Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MEASUREMENT OF JET MASS DISTRIBUTIONS OF BOOSTED W BOSONS

DISSERTATION

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

vorgelegt von

Steffen Albrecht

HAMBURG, 2023

Gutachter/innen der Dissertation:	Dr. Andreas Hinzmann
	Prof. Dr. Johannes Haller
Zusammensetzung der Prüfungskommission:	Prof. Dr. Dieter Horns
	Prof. Dr. Gudrid Moortgat-Pick
	Prof. Dr. Peter Schleper
	Prof. Dr. Johannes Haller
	Dr. Andreas Hinzmann
Vorsitzende/r der Prüfungskommission:	Prof. Dr. Dieter Horns
Datum der Disputation:	22.11.2023
Vorsitzende/r des Fach-Promotionsausschusses Physik:	Prof. Dr. Markus Drescher
Leiter/in des Fachbereichs Physik:	Prof. Dr. Wolfgang J. Parak
Dekan der Fakultät MIN:	Prof. Dr.-Ing. Norbert Ritter

Abstract

In this thesis, the groomed jet mass of hadronically decaying W bosons and top quarks is analyzed in data of proton-proton collisions from the LHC at $\sqrt{s} = 13\,\text{TeV}$. The data were collected by the CMS experiment in the years 2016 to 2018 and correspond to an integrated luminosity of $138\,\text{fb}^{-1}$. Two analyses are presented using events in which the W boson or top quark has a large transverse momentum and thus produces strongly collimated decay products reconstructed as single large-radius jets. Such boosted jets are produced at the LHC at a high rate. They are subject of many Standard Model measurements and beyond the Standard Model searches alike and therefore a precise measurement and understanding of such jets is of wide use. Boosted jets can be identified by studying their substructure, in particular, the jet mass is sensitive as it is a measure of the mass of the W boson or top quark. Previous measurements of jet mass have been carried out for gluon, quark and top jets in dijet, $Z(\ell\ell)$+jet and $t\bar{t}$ samples, while the measurement of W jet mass has not been performed yet and is for the first time pursued in this thesis. The substructure of these jets serves as a valuable tool in various processes such as jet calibration and jet tagging/identification. However, there is still room for enhancement in understanding this substructure due to imperfect modeling of effects arising from perturbative QCD (parton showers) and non-perturbative QCD (hadronization of partons) in simulations. One aspect that is particularly difficult to predict is that large-radius jets encompass soft and wide-angle radiation, which obscures jet substructure variables. To mitigate this issue, jet grooming algorithms are applied here.

In the first analysis, precise correction factors to the simulation for the groomed jet mass scale of W jets and top jets are measured in bins of the transverse momentum from 500 to 1200 GeV of the ungroomed jet in samples of semileptonic $t\bar{t}$ events and fully-hadronic $W(q\bar{q})$ events. The correction factors are measured in a simultaneous fit to data in signal and control regions. The main challenge here is the dominant QCD multijet background in the $W(q\bar{q})$, which is estimated using a sophisticated method from control region data. The signal and control regions are constructed using two different boosted jet tagging approaches to compare their effect on the jet mass scale. Further, the correction factors are studied to estimate the correlation between the jet mass scale and the jet energy scale in the CMS experiment. The final correction factor measurement reaches a high precision of $\sim 1\%$ and shows a residual difference of the jet mass scale between data and simulation of under only 2% in a largely extended range of transverse momentum up to 1200 GeV using a calibration sample with $W(q\bar{q})$ for the first time.

The second analysis is the first measurement of the groomed jet mass distribution of W jets in bins of the transverse momentum on particle-level in data. For the measurement, a two-dimensional maximum likelihood unfolding is performed in fully-hadronic $W(q\bar{q})$ events. The unfolded data and the compared prediction from simulation at LO+MLM accuracy with NLO (QCD and EWK) corrections are in agreement within the uncertainties. The uncertainty on the unfolded data distribution in the W mass peak region is between 30–80% in the range of transverse momentum from 650–1200 GeV. With the planned HL-LHC, higher precision can be reached, however, requiring significantly reduced systematic uncertainties in jet substructure modeling. This study is an important first step towards a measurement of the W mass in the all-jets final state.

Zusammenfassung

In dieser Arbeit wird die Jetmasse unter Anwendung von Grooming-Algorithmen ("groomed" Jetmasse) von hadronisch zerfallenden W-Bosonen und Top-Quarks in Daten von Proton-Proton-Kollisionen am LHC bei $\sqrt{s} = 13\,\text{TeV}$ analysiert. Die Daten wurden vom CMS-Experiment in den Jahren 2016 bis 2018 gesammelt und entsprechen einer integrierten Luminosität von $138\,\text{fb}^{-1}$. Zwei Analysen werden vorgestellt, die Ereignisse verwenden, in denen das W-Boson oder das Top-Quark einen großen Transversalimpuls hat und somit stark kollimierte Zerfallsprodukte erzeugt, und somit als einzelne Jets mit großem Radius rekonstruiert werden. Solche "boosted" Jets werden am LHC in hoher Rate produziert. Sie sind Gegenstand vieler Messungen des Standardmodells und auch von Untersuchungen jenseits des Standardmodells, weshalb eine präzise Messung und ein Verständnis solcher Jets von weitem Nutzen ist. Boosted Jets können anhand ihrer Substruktur identifiziert werden. Insbesondere die Jetmasse ist sensitiv, da sie ein Maß für die Masse des W-Bosons oder des Top-Quarks ist. Frühere Messungen der Jetmasse wurden für Gluon-, Quark- und Top-Jets in Dijet-, $Z(\ell\ell)$+Jet- und $t\bar{t}$-Prozessen durchgeführt, während die Messung der W-Jetmasse noch aussteht und erstmals in dieser Arbeit verfolgt wird. Die Substruktur dieser Jets erweist sich als wertvolles Werkzeug in verschiedenen Prozessen wie Jetkalibrierung und Jet-Identifikation ("tagging"). Dennoch gibt es immer noch Raum für Verbesserungen im Verständnis dieser Substruktur aufgrund der unvollkommenen Modellierung der Effekte, die aus der perturbativen QCD (Parton Showers) und der nicht-perturbativen QCD (Hadronisierung von Partonen) in Simulationen resultieren. Im Kontext dieser Arbeit beinhalten die Jets mit großem Radius nieder