample. The jets are clustered from PF candidates with CHS applied. Loose quality criteria are applied on the jets. Four-momenta of leptons are subtracted from the jets to avoid a double counting of the lepton energy. The four-vector of a muon is subtracted from the four-vector of the jet if the distance between the muon and the jet is smaller than 0.4, the muon multiplicity of the jet is different from zero, and the muon energy fraction is compatible with the hypothesis that a muon is clustered to the jet. Jet energy corrections are undone for the lepton cleaning and reapplied afterwards. The AK4 jets are corrected with respective JECs derived within CMS. A JER smearing is applied to correct for differences in the JER between data and simulation. Jets are identified as b jets by the CSVv2 algorithm with a medium working point corresponding to a mistag rate of $\sim 1\,\%$. Scale factors to correct for differences in the b tagging efficiency between data and simulation have been studied within CMS and are applied in this analysis. The AK4 jets are considered for $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$.

AK8 jets are used for the evaluation of the CMSTopTagger v2 with CHS and PUPPI pileup removal. These jets have been described in section 4.4. The HOTVR tagger uses its own jets described in sections 4.4 and 7.3. The HOTVR jets are only studied with PUPPI pileup removal.

The p_T distribution of the top quarks at the generator level is reweighted for all $\mathrm{t\bar{t}}$ samples based on measurements of the top quark p_T spectrum in CMS. The reweighting is needed because the top quark p_T spectrum is measured to be softer in data compared to the simulation. It is recommended in CMS for analyses searching for new physics which are the standard use cases for the scale factors that are measured in this section. The reweighting is applied with event weights defined as

$$w = \sqrt{(\exp 0.0615 - 0.0005 p_\mathrm{T}^\mathrm{t})(\exp 0.0615 - 0.0005 p_\mathrm{T}^{\bar{\mathrm{t}}})}, \tag{7.1}$$

where p_T^t and $p_\mathrm{T}^{\bar{\mathrm{t}}}$ are the transverse momenta of the top quark and the anti-top quark. The reweighting is defined up to a p_T of 400 GeV. The factors for $p_\mathrm{T} = 400\,\mathrm{GeV}$ are used for all p_T values larger than 400 GeV.

7.4.2 Categorization of jets from $\mathrm{t\bar{t}}$ events in simulation

The $\mathrm{t\bar{t}}$ simulation is divided into three different categories by matching the three quarks from the hadronic top quark decay to a selected probe jet at the detector level. A quark from the top decay q is matched to an AK8 probe jet if the distance between the quark and the jet is smaller than the distance parameter of the jet, $\Delta R(\mathrm{jet, q}) < 0.8$. The jet is called 'fully-merged' if all three decay products are matched to the probe jet. It is called 'semi-merged' if only two of the three decay products are matched to the jet, and it is called 'not-merged' for all other $\mathrm{t\bar{t}}$ events, including misidentified dileptonic and fully-hadronic $\mathrm{t\bar{t}}$ decays.

The matching for HOTVR jets is performed in a slightly different manner, since HOTVR jets do not have a fixed distance parameter and are less conical compared to AK8 jets. A matching of the top quark decay products to the jet by a simple requirement on the distance to the HOTVR jet was found to be ineffective. A better categorization is obtained by a matching to the subjets instead of the full jet. The distance parameter used for the matching is estimated by $\sqrt{A_\mathrm{subjet}/\pi}$, where A_subjet is the area of the respective subjet in the η-ϕ plane. A HOTVR probe jet is now called 'fully-merged' if all three quarks from the top quark are matched to at least one subjet, 'semi-merged' if just two quarks are matched, and 'not-merged' in any other case.

7.4.3 Event selection

The event selection targets a pure selection of high-momentum top quarks in the muon+jets decay channel. It is therefore very similar to the selection described in section 6.6 and similar to the selection in the case of 2015 data described in reference [174].

Events are triggered by a single-muon trigger requiring one muon with $p_{\mathrm{T}} > 50\,\mathrm{GeV}$ and $|\eta| < 2.4$. Scale factors to correct for differences in the trigger efficiency between data and MC are provided by CMS and applied in this analysis.

Events are selected if they contain exactly one muon with $p_{\mathrm{T}} > 55\,\mathrm{GeV}$ and $|\eta| < 2.4$. At least two AK4 jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$ are required. A $p_{\mathrm{T}}^{\mathrm{miss}}$ larger than $50\,\mathrm{GeV}$ and a p_{T} of the leptonic W boson, built from the vectorial sum of the lepton p_{T} and $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$, larger than $150\,\mathrm{GeV}$ are required to reduce the QCD multijet background. A two-dimensional isolation criterion is applied on the muon similar to the analysis in section 6.6 to further reduce the QCD multijet background. It requires a minimum distance between the muon and the closest jet ΔR_{min} to be lager than 0.4 in a logical “or” with the requirement on the perpendicular component of the lepton momentum with respect to the closest jet ($p_{\mathrm{T,rel}}$) to be larger than $25\,\mathrm{GeV}$. At least one AK4 jet is required to be b tagged to reduce the W+jets and QCD multijet backgrounds.

7.4.4 Tag-and-probe method

The probe jets that are used to study the top tagging efficiencies are selected by a tag-and-probe method using events of the $\mathrm{t\bar{t}}$ selection described above. The tag-and-probe selection is based on the work in reference [174]. The event is divided into two hemispheres using the distance to the selected muon in ϕ. All objects within $\Delta\phi < 2/3\,\pi$ are associated to the leptonic hemisphere and all objects within $\Delta\phi > 2/3\,\pi$ are associated to the hadronic hemisphere. An event is tagged if at least one AK4 jet in the leptonic hemisphere is identified as a b jet. The probe jet is defined as the AK8 or HOTVR jet with the highest p_{T} in the hadronic hemisphere depending on the algorithm that is studied. A loose selection on the Soft Drop mass is applied on AK8 jets, requiring $m_{\mathrm{SD}} > 10\,\mathrm{GeV}$. A sketch of the tag-and-probe selection can be found in figure 7.7.

Some control distributions comparing data and simulation after the full selection are shown for AK8 jets in figure 7.8 with PUPPI and in figure 7.9 with CHS. In these distributions the $\mathrm{t\bar{t}}$ simulation is scaled to the data to allow a better shape comparison.

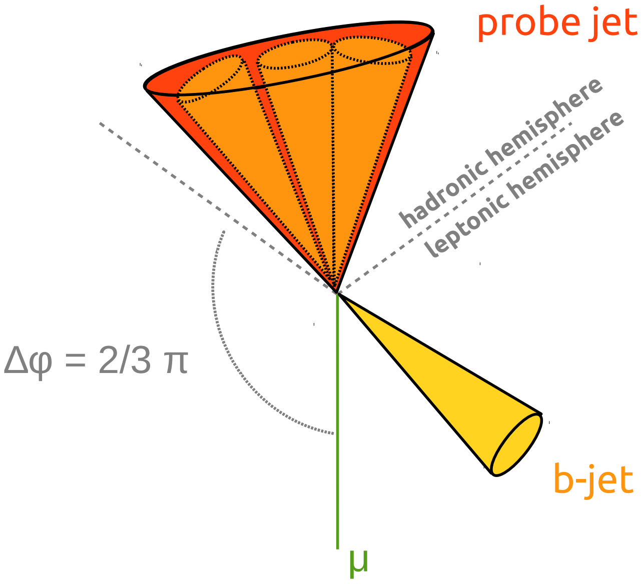

Figure 7.7: Sketch of the tag-and-probe method to select the probe jets for top tagging studies in data.

The scaling is needed because of a softer top quark p_{T} spectrum in data compared to simulation leading to a more $\mathrm{t\bar{t}}$ events in simulation for a high top quark p_{T}. This effect is not fully covered by the reweighting of the top quark p_{T} spectrum for high momenta larger than 400 GeV. The top quark p_{T} spectrum was measured in references [150–154] for different ranges in the top quark p_{T}. Both figures show a trend in the jet-p_{T} spectrum which is expected because of the softer top quark p_{T} spectrum in data. The η and the Soft Drop mass distributions are consistent between data and simulation within the uncertainties. The N-subjettiness distribution shows a small disagreement for high values where light jets dominate. This discrepancy was also observed in a measurement in resolved $\mathrm{t\bar{t}}$ decays in reference [185]. The same distributions for HOTVR jets can be found in figure 7.10 showing a similar trend in the jet-p_{T} distribution and the same disagreement for high values of the N-subjettiness distribution. Additional distributions for HOTVR jets showing the $f_{p_{\mathrm{T}}}$ distribution, the number of subjets, and the minimum pairwise mass can be found in figure 7.11. They show good agreement between data and simulation.

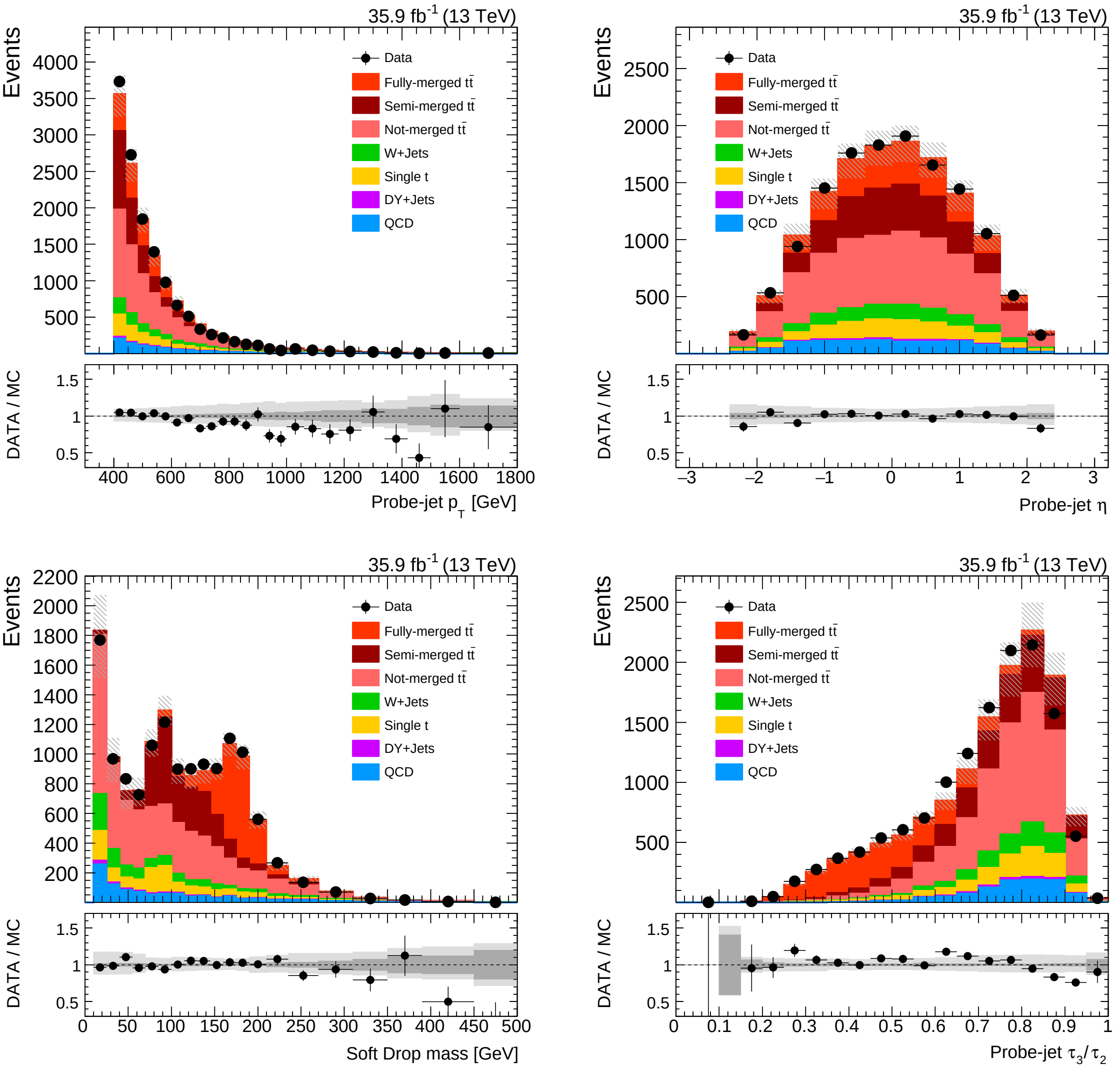

Figure 7.8: Control distributions in data and simulation for AK8 jets with PUPPI and a jet p_T larger than 400 GeV. The full event selection is applied. The jet-p_T distribution is shown on the top left, the η distribution on the top right, the Soft Drop mass distribution on the bottom left, and the N-subjettiness ratio on the bottom right. The $\mathrm{t\bar{t}}$ simulation is scaled to the data. The data is shown as black points and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution. The dark gray area shows the statistical uncertainty on the simulation, and the light gray area shows the total uncertainty including systematic uncertainties.

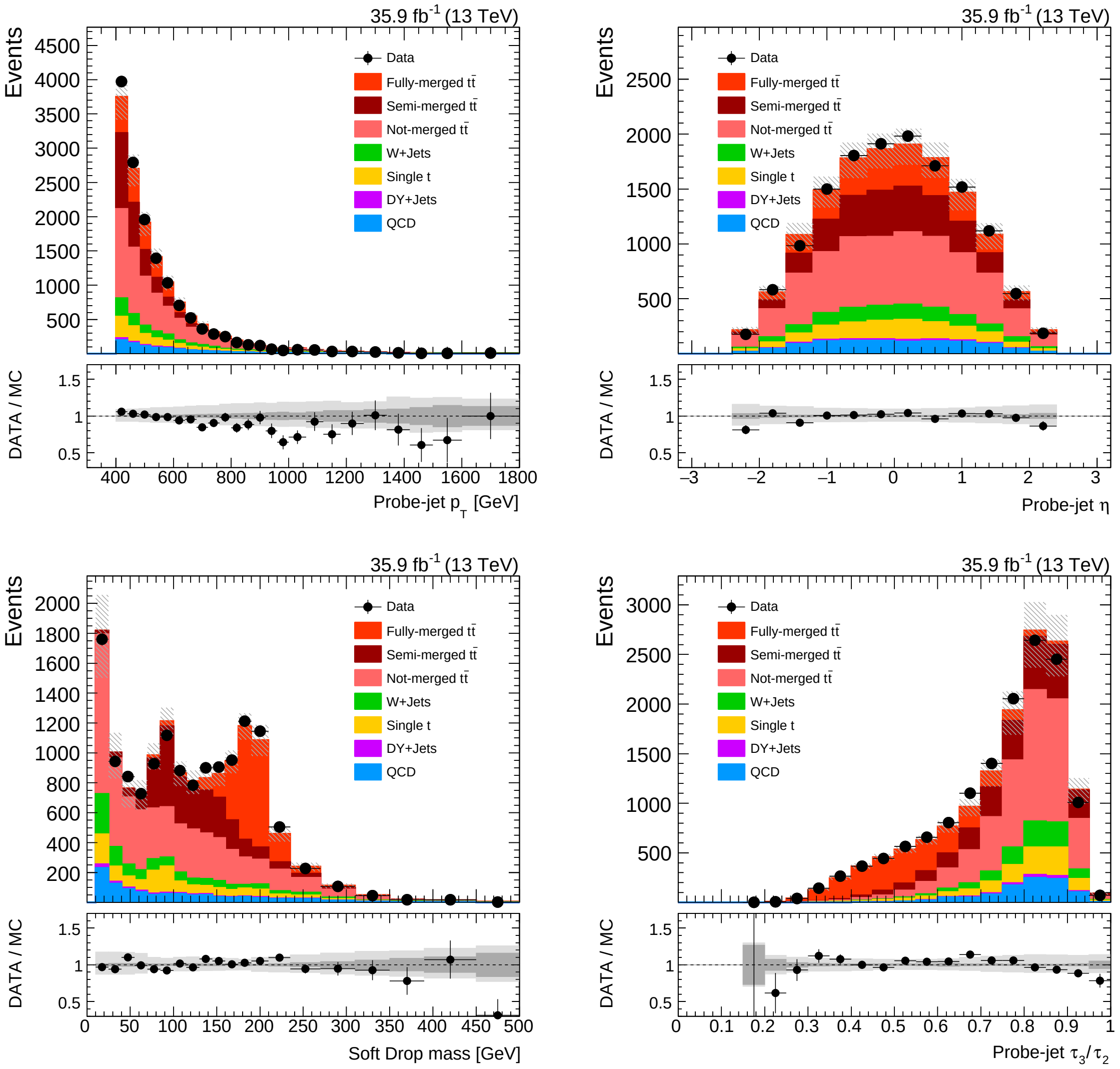

Figure 7.9: Control distributions in data and simulation for AK8 jets with CHS and a jet p_{T} larger than 400 GeV. The full event selection is applied. The jet-p_{T} distribution is shown on the top left, the η distribution on the top right, the Soft Drop mass distribution on the bottom left, and the N-subjettiness ratio on the bottom right. The $t\bar{t}$ simulation is scaled to the data. The data is shown as black points and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution. The dark gray area shows the statistical uncertainty on the simulation, and the light gray area shows the total uncertainty including systematic uncertainties.

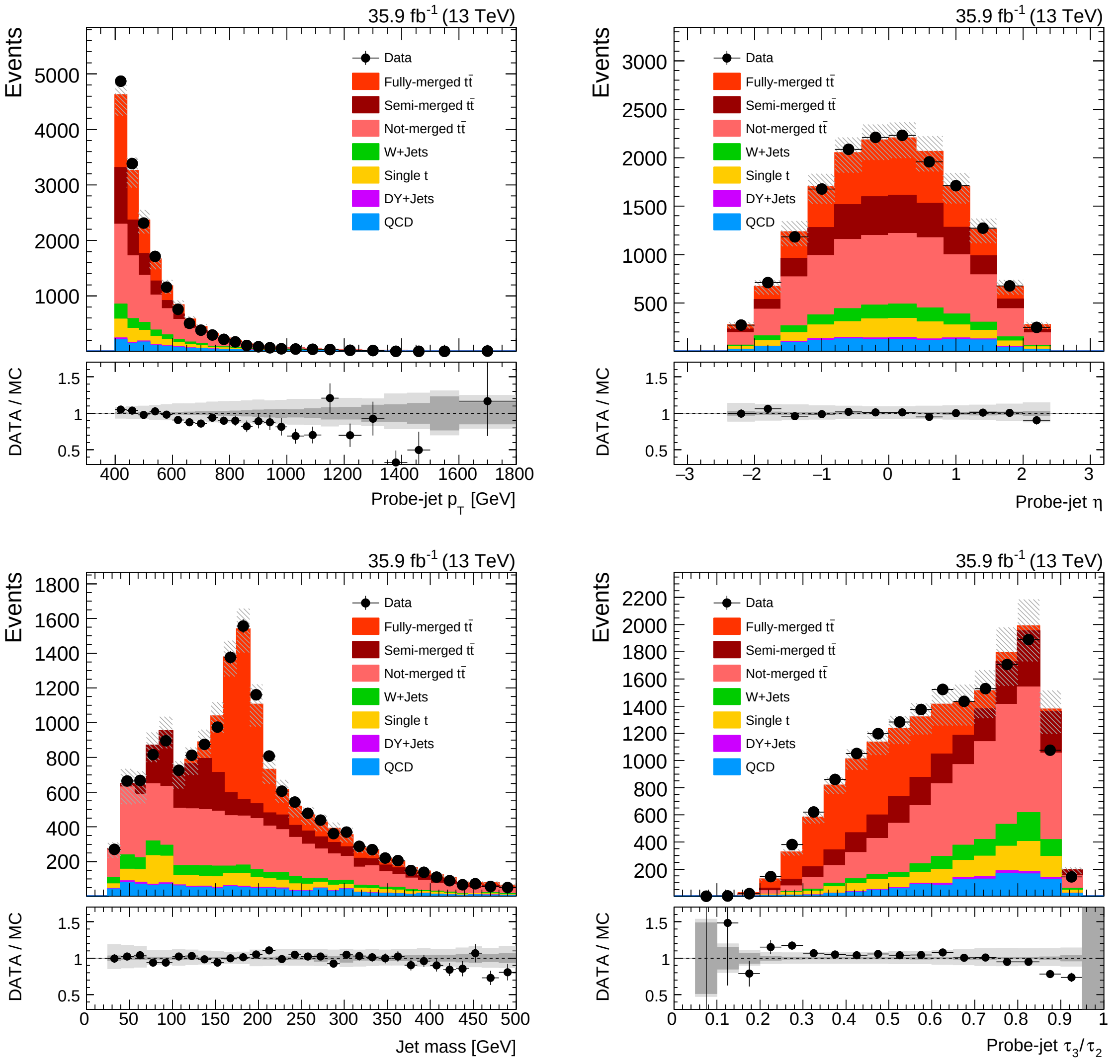

Figure 7.10: Control distributions in data and simulation for HOTVR jets with PUPPI and a jet p_{T} larger than 400 GeV. The full event selection is applied. The jet-p_{T} distribution is shown on the top left, the η distribution on the top right, the jet-mass distribution on the bottom left, and the N-subjettiness ratio on the bottom right. The t$\bar{\mathrm{t}}$ simulation is scaled to the data. The data is shown as black points and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution. The dark gray area shows the statistical uncertainty on the simulation, and the light gray area shows the total uncertainty including systematic uncertainties.

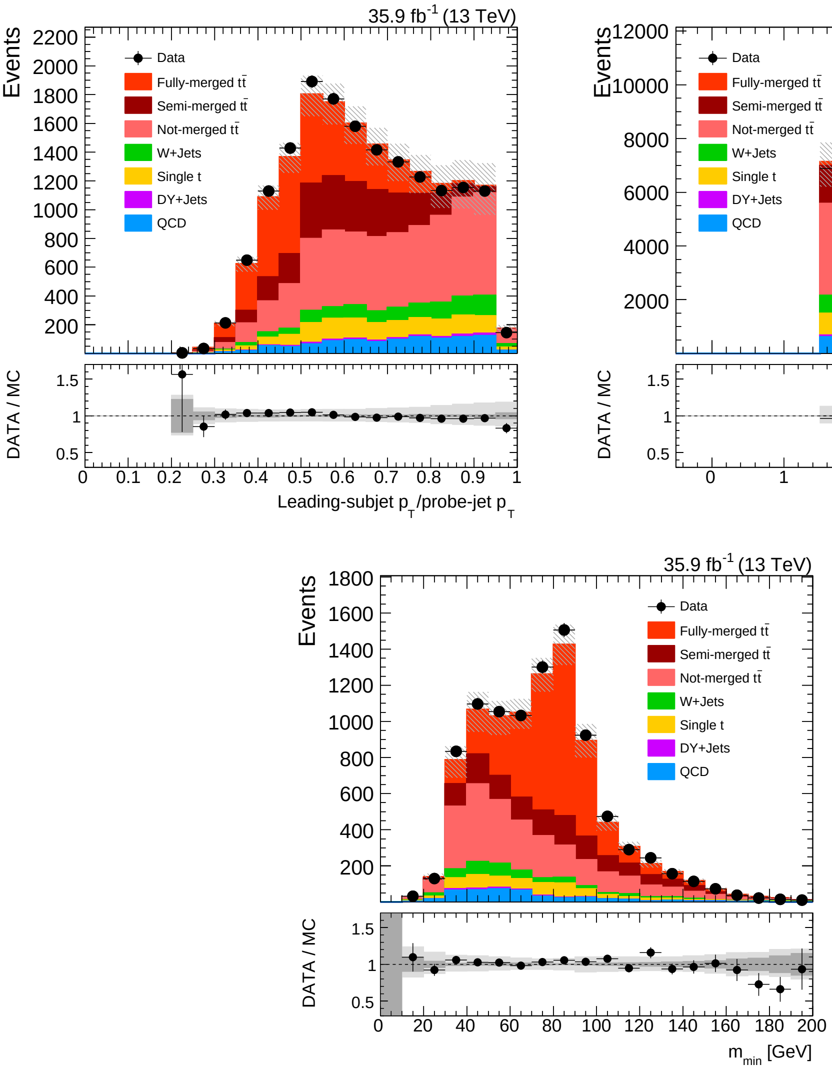

Figure 7.11: Control distributions in data and simulation for HOTVR jets with PUPPI and a jet p_{T} larger than 400 GeV. The full event selection is applied. The $f_{p_{\mathrm{T}}}$ distribution is shown on the top left, the number of subjets on the top right, and the minimum pairwise mass m_{min} on the bottom. The $\mathrm{t\bar{t}}$ simulation is scaled to the data. The data is shown as black points and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution. The dark gray area shows the statistical uncertainty on the simulation, and the light gray area shows the total uncertainty including systematic uncertainties.

7.4.5 Systematic uncertainties

This section includes a list of systematic uncertainties considered for the efficiency measurement. The following list includes only systematic effects that have a significant effect on the top tagging efficiency.

- The jet energy corrections on the AK4 jets and on the probe jets are varied simultaneously within the uncertainties. The uncertainties on the energy corrections on the probe-jet subjets are treated as fully correlated. It was checked that the fully-correlated treatment is the conservative choice.
- The jet energy resolution smearing is varied within the uncertainties simultaneously on the AK4 jets and on the AK8 probe jets.
- The uncertainty on the PDFs is estimated by a reweighting of the events by a weight for each of the 100 eigenvectors of the PDF set on $\mathrm{t\bar{t}}$ simulation. The RMS of resulting variations in each bin of the jet-mass distribution is taken as the PDF uncertainty on the jet-mass distribution.
- The renormalization and factorization scales μ_r and μ_f are varied independently by factors of 0.5 and 2.
- An uncertainty on the parton shower is estimated by a comparison of a $\mathrm{t\bar{t}}$ sample simulated with POWHEG +HERWIG++ with the default POWHEG +PYTHIA sample. The POWHEG +HERWIG++ sample is used as a systematic template in the following fits.

Variations of the muon ID scale factors, the trigger scale factors, and the b tagging scale factors within uncertainties have been studied but they are not used in the fits in section 7.4.6 because they show no significant effect on the efficiency measurement.

7.4.6 Template fits

The new aspect of this analysis with respect to previous studies of top tagging efficiencies is a simultaneous extraction of efficiencies and resulting scale factors for the three different jet categories of the simulated $\mathrm{t\bar{t}}$ events as defined above. The simulated templates for the different categories are fitted to the data using a maximum-likelihood method in the THETA framework [186]. The fit is performed in the Soft Drop mass or HOTVR jet-mass distribution in a pass and a fail region, passing or failing the top tag. The normalizations of the templates for the different categories in the pass and in the fail region are fitted and efficiencies and scale factors as a function of the probe-jet p_{T} for each $\mathrm{t\bar{t}}$ category are calculated from the number of events in the pass and fail regions. Previous measurements of top tagging scale factors in references [5] and [174] derived inclusive scale factors for all categories. The inclusive treatment might lead to a dependence on the phase space in which the efficiencies and scale factors are measured because the different categories have different efficiencies and therefore potentially different scale factors. The template fit method was chosen to be less dependent on the chosen phase space. Furthermore, it provides the possibility to constrain some systematic effects by the fit. For example, variations of the jet mass due to variations of the jet energy scale within uncertainties can be directly constrained by the data.

Maximum likelihood fits with the Theta framework

The simulated templates of the jet-mass distribution for the different $\mathrm{t\bar{t}}$ contributions are fitted to the data using a simultaneous maximum-likelihood fit in a pass and fail region in Theta. No selection criterion on the jet mass is included at this stage to use the full potential of the shape differences between the different categories in the fit. This leads to better constraints on the 'semi-merged' and 'not-merged' contributions. An estimation of the efficiency of the mass selection criterion is performed at a later stage. The normalization of each $\mathrm{t\bar{t}}$ category is determined separately in the pass and fail regions and for each category. Several fits are performed in different bins of the probe-jet p_{T}.

Systematic uncertainties are included in the fits by simulated templates with systematic variations. A template morphing is used in the fits introducing one nuisance parameter per systematic uncertainty. The normalizations of all systematic $\mathrm{t\bar{t}}$ simulation templates are scaled to the nominal $\mathrm{t\bar{t}}$ template before they are handed to the fit. This normalization is done because uncertainties affecting the overall $\mathrm{t\bar{t}}$ normalization do not contribute to the uncertainties on the tagging efficiency. Only differences in the normalization between the pass and the fail region have an influence on the tagging efficiency.

The template morphing and the scaling of the different contributions with respect to the post-fit nuisance parameters is applied to all simulation templates as done within the Theta framework. For each template morphing a small shift is calculated in each bin. The uncertainty on the shift in each bin is estimated by a Gaussian error propagation of the uncertainty on the respective nuisance parameter to the shift. A covariance matrix is built for each systematic uncertainty assuming each systematic uncertainty is fully correlated or anti-correlated between different bins depending on the direction of the shifts in the respective bins. All covariance matrices are added to the full uncertainty.

The numbers of events in the pass and fail regions are obtained by the integral over the respective region. The uncertainties on these numbers are obtained by an integral over the respective covariance matrices.

Figure 7.12 shows the distributions of the Soft Drop mass in data and in simulation in the pass region (top) and in the fail region (bottom). The figures on the left show the distributions before the maximum-likelihood fit and the figures on the right show the same distributions after the fits. The $\mathrm{t\bar{t}}$ simulation is scaled to the data before the fits in each region in order to obtain good starting values for the fit. All figures show selected AK8 probe jets with PUPPI applied for $400 < p_{\mathrm{T}} < 480\,\mathrm{GeV}$. Figure 7.13 shows the same for AK8 jets with CHS and figures 7.14 and 7.15 show HOTVR jets with PUPPI for $300 < p_{\mathrm{T}} < 400\,\mathrm{GeV}$ and for $p_{\mathrm{T}} > 600\,\mathrm{GeV}$. In general, the post-fit distributions show a better agreement between data and simulation as expected. The uncertainties on the plots before the fit do not contain uncertainties on the normalization of the $\mathrm{t\bar{t}}$ contributions. The normalizations of the individual contributions are not known and only constrained by the fit. The constrained uncertainties on the normalization of the different contributions are included in the systematic uncertainties after the fit. Therefore it is possible that the uncertainties in the plots after the fit are larger compared to the uncertainties before the fit.

Post-fit nuisance parameters for one example fit are shown in figure 7.16. Each nuisance parameter starts at zero with an uncertainty of one. The uncertainties of the nuisance parameters before the fit are shown as green (1σ) and yellow (2σ) bands. A post-fit value of zero with an uncertainty consistent with the 1σ band indicates that the parameter can not be constrained by the fit. Post-fit uncertainties can be significantly constrained by the fit if the fit is sensitive to the respective effect. The nuisance parameters on the normalizations of some of the $\mathrm{t\bar{t}}$ contributions are strongly constrained by the fit. This is the case because the priors are chosen large and the main purpose of the fits is to constrain the normalizations of the $\mathrm{t\bar{t}}$ categories. The JEC uncertainty is also constrained by the fit because the variation of the JECs within the uncertainties shift the top quark mass peak. The uncertainty on the shower model is constrained because the shape of the jet-mass distribution is different with HERWIG compared to PYTHIA and not consistent with data. The rest of the nuisance parameters can not be constrained by the fit.

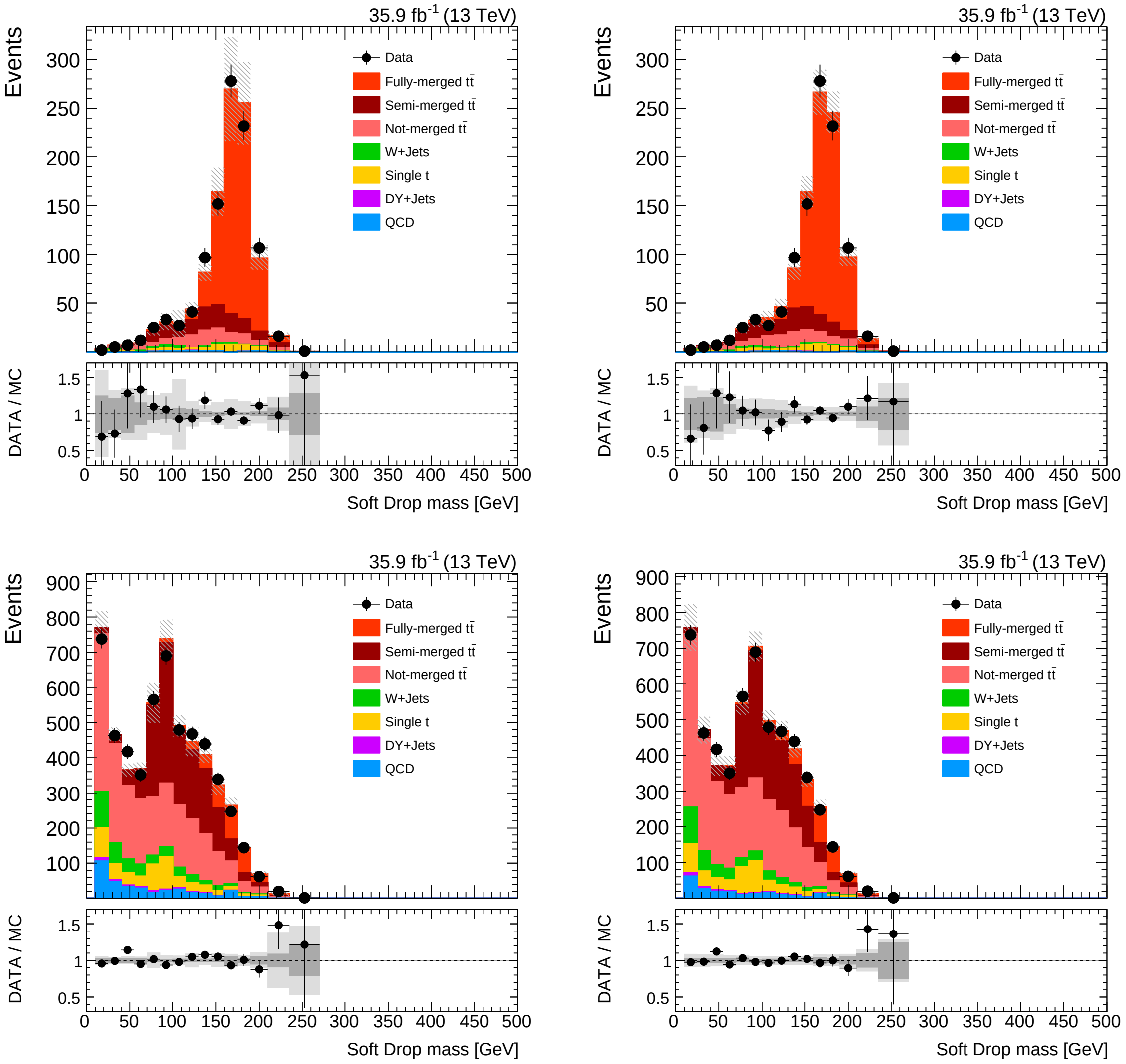

Figure 7.12: Soft Drop mass distributions for AK8 probe jets with PUPPI applied for $400 < p_{\mathrm{T}} < 480\,\mathrm{GeV}$. Distributions for the pass region are shown on the top and for the fail region on the bottom. The distributions on the left are shown before the maximum-likelihood fit and the distributions on the right after the fit. Data is shown as black dots with vertical bars showing the statistical uncertainties on the data. Simulation is shown as filled histograms with the hatched area showing the total uncertainty on the simulation. A ratio of data divided by simulation is shown under each distribution. The dark gray band shows the statistical uncertainty on the simulation and the light gray band the total uncertainty.

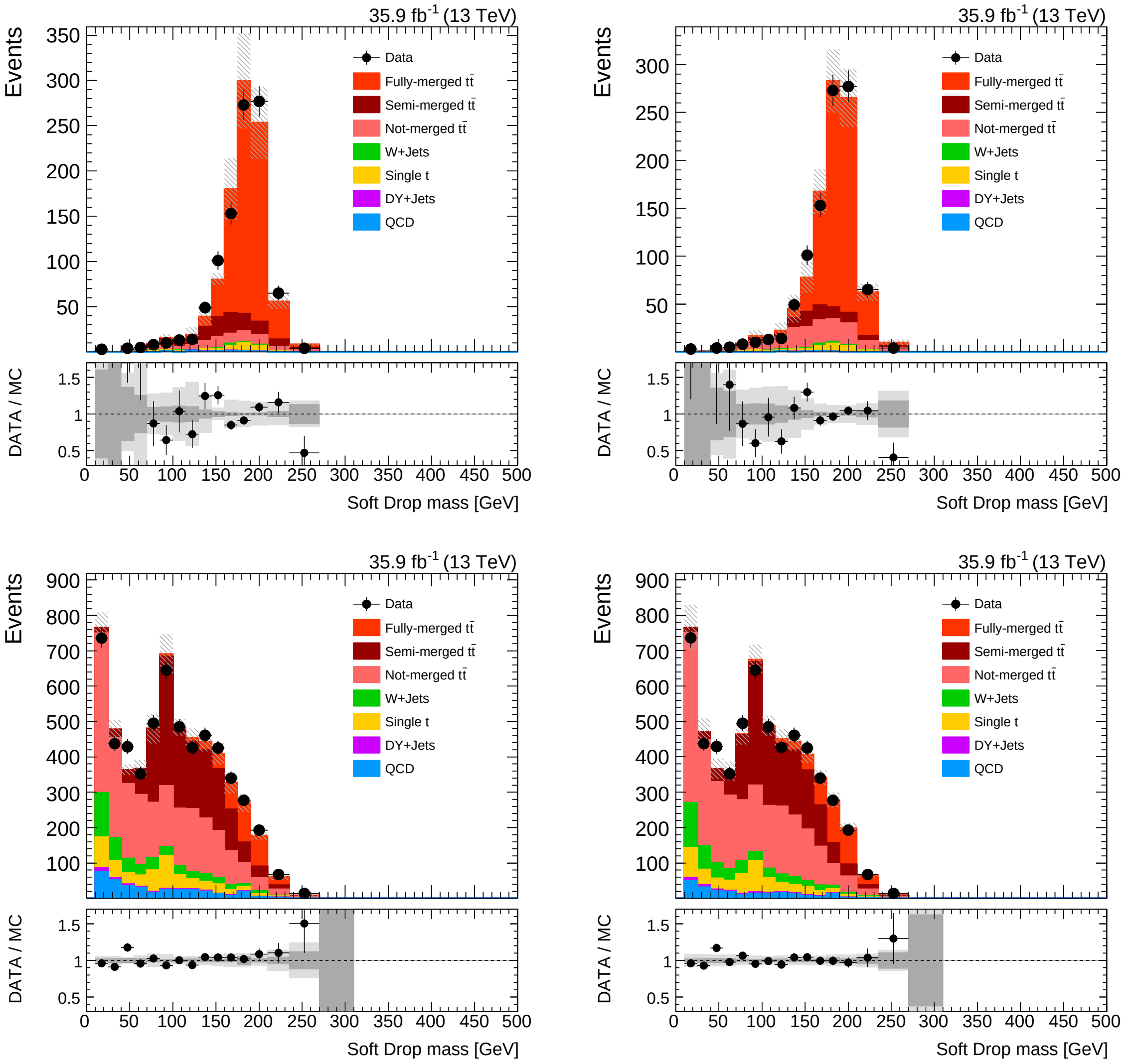

Figure 7.13: Soft Drop mass distributions for AK8 probe jets with CHS applied for $400 < p_{\mathrm{T}} < 480\,\mathrm{GeV}$. Distributions for the pass region are shown on the top and for the fail region on the bottom. The distributions on the left are shown before the maximum-likelihood fit and the distributions on the right after the fit. Data is shown as black dots with vertical bars showing the statistical uncertainties on the data. Simulation is shown as filled histograms with the hatched area showing the total uncertainty on the simulation. A ratio of data divided by simulation is shown under each distribution. The dark gray band shows the statistical uncertainty on the simulation and the light gray band the total uncertainty.

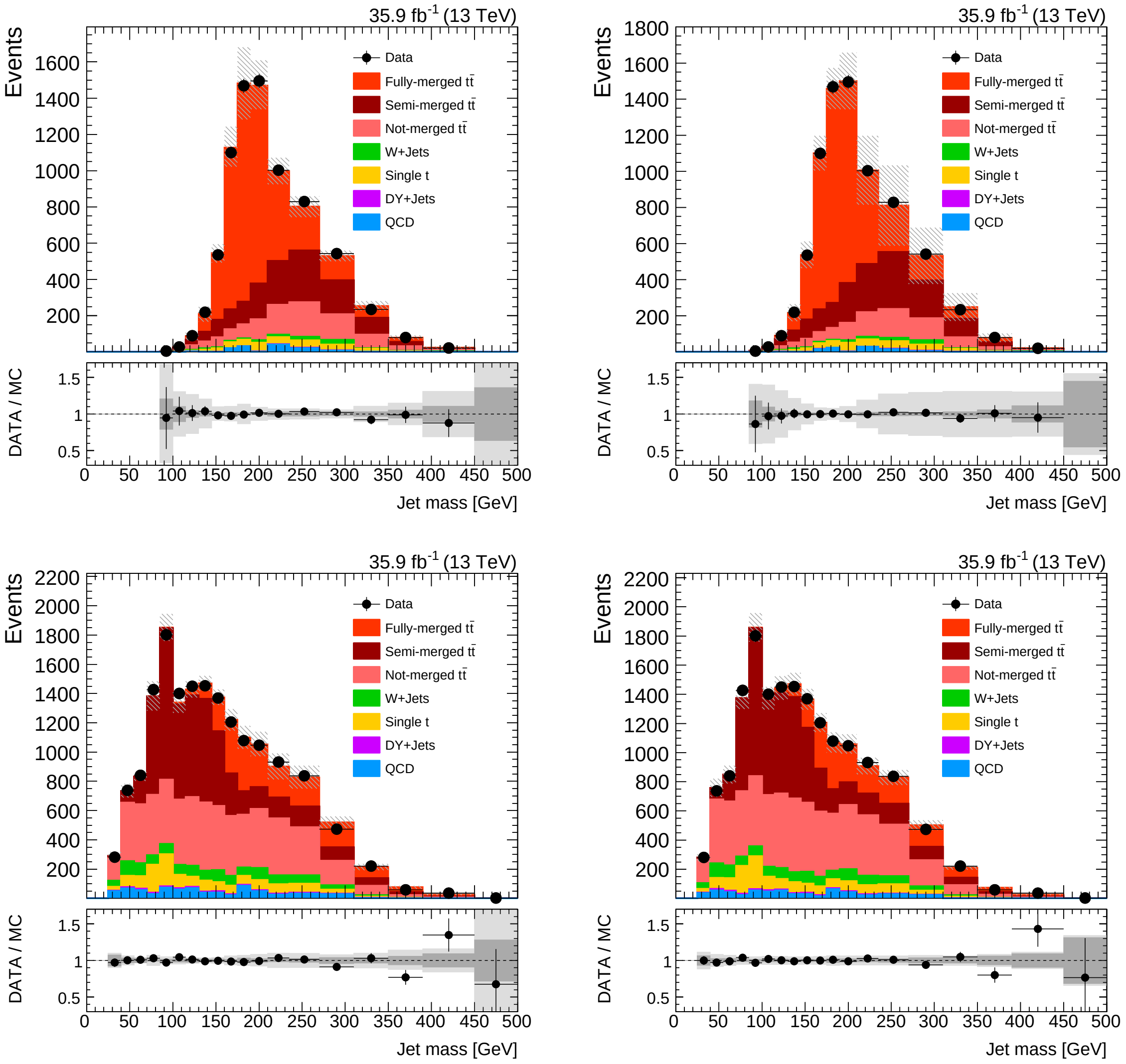

Figure 7.14: Jet-mass distributions for HOTVR probe jets with PUPPI applied for $300 < p_{\mathrm{T}} < 400\,\mathrm{GeV}$. Distributions for the pass region are shown on the top and for the fail region on the bottom. The distributions on the left are shown before the maximum-likelihood fit and the distributions on the right after the fit. Data is shown as black dots with vertical bars showing the statistical uncertainties on the data. Simulation is shown as filled histograms with the hatched area showing the total uncertainty on the simulation. A ratio of data divided by simulation is shown under each distribution. The dark gray band shows the statistical uncertainty on the simulation and the light gray band the total uncertainty.

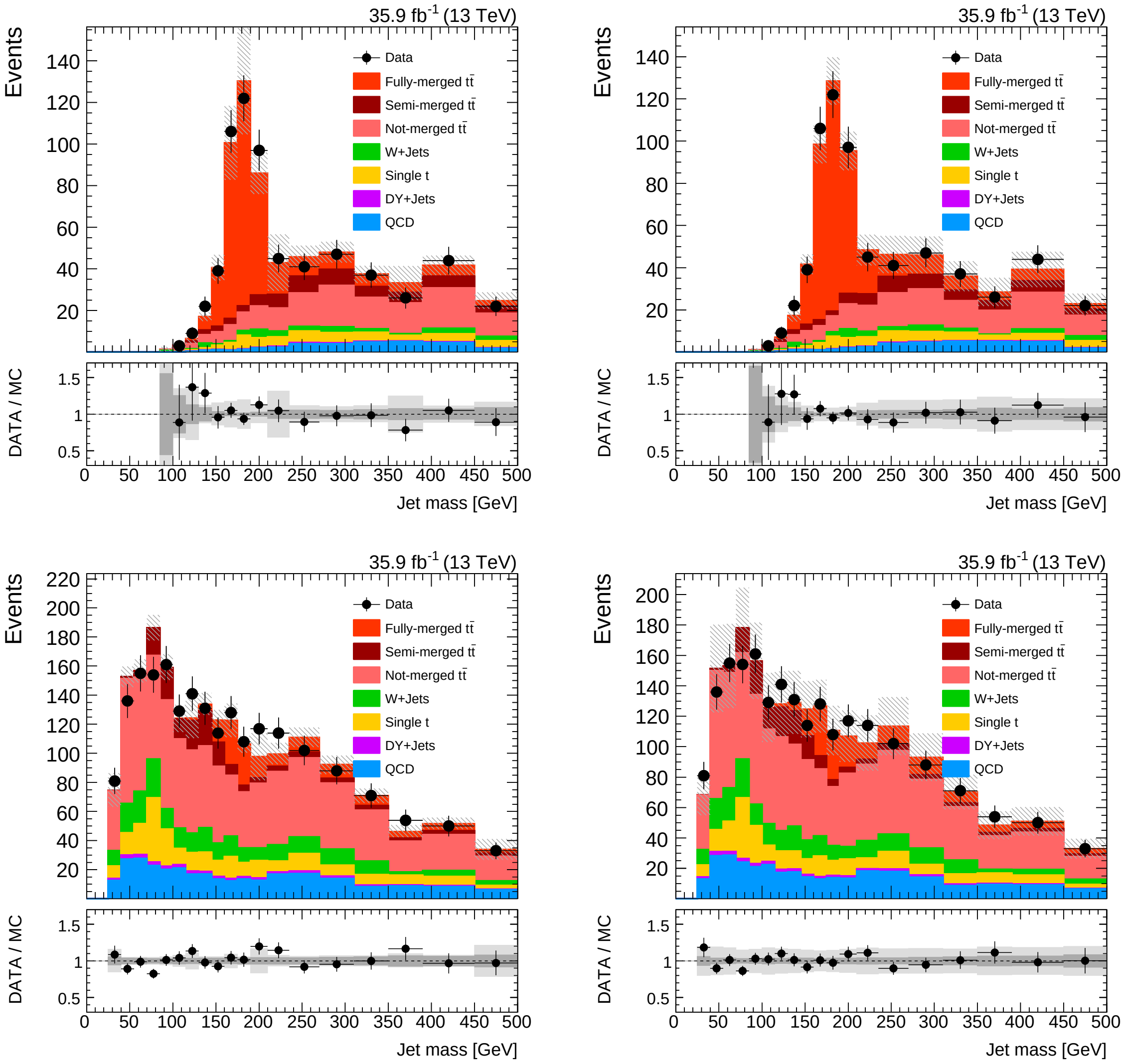

Figure 7.15: Jet-mass distributions for HOTVR probe jets with PUPPI applied for $p_\mathrm{T} > 600\,\mathrm{GeV}$. Distributions for the pass region are shown on the top and for the fail region on the bottom. The distributions on the left are shown before the maximum-likelihood fit and the distributions on the right after the fit. Data is shown as black dots with vertical bars showing the statistical uncertainties on the data. Simulation is shown as filled histograms with the hatched area showing the total uncertainty on the simulation. A ratio of data divided by simulation is shown under each distribution. The dark gray band shows the statistical uncertainty on the simulation and the light gray band the total uncertainty.

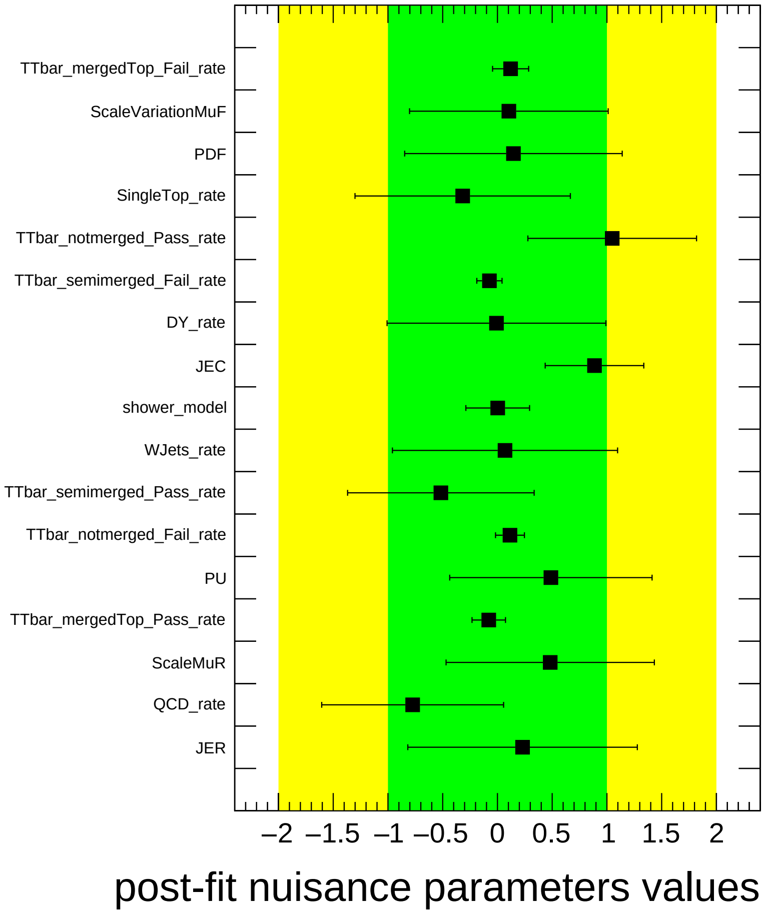

Figure 7.16: Post-fit nuisance parameters for one example fit for the CMSTopTagger v2 with CHS, with an N-subjettiness selection of $\tau_3/\tau_2 < 0.57$, and no subjet b tagging requirement in a p_T range between 400 and 480 GeV

7.4.7 Efficiencies and scale factors

The top tagging efficiency without the mass selection criterion is calculated before and after the likelihood fits. It is calculated as $\epsilon = N_{\text{pass}}/(N_{\text{pass}} + N_{\text{fail}})$, where N_{pass} is the number of events in the pass region and N_{fail} is the number of events in the fail region. The efficiency in simulation is evaluated on the unscaled pre-fit distributions and the efficiency in data by fitting the distributions to the data. The statistical uncertainty on the data efficiency is evaluated in a fit with statistical uncertainties only and applied to the final result. The data efficiency with full uncertainties is evaluated in a fit with all uncertainties included. The uncertainties on the numbers of events in the pass and the fail region are propagated to the efficiency by Gaussian error propagation. Figure 7.17 shows the efficiency for one working point of the CMSTopTagger v2 with AK8 PUPPI jets and for HOTVR for the fully-merged contribution as a function of the probe-jet p_{T}. The efficiency measurement in data is consistent with the one in simulation.

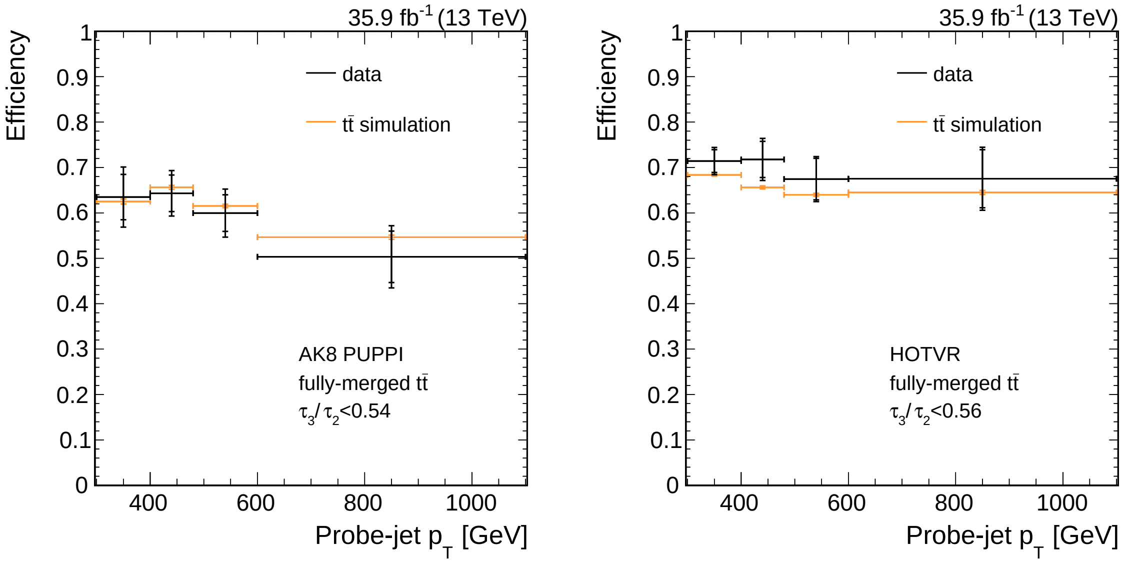

Figure 7.17: Efficiency of the CMSTopTagger v2 without the mass selection criterion as a function of the probe-jet p_{T} for a working point with $\tau_3/\tau_2 < 0.54$ and no subjet b-tagging requirement (left) and the same for the HOTVR selection without the mass selection criterion (right). The efficiency in simulation is evaluated with the unscaled simulated distributions and the efficiency in data is calculated from the fitted distributions. The vertical bars on the simulation efficiencies show the statistical uncertainties on the simulation. The inner bars on the data points show the fit uncertainties from a fit with statistical uncertainties only and the outer error bars the fit uncertainty from a fit with all uncertainties. The horizontal bars show the bin width.

Mass selection efficiency

An extra mass selection efficiency is evaluated in the pass region after the fit. It should serve as an estimate on the effect of the modeling of the jet mass in simulation after the compensation of the normalization differences. The mass selection efficiency is calculated as $\epsilon = N_{\text{pass, mass cut}}/N_{\text{pass}}$, where $N_{\text{pass, mass cut}}$ is the number of events in the pass region that also pass the mass selection and N_{pass} is the total number of events in the pass region. The efficiency in data is evaluated directly from the data distribution and the efficiency in simulation from all post-fit simulation samples combined. This leads to one inclusive efficiency for all simulation templates. Contributions of the fit uncertainties that influence the overall normalization in simulation are not considered in the uncertainty on the efficiency in simulation. Distributions of the efficiency as a function of the probe-jet p_{T} are shown in figure 7.18 for one example working point of the CMSTopTagger v2 with PUPPI and for HOTVR.

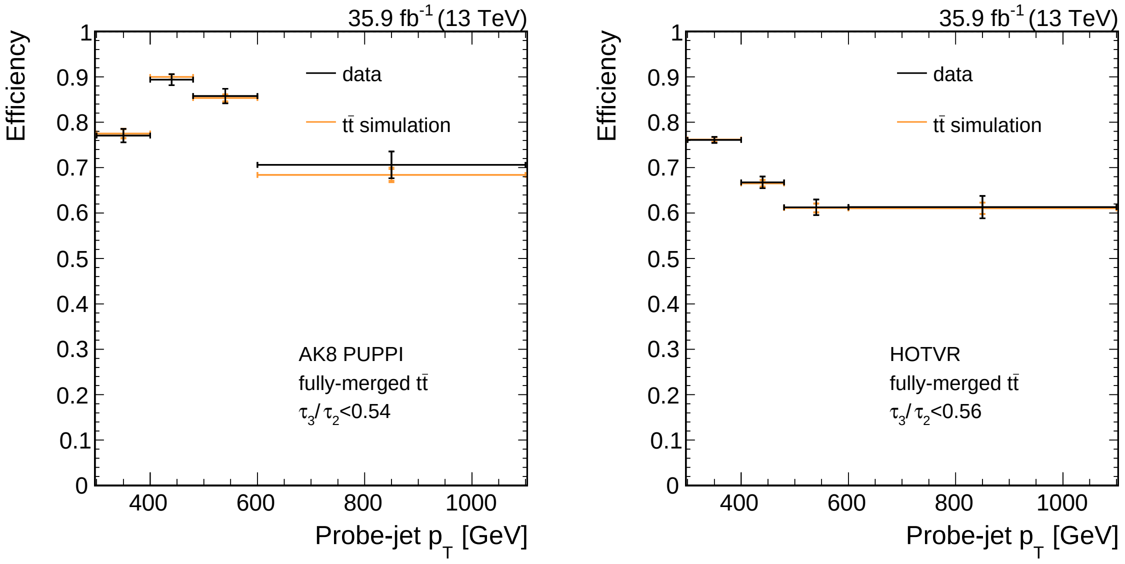

Figure 7.18: Efficiency of the jet-mass selection as a function of the probe-jet p_{T} for one working point of the CMSTopTagger v2 (left) and for HOTVR (right).

Data-to-Simulation scale factors

Data-to-simulation scale factors are evaluated to correct for differences in the top tagging efficiency between data and simulation. The two efficiency contributions without the mass selection and for the mass selection only are combined for the final scale-factor computation by $\epsilon = \epsilon_{\text{tag (no mass)}}\epsilon_{\text{mass}}$. The scale factor is then obtained by dividing the efficiency in data by the efficiency in simulation, $sf = \epsilon_{\text{data}}/\epsilon_{\text{simulation}}$.

Table 7.4.7 shows the top tagging efficiency for the fully-merged contribution in simulation as additional information on the working points. The efficiency in this table is calculated with a cut-and-count method to account properly for the mass selection efficiency. The efficiency of the different working points is shown inclusive in p_T for $p_\mathrm{T} > 400\,\mathrm{GeV}$. The efficiencies in this table are not directly comparable to the efficiency studies in section 7.3 because the signal definition is different. In section 7.3 the efficiency is defined with respect to all boosted top quarks and here with respect to already correctly reconstructed jets. The probability of reconstructing the full top quark decay within the jets is higher for HOTVR compared to the CMSTopTagger v2. A comparison between the two taggers on the basis of the efficiencies shown in this table might not be a fair comparison.

Table 7.2: Top tagging efficiency for the fully-merged contribution in simulation for all taggers and working points. The efficiency is calculated inclusive in p_T for $p_\mathrm{T} > 400\,\mathrm{GeV}$. A cut-and-count method was used to obtain these efficiencies.

Tagger	Working point	Efficiency
CMSTopTagger v2 PUPPI	$\tau_3/\tau_2 < 0.4$	27 %
	$\tau_3/\tau_2 < 0.4$ + subjet b tag	24 %
	$\tau_3/\tau_2 < 0.46$	41 %
	$\tau_3/\tau_2 < 0.46$ + subjet b tag	35 %
	$\tau_3/\tau_2 < 0.54$	58 %
	$\tau_3/\tau_2 < 0.54$ + subjet b tag	50 %
	$\tau_3/\tau_2 < 0.65$	75 %
	$\tau_3/\tau_2 < 0.65$ + subjet b tag	64 %
	$\tau_3/\tau_2 < 0.80$	88 %
	$\tau_3/\tau_2 < 0.80$ + subjet b tag	75 %
CMSTopTagger v2 CHS	$\tau_3/\tau_2 < 0.50$	38 %
	$\tau_3/\tau_2 < 0.50$ + subjet b tag	33 %
	$\tau_3/\tau_2 < 0.57$	55 %
	$\tau_3/\tau_2 < 0.57$ + subjet b tag	48 %
	$\tau_3/\tau_2 < 0.67$	74 %
	$\tau_3/\tau_2 < 0.67$ + subjet b tag	64 %
	$\tau_3/\tau_2 < 0.81$	88 %
	$\tau_3/\tau_2 < 0.81$ + subjet b tag	75 %
HOTVR	$\tau_3/\tau_2 < 0.56$	53 %

Figure 7.19 shows top tagging scale factors as a function of the probe-jet p_T for the 'fully-merged', the 'semi-merged', and the 'not-merged' categories. The scale factors are shown for the CMSTopTagger v2 with AK8 PUPPI jets for an example working point corresponding to a selection on $\tau_3/\tau_2 < 0.54$ and no subjet b-tagging requirement. All scale factors are consistent with one within the uncertainties. The uncertainties on the 'semi-merged' and 'not-merged' categories are significantly larger compared to the 'fully-merged' contribution because their contribution to the pass region is small and they have

a similar shape. They are less constrained by the fits.

Figure 7.20 shows a comparison of all measured top tagging scale factors for the CMSTopTagger v2 with AK8 PUPPI jets for different working points of the N-subjettiness selection without subjet b tagging on the top and with subjet b tagging on the bottom. The same is shown for the CMSTopTagger v2 with AK8 CHS jets in figure 7.21 and for HOTVR PUPPI jets in figure 7.22. Most of the scale factors are well consistent with unity.

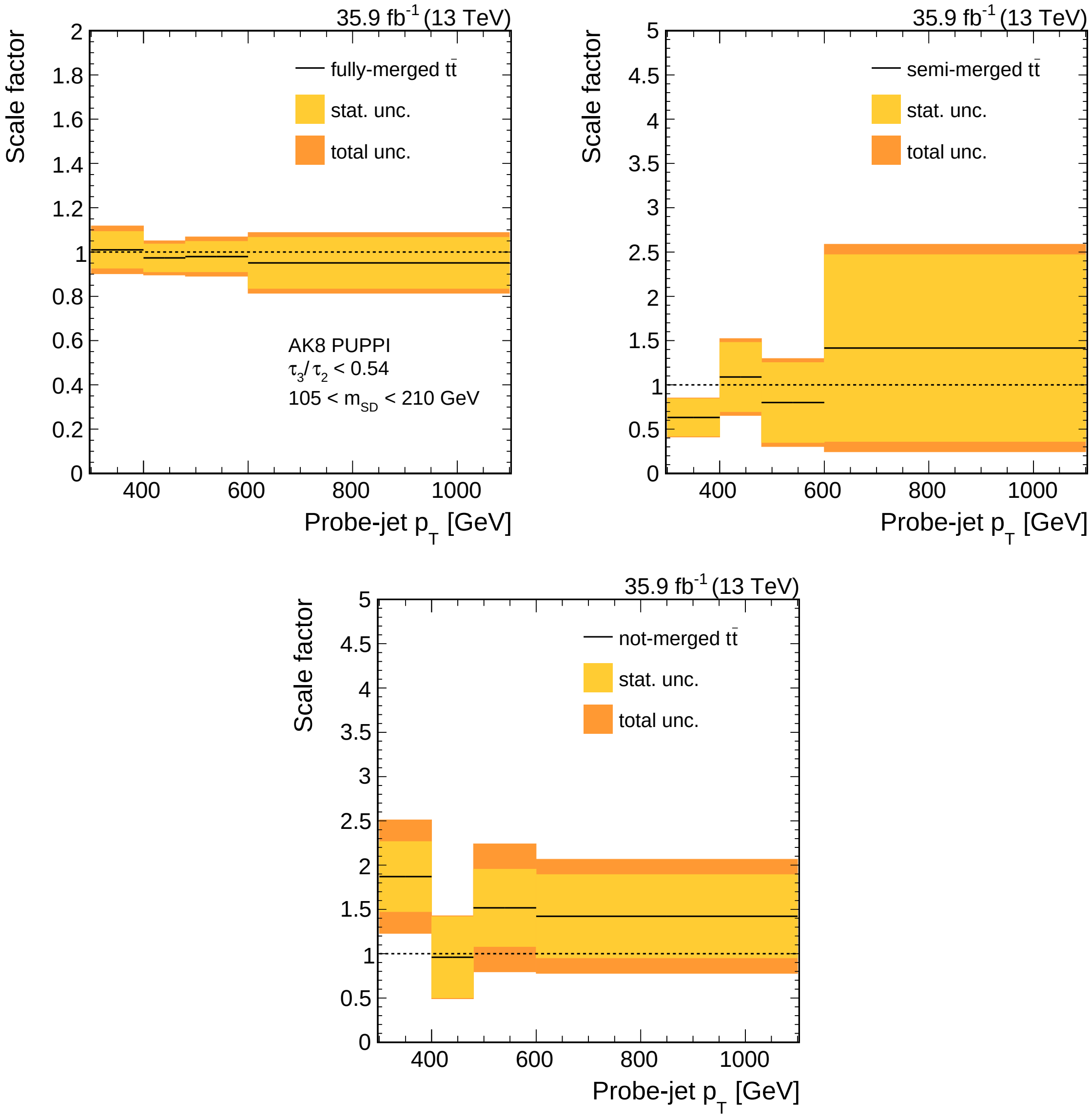

Figure 7.19: Data-to-simulation scale factors for the CMSTopTagger v2 with AK8 PUPPI jets as a function of the probe-jet p_{T}. The scale factors are obtained for an example working point corresponding to a selection of $\tau_3/\tau_2 < 0.54$ and no requirement on subjet b tagging. The figure on the top left shows the scale factors for fully-merged $\mathrm{t\bar{t}}$ decays, the one on the top right for semi-merged $\mathrm{t\bar{t}}$ decays, and the figure on the bottom shows the scale factors for the not-merged category.

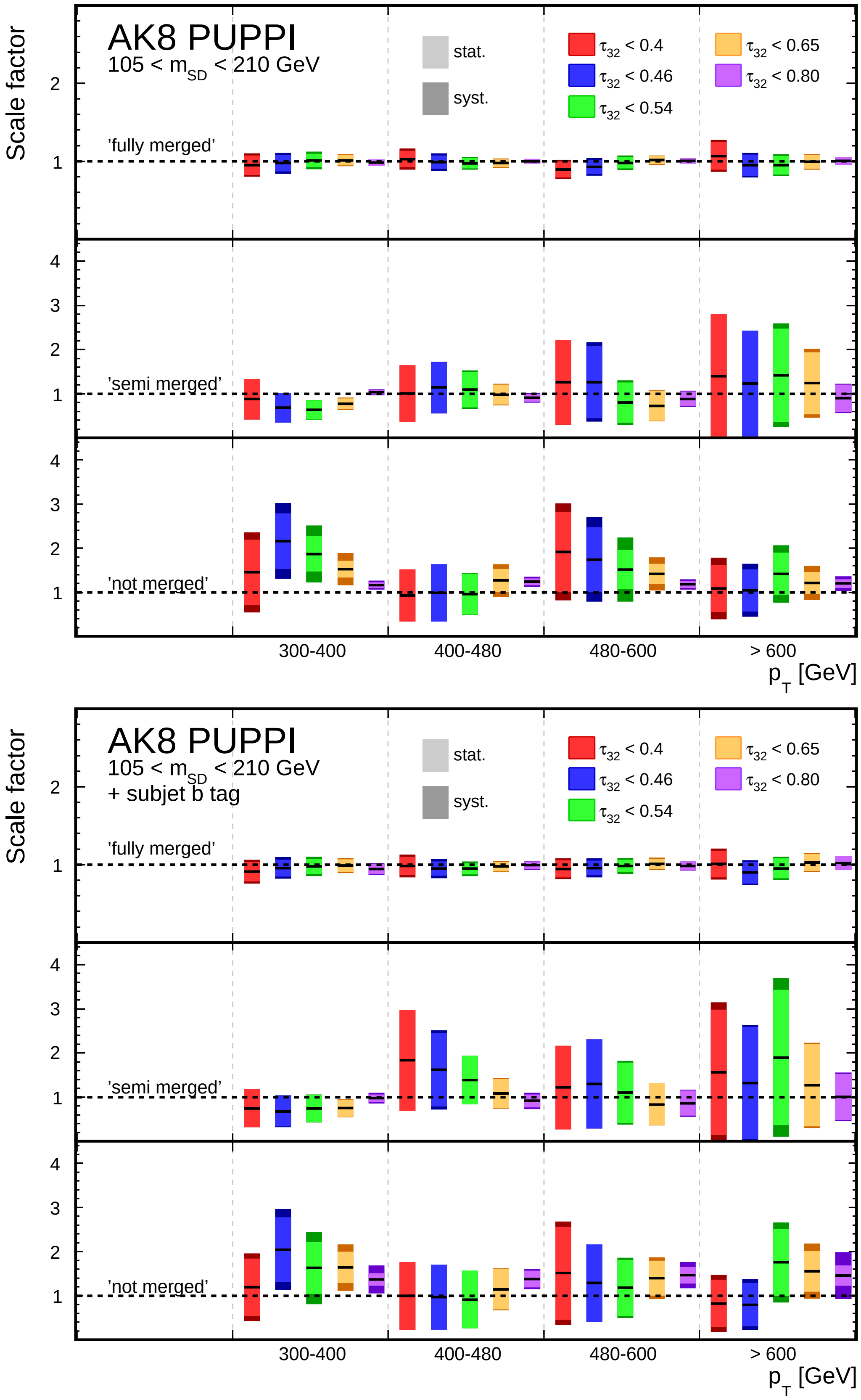

Figure 7.20: Overview of all measured data-to-simulation scale factors for the CMSTopTagger v2 with AK8 PUPPI jets. Scale factors for several working points are shown without subjet b tagging (top) and with a subjet b tag (bottom).

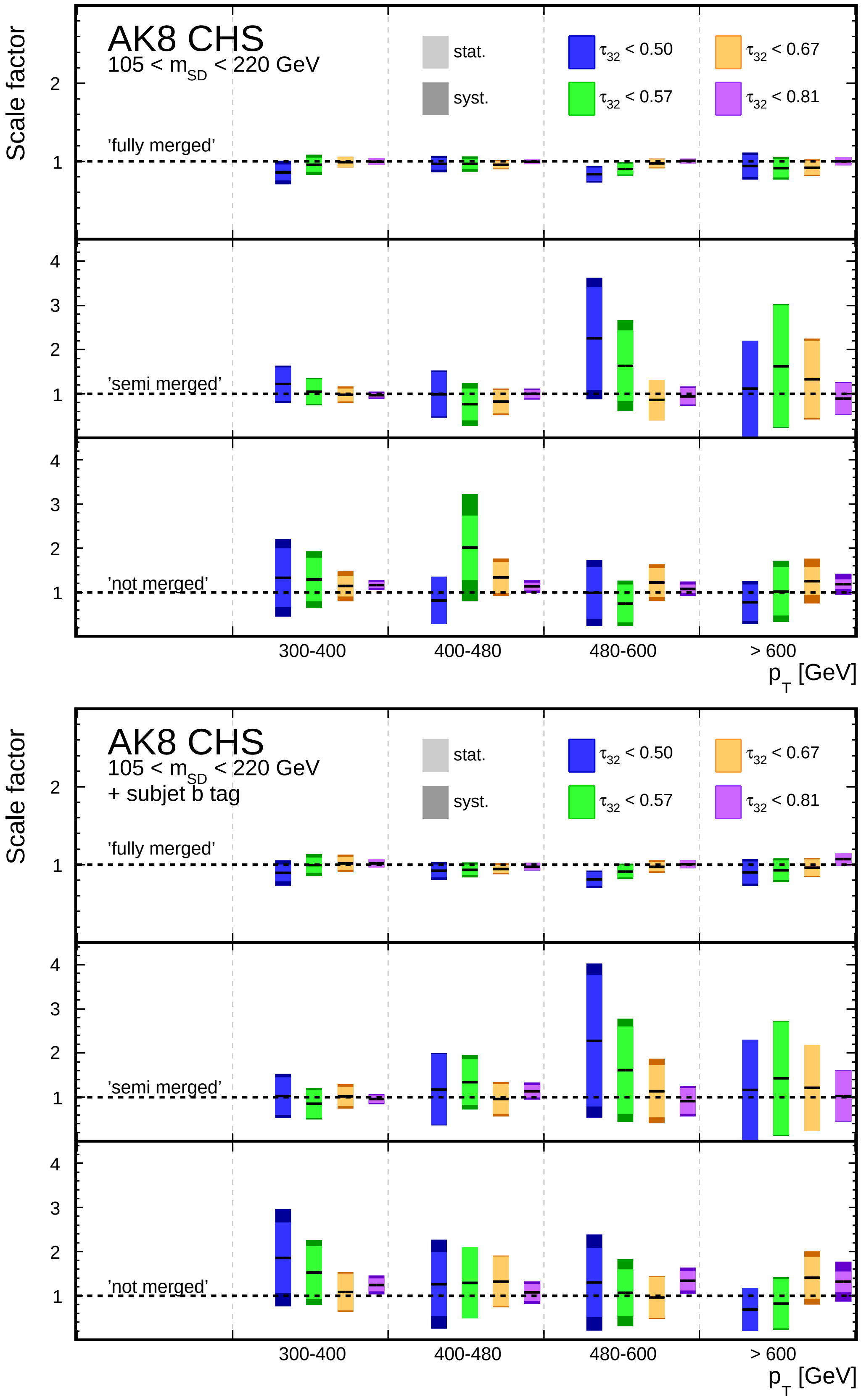

Figure 7.21: Overview of all measured data-to-simulation scale factors for the CMSTopTagger v2 with AK8 CHS jets. Scale factors for several working points are shown without subjet b tagging (top) and with a subjet b tag (bottom).

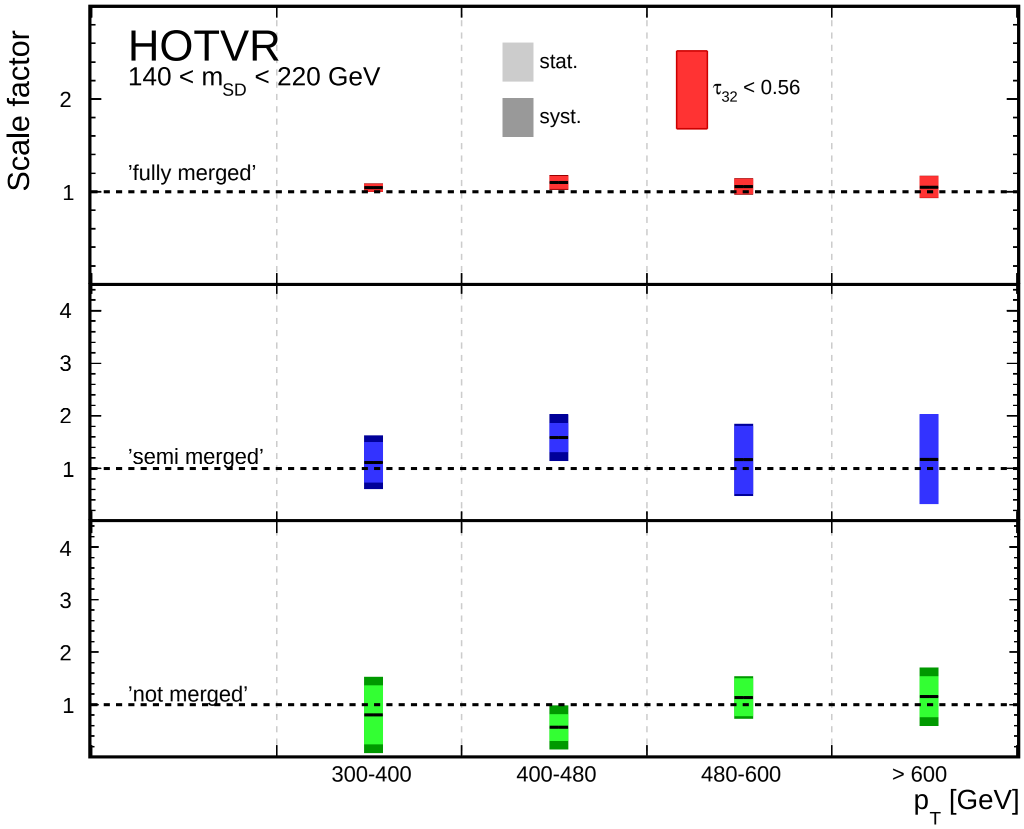

Figure 7.22: Overview of all measured data-to-simulation scale factors for the HOTVR algorithm with PUPPI. Scale factors for different categories of the $t\bar{t}$ simulation are sown in the different rows.

7.5 Mistag rate in 2016 data

In addition to the signal efficiency and scale factors, the mistag rate and corresponding data-to-simulation scale factors are studied in the 2016 data set. The mistag rates for the different top tagging algorithms are studied on a sample enriched with QCD multijet production. The efficiency in data is compared to different simulations of QCD multijet production and data-to-simulation scale factors are given for the comparison of the data to each simulation.

The studies in this section have been performed within the scope of studies towards a comparison with other top tagging algorithms in CMS in reference [134]. The event selection selection is similar to the selection in this reference.

7.5.1 Object definitions

The objects used for the mistag rate studies are similar to those used in the study of signal efficiency in 2016 data. The main difference is the definition of leptons used in the definition of a lepton veto. The p_{T} threshold for leptons can be chosen much lower in order to obtain a strong veto against leptons and to reduce the influence of processes like $\mathrm{t\bar{t}}$ or W+jets production.

Muons are identified with loose identification criteria [183] corresponding to the following selection: the muon candidate is identified as a particle-flow muon and is either identified as a global muon or as a tracker muon. On top of the loose identification selection candidates are further required to fulfill two requirements on the distance to the interaction point of $|d_{xy}| < 0.2$ and $|d_z| < 0.5$. They are required to pass an isolation criterion of miniIso/$p_{\mathrm{T}} < 0.2$, where the relative mini isolation is related the energy collected within a certain distance to the muon candidate decreasing with the p_{T} of the muon. Muons candidates are considered for $p_{\mathrm{T}} > 5\,\mathrm{GeV}$ and $|\eta| < 2.4$.

Isolated electron candidates are required to fulfill the cut-based veto identification requirements [144]. The isolation is applied again by a selection on the relative mini isolation miniIso/$p_{\mathrm{T}} < 0.1$. Electron candidates are considered for $p_{\mathrm{T}} > 5\,\mathrm{GeV}$ and $|\eta| < 2.4$.

The AK4 jets are defined in the same way as they are defined for the signal efficiency studies in section 7.4.1. They are considered for $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.4$. Large jets are either AK8 jets with CHS or PUPPI applied or HOTVR jets with PUPPI, depending on the tagging algorithm that is studied. All these jets are defined in the same way as in sections 7.4.1 and 7.3.

7.5.2 Event selection

The event selection aims on a pure selection of events from QCD multijet production. A veto is applied on electrons and muons in the event to suppress leptonic top quark or W boson decays. At least two AK4 jets are required and a value of $H_\mathrm{T} > 1000\,\mathrm{GeV}$, where H_T is the scalar sum of the p_T of the corrected AK4 jets. The probe jet is chosen to be the large jet with the highest p_T in the event (leading jet). Figure 7.23 shows a few distributions for AK8 jets with PUPPI after the full selection and a requirement on the probe-jet p_T of $p_\mathrm{T} > 200\,\mathrm{GeV}$. The p_T distribution, the η distribution, the Soft Drop mass distribution, and the distribution of the N-subjettiness ratio τ_3/τ_2 are shown for the data compared to different simulations of QCD multijet production with MadGraph +pythia, herwig++, and pythia stacked with the background contributions from W+jets production, Z+jets production, $\mathrm{t\bar{t}}$ production, and single top quark production. Each QCD sample is normalized to the data to allow a shape comparison. Ratios are shown beneath each figure with respect to the default QCD simulation with MadGraph +pythia. The shape of the p_T and the Soft Drop Mass are reasonably well described by the pythia simulation and worse for MadGraph +pythia and herwig++. The N-subjettiness distribution is not well described by any of the samples. herwig++ describes the shape for high values best, the pythia simulations show a better description for low values.

7.5.3 Efficiencies and scale factors

The efficiency is in this case evaluated by a simple cut-and-count method. It is calculated by $\epsilon = N_\mathrm{tagged}/N_\mathrm{probe\ jets}$, where N_tagged is the number of tagged probe jets and $N_\mathrm{probe\ jets}$ is the number of all probe jets. It is evaluated for each QCD multijet sample and from data after the subtraction of the relevant background processes from simulation. The scale factors are obtained by dividing the efficiency in data by the efficiency in simulation $sf = \epsilon_\mathrm{data}/\epsilon_\mathrm{simulation}$. The results are top tagging mistag rates in data and simulation as well as data-to-simulation scale factors as a function of the probe-jet p_T. Figure 7.24 shows the efficiency in data compared to the different simulation samples for one example working point of the CMSTopTagger v2 with AK8 PUPPI and for HOTVR. The data-to-simulation scale factors for the different simulations are shown beneath the respective efficiency plots. The PYTHIA simulation describes the data best with scale factors closest to one. All scale factors show a constant behavior as a function of the probe-jet p_{T}.

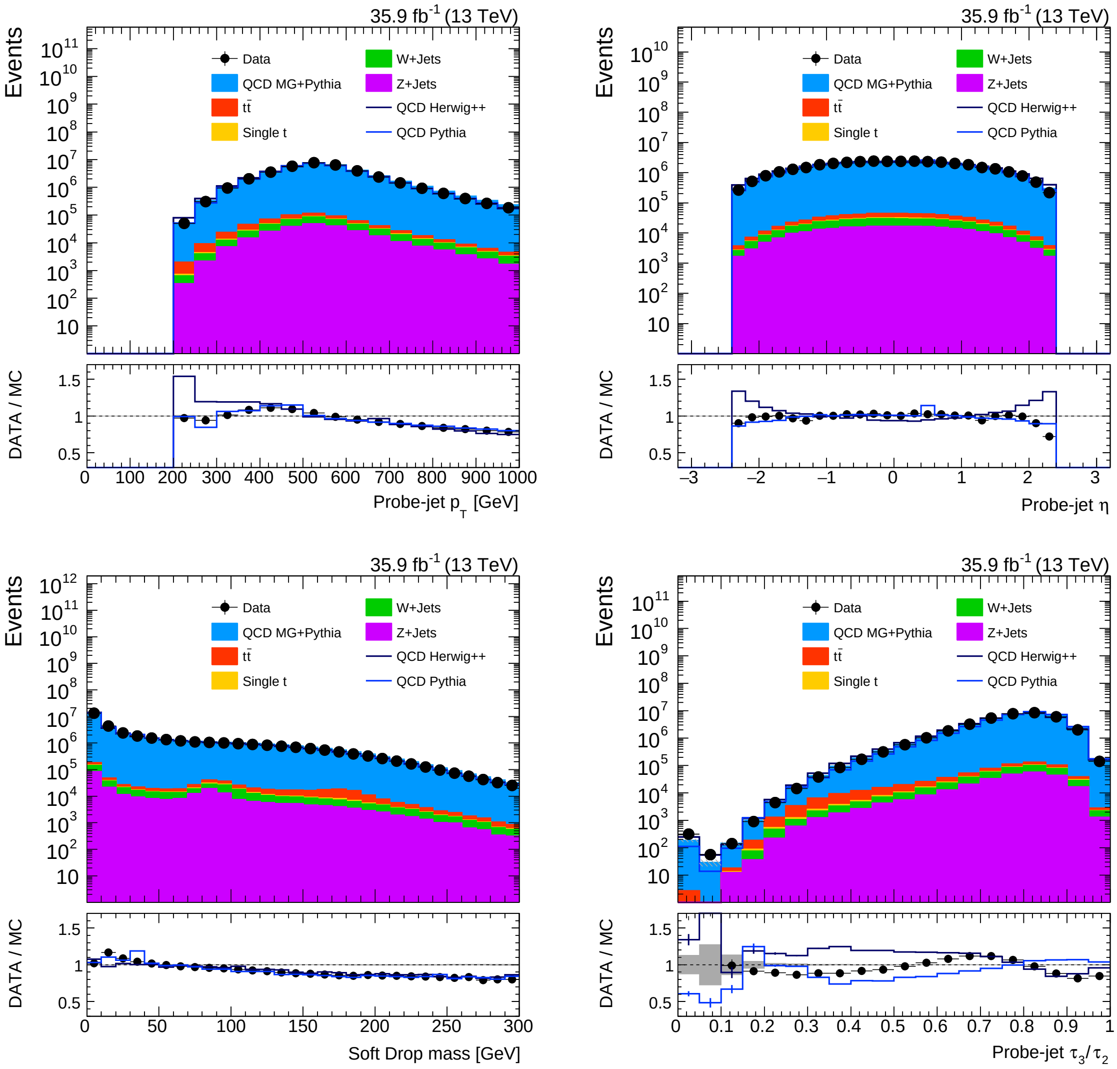

Figure 7.23: Distributions of different probe-jet properties after the full selection described above and a requirement on the probe-jet $p_{\mathrm{T}} > 200\,\mathrm{GeV}$. The p_{T} distribution is shown on the top left, the η distribution on the top right, the Soft Drop mass distribution on the bottom left, and N-subjettiness ratio τ_3/τ_2 on the bottom right. Data is shown as black points with vertical bars showing the statistical uncertainty. The horizontal bars show the bin width. Simulation is shown as filled histograms with the hatched area showing the statistical uncertainty on the simulation. Two additional QCD simulations are shown as lines. They are also stacked with the background samples like the default QCD simulation to allow a good comparison. All QCD samples are normalized to the data to allow a shape-comparison only. A ratio is shown under each figure with respect to the default QCD simulation with MadGraph +pythia. The gray area shows the statistical uncertainty on the simulation.

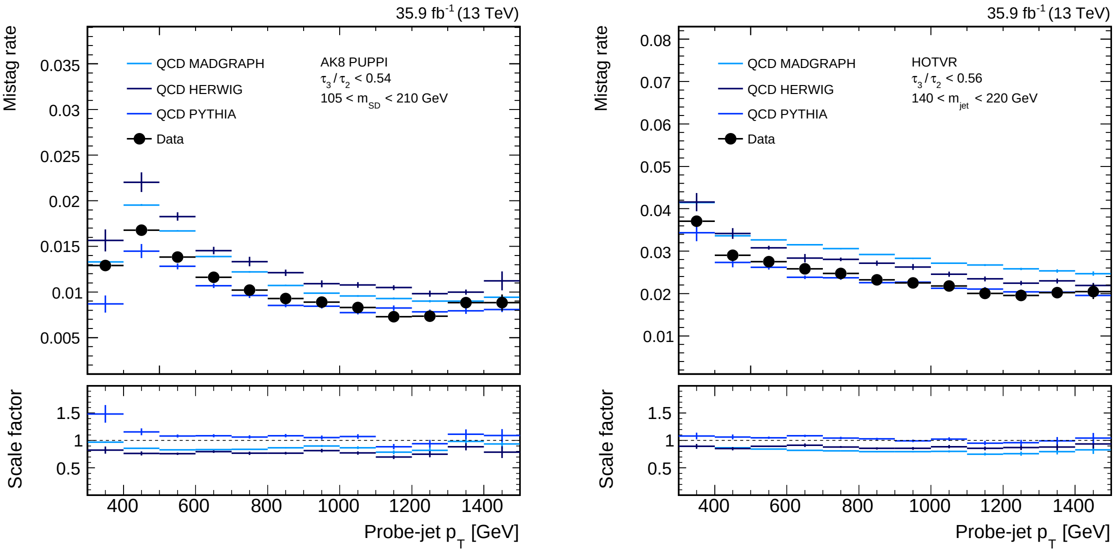

Figure 7.24: Mistag rates as a function of the probe-jet p_T for data and different QCD multijet simulations with MadGraph+pythia, pythia, and herwig++. The efficiency is shown in each bin with a vertical bar showing the statistical uncertainty. The horizontal bars show the bin width. The mistag rate for one example working point of the CMSTopTagger v2 for $\tau_3/\tau_2 < 0.54$ and no subjet b tag is shown on the left and the mistag rate for the HOTVR algorithm is shown on the right. Data-to-simulation scale factors are shown below the respective mistag rates for each of the different simulations.

7.6 Validation in 2017 data

Top tagging efficiencies and data-to-simulation scale factors are measured in data collected in the year 2017 for the CMSTopTagger v2 with CHS and PUPPI. The measurement follows the procedure of the 2016 studies described in section 7.4 with only small differences discussed in the following section.

7.6.1 Differences with respect to the 2016 studies

One of the most important differences with respect to the 2016 studies is a new PYTHIA tune that is used for the $\mathrm{t\bar{t}}$ simulation. The CP5 tune was used in the production of the 2017 MC samples and replaced the CUETP8M2T4 tune from the 2016 production.

The only difference in the event selection is a change of the algorithm used for the b tagging of AK4 jets and AK8 subjets. The DeepCSV algorithm is used instead of the CSVv2 algorithm leading to a better signal efficiency at the same mistag rate. All objects are defined in the same way as for the previous studies in section 7.4. The selection criteria in the event selection are the same. All data-to-simulation scale factors are updated with most recent recommendation provided in CMS at the time these studies where performed.

The systematic uncertainty on the parton shower is evaluated by a variation of the parton-shower cutoff scale for final-state radiation (FSR) by factors of 0.5 and 2 instead of a comparison of PYTHIA with HERWIG++ which was used for the studies with 2016 data.

Figure 7.25 shows distributions of the probe-jet p_T, the probe-jet η, the Soft Drop mass, and the N-subjettiness ratio τ_3/τ_2 after the full selection for AK8 jets clustered with PUPPI. Figure 7.26 shows the same distributions with CHS jets. The $\mathrm{t\bar{t}}$ simulation is scaled to the number of events in data to compensate the normalization difference between data and simulation because of a softer top quark p_T spectrum in data which is not fully compensated by the reweighting of the top quark p_T spectrum for high momenta larger than 400 GeV. The N-subjettiness distribution shows a worse description by the simulation compared to the 2016 studies in figures 7.8 and 7.9. All other distributions show a reasonable agreement between data and simulation.

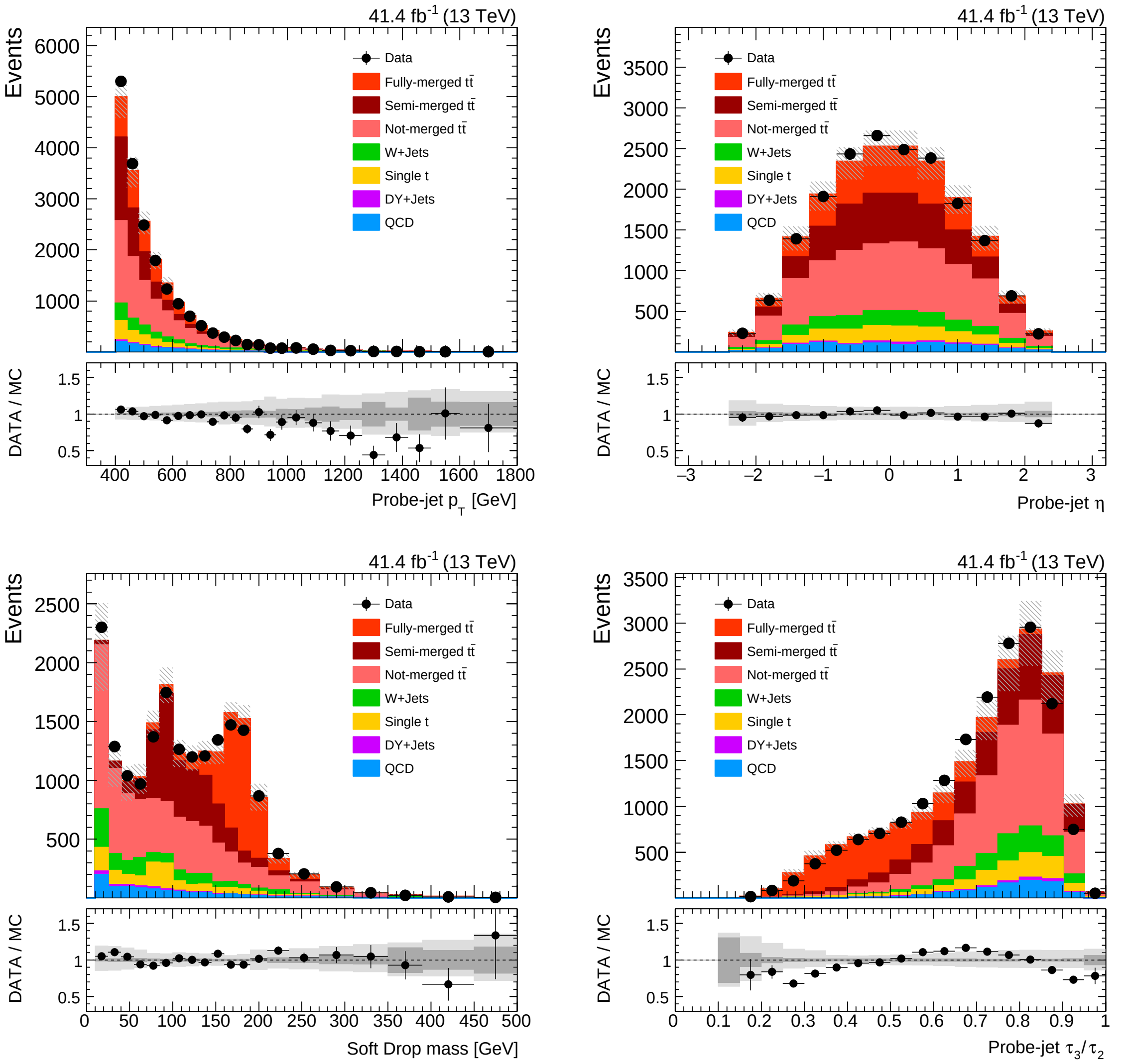

Figure 7.25: Control distributions in data and simulation for AK8 jets with PUPPI and a jet-p_T larger than 400 GeV. The full event selection is applied. The jet-p_T distribution is shown on the top left, the η distribution on the top right, the Soft Drop mass distribution on the bottom left, and the N-subjettiness ratio on the bottom right. The $\mathrm{t\bar{t}}$ simulation is scaled to the data. The data is shown as black points and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution. The dark gray area shows the statistical uncertainty on the simulation, and the light gray area shows the total uncertainty including systematic uncertainties.

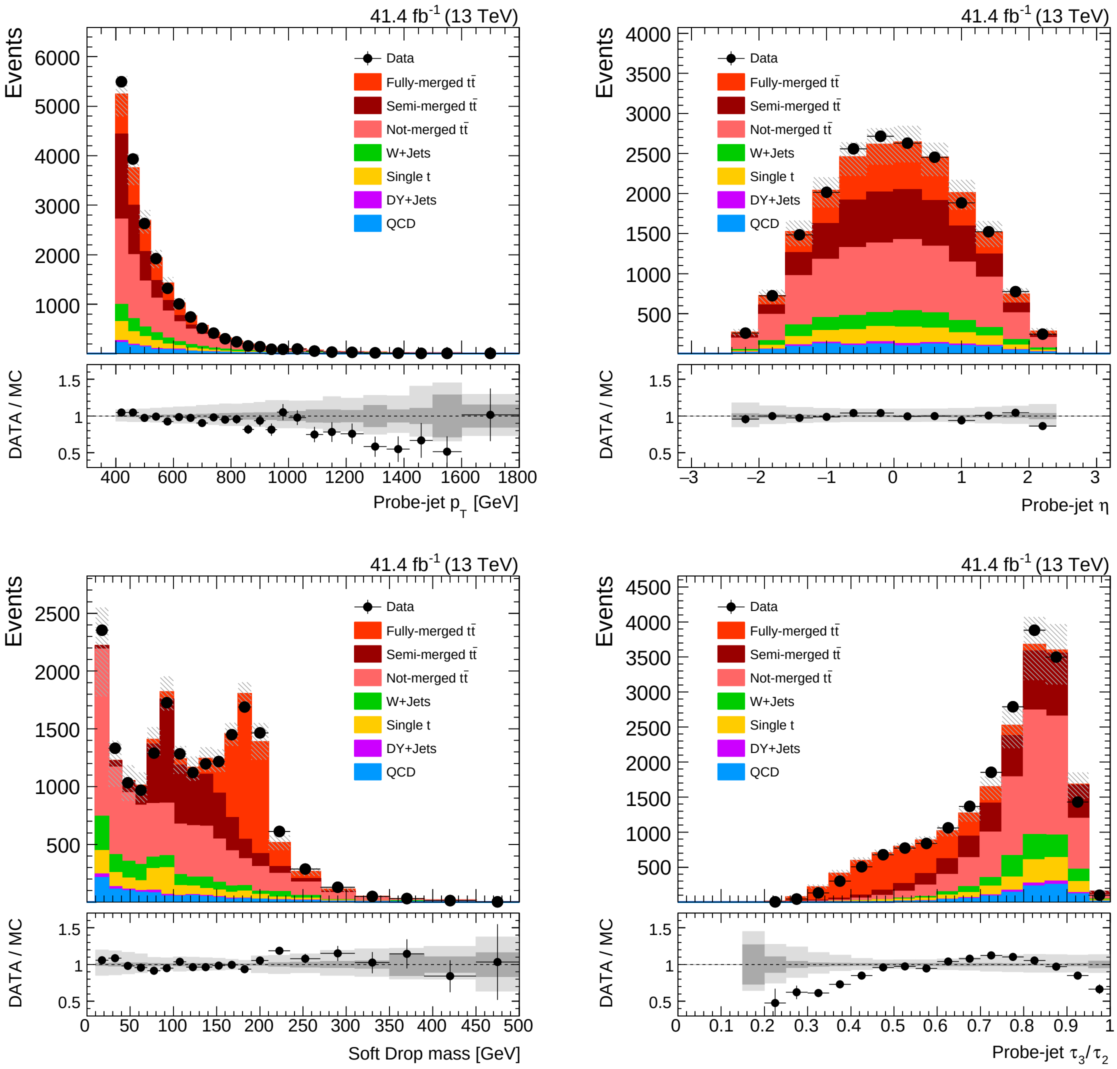

Figure 7.26: Control distributions in data and simulation for AK8 jets with CHS and a jet-p_T larger than 400 GeV. The full event selection is applied. The jet-p_T distribution is shown on the top left, the η distribution on the top right, the Soft Drop mass distribution on the bottom left, and the N-subjettiness ratio on the bottom right. The $\mathrm{t\bar{t}}$ simulation is scaled to the data. The data is shown as black points and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution. The dark gray area shows the statistical uncertainty on the simulation, and the light gray area shows the total uncertainty including systematic uncertainties.

7.6.2 Parton-shower modeling

The worse description of τ_3/τ_2 in the simulation samples for 2017 compared to the 2016 studies motivates studies on the influence of the parton-shower cutoff scales and the new PYTHIA tune on the shape of the τ_3/τ_2 distribution.

Parton-shower cutoff scales

The influence of the parton shower on the top tagging efficiency is studied in the 2017 data by reweighting simulated events corresponding to variations of the parton-shower cutoff scales for initial-state radiation (ISR) and final-state radiation (FSR) by factors of 0.5 and 2. Figure 7.27 shows the Soft Drop mass distribution and the τ_3/τ_2 distribution in $\mathrm{t\bar{t}}$ simulation for all variations of the ISR and FSR compared to the default values. A ratio is shown beneath each distribution between each variation and the default sample. The effects of the ISR variations on the Soft Drop mass and the N-subjettiness ratio are negligible and are not considered for the following fits. The τ_3/τ_2 distribution shows a clear dependence on the FSR cutoff scale. The variation of the FSR cutoff scale by factors of 0.5 and 2 is therefore considered as a systematic uncertainty in the following maximum-likelihood fits.

Old versus new PYTHIA tune

The distributions of the N-subjettiness ratio τ_3/τ_2 for 2017 data in figures 7.25 (bottom right) and 7.26 (bottom right) show a worse description by the simulations compared to the 2016 data in figures 7.8 (bottom right) and 7.9 (bottom right). One of the main differences between the 2016 and 2017 studies is a new PYTHIA tune (CP5) used in the $\mathrm{t\bar{t}}$ simulation. An extra sample for lepton+jets $\mathrm{t\bar{t}}$ events was produced with the old PYTHIA tune (CUETP8M2T4) used for the 2016 MC production but with the setup used for 2017 data. This sample is produced for events with a selection on the p_T of the hadronic W boson to be larger than 150 GeV. The requirement on $p_\mathrm{T,W}$ on detector level was increased to $p_\mathrm{T,W} > 250\,\mathrm{GeV}$ to compare the new sample with the data. For the comparison with data the 'fully-merged' and the 'semi-merged' contributions of the $\mathrm{t\bar{t}}$ simulation are replaced with the new sample with the CUETP8M2T4 tune. The 'not-merged' contribution includes a significant number of dileptonic $\mathrm{t\bar{t}}$ decays and is therefore taken from the old sample with the CP5 tune. The exchange of the 'fully-merged' and the 'semi-merged' contribution is sufficient to study how the different tunes

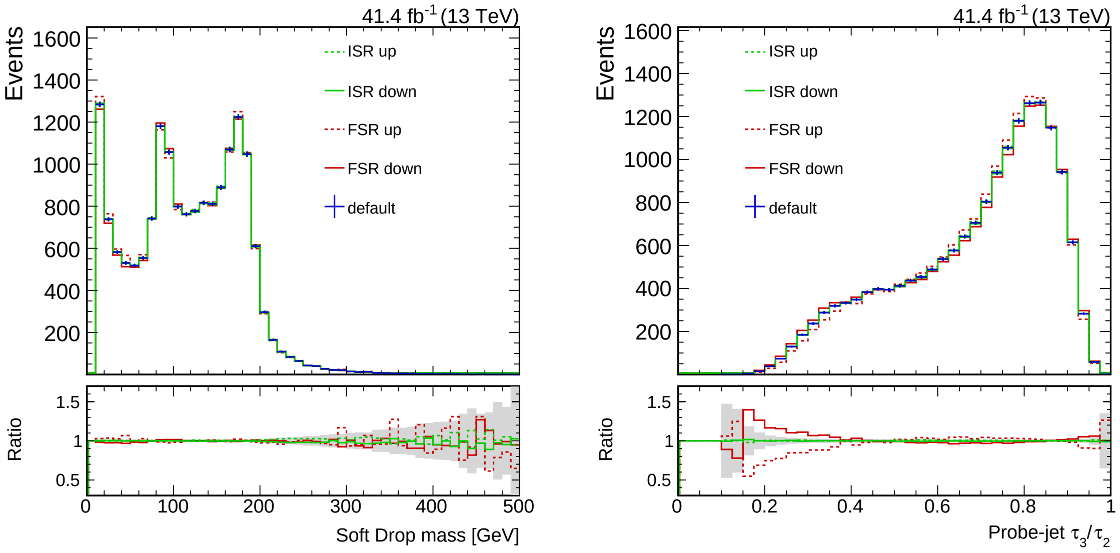

Figure 7.27: Distributions of the Soft Drop mass (left) and the τ_3/τ_2 distribution (right) for the probe jet after the full selection in $t\bar{t}$ events simulated with POWHEG +PYTHIA. The default distributions are shown together with reweighted distributions corresponding to variations of the ISR and FSR cutoff scales by factors of 0.5 and 2. A ratio of each contribution to the default is shown under each distribution. The gray band shows the statistical uncertainty on the default simulation.

influence the N-subjettiness distribution, since these two contributions dominate at low values of τ_3/τ_2 where the largest difference between the 2016 and the 2017 distributions is observed. Figure 7.28 shows the N-subjettiness distributions in data and simulation after the full selection with $p_{\mathrm{T,W}} > 250\,\mathrm{GeV}$. The left distributions show the $t\bar{t}$ simulation with the CP5 tune and the distributions on the right show the case in which the 'fully-merged' and the 'semi-merged' contributions are simulated with the CUETP8M2T4 tune. A maximum-likelihood fit in the jet mass is performed fitting the normalization of the individual $t\bar{t}$ contributions to the data as it is done for the scale factor measurements. The distributions after the fits are shown at the bottom of figure 7.28. Overall the N-subjettiness distributions at low values are better described by the CUETP8M2T4 tune compared to the CP5 tune. This indicates that the worse agreement in the N-subjettiness might be caused to some degree by the change of the PYTHIA tune from CUETP8M2T4 to CP5.

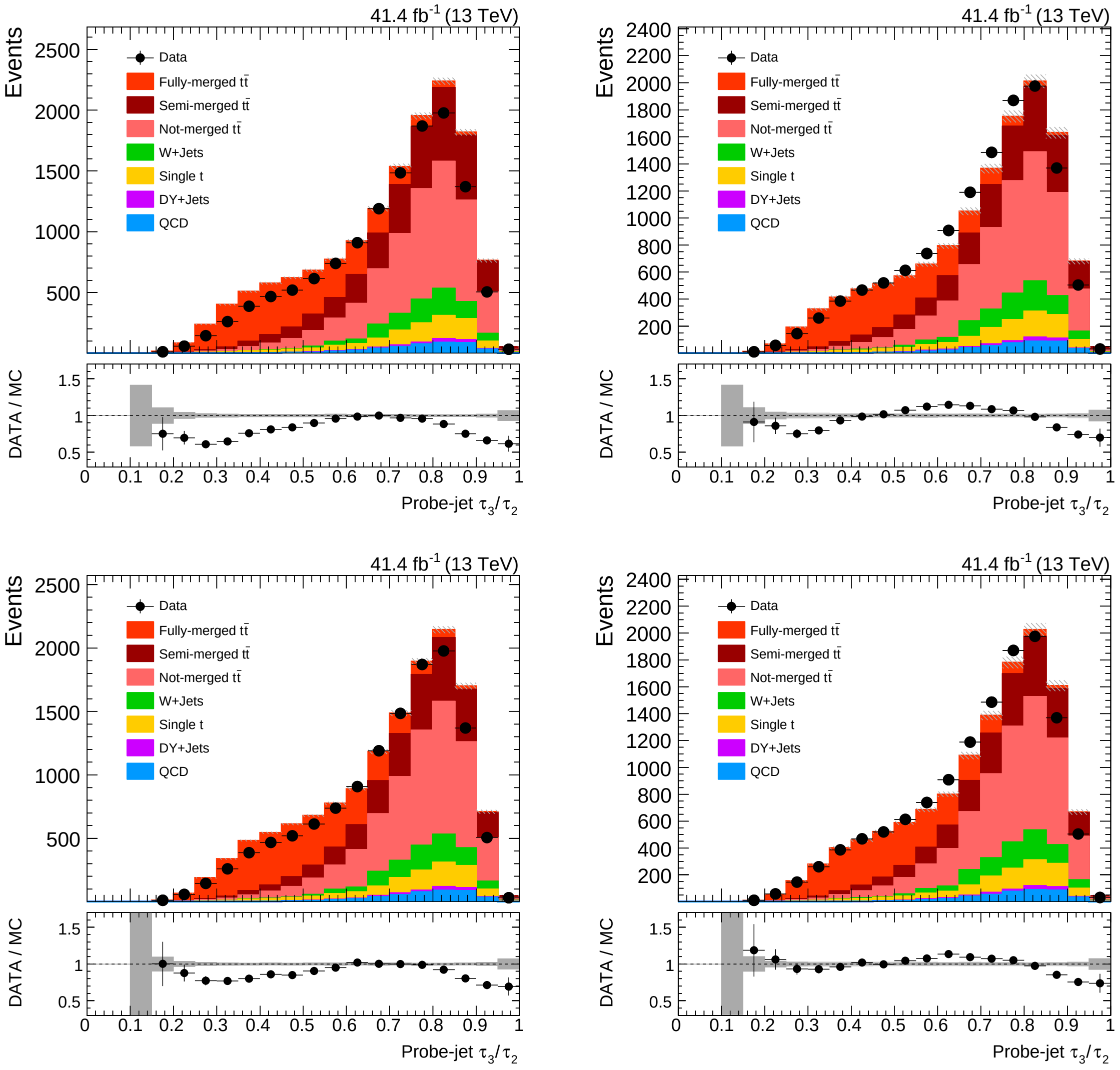

Figure 7.28: N-subjettiness distributions in data and simulation after the full selection with $p_{\mathrm{T,W}} > 250\,\mathrm{GeV}$. In the figures on the left the $\mathrm{t\bar{t}}$ samples are simulated with the CP5 PYTHIA tune. For the distributions on the right the 'fully-merged' and the 'semi-merged' $\mathrm{t\bar{t}}$ contributions are exchanged with a sample simulated with the CUETP8M2T4 PYTHIA tune. In the distributions on the bottom the normalizations of the individual $\mathrm{t\bar{t}}$ contributions are fitted to the data by a maximum-likelihood fit in the Soft Drop mass distribution. In each figure the data is shown as black points and compared to simulation (filled histograms). The statistical uncertainty on the data points is shown by vertical bars. The horizontal bars show the bin width. The hatched region gives the full uncertainty on the MC simulation. A ratio between data and MC is shown below each distribution. The dark gray area shows the statistical uncertainty on the simulation, and the light gray area shows the total uncertainty including fit uncertainties.

7.6.3 Template fits

The top tagging efficiencies are obtained again for different contributions of the $\mathrm{t\bar{t}}$ simulation fitting the different simulation templates to data as it was done in section 7.4. The same fitting setup is used for 2016 and 2017 data to obtain a consistent set of scale factors for both data sets. The only difference is the parton-shower uncertainty that was estimated with a sample simulated with POWHEG +HERWIG++ for the 2016 measurement and is now estimated by a reweighting corresponding to a variation of the FSR cutoff scale by factors of 0.5 and 2. Figure 7.29 shows the Soft Drop mass distribution for the CMSTopTagger v2 with PUPPI before and after the fit in a pass and a fail region passing and failing the requirement on $\tau_3/\tau_2 < 0.54$. All distributions are shown for a p_T of the probe jet between 400 and 480 GeV. Similar distributions for the CMSTopTagger v2 with CHS passing and failing a requirement on $\tau_3/\tau_2 < 0.57$ can be found in figure 7.30. The fitted distributions show a significantly better agreement between data and simulation. In the case of AK8 CHS jets the mass peak in simulation is shifted compared to the peak in data before the fit. The simulation is shifted towards the data within the fit by the template morphing using the systemic templates for variations of the JECs leading to a reasonable agreement after the fit. The fits are performed in four different p_T bins for the five N-subjettiness working points for PUPPI jets and the four working points for CHS jets. All working points are studied with and without a requirement on at least one subjet b tag.

7.6.4 Efficiencies and scale factors

Top tagging efficiencies and data-to-simulation scale factors are calculated in the same way as for the 2016 data in section 7.4. Figure 7.31 shows scale factors as a function of the probe-jet p_T for AK8 jets with CHS for an example working point with $\tau_3/\tau_2 < 0.57$ and no subjet b tagging. The scale factors for the 'fully-merged' contribution are significantly lower than one because of the worse description of the N-subjettiness by the simulation. All measured scale factors are shown in figure 7.32 for PUPPI jets and in figure 7.33 for CHS jets. The scale factors for the 'fully-merged' contribution are lower than one for all working points because the N-subjettiness ratio τ_3/τ_2 is worse described by the simulation compared to the 2016 studies.

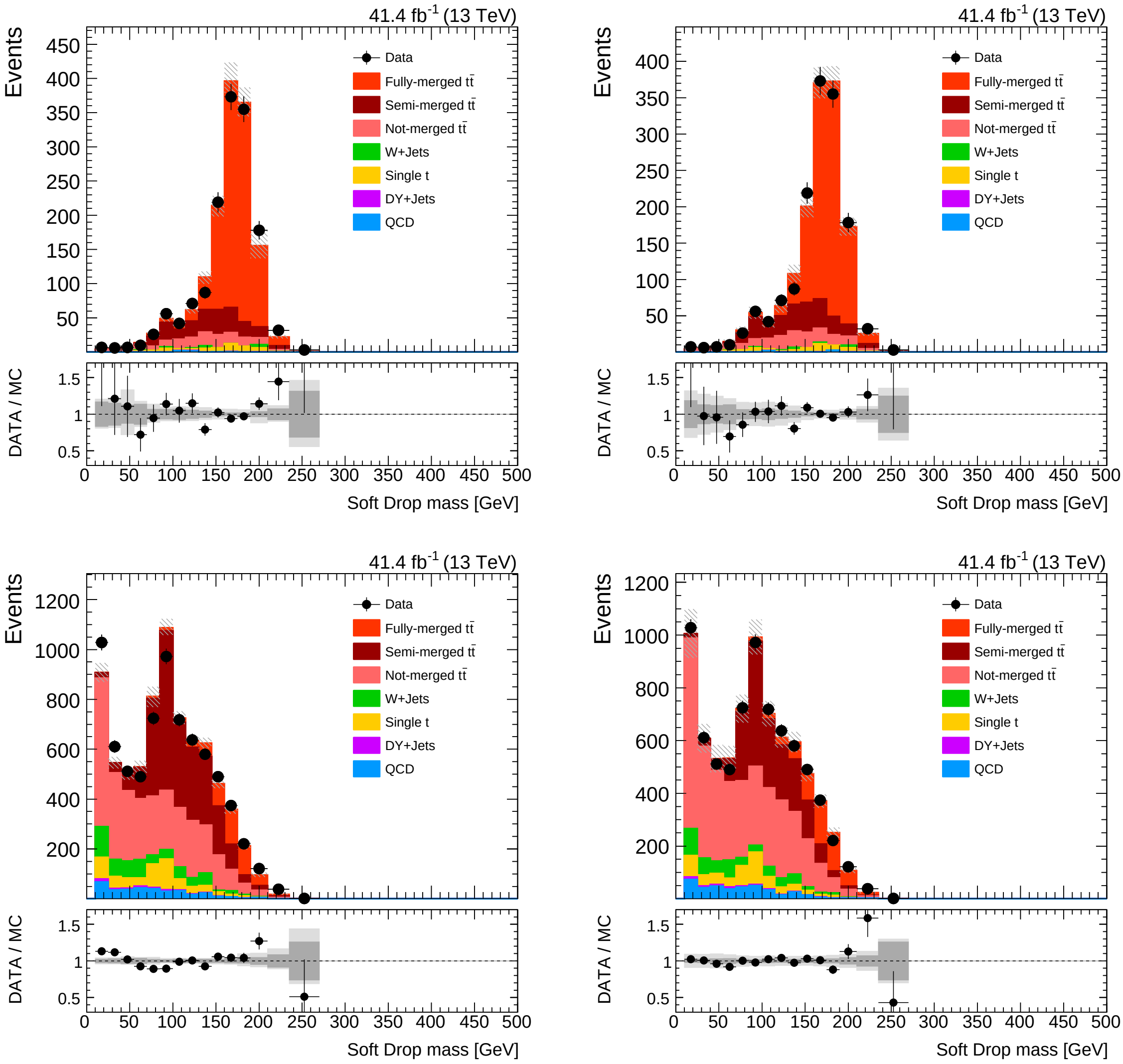

Figure 7.29: Soft Drop mass distributions for AK8 probe jets with PUPPI applied for $400 < p_{\mathrm{T}} < 480\,\mathrm{GeV}$. Distributions for the pass region are shown on the top and for the fail region on the bottom. The distributions on the left are shown before the maximum-likelihood fit and the distributions on the right after the fit. Data is shown as black dots with vertical bars showing the statistical uncertainties on the data. Simulation is shown as filled histograms with the hatched area showing the total uncertainty on the simulation. A ratio of data divided by simulation is shown under each distribution. The dark gray band shows the statistical uncertainty on the simulation and the light gray band the total uncertainty.

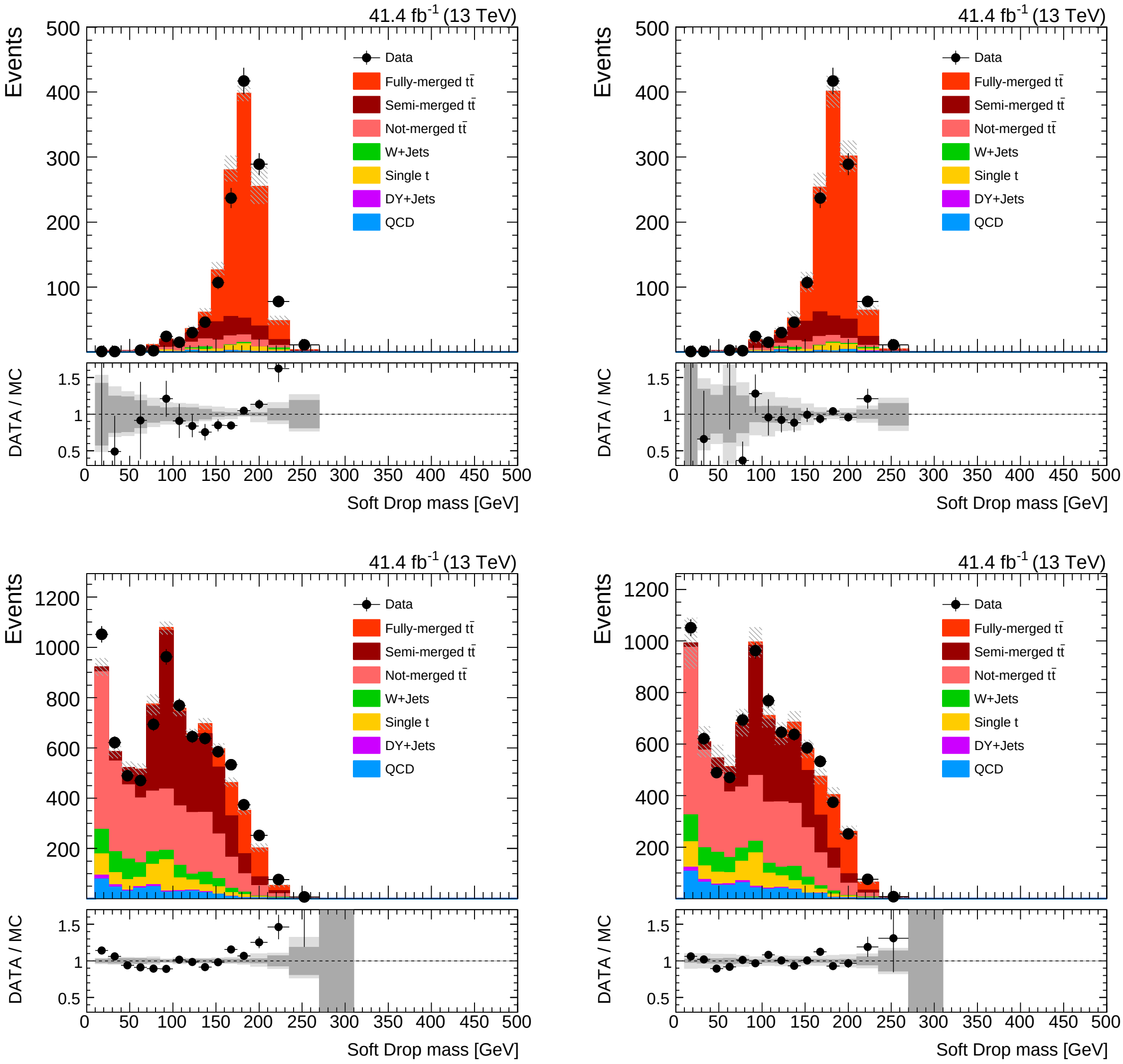

Figure 7.30: Soft Drop mass distributions for AK8 probe jets with CHS applied for $400 < p_{\mathrm{T}} < 480\,\mathrm{GeV}$. Distributions for the pass region are shown on the top and for the fail region on the bottom. The distributions on the left are shown before the maximum-likelihood fit and the distributions on the right after the fit. Data is shown as black dots with vertical bars showing the statistical uncertainties on the data. Simulation is shown as filled histograms with the hatched area showing the total uncertainty on the simulation. A ratio of data divided by simulation is shown under each distribution. The dark gray band shows the statistical uncertainty on the simulation and the light gray band the total uncertainty.

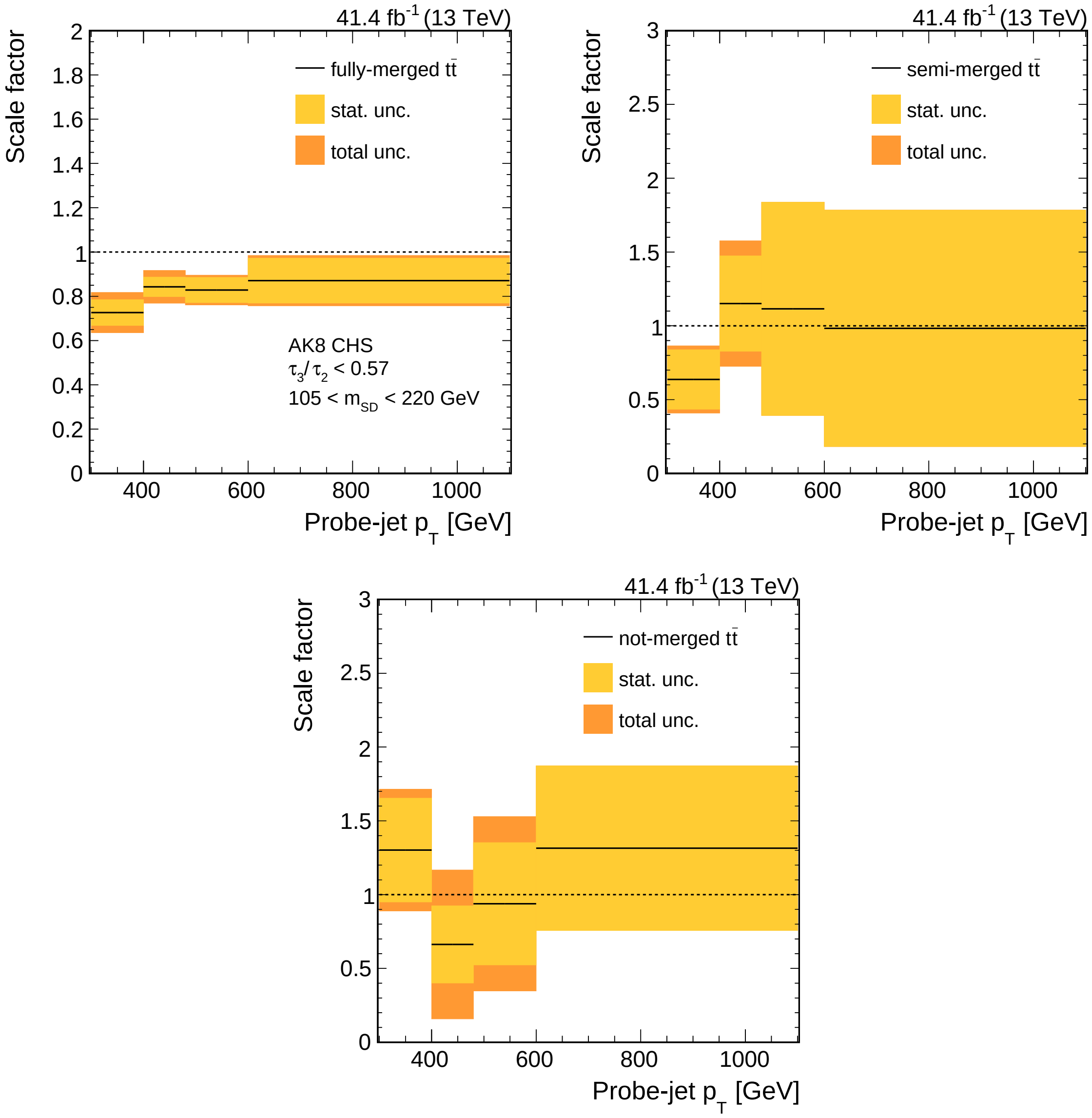

Figure 7.31: Data-to-simulation scale factors for the CMSTopTagger v2 with AK8 CHS jets as a function of the probe-jet p_{T}. The scale factors are obtained for an example working point corresponding to a selection of $\tau_3/\tau_2 < 0.57$ and no subjet b tagging applied. The figure on the top left shows the scale factors for fully-merged $t\bar{t}$ decays, the one on the top right for semi-merged $t\bar{t}$ decays, and the figure on the bottom shows the scale factors for the not-merged category.

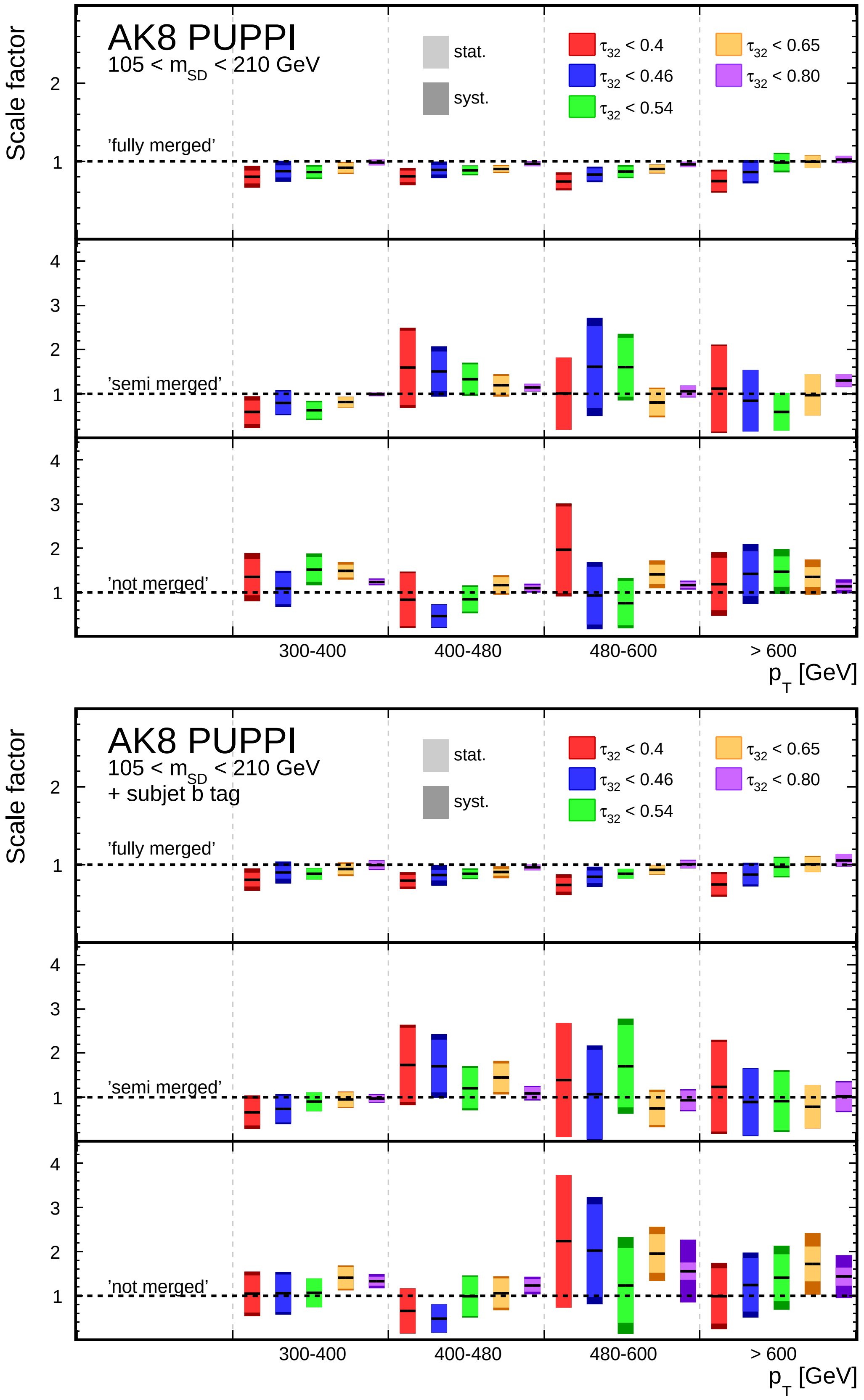

Figure 7.32: Overview of all data-to-simulation scale factors measured in 2017 data for the CMSTopTagger v2 with AK8 PUPPI jets. Scale factors for several working points are shown without subjet b tagging (top) and with a subjet b tag (bottom).

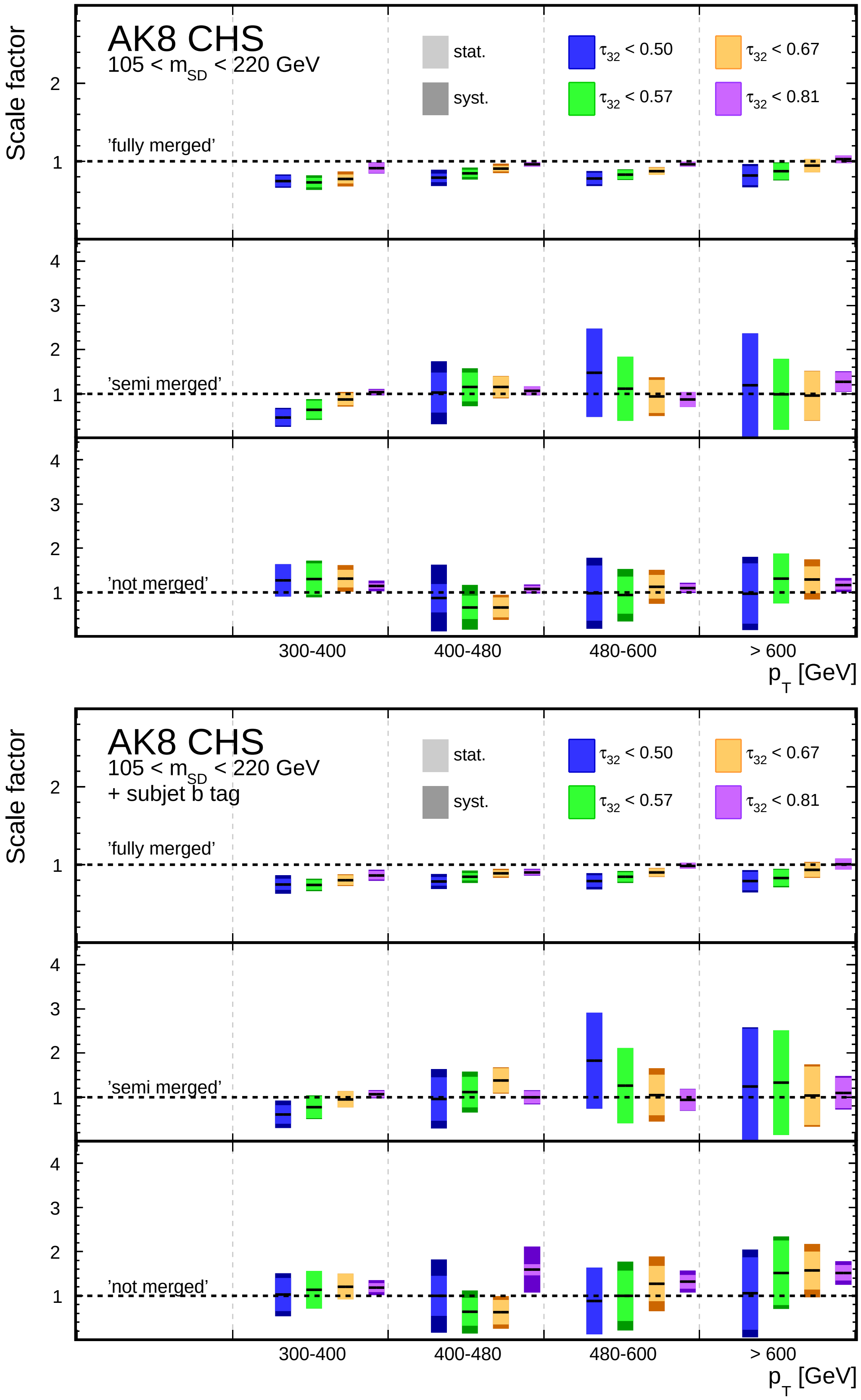

Figure 7.33: Overview of all data-to-simulation scale factors measured in 2017 data for the CMSTopTagger v2 with AK8 CHS jets. Scale factors for several working points are shown without subjet b tagging (top) and with a subjet b tag (bottom).

7.7 Summary

Comparisons of the performance of the CMSTopTagger v2 and HOTVR have been performed in MC simulation reproducing the 2016 data taking conditions. Both taggers are studied for PUPPI and CHS pileup removal. HOTVR uses a jet clustering with a variable distance parameter and is able to reconstruct jets at lower p_{T} with a larger radius. Therefore it can tag top quarks with lower momentum compared to the CMSTopTagger v2. The tagging efficiency of HOTVR is observed to be better than for the CMSTopTagger v2 at low momentum ($300 < p_{\mathrm{T}} < 470\,\mathrm{GeV}$) and very similar for high momentum ($1000 < p_{\mathrm{T}} < 1400\,\mathrm{GeV}$). Jets with PUPPI pileup removal are observed to be less effected by pileup compared to CHS jets.

Data-to-simulation scale factors to correct for differences in the tagging efficiency between data and simulation have been derived as a function of the probe-jet p_{T} and for 'fully-merged', 'semi-merged', and 'not-merged' contributions of the $\mathrm{t\bar{t}}$ simulation. The different contributions are defined by a matching of the top quark decay products from the MC generator to the jets in simulation. The efficiencies in data are obtained by a maximum-likelihood fit of the different contributions to the data in a pass and a fail region. Scale factors for the CMSTopTagger v2 with CHS and PUPPI and for HOTVR with PUPPI have been derived in 2016 data. The scale factors for the 'fully-merged' contributions are consistent with one for both taggers. The CMSTopTagger v2 is also studied in 2017 data with CHS and PUPPI. The scale factors for the 'fully-merged' contribution are significantly lower than one because of a worse description of the N-subjettiness ratio τ_3/τ_2 by the simulation compared to the 2016 studies. Studies show that at least part of the worse description might be related to the change of the parton-shower tune from the 2016 to the 2017 simulation. The measured scale factors for 2016 and 2017 data are used in several CMS publications using top tagging in searches for new physics with boosted top quarks. They are important for those analyses to correct for differences in the top tagging efficiency between data and simulation.

The mistag rate was studied in 2016 data in a region enriched with QCD multijet production. The efficiency as a function of p_{T} in data was compared to different QCD simulations and data-to-simulation scale factors have been derived. The scale factors are different from unity but constant as a function of p_{T}.

8 Conclusion

Two analyses with jet substructure in high-momentum top quark production have been performed within the scope of this thesis. The high center-of-mass energy in pp collisions at the LHC up 13 TeV leads to a large production of high-momentum top quarks from standard model processes and could also lead to a production of very heavy new particles that decay into high-momentum top quarks. Because of the large Lorentz boost the reconstruction of top quarks at very high momentum becomes challenging and top quarks are often reconstructed in large jets that are identified by algorithms using jet-substructure information. Together with the production of other boosted objects this lead to an increasing interest on jet substructure in experiments and in theoretical calculations.

The first analysis in this thesis is the first measurement of the jet-mass distribution in fully-merged top quark decays in data collected at a center-of-mass energy of 8 TeV. The jet mass is an important jet-substructure variable used in many top tagging algorithms to identify jets that contain a fully-merged top quark decay. The data is corrected to the particle level and compared to different MC simulations. The shape of the jet-mass distribution is consistent between data and simulation. The measured data can be compared to new simulations and to analytic calculations once they are available for the measured phase space. This might help to improve the understanding of the underlying physics of jet substructure. The peak position of the measured distribution is sensitive to the mass of the top quark which is a fundamental parameter of the standard model. A first extraction of the top quark mass in the highly boosted regime was performed by a comparison with MC simulation leading to a value of $m_{\mathrm{t}} = 170.8 \pm 9.0$ GeV. A future comparison with analytic calculations could lead to an extraction of a well-defined top quark mass. The uncertainties in this measurement are still dominated by statistical uncertainties and large improvements are expected with 13 TeV data.

The second analysis includes studies of the performance of two top tagging algorithms, the CMSTopTagger v2 and the HOTVR algorithm, in data and simulation. Top tagging is an important tool for many analysis searching for heavy new particles decaying into top quarks with high momentum. A good understanding of the performance in data and simulation is therefore important. The HOTVR algorithm shows a better performance at low p_{T} and a performance comparable to the CMSTopTagger v2 at high p_{T}. Both algorithms show a more stable behavior against pileup for jets clustered with PUPPI pileup removal compared to CHS jets. A new method is used in this thesis to measure the top tagging data-to-simulation scale factors in 2016 and 2017 data at a center-of-mass energy of 13 TeV. The scale factors are used in many CMS analysis with boosted top quarks in the final state to correct for differences in the top tagging efficiency between data and simulation. They are measured as a function of the jet-p_{T} for a 'fully-merged', a 'semi-merged', and a 'not-merged' contribution using a template fit method. This method leads to less dependence of the scale factors on the measurement phase space compared to previous methods. The scale factors for the 'fully-merged' category are consistent with one for the 2016 data and significantly lower for 2017 data because of a worse description of the N-subjettiness distribution by the simulation. The mistag rates have been studied in 2016 data in a phase space enriched with QCD multijet production using a simple cut-and-count method.

A Additional material for the measurement of the top-jet mass

A.1 Unfolding model dependence tests

This section includes additional model-dependence tests on the unfolding as described in section 6.7.4. These tests are the basis for the estimation of the model-dependent uncertainties applied to the final measurement. Figures A.1 and A.2 show an unfolding of pseudo-data simulated with POWHEG +PYTHIA with renormalization and factorization scales μ_r and μ_f scaled by a factors of 0.5 and 2 for $m_{\mathrm{t\bar{t}}} > 700\,\mathrm{GeV}$. The pseudo-data is unfolded with the default POWHEG +PYTHIA simulation also for $m_{\mathrm{t\bar{t}}} > 700\,\mathrm{GeV}$. Figure A.3 shows an unfolding of POWHEG +PYTHIA simulated pseudo-data with a reweighed top quark p_{T} spectrum unfolded with the default simulation. The influence on a different parton-shower model is tested in figure A.4, where pseudo-data simulated with MC@NLO +HERWIG is unfolded with the POWHEG +PYTHIA simulation.

The influence of different top quark masses used in the simulation is tested with MadGraph +PYTHIA. Simulated pseudo-data with different top quark masses are unfolded with the central MadGraph +PYTHIA sample with a top quark mass of 172.5 GeV. The masses of the different samples vary from the central value by ±1 GeV, ±3 GeV, and ±6 GeV. The tests are shown in figures A.5-A.10.

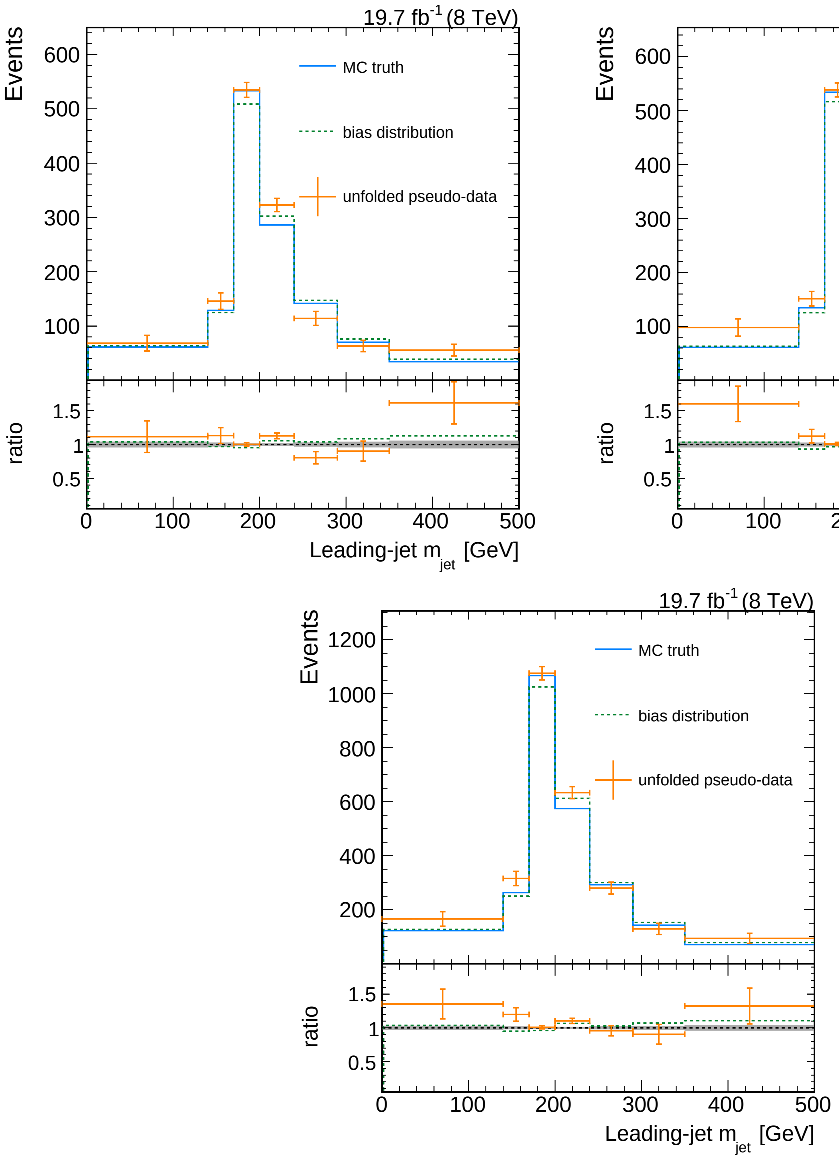

Figure A.1: Distribution of the leading-jet mass for unfolded pseudo-data simulated with POWHEG +PYTHIA with renormalization and factorization scales μ_r and μ_f scaled by a factor of 2 and unfolded with a response matrix evaluated with the default POWHEG +PYTHIA sample. The unfolding is performed for $m_{t\bar{t}} > 700\,\text{GeV}$. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

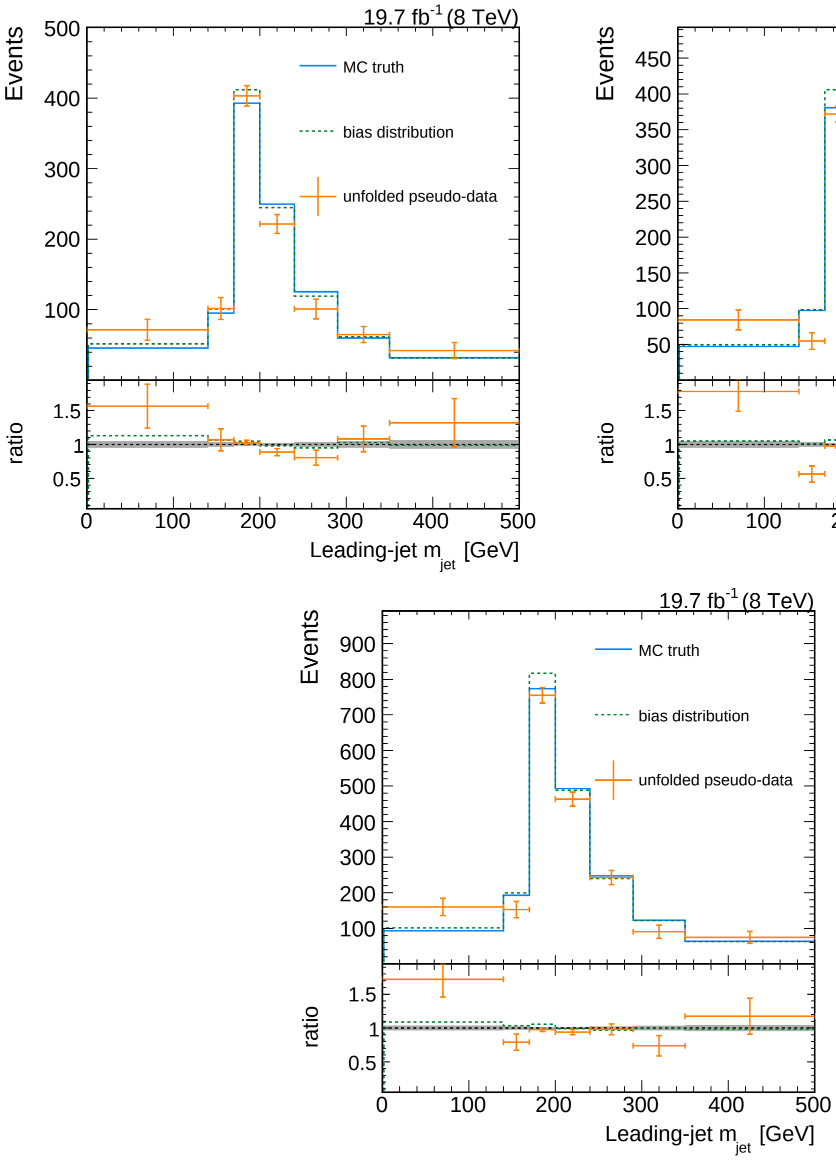

Figure A.2: Distribution of the leading-jet mass for unfolded pseudo-data simulated with POWHEG +PYTHIA with renormalization and factorization scales μ_r and μ_f scaled by a factor of 0.5 and unfolded with a response matrix evaluated with the default POWHEG +PYTHIA sample. The unfolding is performed for $m_{t\bar{t}} > 700\,\text{GeV}$. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

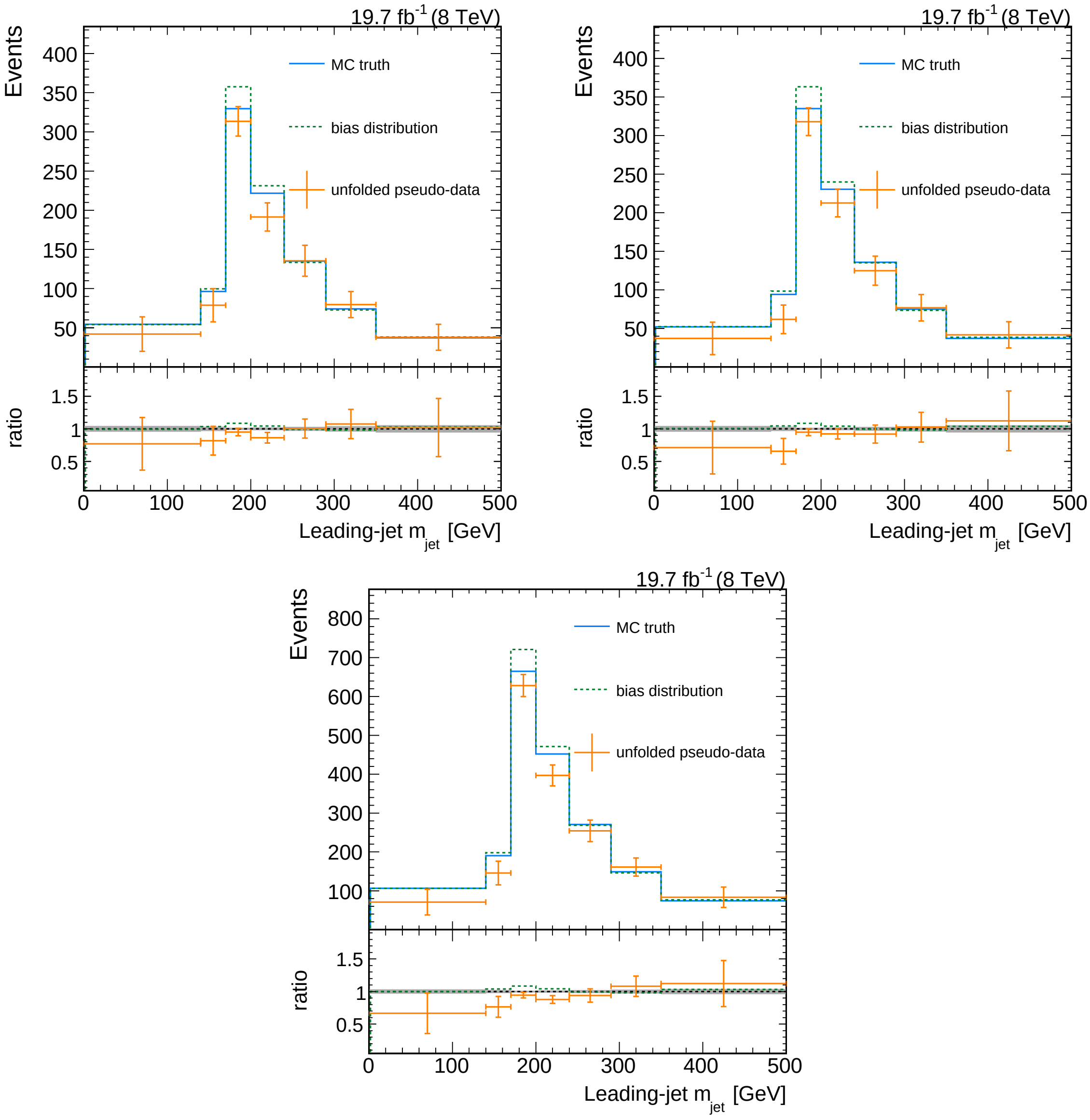

Figure A.3: Distribution of the leading-jet mass for unfolded pseudo-data simulated with POWHEG +PYTHIA with a reweighted top quark p_{T} spectrum and unfolded with a response matrix evaluated with POWHEG +PYTHIA. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

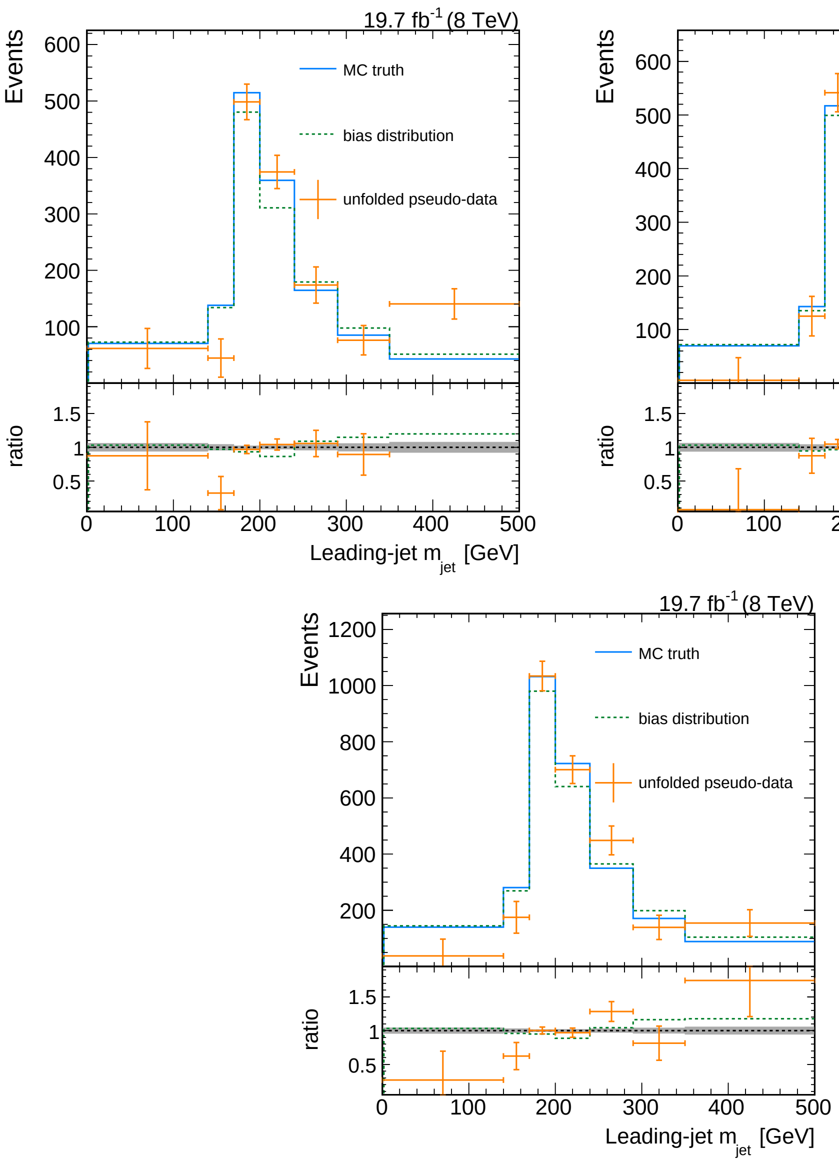

Figure A.4: Distribution of the leading-jet mass for unfolded pseudo-data simulated with MC@NLO +HERWIG and unfolded with a response matrix evaluated with POWHEG +PYTHIA. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

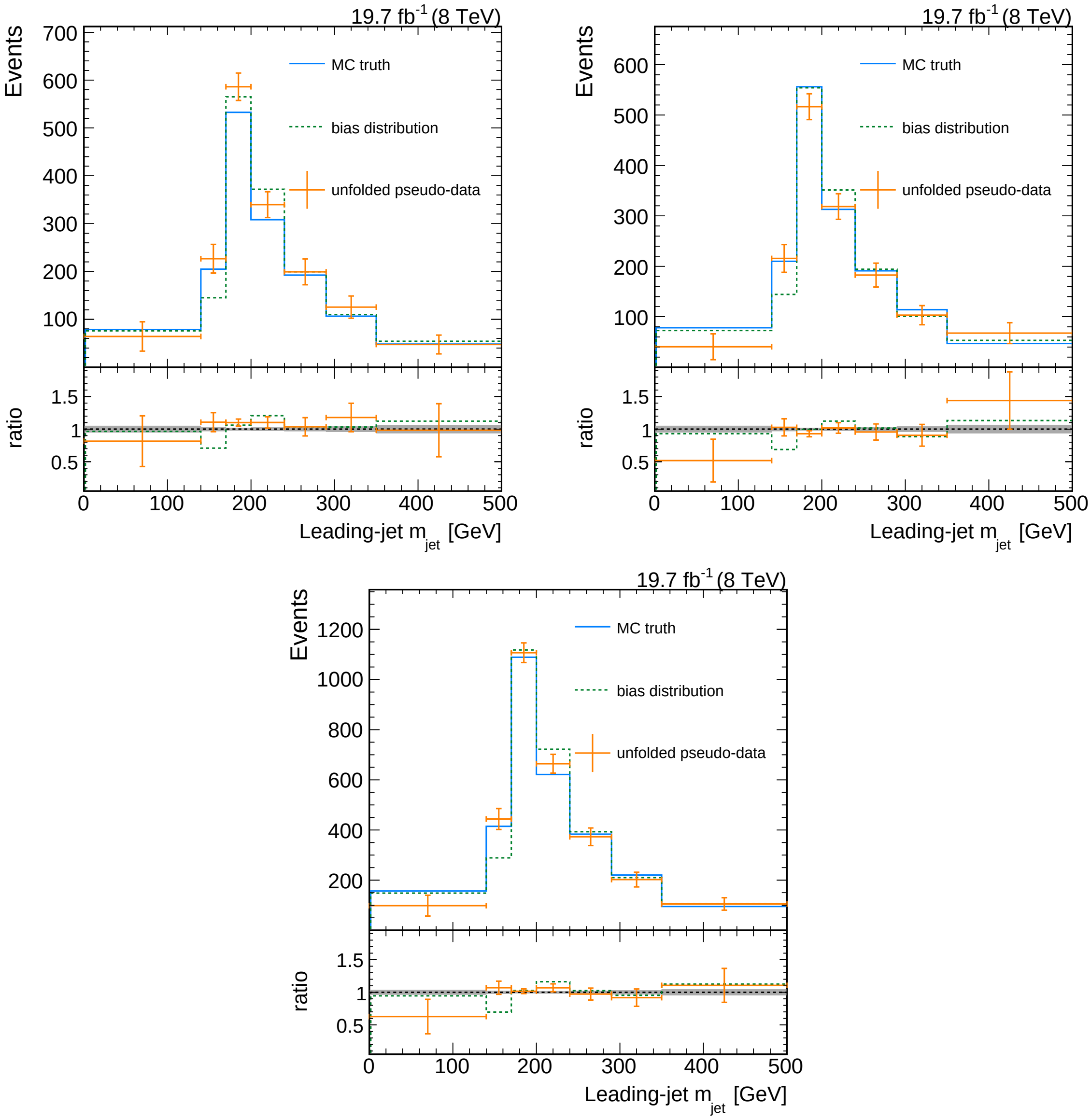

Figure A.5: Distribution of the leading-jet mass for unfolded pseudo-data simulated with MadGraph +pythia and a top quark mass of 166.5 GeV. The pseudo-data is unfolded with a response matrix evaluated with MadGraph+pythia with a top quark mass of 172.5 GeV. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

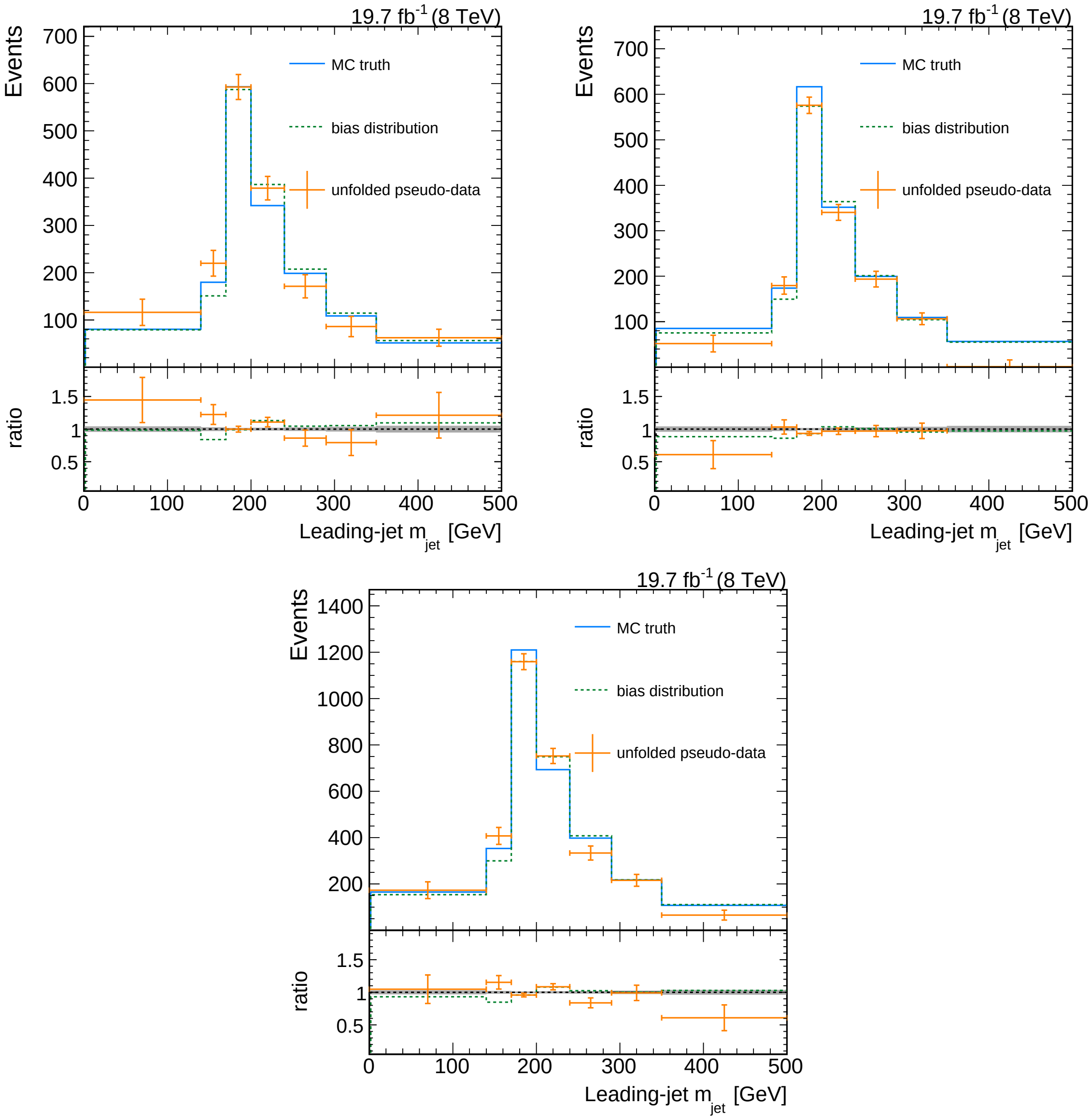

Figure A.6: Distribution of the leading-jet mass for unfolded pseudo-data simulated with MadGraph+pythia and a top quark mass of 169.5 GeV. The pseudo-data is unfolded with a response matrix evaluated with MadGraph+pythia with a top quark mass of 172.5 GeV. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

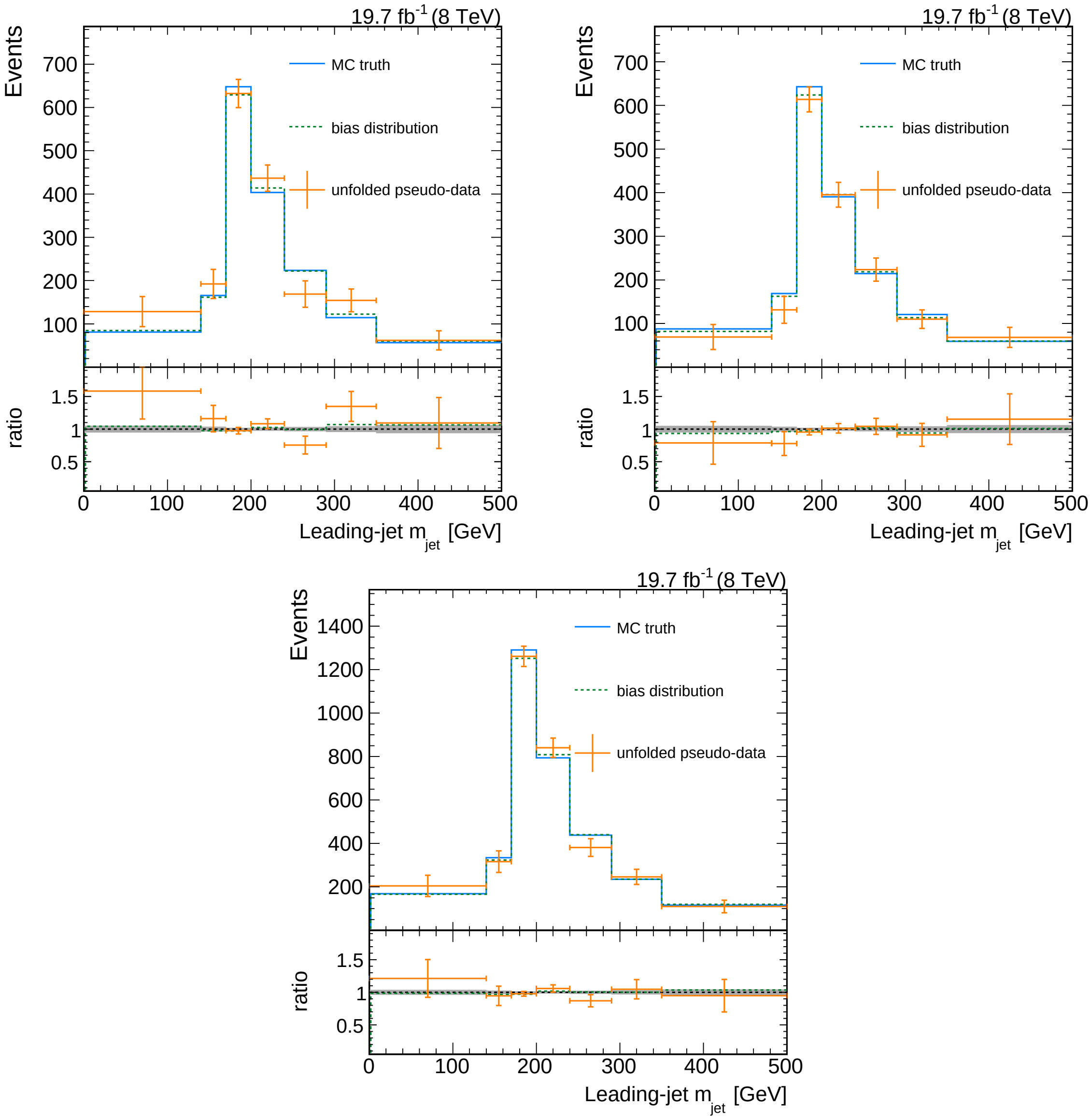

Figure A.7: Distribution of the leading-jet mass for unfolded pseudo-data simulated with MadGraph+pythia and a top quark mass of 171.5 GeV. The pseudo-data is unfolded with a response matrix evaluated with MadGraph+pythia with a top quark mass of 172.5 GeV. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

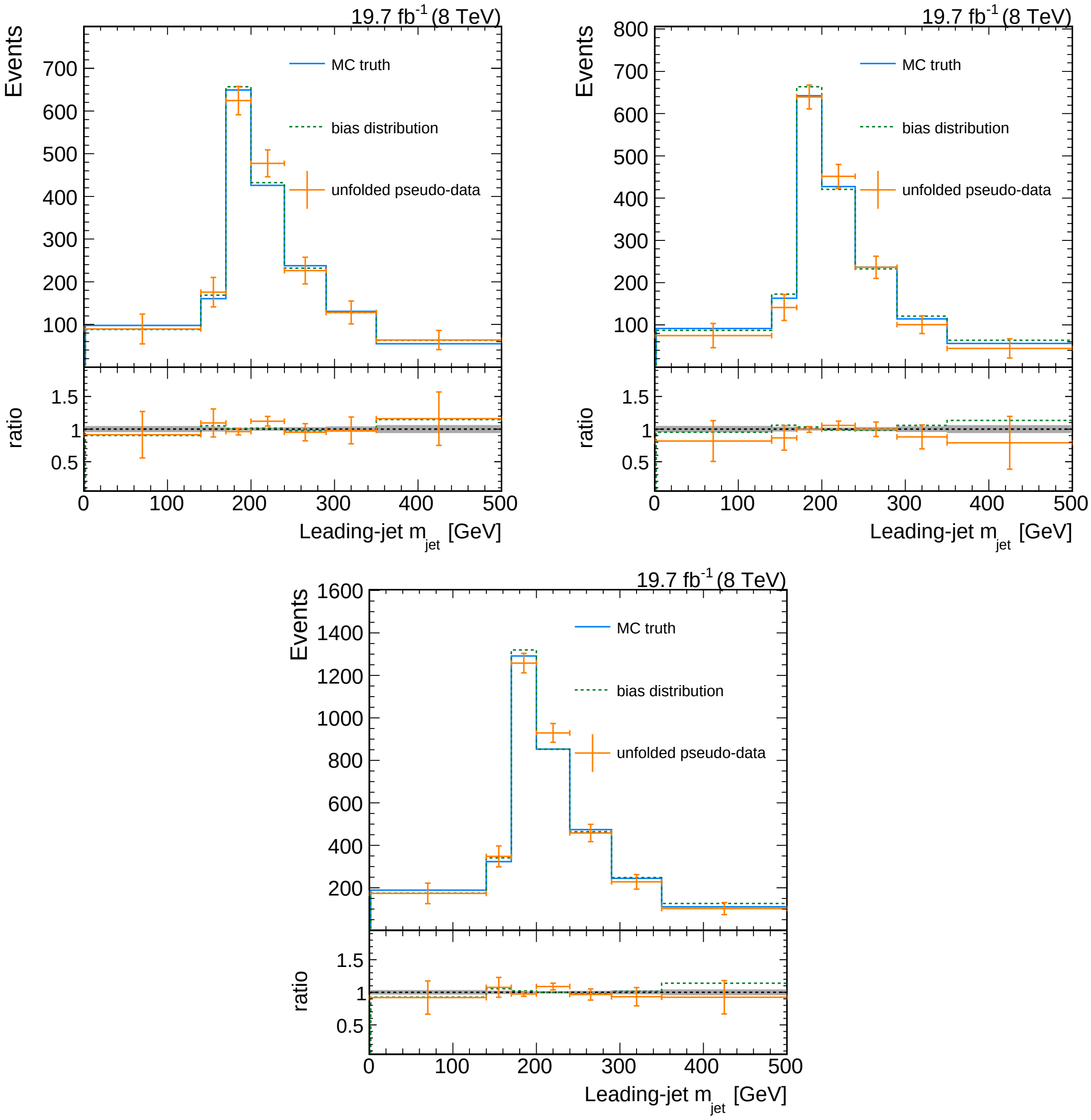

Figure A.8: Distribution of the leading-jet mass for unfolded pseudo-data simulated with MadGraph+pythia and a top quark mass of 173.5 GeV. The pseudo-data is unfolded with a response matrix evaluated with MadGraph+pythia with a top quark mass of 172.5 GeV. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

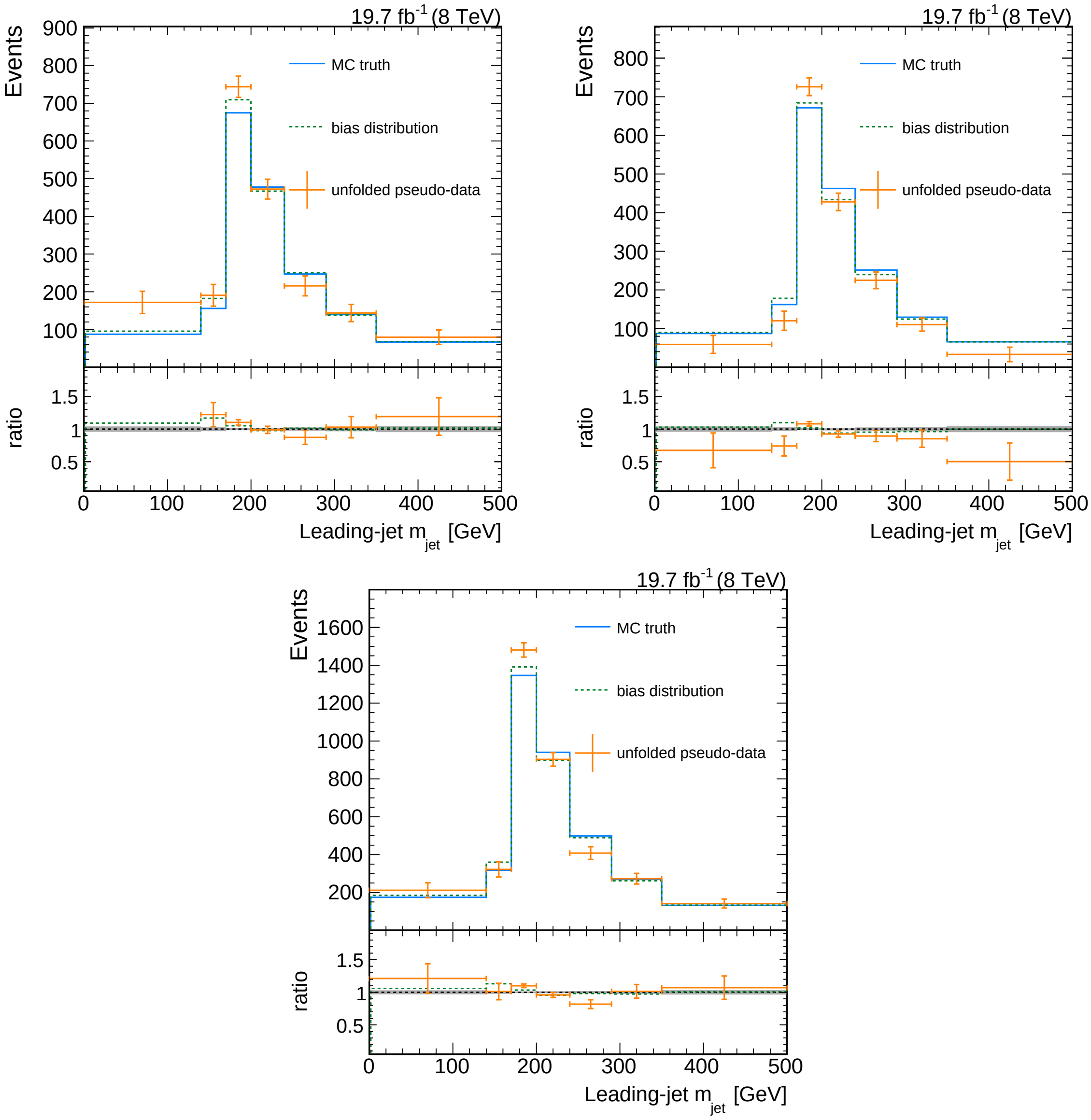

Figure A.9: Distribution of the leading-jet mass for unfolded pseudo-data simulated with MadGraph +pythia and a top quark mass of 175.5 GeV. The pseudo-data is unfolded with a response matrix evaluated with MadGraph+pythia with a top quark mass of 172.5 GeV. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

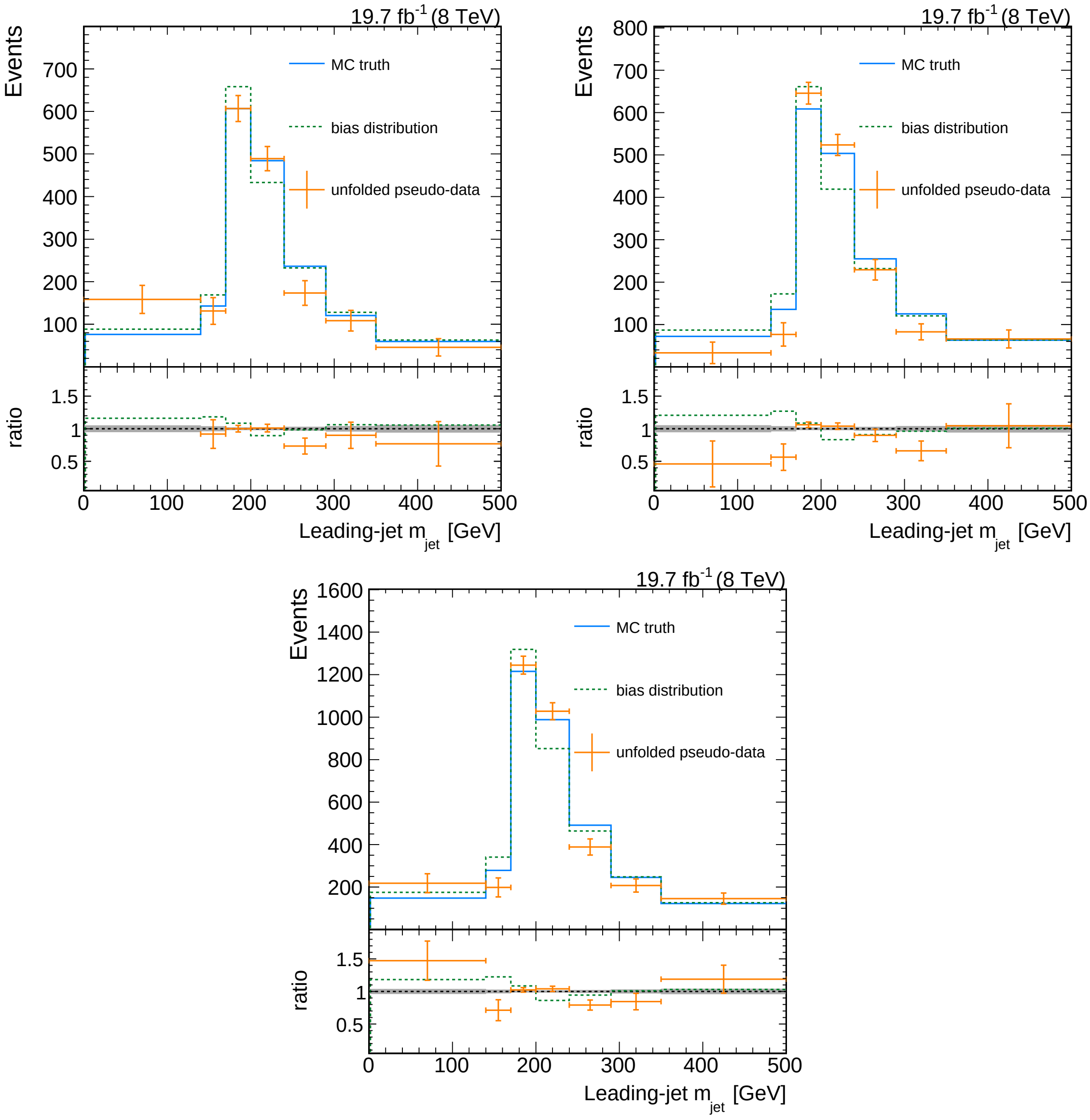

Figure A.10: Distribution of the leading-jet mass for unfolded pseudo-data simulated with MadGraph+pythia and a top quark mass of 178.5 GeV. The pseudo-data is unfolded with a response matrix evaluated with MadGraph +pythia with a top quark mass of 172.5 GeV. The unfolded pseudo-data is compared to the respective particle-level distribution and to the bias distribution used in the unfolding. The figure on the top left shows the electron channel, the one on the top right shows the muon channel, and the figure on the bottom shows an unfolding in the combination of both channels.

A.2 Purity, stability, and reconstruction efficiency

Purity and stability are two properties that are often studied for an unfolding to decide on the optimal bin width for the measurement. They are defined as

$$\text{purity} = \frac{N_{\text{rec,gen}}}{N_{\text{rec}}} \quad \text{and} \tag{A.1}$$

$$\text{stability} = \frac{N_{\text{rec,gen}}}{N_{\text{gen}}}, \tag{A.2}$$

where $N_{\text{rec,gen}}$ is the number of events that are reconstructed and generated in the same bin and N_{rec} and N_{gen} are the number of reconstructed and the number of generated events in a certain bin. The purity and stability should give an estimate on the corrections in the unfolding process and should be sufficiently high. In the case of the unfolding in this measurement the binning at the detector level is different to the binning at the particle level which makes the definition difficult. For the distributions of the purity and stability in figure A.11 the particle-level binning is used also at the detector level. The distributions are derived with the default $t\bar{t}$ simulation with POWHEG +PYTHIA. Low values are observed for both quantities rising with increasing jet mass. This effect can be explained with the observation of low masses being reconstructed significantly higher in figure 6.9. This leads to the effect that events generated with low mass are more likely reconstructed in the next bin than in the bin in which they are generated. This results in low purity values at low masses. For the unfolding in TUnfold it should not be important

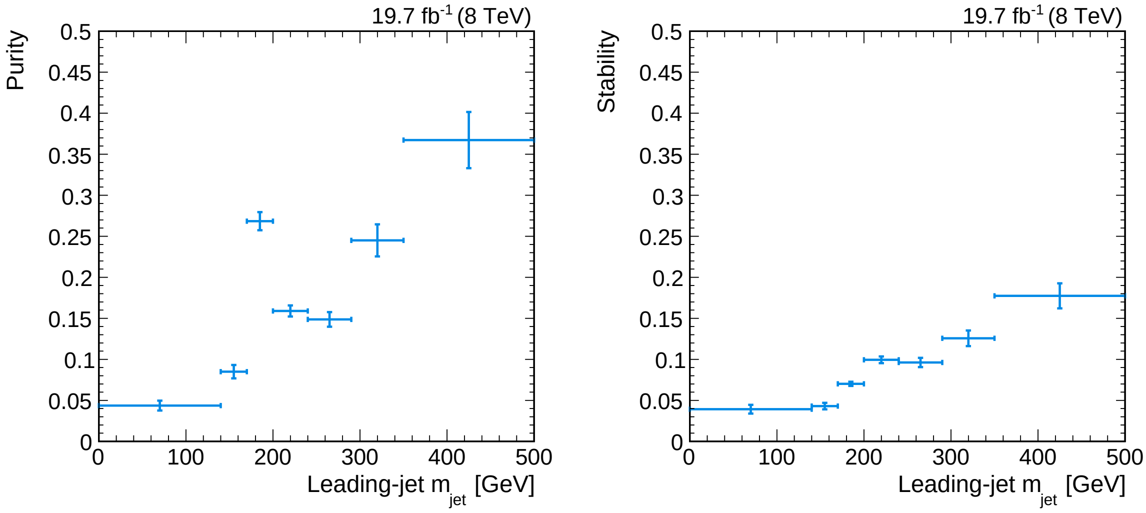

Figure A.11: Purity and stability for the unfolding described in chapter 6. The same binning is used for the particle and the detector level to evaluate these properties. They are estimated with the default $t\bar{t}$ simulation with POWHEG +PYTHIA.

in which bin the events are reconstructed as long as they are not spread over too many bins. The unfolding should be able to correct for the shift between the particle and the detector level. For this reason the width of the reconstruction resolution in figure 6.9 is used to define the binning at the particle level instead of purity and stability.

The stability values are even lower compared to the purity because it is influenced by the reconstruction efficiency shown in figure A.12. The reconstruction efficiency is defined as the number of events that are generated in one bin and reconstructed in any other bin divided by the number of all generated events.

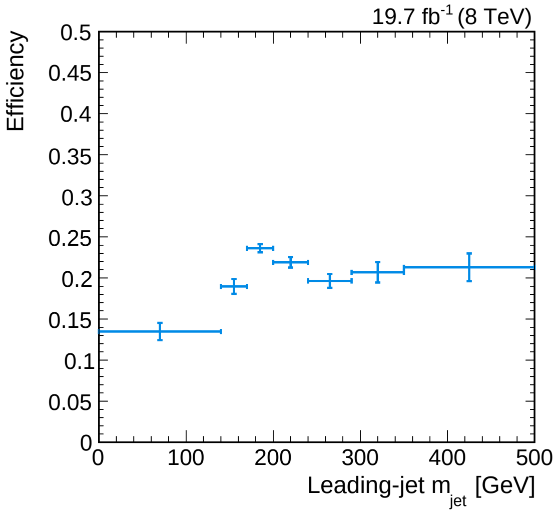

Figure A.12: Reconstruction efficiency for the unfolding described in chapter 6. The efficiency is estimated with the default $t\bar{t}$ simulation with POWHEG +PYTHIA.

A.3 Estimation of the jet-mass scale

This section includes studies on the jet-mass scale of the CA12 jets used in chapter 6. These studies are important since the mass of the CA12 jets have never been studied in CMS before and it is not known how well the jet mass is described by the simulation. Corrections might be needed to account for differences of the jet-mass scale between data and simulation.

The jet-mass scale is studied in a phase space complimentary to the measurement phase space on jets that include a hadronic W boson decay in $t\bar{t}$ decays. Since the W boson mass is well known it serves as a good reference to test if the jet-mass scale is well described.

The studies that are presented in this section are based on studies on the jet-mass scale in the bachelor thesis of Malte Stender in reference [187].

A.3.1 Event selection and reconstruction

The study is performed in $\mathrm{t\bar{t}}$ decays in the muon+jets decay channel. Events are selected by the same non-isolated single-muon trigger as it was also used in section 6.6 requiring a muon with $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ and $|\eta| < 2.1$. The following event selection is applied.

- The event is required to include exactly one muon with $p_{\mathrm{T}} > 45\,\mathrm{GeV}$ and $|\eta| < 2.1$.
- A veto is set on additional electron candidates.
- At least two AK5 jets are required with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$.
- A two-dimensional isolation criterion is applied as described in section 6.6.

The reconstruction of the $\mathrm{t\bar{t}}$ system and the selection of the W-jet candidate is now based on a definition of a hadronic and a leptonic hemisphere. It starts with the reconstruction of the leptonic top quark. The closest AK5 jet to the muon with a $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.1$ passing a CSV medium b tag is considered as the b jet from the leptonic top quark decay. The neutrino is reconstructed from $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ and the muon assuming they originate from a W boson with a mass of $m_{\mathrm{W}} = 80\,\mathrm{GeV}$. The leptonic top quark is obtained by a combination of the four-momenta of the muon, the b jet, and neutrino candidates.

Objects with a distance to the leptonic top candidate smaller than $\Delta R < \pi/2$ are associated to the leptonic hemisphere and objects with $\Delta R > \pi/2$ are associated to the hadronic hemisphere. Events with more than one b jet in the leptonic hemisphere or masses of the leptonic top quark candidate larger than 230 GeV are not considered in the following.

The reconstruction continues with the definition of a b jet from the hadronic top quark decay as the leading AK5 jet in the hadronic hemisphere with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.1$ passing a loose CSV b tag. Two additional AK5 jets in the hadronic hemisphere with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and $|\eta| < 2.4$ serve as candidates for the two light quarks from the W boson decay. Events with more than two AK5 jets with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and $|\eta| < 2.4$ in the hadronic hemisphere are rejected. The hadronic W-jet candidates is finally defined as a CA12 jet in the hadronic hemisphere with $90 < p_{\mathrm{T}} < 260\,\mathrm{GeV}$, a distance to the two light-quark jets smaller than $\Delta R < 1.2$, and a distance to the b jet larger than 1.2 but smaller than 2.1.

Events are only kept if the full reconstruction described above was successful. The distribution of the invariant mass of the hadronic W-jet candidate is shown in figure A.13 in data and simulation. The $t\bar{t}$ simulation is divided into a matched and mismatched category. Events are called matched if the distances between the W-jet candidate and the two quarks from the top quark decay at the generator level are smaller than 1.0. Otherwise they are called mismatched. The matched events show a clear peak at about 100 GeV connected to the W boson mass and shifted to higher values by additional radiation. The distribution for the mismatched jets peaks at a much lower value.

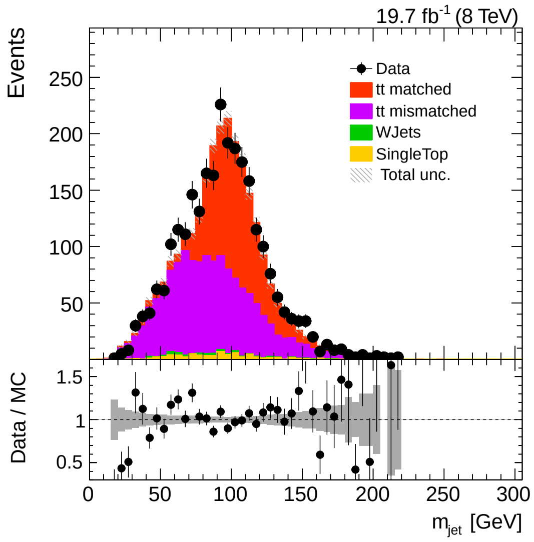

Figure A.13: Distribution of the invariant mass of the hadronic W-jet candidate. The data is shown as black dots with vertical bars showing the statistical uncertainty on the data. Simulation is shown as filled histograms. The hatched region shows the statistical uncertainty on the simulation. Below the mass distribution a ratio of data divided by simulation is shown. The gray area shows the statistical uncertainty on the simulation.

A.3.2 Fits

The next step is a determination of the peak position of the jet-mass spectrum for fully-merged W jets in simulation and data. The matched events in figure A.13 are therefore considered as signal, while the mismatched $t\bar{t}$ events and the background processes are considered as background. A simultaneous fit of the signal and background contributions is performed. The signal is described by a mirrored Crystal Ball function. The Crystal Ball function was originally developed to describe a distribution with a Gaussian core and a tail to lower values due to radiation loss. In this case the Crystal Ball function is mirrored to describe a tail to higher masses because of additional radiation. The background is modeled by an exponential function of the from:

$$f(x) = a(x-b)^c e^{d(x-b)^f}. \tag{A.3}$$

The signal and background contributions are first fitted separately in simulation. The background-only fit is performed with a background function and a Crystal Ball function. This is needed because the separation of signal and background by the matching is not perfect and there is some signal left in the background simulation.

The fitted parameters of the separate fits are used as input for the combined fit. The combined fit is performed with one background and one signal function. Figure A.14 shows the jet-mass distribution of the W-jet candidate in simulation on the left and in data on the right. The distributions are shown together with the fitted functions from the combined fit as solid lines and the background contributions of the full fit functions as dashed lines. The peak positions of the signal contribution in simulation and in data are found by the fit to be:

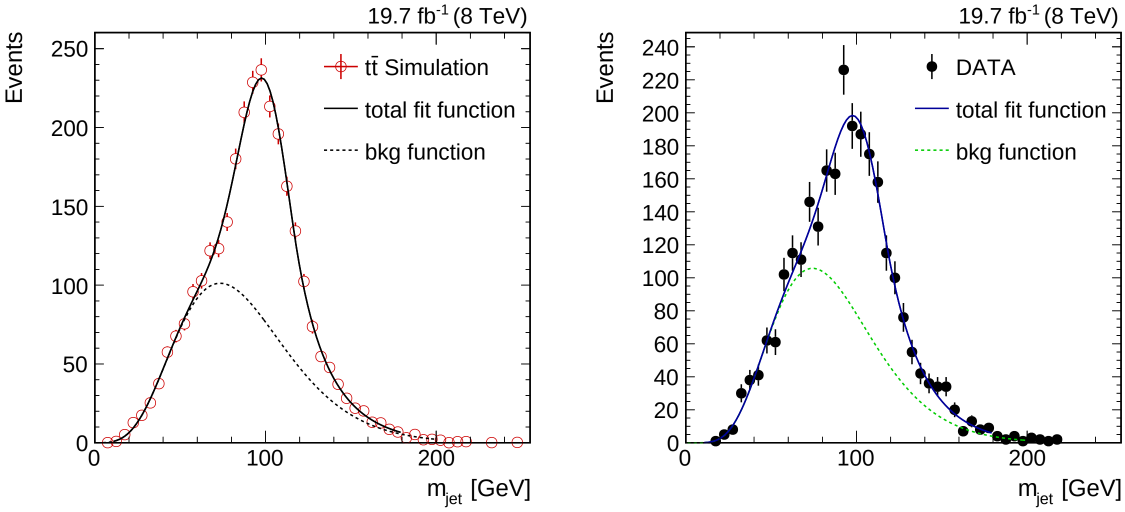

Figure A.14: Distribution of the invariant mass of the hadronic W-jet candidate in simulation (left) and in data (right). The distributions are shown together with the fitted functions from the combined fit as solid lines and the background contribution of the full fit function as dashed lines.

$$\begin{aligned} \mu_{\text{Data}} &= 101.3 \pm 1.1\,\text{GeV} \\ \mu_{\text{MC}} &= 99.8 \pm 0.5\,\text{GeV}, \end{aligned}$$

where only statistical uncertainties are considered. The peak positions in data and simulation are consistent within the statistical uncertainties. Also the fitted width of the signal contribution is consistent in data and simulation with

$$\begin{aligned}\sigma_{\mathrm{Data}} &= 16.0 \pm 2.5\,\mathrm{GeV}\\ \sigma_{\mathrm{MC}} &= 15.0 \pm 0.8\,\mathrm{GeV}.\end{aligned}$$

Therefore no corrections on the jet-mass scale or the jet-mass resolution are applied in the measurement of the top quark mass peak in chapter 6. The jet-mass scale is still varied by 1.5% corresponding to the difference in the peak position between data and simulation.

A.4 Jet energy correction studies

This section includes studies for the jet energy corrections applied to the CA12 jets. The four-vectors of the CA12 jets are corrected with corrections derived within CMS for AK7 jets because no corrections for CA12 jets have been derived and the AK7 jets are the most similar to the CA12 jets with available corrections. Studies are following to test if the uncertainty on the JECs are large enough to cover the possible difference between the AK7 jets and the CA12 jets.

In the ideal case the JECs should correct the jet p_{T} at the reconstruction level to match the p_{T} at the particle level. The relative difference in the transverse momentum between the reconstruction and particle level $\left(p_{\mathrm{T}}^{\mathrm{rec}} - p_{\mathrm{T}}^{\mathrm{gen}}\right)/p_{\mathrm{T}}^{\mathrm{gen}}$ is studied in different bins of the particle-level $p_{\mathrm{T}}^{\mathrm{gen}}$. The value for $p_{\mathrm{T}}^{\mathrm{gen}}$ is obtained from the particle-level jet closest to the reconstruction-level jet. The peak positions of the resulting distributions are derived by the mean from ± three bins around the peak bin. The mean difference is shown as a function of $p_{\mathrm{T}}^{\mathrm{gen}}$ on the left in figure A.15. The uncertainties shown on the values are evaluated by the mean of the distribution in the case of up or down variations of the JECs. The uncertainties on the JECs are now increased to be consistent with zero shown in the right distribution of figure A.15. These uncertainties are used in the final measurement. The uncertainty for $300 < p_{\mathrm{T}}^{\mathrm{gen}} < 350\,\mathrm{GeV}$ is also used for all jets with $p_{\mathrm{T}}^{\mathrm{gen}} < 300\,\mathrm{GeV}$.

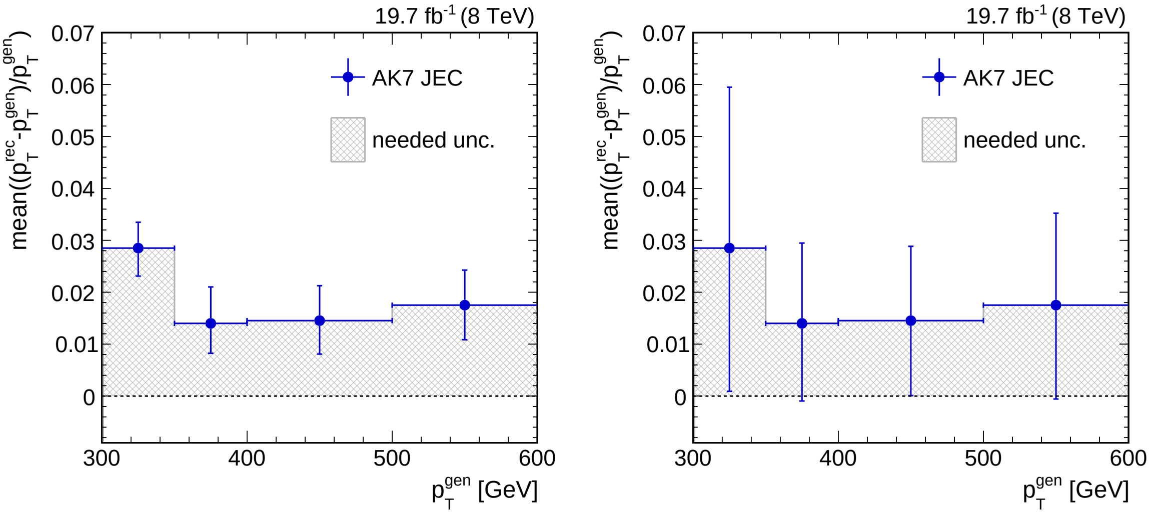

Figure A.15: Mean relative difference in the jet p_{T} between the reconstruction level and the particle level $\left(p_{\mathrm{T}}^{\mathrm{rec}} - p_{\mathrm{T}}^{\mathrm{gen}}\right)/p_{\mathrm{T}}^{\mathrm{gen}}$ as a function of $p_{\mathrm{T}}^{\mathrm{gen}}$. The uncertainties shown on the values in the left distributions result from variations of the AK7 correction within their uncertainties derived by CMS. The uncertainties on the right are increased to be consistent with zero. These uncertainties are used in the final measurement.

A.5 Uncertainties on the cross-section measurement

This section includes additional uncertainty tables for the cross-section measurements presented in section 6.10. Tables A.2 and A.1 provide more details on individual uncertainties for the cross-section measurements in tables 6.2 and 6.3. Covariance matrices for these measurements with statistical uncertainties only and with the full uncertainties are shown in tables A.3 to A.6. The respective correlation coefficients are shown in figures A.16 and A.17.

Table A.1: More details on the different uncertainty sources for each bin of the differential cross-section measurement presented in table 6.2 in section 6.10.

Range in m_{jet} [GeV]	140–170	170–200	200–240	240–290	290–350
Statistical uncertainties [%]					
Input	45	12	18	28	240
Response matrix	26	5.7	10	18	170
Background subtraction	13	2.2	2.8	4.3	41
Total statistical uncertainty [%]	54	13	21	34	300
Systematic uncertainties [%]					
JER	0.97	0.3	1.7	0.89	4.0
JEC	13	7.7	14	18	7.1
Pileup	4.9	2.6	2.1	5.4	0.92
B tag	2.2	1.4	1.6	1.7	3.6
Mass scale	37	3.6	4.7	5.5	11
Trigger	1.1	0.66	0.87	0.75	1.5
Background subtraction	5.6	0.56	2.5	2.2	20
Luminosity	2.6	2.6	2.6	2.6	2.6
Total systematic uncertainty [%]	40	9.4	16	20	25
Model uncertainties [%]					
Scale	21	2.4	6.0	2.0	26
Parton shower+MC	38	0.2	3.0	28	19
Choice of m_{t}	29	10	8.9	21	16
PDF	4.4	0.69	0.89	3.6	2.3
Total model uncertainty [%]	52	10	11	35	36
Total uncertainty [%]	85	19	28	53	300

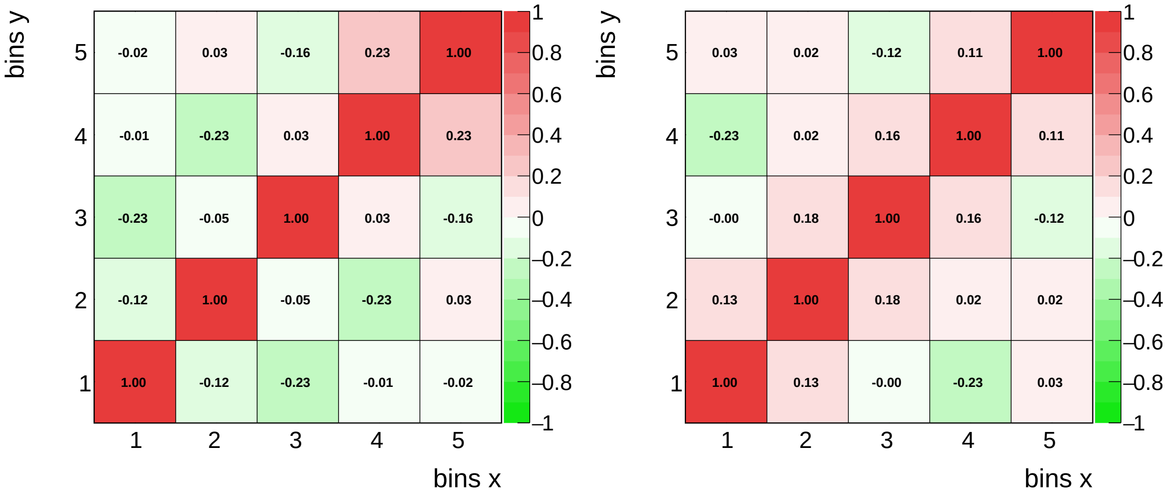

Figure A.16: Correlation coefficients for the differential cross section presented in figure 6.20 and table 6.2 in section 6.10. The left figure shows the correlation coefficients with statistical uncertainties only and the figure on the right with full uncertainties.

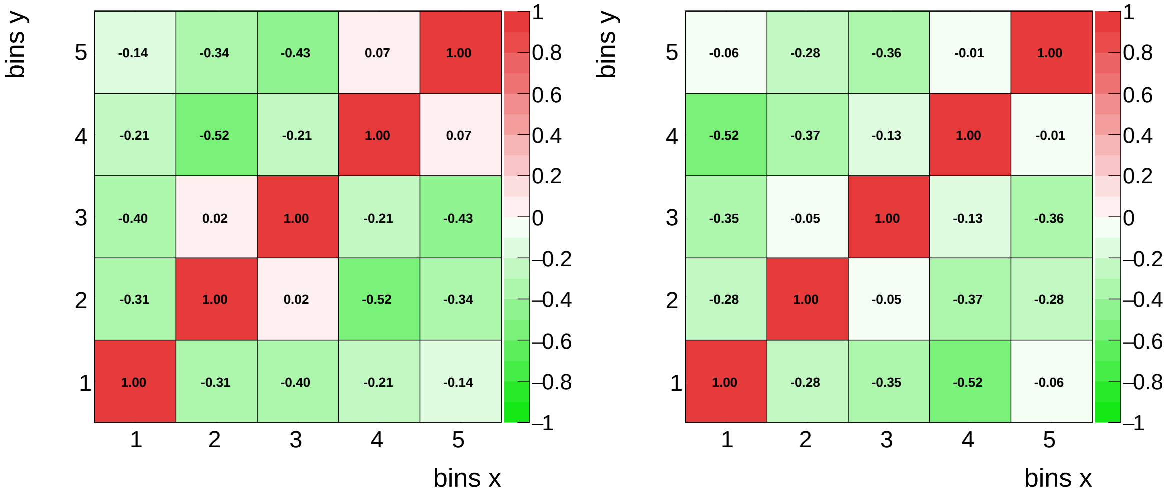

Figure A.17: Correlation coefficients for the normalized differential cross section presented in figure 6.22 and table 6.3 in section 6.10. The left figure shows the correlation coefficients with statistical uncertainties only and the figure on the right with full uncertainties.

Table A.2: More details on the different uncertainty sources for each bin of the normalized differential cross-section measurement presented in table 6.3 in section 6.10

Range in m_{jet} [GeV]	140–170	170–200	200–240	240–290	290–350
Statistical uncertainties [%]					
Input	43	13	18	24	240
Response matrix	25	7.2	10	16	170
Background subtraction	12	2.6	3.2	3.7	40
Total statistical uncertainty [%]	51	15	21	29	290
Systematic uncertainties [%]					
JER	1.7	0.38	0.97	0.21	3.3
JEC	1.0	4.0	2.3	6.8	19
Pileup	4.8	2.5	2.2	5.5	1.0
B tagging	0.61	0.26	0.01	0.03	2.0
Mass scale	34	0.19	8.5	9.3	7.0
Trigger	0.28	0.13	0.07	0.04	0.67
Background subtraction	4.4	1.6	1.3	0.25	18
Total systematic uncertainty [%]	34	4.9	9.2	13	27
Model uncertainties [%]					
Scale	16	2.7	1.1	5.3	26
Parton shower	37	0.53	2.7	29	18
Choice of m_{t}	26	8.1	9.3	18	17
PDF	3.6	1.3	1.2	3.0	2.0
Total model uncertainty [%]	48	8.6	9.8	34	36
Total uncertainty [%]	78	18	25	47	300

Table A.3: Covariance matrix for the differential cross-section measurement presented in figure 6.20 and table 6.2 in section 6.10. Only statistical uncertainties are shown in [fb^2]. The covariance matrix is published in [3].

Bin	1	2	3	4	5
1	+40.1309	−4.3127	−7.9546	−0.2265	−0.6234
2		+31.6527	−1.5335	−8.0645	+0.7645
3			+30.7374	+0.9627	−4.5225
4				+38.1138	+7.3090
5					+26.1656

Table A.4: Covariance matrix for the differential cross-section measurement presented in figure 6.20 and table 6.2 in section 6.10. The full uncertainties are shown in [fb^2]. The covariance matrix is published in [3].

Bin	1	2	3	4	5
1	+100.3776	+10.4024	−0.2932	−22.5348	+1.6248
2		+66.1185	+11.0437	+1.3971	+0.7563
3			+57.3797	+11.9955	−4.6654
4				+93.7987	+5.2878
5					+26.7189

Table A.5: Covariance matrix for the normalized differential cross-section measurement presented in figure 6.22 and table 6.3 in section 6.10. Only statistical uncertainties are shown in units of [10^{-4}]. The covariance matrix is published in [3].

Bin	1	2	3	4	5
1	+35.0451	−11.2095	−12.9678	−6.7079	−4.1598
2		+38.3103	+0.6831	−17.1837	−10.6001
3			+30.1347	−6.0193	−11.8308
4				+28.1469	+1.7640
5					+24.8267

Table A.6: Covariance matrix for the normalized differential cross-section measurement presented in figure 6.22 and table 6.3 in section 6.10. The full uncertainties are shown in units of [10^{-4}]. The covariance matrix is published in [3].

Bin	1	2	3	4	5
1	+83.2384	−18.8992	−21.0414	−40.7114	−2.5865
2		+55.5641	−2.6015	−23.7075	−10.3559
3			+43.1286	−7.4742	−12.0116
4				+72.3585	−0.4654
5					+25.4194

A.6 Top quark mass calibration

This section includes studies of the extraction procedure of the top quark mass using simulated pseudo-data and studies for a possible calibration of the measured top quark mass in data.

The top quark mass is extracted from simulated pseudo-data in the same way as it is done in data described in section 6.11. The pseudo-data is simulated with MadGraph +pythia and different values of the top quark mass used in the simulation. The pseudo-data is unfolded with the default $\mathrm{t\bar{t}}$ simulation with powheg +pythia and the top quark mass is extracted from the unfolded distributions. Figure A.18 shows the measured values from the different pseudo-data samples against the top quark mass used in the simulation. In the optimal case the extraction would return perfectly the mass used in simulation and all points would lie on the diagonal. Figure A.18 however shows a bias that can be used to calibrate the mass measurement in data.

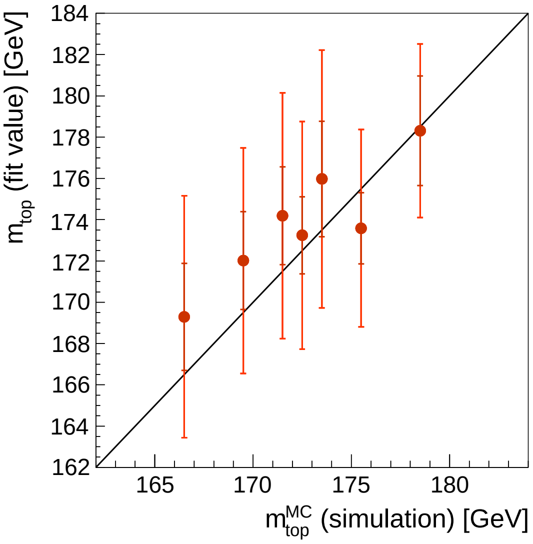

Figure A.18: Top quark mass extracted from unfolded pseudo-data simulated with MadGraph +pythia and different values of m_t. The pseudo-data is unfolded with the default powheg +pythia sample. The measured values are shown as a function of the value of m_t used in the simulation. The inner error bars show the statistical uncertainty and the outer error bar the full uncertainty including systematic and model uncertainties.

In the case of a mass calibration the model uncertainty on the choice of the top quark mass in simulation applied to the measured cross sections is removed to avoid a double counting of uncertainties. The mass extraction from pseudo-data is redone without this uncertainty shown in figure A.19. A linear function is fitted to the points using the

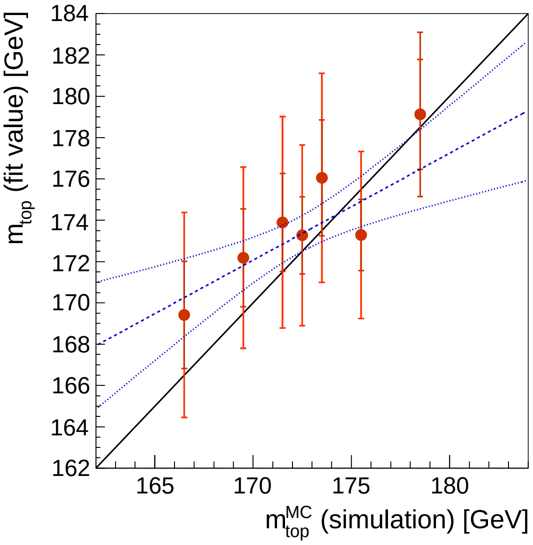

Figure A.19: Top quark mass extracted from unfolded pseudo-data simulated with MadGraph +pythia and different values of m_t. The extraction is performed without the model uncertainty on the choice of m_t in the unfolding. The pseudo-data is unfolded with the default powheg +pythia sample. The measured values are shown as a function of the value of m_t used in the simulation. The inner error bars show the statistical uncertainty and the outer error bar the full uncertainty including systematic and model uncertainties. A linear function is fitted to the points using statistical uncertainties only and shown as a dashed line. The fine dashed lines show the uncertainty on the fit function.

statistical uncertainties only.

Without the model uncertainty on m_t on the differential cross section a top quark mass of

$$m_t = 170.5 \pm 8.6\,\text{GeV} \tag{A.4}$$

is extracted from the data. This mass can now be calibrated with the calibration function in figure A.19. This leads to a top quark mass of

$$m_t = 167.0^{+16.7}_{-7.1}\,\text{GeV}, \tag{A.5}$$

with additional uncertainties from the calibration of +2.4 and −5.0 GeV. The calibration leads to a lower value of the measured mass and a largely asymmetric uncertainty.

The calibration has the disadvantage that it is not reproducible using the published cross-section measurement. This however is a very important aspect of this mass determination because it should serve as a demonstration on the sensitivity that can be reached in future studies using the cross-section measurement directly. The calibration further leads to a

]

very large upper uncertainty reaching far in a region where no simulated samples exist and it is not clear how trustworthy the extrapolation into this region is.

For these reasons the calibration is not used on the final result of this measurement but instead the model uncertainty for the choice of m_t in simulation is used in the mass extraction. It was finally tested that the mass uncertainty is sufficient to cover the bias observed in figure A.19. This is tested evaluating the calibration function again using statistical plus mass uncertainties shown in figure A.20. The resulting calibration function is consistent with the diagonal within the uncertainties.

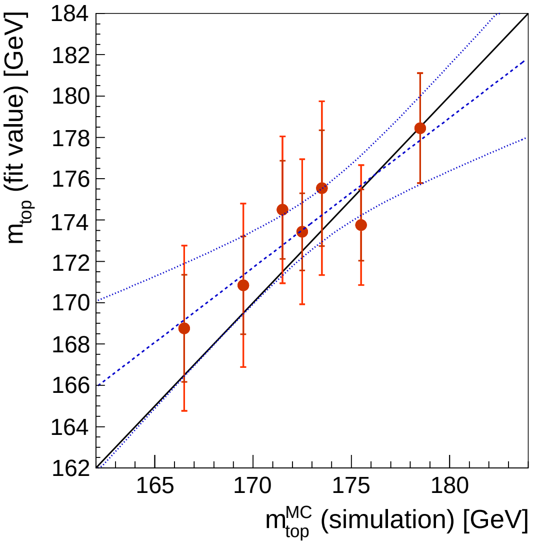

Figure A.20: Top quark mass extracted from unfolded pseudo-data simulated with MadGraph +pythia and different values of m_t. The extraction is performed with only the statistical uncertainties and the the model uncertainty on the choice of m_t. The pseudo-data is unfolded with the default powheg +pythia sample. The measured values are shown as a function of the value of m_t used in the simulation. The inner error bars show the statistical uncertainty and the outer error bar the full uncertainty including systematic and model uncertainties. A linear function is fitted to the points using statistical uncertainties and the uncertainties on the choice of m_t. It is shown as a dashed line. The fine dashed lines show the uncertainty on the fit function.

Bibliography

[1] D. J. Griffiths, *Introduction to elementary particles; 2nd rev. version.* Physics textbook. Wiley, New York, NY, 2008.

[2] **Particle Data Group** Collaboration, “Review of Particle Physics”, *Phys. Rev. D* **98** (Aug, 2018) 030001, doi:10.1103/PhysRevD.98.030001.

[3] **CMS** Collaboration, “Measurement of the jet mass in highly boosted $\mathrm{t\bar{t}}$ events from pp collisions at $\sqrt{s} = 8$ TeV”, *Eur. Phys. J.* **C77** no. 7, (2017) 467, doi:10.1140/epjc/s10052-017-5030-3, `arXiv:1703.06330`.

[4] T. Dreyer, “Studies for the measurement of the jet mass in fully merged hadronic top quark decays”, Master’s thesis, Universität Hamburg, 2015.

[5] **CMS** Collaboration, “Top Tagging with New Approaches”, CMS Physics Analysis Summary CMS-PAS-JME-15-002, 2016. `https://cds.cern.ch/record/2126325`.

[6] T. Lapsien, R. Kogler, and J. Haller, “A new tagger for hadronically decaying heavy particles at the LHC”, *Eur. Phys. J. C* **76** (2016) 600, doi:10.1140/epjc/s10052-016-4443-8, `arXiv:1606.04961`.

[7] Cush, “Standard Model of Elementary Particles.” `https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg`. (accessed August 21, 2019).

[8] **CMS** Collaboration, “Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV”, *Eur. Phys. J. C* **75** (2015) 212, doi:10.1140/epjc/s10052-015-3351-7, `arXiv:1412.8662`.

[9] **ATLAS** Collaboration, “Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment”, *Eur. Phys. J.* **C76** no. 1, (2016) 6, doi:10.1140/epjc/s10052-015-3769-y, `arXiv:1507.04548`.

[10] **ATLAS, CMS** Collaboration, "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV", *JHEP* **08** (2016) 045, doi:10.1007/JHEP08(2016)045, `arXiv:1606.02266`.

[11] **CMS** Collaboration, "Combined measurements of Higgs boson couplings in proton–proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$", *Eur. Phys. J.* **C79** no. 5, (2019) 421, doi:10.1140/epjc/s10052-019-6909-y, `arXiv:1809.10733`.

[12] **ATLAS** Collaboration, "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector", *Phys. Lett.* **B784** (2018) 173–191, doi:10.1016/j.physletb.2018.07.035, `arXiv:1806.00425`.

[13] **ATLAS** Collaboration, "Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", *Phys. Rev.* **D99** (2019) 072001, doi:10.1103/PhysRevD.99.072001, `arXiv:1811.08856`.

[14] **ATLAS** Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B* **716** (2012) 1, doi:10.1016/j.physletb.2012.08.020, `arXiv:1207.7214`.

[15] **CMS** Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys. Lett.* **B716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, `arXiv:1207.7235`.

[16] **Planck** Collaboration, "Planck 2015 results. XIII. Cosmological parameters", *Astron. Astrophys.* **594** (2016) A13, doi:10.1051/0004-6361/201525830, `arXiv:1502.01589`.

[17] **Planck** Collaboration, "Planck 2018 results. VI. Cosmological parameters", `arXiv:1807.06209`.

[18] A. Riotto, "Theories of baryogenesis", in *Proceedings, Summer School in High-energy physics and cosmology: Trieste, Italy, June 29-July 17, 1998*, pp. 326–436. 1998. `arXiv:hep-ph/9807454`.

[19] **D0** Collaboration, "Observation of the top quark", *Phys. Rev. Lett.* **74** (1995) 2632–2637, doi:10.1103/PhysRevLett.74.2632, `arXiv:hep-ex/9503003`.

[20] **CDF** Collaboration, "Observation of top quark production in $\bar{p}p$ collisions", *Phys. Rev. Lett.* **74** (1995) 2626–2631, doi:10.1103/PhysRevLett.74.2626, `arXiv:hep-ex/9503002`.

[21] ATLAS, CDF, CMS and D0 Collaborations, “First combination of Tevatron and LHC measurements of the top-quark mass.” 2014.

[22] **CMS** Collaboration, “Measurement of the top quark mass using proton-proton data at $\sqrt{s} = 7$ and 8 TeV”, *Phys. Rev. D* **93** (2016) 072004, doi:10.1103/PhysRevD.93.072004, arXiv:1509.04044.

[23] J. Haller, A. Hoecker, R. Kogler, K. Mönig, T. Peiffer, and J. Stelzer, “Update of the global electroweak fit and constraints on two-Higgs-doublet models”, *Eur. Phys. J.* **C78** no. 8, (2018) 675, doi:10.1140/epjc/s10052-018-6131-3, arXiv:1803.01853.

[24] **NNPDF** Collaboration, “Parton distributions for the LHC Run II”, *JHEP* **04** (2015) 040, doi:10.1007/JHEP04(2015)040, arXiv:1410.8849.

[25] M. Czakon and A. Mitov, “Top++: A program for the calculation of the top-pair cross-section at hadron colliders”, *Comput. Phys. Commun.* **185** (2014) 2930, doi:10.1016/j.cpc.2014.06.021, arXiv:1112.5675.

[26] G. Degrassi, S. Di Vita, J. Elias-Miro, J. R. Espinosa, G. F. Giudice, G. Isidori, and A. Strumia, “Higgs mass and vacuum stability in the Standard Model at NNLO”, *JHEP* **08** (2012) 098, doi:10.1007/JHEP08(2012)098, arXiv:1205.6497.

[27] **ATLAS** Collaboration, “Measurement of the top quark mass in the $t\bar{t} \rightarrow$ lepton+jets channel from $\sqrt{s} = 8$ TeV ATLAS data and combination with previous results”, *Eur. Phys. J.* **C79** no. 4, (2019) 290, doi:10.1140/epjc/s10052-019-6757-9, arXiv:1810.01772.

[28] A. H. Hoang and I. W. Stewart, “Top mass measurements from jets and the Tevatron top-quark mass”, *Nucl. Phys. Proc. Suppl.* **185** (2008) 220, doi:10.1016/j.nuclphysbps.2008.10.028, arXiv:0808.0222.

[29] I. I. Y. Bigi, M. A. Shifman, N. G. Uraltsev, and A. I. Vainshtein, “The Pole mass of the heavy quark. Perturbation theory and beyond”, *Phys. Rev.* **D50** (1994) 2234–2246, doi:10.1103/PhysRevD.50.2234, arXiv:hep-ph/9402360.

[30] M. Beneke and V. M. Braun, “Heavy quark effective theory beyond perturbation theory: Renormalons, the pole mass and the residual mass term”, *Nucl. Phys.* **B426** (1994) 301–343, doi:10.1016/0550-3213(94)90314-X, arXiv:hep-ph/9402364.

[31] A. H. Hoang, A. Jain, I. Scimemi, and I. W. Stewart, “Infrared Renormalization Group Flow for Heavy Quark Masses”, *Phys. Rev. Lett.* **101** (2008) 151602, doi:10.1103/PhysRevLett.101.151602, arXiv:0803.4214.

[32] A. H. Hoang, A. Jain, C. Lepenik, V. Mateu, M. Preisser, I. Scimemi, and I. W. Stewart, "The MSR mass and the $\mathcal{O}\left(\Lambda_{\mathrm{QCD}}\right)$ renormalon sum rule", *JHEP* **04** (2018) 003, doi:10.1007/JHEP04(2018)003, `arXiv:1704.01580`.

[33] A. J. Barr and C. G. Lester, "A Review of the Mass Measurement Techniques proposed for the Large Hadron Collider", *J. Phys.* **G37** (2010) 123001, doi:10.1088/0954-3899/37/12/123001, `arXiv:1004.2732`.

[34] C. G. Lester and D. J. Summers, "Measuring masses of semiinvisibly decaying particles pair produced at hadron colliders", *Phys. Lett.* **B463** (1999) 99–103, doi:10.1016/S0370-2693(99)00945-4, `arXiv:hep-ph/9906349`.

[35] H.-C. Cheng and Z. Han, "Minimal Kinematic Constraints and m(T2)", *JHEP* **12** (2008) 063, doi:10.1088/1126-6708/2008/12/063, `arXiv:0810.5178`.

[36] **CMS** Collaboration, "Measurement of the top quark mass in the dileptonic $t\bar{t}$ decay channel using the mass observables $M_{b\ell}$, M_{T2}, and $M_{b\ell\nu}$ in pp collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev.* **D96** no. 3, (2017) 032002, doi:10.1103/PhysRevD.96.032002, `arXiv:1704.06142`.

[37] S. Ferrario Ravasio, T. Ježo, P. Nason, and C. Oleari, "A theoretical study of top-mass measurements at the LHC using NLO+PS generators of increasing accuracy", *Eur. Phys. J.* **C78** no. 6, (2018) 458, doi:10.1140/epjc/s10052-018-5909-7, `arXiv:1801.03944`.

[38] **CMS** Collaboration, "Measurement of the $\mathrm{t\bar{t}}$ production cross section in the $\mathrm{e}\mu$ channel in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV", *JHEP* **08** (2016) 029, doi:10.1007/JHEP08(2016)029, `arXiv:1603.02303`.

[39] **ATLAS** Collaboration, "Measurement of the $\mathrm{t\bar{t}}$ production cross-section using $\mathrm{e}\mu$ events with b-tagged jets in pp collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS detector", *Eur. Phys. J. C* **74** (2014) 3109, doi:10.1140/epjc/s10052-014-3109-7, `arXiv:1406.5375`.

[40] S. Fleming, A. H. Hoang, S. Mantry, and I. W. Stewart, "Jets from massive unstable particles: Top-mass determination", *Phys. Rev. D* **77** (2008) 074010, doi:10.1103/PhysRevD.77.074010, `arXiv:hep-ph/0703207`.

[41] S. Fleming, A. H. Hoang, S. Mantry, and I. W. Stewart, "Top jets in the peak region: Factorization analysis with next-to-leading-log resummation", *Phys. Rev. D* **77** (2008) 114003, doi:10.1103/PhysRevD.77.114003, `arXiv:0711.2079`.

[42] A. H. Hoang, A. Pathak, P. Pietrulewicz, and I. W. Stewart, "Hard matching for boosted tops at two loops", *JHEP* **12** (2015) 059, doi:10.1007/JHEP12(2015)059, `arXiv:1508.04137`.

[43] C. W. Bauer, S. Fleming, and M. E. Luke, "Summing Sudakov logarithms in $B \to X_s + \gamma$ in effective field theory", *Phys. Rev. D* **63** (2000) 014006, doi:10.1103/PhysRevD.63.014006, `arXiv:hep-ph/0005275`.

[44] C. W. Bauer, S. Fleming, D. Pirjol, and I. W. Stewart, "An effective field theory for collinear and soft gluons: Heavy to light decays", *Phys. Rev. D* **63** (2001) 114020, doi:10.1103/PhysRevD.63.114020, `arXiv:hep-ph/0011336`.

[45] C. W. Bauer and I. W. Stewart, "Invariant operators in collinear effective theory", *Phys. Lett. B* **516** (2001) 134, doi:10.1016/S0370-2693(01)00902-9, `arXiv:hep-ph/0107001`.

[46] C. W. Bauer, D. Pirjol, and I. W. Stewart, "Soft-collinear factorization in effective field theory", *Phys. Rev. D* **65** (2002) 054022, doi:10.1103/PhysRevD.65.054022, `arXiv:hep-ph/0109045`.

[47] M. Butenschoen, B. Dehnadi, A. H. Hoang, V. Mateu, M. Preisser, and I. W. Stewart, "Top quark mass calibration for Monte Carlo event generators", *Phys. Rev. Lett.* **117** (2016) 232001, doi:10.1103/PhysRevLett.117.232001, `arXiv:1608.01318`.

[48] I. W. Stewart, F. J. Tackmann, and W. J. Waalewijn, "N-Jettiness: An Inclusive Event Shape to Veto Jets", *Phys. Rev. Lett.* **105** (2010) 092002, doi:10.1103/PhysRevLett.105.092002, `arXiv:1004.2489`.

[49] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, "An Introduction to PYTHIA 8.2", *Comput. Phys. Commun.* **191** (2015) 159–177, doi:10.1016/j.cpc.2015.01.024, `arXiv:1410.3012`.

[50] A. H. Hoang, S. Mantry, A. Pathak, and I. W. Stewart, "Extracting a Short Distance Top Mass with Light Grooming", `arXiv:1708.02586`.

[51] **CMS** Collaboration, "Measurement of the jet mass distribution in highly boosted top quark decays in pp collisions at $\sqrt{s} = 13$ TeV", CMS Physics Analysis Summary CMS-PAS-TOP-19-005, 2019. `https://cds.cern.ch/record/2682624`.

[52] A. J. Larkoski, I. Moult, and B. Nachman, "Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning", `arXiv:1709.04464`.

[53] S. Catani, G. Turnock, and B. R. Webber, "Heavy jet mass distribution in e+ e- annihilation", *Phys. Lett.* **B272** (1991) 368–372, doi:10.1016/0370-2693(91)91845-M.

[54] S. Catani, L. Trentadue, G. Turnock, and B. R. Webber, "Resummation of large logarithms in e+ e- event shape distributions", *Nucl. Phys.* **B407** (1993) 3–42, doi:10.1016/0550-3213(93)90271-P.

[55] Y.-T. Chien and M. D. Schwartz, "Resummation of heavy jet mass and comparison to LEP data", *JHEP* **08** (2010) 058, doi:10.1007/JHEP08(2010)058, `arXiv:1005.1644`.

[56] M. Dasgupta, A. Fregoso, S. Marzani, and A. Powling, "Jet substructure with analytical methods", *Eur. Phys. J.* **C73** no. 11, (2013) 2623, doi:10.1140/epjc/s10052-013-2623-3, `arXiv:1307.0013`.

[57] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, "Towards an understanding of jet substructure", *JHEP* **09** (2013) 029, doi:10.1007/JHEP09(2013)029, `arXiv:1307.0007`.

[58] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft Drop", *JHEP* **05** (2014) 146, doi:10.1007/JHEP05(2014)146, `arXiv:1402.2657`.

[59] C. Frye, A. J. Larkoski, M. D. Schwartz, and K. Yan, "Precision physics with pile-up insensitive observables", `arXiv:1603.06375`.

[60] C. Frye, A. J. Larkoski, M. D. Schwartz, and K. Yan, "Factorization for groomed jet substructure beyond the next-to-leading logarithm", *JHEP* **07** (2016) 064, doi:10.1007/JHEP07(2016)064, `arXiv:1603.09338`.

[61] **CMS** Collaboration, "Measurements of the differential jet cross section as a function of the jet mass in dijet events from proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **11** (2018) 113, doi:10.1007/JHEP11(2018)113, `arXiv:1807.05974`.

[62] J. Thaler and K. Van Tilburg, "Identifying boosted objects with N-subjettiness", *JHEP* **03** (2011) 015, doi:10.1007/JHEP03(2011)015, `arXiv:1011.2268`.

[63] J. Thaler and K. Van Tilburg, "Maximizing boosted top identification by minimizing N-subjettiness", *JHEP* **02** (2012) 093, doi:10.1007/JHEP02(2012)093, `arXiv:1108.2701`.

[64] A. J. Larkoski and J. Thaler, "Unsafe but Calculable: Ratios of Angularities in Perturbative QCD", *JHEP* **09** (2013) 137, doi:10.1007/JHEP09(2013)137, `arXiv:1307.1699`.

[65] G. P. Salam, L. Schunk, and G. Soyez, "Dichroic subjettiness ratios to distinguish colour flows in boosted boson tagging", *JHEP* **03** (2017) 022, doi:10.1007/JHEP03(2017)022, `arXiv:1612.03917`.

[66] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, "Jet substructure as a new Higgs search channel at the LHC", *Phys. Rev. Lett.* **100** (2008) 242001, doi:10.1103/PhysRevLett.100.242001, `arXiv:0802.2470`.

[67] D. Krohn, J. Thaler, and L.-T. Wang, "Jet Trimming", *JHEP* **02** (2010) 084, doi:10.1007/JHEP02(2010)084, `arXiv:0912.1342`.

[68] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, "Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches", *Phys. Rev.* **D81** (2010) 094023, doi:10.1103/PhysRevD.81.094023, `arXiv:0912.0033`.

[69] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, "Better jet clustering algorithms", *JHEP* **08** (1997) 001, doi:10.1088/1126-6708/1997/08/001, `arXiv:hep-ph/9707323`.

[70] M. Wobisch and T. Wengler, "Hadronization corrections to jet cross sections in deep-inelastic scattering", in *Monte Carlo generators for HERA physics, Hamburg, Germany*. 1998. `arXiv:hep-ph/9907280`.

[71] J. A. Aguilar-Saavedra, R. Benbrik, S. Heinemeyer, and M. Pérez-Victoria, "Handbook of vectorlike quarks: Mixing and single production", *Phys. Rev.* **D88** no. 9, (2013) 094010, doi:10.1103/PhysRevD.88.094010, `arXiv:1306.0572`.

[72] **ATLAS** Collaboration, "Search for pair production of heavy vector-like quarks decaying into hadronic final states in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", *Phys. Rev.* **D98** no. 9, (2018) 092005, doi:10.1103/PhysRevD.98.092005, `arXiv:1808.01771`.

[73] **ATLAS** Collaboration, "Search for pair production of heavy vector-like quarks decaying into high-p_T W bosons and top quarks in the lepton-plus-jets final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", *JHEP* **08** (2018) 048, doi:10.1007/JHEP08(2018)048, `arXiv:1806.01762`.

[74] **ATLAS** Collaboration, "Search for pair production of up-type vector-like quarks and for four-top-quark events in final states with multiple b-jets with the ATLAS detector", *JHEP* **07** (2018) 089, doi:10.1007/JHEP07(2018)089, `arXiv:1803.09678`.

[75] **ATLAS** Collaboration, "Search for pair production of vector-like top quarks in events with one lepton, jets, and missing transverse momentum in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector", *JHEP* **08** (2017) 052, doi:10.1007/JHEP08(2017)052, `arXiv:1705.10751`.

[76] **CMS** Collaboration, “Search for vector-like quarks in events with two oppositely charged leptons and jets in proton-proton collisions at $\sqrt{s} = 13$ TeV”, *Eur. Phys. J.* **C79** no. 4, (2019) 364, doi:10.1140/epjc/s10052-019-6855-8, `arXiv:1812.09768`.

[77] **CMS** Collaboration, “Search for a heavy resonance decaying to a top quark and a vector-like top quark in the lepton+jets final state in pp collisions at $\sqrt{s} = 13$ TeV”, *Eur. Phys. J.* **C79** no. 3, (2019) 208, doi:10.1140/epjc/s10052-019-6688-5, `arXiv:1812.06489`.

[78] **CMS** Collaboration, “Search for a W’ boson decaying to a vector-like quark and a top or bottom quark in the all-jets final state”, *JHEP* **03** (2019) 127, doi:10.1007/JHEP03(2019)127, `arXiv:1811.07010`.

[79] **CMS** Collaboration, “Search for single production of vector-like quarks decaying to a top quark and a W boson in proton-proton collisions at $\sqrt{s} = 13$ TeV”, *Eur. Phys. J.* **C79** (2019) 90, doi:10.1140/epjc/s10052-019-6556-3, `arXiv:1809.08597`.

[80] **CMS** Collaboration, “Search for vector-like T and B quark pairs in final states with leptons at $\sqrt{s} = 13$ TeV”, *JHEP* **08** (2018) 177, doi:10.1007/JHEP08(2018)177, `arXiv:1805.04758`.

[81] J. C. Pati and A. Salam, “Erratum: Lepton number as the fourth "color"”, *Phys. Rev. D* **11** (Feb, 1975) 703–703, doi:10.1103/PhysRevD.11.703.2. `https://link.aps.org/doi/10.1103/PhysRevD.11.703.2`.

[82] H. Georgi and S. L. Glashow, “Unity of All Elementary-Particle Forces”, *Phys. Rev. Lett.* **32** (Feb, 1974) 438–441, doi:10.1103/PhysRevLett.32.438. `https://link.aps.org/doi/10.1103/PhysRevLett.32.438`.

[83] S. Chakdar, T. Li, S. Nandi, and S. K. Rai, “Unity of elementary particles and forces for the third family”, *Phys. Lett.* **B718** (2012) 121–124, doi:10.1016/j.physletb.2012.10.021, `arXiv:1206.0409`.

[84] B. Schrempp and F. Schrempp, “Light leptoquarks”, *Physics Letters B* **153** no. 1, (1985) 101 – 107, doi:https://doi.org/10.1016/0370-2693(85)91450-9.

[85] B. Gripaios, “Composite Leptoquarks at the LHC”, *JHEP* **02** (2010) 045, doi:10.1007/JHEP02(2010)045, `arXiv:0910.1789`.

[86] E. Farhi and L. Susskind, “Technicolour”, *Physics Reports* **74** no. 3, (1981) 277 – 321, doi:https://doi.org/10.1016/0370-1573(81)90173-3.

[87] **CMS** Collaboration, “Search for third-generation scalar leptoquarks decaying to a top quark and a τ lepton at $\sqrt{s} = 13$ TeV”, *Eur. Phys. J.* **C78** (2018) 707, doi:10.1140/epjc/s10052-018-6143-z, `arXiv:1803.02864`.

[88] **CMS** Collaboration, “Search for leptoquarks coupled to third-generation quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV”, *Phys. Rev. Lett.* **121** no. 24, (2018) 241802, doi:10.1103/PhysRevLett.121.241802, `arXiv:1809.05558`.

[89] **ATLAS** Collaboration, “Searches for third-generation scalar leptoquarks in $\sqrt{s} =$ 13 TeV pp collisions with the ATLAS detector”, `arXiv:1902.08103`.

[90] J. L. Rosner, “Prominent decay modes of a leptophobic Z'”, *Phys. Lett.* **B387** (1996) 113–117, doi:10.1016/0370-2693(96)01022-2, `arXiv:hep-ph/9607207`.

[91] K. R. Lynch, E. H. Simmons, M. Narain, and S. Mrenna, “Finding Z' bosons coupled preferentially to the third family at LEP and the Tevatron”, *Phys. Rev.* **D63** (2001) 035006, doi:10.1103/PhysRevD.63.035006, `arXiv:hep-ph/0007286`.

[92] M. Carena, A. Daleo, B. A. Dobrescu, and T. M. P. Tait, “Z' gauge bosons at the Tevatron”, *Phys. Rev.* **D70** (2004) 093009, doi:10.1103/PhysRevD.70.093009, `arXiv:hep-ph/0408098`.

[93] K. Agashe, A. Belyaev, T. Krupovnickas, G. Perez, and J. Virzi, “LHC Signals from Warped Extra Dimensions”, *Phys. Rev.* **D77** (2008) 015003, doi:10.1103/PhysRevD.77.015003, `arXiv:hep-ph/0612015`.

[94] **CMS** Collaboration, “Search for resonant $\mathrm{t\bar{t}}$ production in proton-proton collisions at $\sqrt{s} = 13$ TeV”, *JHEP* **04** (2019) 031, doi:10.1007/JHEP04(2019)031, `arXiv:1810.05905`.

[95] **ATLAS** Collaboration, “Search for heavy particles decaying into top-quark pairs using lepton-plus-jets events in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”, *Eur. Phys. J.* **C78** no. 7, (2018) 565, doi:10.1140/epjc/s10052-018-5995-6, `arXiv:1804.10823`.

[96] **ATLAS** Collaboration, “Search for heavy particles decaying into a top-quark pair in the fully hadronic final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”, `arXiv:1902.10077`.

[97] P. Nason, “A new method for combining NLO QCD with shower Monte Carlo algorithms”, *JHEP* **11** (2004) 040, doi:10.1088/1126-6708/2004/11/040, `arXiv:hep-ph/0409146`.

[98] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with parton shower simulations: The POWHEG method”, *JHEP* **11** (2007) 070, doi:10.1088/1126-6708/2007/11/070, `arXiv:0709.2092`.

[99] S. Alioli, P. Nason, C. Oleari, and E. Re, “A general framework for implementing NLO calculations in shower Monte Carlo programs: The POWHEG BOX”, *JHEP* **06** (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.

[100] S. Alioli, P. Nason, C. Oleari, and E. Re, “NLO single-top production matched with shower in POWHEG: s- and t-channel contributions”, *JHEP* **09** (2009) 111, doi:10.1088/1126-6708/2009/09/111, arXiv:0907.4076. [Erratum: *JHEP* **02** (2010) 011].

[101] E. Re, “Single-top Wt-channel production matched with parton showers using the POWHEG method”, *Eur. Phys. J. C* **71** (2011) 1547, doi:10.1140/epjc/s10052-011-1547-z, arXiv:1009.2450.

[102] S. Frixione and B. R. Webber, “Matching NLO QCD computations and parton shower simulations”, *JHEP* **06** (2002) 029, doi:10.1088/1126-6708/2002/06/029, arXiv:hep-ph/0204244.

[103] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”, *JHEP* **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[104] T. Sjöstrand, S. Mrenna, and P. Skands, “PYTHIA 6.4 physics and manual”, *JHEP* **05** (2006) 026, doi:10.1088/1126-6708/2006/05/026, arXiv:hep-ph/0603175.

[105] G. Corcella, I. G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, P. Richardson, M. H. Seymour, and B. R. Webber, “HERWIG 6: An event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)”, *JHEP* **01** (2001) 010, doi:10.1088/1126-6708/2001/01/010, arXiv:hep-ph/0011363.

[106] **GEANT4** Collaboration, “GEANT4 – a simulation toolkit”, *Nucl. Instrum. Meth. A* **506** (2003) 250, doi:10.1016/S0168-9002(03)01368-8.

[107] L. R. Evans and P. Bryant, “LHC Machine”, *JINST* **3** (2008) S08001. 164 p. http://cds.cern.ch/record/1129806. This report is an abridged version of the LHC Design Report (CERN-2004-003).

[108] F. Marcastel, “CERN’s Accelerator Complex. La chaîne des accélérateurs du CERN”,. https://cds.cern.ch/record/1621583. General Photo.

[109] **ATLAS** Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *JINST* **3** (2008) S08003, doi:10.1088/1748-0221/3/08/S08003.

[110] **CMS** Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[111] **ALICE** Collaboration, "The ALICE experiment at the CERN LHC", *JINST* **3** (2008) S08002, doi:10.1088/1748-0221/3/08/S08002.

[112] **LHCb** Collaboration, "The LHCb Detector at the LHC", *JINST* **3** (2008) S08005, doi:10.1088/1748-0221/3/08/S08005.

[113] **CMS** Collaboration. `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`.

[114] **CMS** Collaboration. `https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/RetrieveFile?docid=11514&version=1&filename=cms_120918_03.png`.

[115] **CMS** Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* **12** no. 10, (2017) P10003, doi:10.1088/1748-0221/12/10/P10003, `arXiv:1706.04965`.

[116] **CMS** Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014) P10009, doi:10.1088/1748-0221/9/10/P10009, `arXiv:1405.6569`.

[117] R. Frühwirth, W. Waltenberger, and P. Vanlaer, "Adaptive Vertex Fitting", Tech. Rep. CMS-NOTE-2007-008, CERN, Geneva, Mar, 2007. `https://cds.cern.ch/record/1027031`.

[118] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual", *Eur. Phys. J. C* **72** (2012) 1896, doi:10.1140/epjc/s10052-012-1896-2, `arXiv:1111.6097`.

[119] S. D. Ellis and D. E. Soper, "Successive combination jet algorithm for hadron collisions", *Phys. Rev.* **D48** (1993) 3160–3166, doi:10.1103/PhysRevD.48.3160, `arXiv:hep-ph/9305266`.

[120] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm", *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, `arXiv:0802.1189`.

[121] D. Krohn, J. Thaler, and L.-T. Wang, "Jets with Variable R", *JHEP* **06** (2009) 059, doi:10.1088/1126-6708/2009/06/059, `arXiv:0903.0392`.

[122] Fastjet contrib. `https://fastjet.hepforge.org/contrib/`.

[123] D. Bertolini, P. Harris, M. Low, and N. Tran, "Pileup Per Particle Identification", *JHEP* **10** (2014) 059, doi:10.1007/JHEP10(2014)059, arXiv:1407.6013.

[124] **CMS** Collaboration, "Identification of b-quark jets with the CMS experiment", *JINST* **8** (2013) P04013, doi:10.1088/1748-0221/8/04/P04013, arXiv:1211.4462.

[125] **CMS** Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** no. 05, (2018) P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[126] S. Schmitt, "TUnfold: An algorithm for correcting migration effects in high energy physics", *JINST* **7** (2012) T10003, doi:10.1088/1748-0221/7/10/T10003, arXiv:1205.6201.

[127] **CMS** Collaboration, "A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging", CMS Physics Analysis Summary CMS-PAS-JME-09-001, Jul, 2009. https://cds.cern.ch/record/1194489.

[128] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, "Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks", *Phys. Rev. Lett.* **101** (2008) 142001, doi:10.1103/PhysRevLett.101.142001, arXiv:0806.0848.

[129] T. Plehn, G. P. Salam, and M. Spannowsky, "Fat Jets for a Light Higgs", *Phys. Rev. Lett.* **104** (2010) 111801, doi:10.1103/PhysRevLett.104.111801, arXiv:0910.5472.

[130] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, "Stop Reconstruction with Tagged Tops", *JHEP* **10** (2010) 078, doi:10.1007/JHEP10(2010)078, arXiv:1006.2833.

[131] G. Kasieczka, T. Plehn, T. Schell, T. Strebler, and G. P. Salam, "Resonance Searches with an Updated Top Tagger", *JHEP* **06** (2015) 203, doi:10.1007/JHEP06(2015)203, arXiv:1503.05921.

[132] J. S. Conway, R. Bhaskar, R. D. Erbacher, and J. Pilot, "Identification of High-Momentum Top Quarks, Higgs Bosons, and W and Z Bosons Using Boosted Event Shapes", *Phys. Rev.* **D94** no. 9, (2016) 094027, doi:10.1103/PhysRevD.94.094027, arXiv:1606.06859.

[133] **CMS** Collaboration, "Search for dark matter in events with energetic, hadronically decaying top quarks and missing transverse momentum at $\sqrt{s} = 13$ TeV", *JHEP* **06** (2018) 027, doi:10.1007/JHEP06(2018)027, arXiv:1801.08427.

[134] **CMS** Collaboration, "Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment", CMS Physics Analysis Summary CMS-PAS-JME-18-002, 2019. `https://cds.cern.ch/record/2683870`.

[135] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, "Deep-learned Top Tagging with a Lorentz Layer", *SciPost Phys.* **5** no. 3, (2018) 028, doi:10.21468/SciPostPhys.5.3.028, `arXiv:1707.08966`.

[136] P. Artoisenet, R. Frederix, O. Mattelaer, and R. Rietkerk, "Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations", *JHEP* **03** (2013) 015, doi:10.1007/JHEP03(2013)015, `arXiv:1212.3460`.

[137] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani, "Matching matrix elements and shower evolution for top-quark production in hadronic collisions", *JHEP* **01** (2007) 013, doi:10.1088/1126-6708/2007/01/013, `arXiv:hep-ph/0611129`.

[138] P. M. Nadolsky, H.-L. Lai, Q.-H. Cao, J. Huston, J. Pumplin, D. Stump, W.-K. Tung, and C.-P. Yuan, "Implications of CTEQ global analysis for collider observables", *Phys. Rev. D* **78** (2008) 013004, doi:10.1103/PhysRevD.78.013004, `arXiv:0802.0007`.

[139] H.-L. Lai, M. Guzzi, J. Huston, Z. Li, P. M. Nadolsky, J. Pumplin, and C.-P. Yuan, "New parton distributions for collider physics", *Phys. Rev. D* **82** (2010) 074024, doi:10.1103/PhysRevD.82.074024, `arXiv:1007.2241`.

[140] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, P. M. Nadolsky, and W. K. Tung, "New generation of parton distributions with uncertainties from global QCD analysis", *JHEP* **07** (2002) 012, doi:10.1088/1126-6708/2002/07/012, `arXiv:hep-ph/0201195`.

[141] **CMS** Collaboration, "Study of the underlying event at forward rapidity in pp collisions at $\sqrt{s} = 0.9$, 2.76, and 7 TeV", *JHEP* **04** (2013) 072, doi:10.1007/JHEP04(2013)072, `arXiv:1302.2394`.

[142] **CMS** Collaboration, "Event generator tunes obtained from underlying event and multiparton scattering measurements", *Eur. Phys. J. C* **76** (2016) 155, doi:10.1140/epjc/s10052-016-3988-x, `arXiv:1512.00815`.

[143] **CMS** Collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV", *JINST* **7** (2012) P10002, doi:10.1088/1748-0221/7/10/P10002, `arXiv:1206.4071`.

[144] **CMS** Collaboration, “Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV”, *JINST* **10** (2015) P06005, doi:10.1088/1748-0221/10/06/P06005, `arXiv:1502.02701`.

[145] **CMS** Collaboration, “Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV”, *JINST* **12** no. 02, (2017) P02014, doi:10.1088/1748-0221/12/02/P02014, `arXiv:1607.03663`.

[146] **CMS** Collaboration, “Performance of the CMS missing transverse momentum reconstruction in pp data at $\sqrt{s} = 8$ TeV”, *JINST* **10** (2015) P02006, doi:10.1088/1748-0221/10/02/P02006, `arXiv:1411.0511`.

[147] A. H. Hoang, S. Plätzer, and D. Samitz, “On the Cutoff Dependence of the Quark Mass Parameter in Angular Ordered Parton Showers”, *JHEP* **10** (2018) 200, doi:10.1007/JHEP10(2018)200, `arXiv:1807.06617`.

[148] A. Hornig, Y. Makris, and T. Mehen, “Jet Shapes in Dijet Events at the LHC in SCET”, *JHEP* **04** (2016) 097, doi:10.1007/JHEP04(2016)097, `arXiv:1601.01319`.

[149] **CMS** Collaboration, “Search for resonant $t\bar{t}$ production in proton-proton collisions at $\sqrt{s} = 8$ TeV”, *Phys. Rev. D* **93** (2016) 012001, doi:10.1103/PhysRevD.93.012001, `arXiv:1506.03062`.

[150] **ATLAS** Collaboration, “Measurements of top-quark pair differential cross-sections in the lepton+jets channel in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector”, *Eur. Phys. J.* **C76** no. 10, (2016) 538, doi:10.1140/epjc/s10052-016-4366-4, `arXiv:1511.04716`.

[151] **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. C* **75** (2015) 542, doi:10.1140/epjc/s10052-015-3709-x, `arXiv:1505.04480`.

[152] **CMS** Collaboration, “Measurement of the $\mathrm{t\bar{t}}$ production cross section in the all-jets final state in pp collisions at $\sqrt{s} = 8$ TeV”, *Eur. Phys. J. C* **76** (2016) 128, doi:10.1140/epjc/s10052-016-3956-5, `arXiv:1509.06076`.

[153] **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”, *Phys. Rev. D* **93** (2016) 032009, doi:10.1103/PhysRevD.93.032009, `arXiv:1510.03818`.

[154] **CMS** Collaboration, “Measurement of the integrated and differential $\mathrm{t\bar{t}}$ production cross sections for high-p_{T} top quarks in pp collisions at $\sqrt{s} = 8$ TeV”, *Phys. Rev. D* **94** (2016) 072002, doi:10.1103/PhysRevD.94.072002, `arXiv:1605.00116`.

[155] G. Antchev *et al.*, "First measurement of the total proton-proton cross section at the LHC energy of $\sqrt{s} = 7$ TeV", *Europhys. Lett.* **96** (2011) 21002, doi:10.1209/0295-5075/96/21002, `arXiv:1110.1395`.

[156] **CMS** Collaboration, "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 7$ TeV", *Phys. Lett.* **B722** (2013) 5–27, doi:10.1016/j.physletb.2013.03.024, `arXiv:1210.6718`.

[157] **CMS** Collaboration, "Performance of b tagging at $\sqrt{s} = 8$ TeV in multijet, $\mathrm{t\bar{t}}$ and boosted topology events", CMS Physics Analysis Summary CMS-PAS-BTV-13-001, 2013. `http://cdsweb.cern.ch/record/1581306`.

[158] **CMS** Collaboration, "Measurement of the production cross section of a W boson in association with two b jets in pp collisions at $\sqrt{s} = 8\,\mathrm{TeV}$", *Eur. Phys. J.* **C77** no. 2, (2017) 92, doi:10.1140/epjc/s10052-016-4573-z, `arXiv:1608.07561`.

[159] **CMS** Collaboration, "Observation of the associated production of a single top quark and a W boson in pp collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. Lett.* **112** (2014) 231802, doi:10.1103/PhysRevLett.112.231802, `arXiv:1401.2942`.

[160] **CMS** Collaboration, "CMS luminosity based on pixel cluster counting - summer 2013 update", CMS Physics Analysis Summary CMS-PAS-LUM-13-001, 2013. `http://cds.cern.ch/record/1598864`.

[161] M. Beneke, P. Falgari, S. Klein, and C. Schwinn, "Hadronic top-quark pair production with NNLL threshold resummation", *Nucl. Phys. B* **855** (2012) 695, doi:10.1016/j.nuclphysb.2011.10.021, `arXiv:1109.1536`.

[162] M. Cacciari, M. Czakon, M. Mangano, A. Mitov, and P. Nason, "Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation", *Phys. Lett. B* **710** (2012) 612, doi:10.1016/j.physletb.2012.03.013, `arXiv:1111.5869`.

[163] P. Bärnreuther, M. Czakon, and A. Mitov, "Percent-level-precision physics at the Tevatron: Next-to-leading order QCD corrections to $\mathrm{q\bar{q}} \to \mathrm{t\bar{t}} + X$", *Phys. Rev. Lett.* **109** (2012) 132001, doi:10.1103/PhysRevLett.109.132001, `arXiv:1204.5201`.

[164] M. Czakon and A. Mitov, "NNLO corrections to top-pair production at hadron colliders: The all-fermionic scattering channels", *JHEP* **12** (2012) 054, doi:10.1007/JHEP12(2012)054, `arXiv:1207.0236`.

[165] M. Czakon and A. Mitov, "NNLO corrections to top pair production at hadron colliders: The quark-gluon reaction", *JHEP* **01** (2013) 080, doi:10.1007/JHEP01(2013)080, `arXiv:1210.6832`.

[166] M. Czakon, P. Fiedler, and A. Mitov, "Total top-quark pair-production cross section at hadron colliders through $O(\alpha_S^4)$", *Phys. Rev. Lett.* **110** (2013) 252004, doi:10.1103/PhysRevLett.110.252004, arXiv:1303.6254.

[167] M. Czakon, D. Heymes, and A. Mitov, "High-precision differential predictions for top-quark pairs at the LHC", *Phys. Rev. Lett.* **116** (2016) 082003, doi:10.1103/PhysRevLett.116.082003, arXiv:1511.00549.

[168] CDF and D0 Collaborations, "Combination of the top-quark mass measurements from the Tevatron collider", *Phys. Rev. D* **86** (2012) 092003, doi:10.1103/PhysRevD.86.092003, arXiv:1207.1069.

[169] **ATLAS** Collaboration, "Measurement of the top quark mass in the $t\bar{t} \rightarrow$ lepton+jets and $t\bar{t} \rightarrow$ dilepton channels using $\sqrt{s} = 7$ TeV ATLAS data", *Eur. Phys. J. C* **75** (2015) 330, doi:10.1140/epjc/s10052-015-3544-0, arXiv:1503.05427.

[170] **ATLAS** Collaboration, "Determination of the top-quark pole mass using $\mathrm{t\bar{t}}$ + 1-jet events collected with the ATLAS experiment in 7 TeV pp collisions", *JHEP* **10** (2015) 121, doi:10.1007/JHEP10(2015)121, arXiv:1507.01769.

[171] **CMS** Collaboration, "Measurement of the top quark mass using charged particles in pp collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. D* **93** (2016) 092006, doi:10.1103/PhysRevD.93.092006, arXiv:1603.06536.

[172] **ATLAS** Collaboration, "Measurement of the top quark mass in the $t\bar{t} \rightarrow$ dilepton channel from $\sqrt{s} = 8$ TeV ATLAS data", *Phys. Lett. B* **761** (2016) 350, doi:10.1016/j.physletb.2016.08.042, arXiv:1606.02179.

[173] I. W. Stewart, F. J. Tackmann, J. Thaler, C. K. Vermilion, and T. F. Wilkason, "XCone: N-jettiness as an Exclusive Cone Jet Algorithm", *JHEP* **11** (2015) 072, doi:10.1007/JHEP11(2015)072, arXiv:1508.01516.

[174] **CMS** Collaboration, "Jet algorithms performance in 13 TeV data", CMS Physics Analysis Summary CMS-PAS-JME-16-003, 2017. https://cds.cern.ch/record/2256875.

[175] R. Frederix and S. Frixione, "Merging meets matching in MC@NLO", *JHEP* **12** (2012) 061, doi:10.1007/JHEP12(2012)061, arXiv:1209.6215.

[176] J. Alwall *et al.*, "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions", *Eur. Phys. J.* **C53** (2008) 473–500, doi:10.1140/epjc/s10052-007-0490-5, arXiv:0706.2569.

[177] **CMS** Collaboration, "Investigations of the impact of the parton shower tuning in Pythia 8 in the modelling of $t\bar{t}$ at $\sqrt{s} = 8$ and 13 TeV", CMS Physics Analysis Summary CMS-PAS-TOP-16-021, 2016. `https://cds.cern.ch/record/2235192`.

[178] P. Skands, S. Carrazza, and J. Rojo, "Tuning PYTHIA 8.1: the Monash 2013 Tune", *Eur. Phys. J.* **C74** no. 8, (2014) 3024, doi:10.1140/epjc/s10052-014-3024-y, `arXiv:1404.5630`.

[179] M. Bahr *et al.*, "Herwig++ Physics and Manual", *Eur. Phys. J.* **C58** (2008) 639–707, doi:10.1140/epjc/s10052-008-0798-9, `arXiv:0803.0883`.

[180] M. H. Seymour and A. Siodmok, "Constraining MPI models using σ_{eff} and recent Tevatron and LHC Underlying Event data", *JHEP* **10** (2013) 113, doi:10.1007/JHEP10(2013)113, `arXiv:1307.5015`.

[181] **CMS** Collaboration, "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements", CMS Physics Analysis Summary CMS-PAS-GEN-17-001, 2018. `http://cds.cern.ch/record/2636284`.

[182] **NNPDF** Collaboration, "Parton distributions from high-precision collider data", *Eur. Phys. J.* **C77** no. 10, (2017) 663, doi:10.1140/epjc/s10052-017-5199-5, `arXiv:1706.00428`.

[183] **CMS** Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **13** no. 06, (2018) P06015, doi:10.1088/1748-0221/13/06/P06015, `arXiv:1804.04528`.

[184] **CMS** Collaboration, "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector", *JINST* **14** no. 07, (2019) P07004, doi:10.1088/1748-0221/14/07/P07004, `arXiv:1903.06078`.

[185] **CMS** Collaboration, "Measurement of jet substructure observables in $t\bar{t}$ events from proton-proton collisions at $\sqrt{s} = 13$TeV", *Phys. Rev.* **D98** no. 9, (2018) 092014, doi:10.1103/PhysRevD.98.092014, `arXiv:1808.07340`.

[186] J. Ott, "THETA–A framework for template-based modeling and inference", 2010. `http://www-ekp.physik.uni-karlsruhe.de/~ott/theta/theta-auto/`.

[187] M. Stender, "Kalibrierung der Jetmasse in CMS", Bachelor's thesis, Universität Hamburg, 2015.

Danksagung

Hiermit möchte ich mich bei allen Menschen bedanken, die mich während meiner Arbeit unterstützt haben.

Zu aller erst möchte ich mich bei Johannes Haller bedanken, der mir diese Arbeit ermöglicht hat. Ich bedanke mich für die Betreuung der Arbeit, die konstruktiven Rücksprachen während der wöchentlichen Gruppentreffen und für die Unterstützung bei der Veröffentlichung der ersten Analyse. Darüber hinaus bedanke ich mich dafür, dass ich die Möglichkeit bekommen habe an internationalen Sommerschulen und Konferenzen teilnehmen zu können.

Elisabetta Gallo danke ich dafür, dass sie sich bereiterklärt hat, die Arbeit als Zweitgutachterin zu lesen.

Ich danke den Mitgliedern der Prüfungskommission, Johannes Haller, Elisabetta Gallo, Günter Sigl, Roman Kogler und Christian Schwanenberger für die Beurteilung der Disputation.

Besonderer Dank gebührt Roman Kogler für die intensive Betreuung der Arbeit, für viele konstruktive Diskussionen zu den Analysen und für die große Hilfe bei der Veröffentlichung der ersten Analyse.

Vielen Dank an Daniel Gonzalez, mit dem ich die gesamte Zeit am Institut ein Büro geteilt habe, für viele interessante Diskussionen während der Kaffeepausen.

Der gesamten Gruppe danke ich für das gute Arbeitsklima und für die gute Zusammenarbeit.

Zu guter Letzt möchte ich mich bei meiner Familie bedanken und ganz besonders bei meiner Mutter Ina Dreyer für die moralische Unterstützung während der letzten Jahre.

Eidesstattliche Versicherung / Declaration on oath

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben.

Die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium.

Die Dissertation wurde in der vorgelegten oder einer ähnlichen Form nicht schon einmal ineinem früheren Promotionsverfahren angenommen oder als ungenügend beurteilt.

Hamburg, den

Unterschrift der Doktorandin / des Doktoranden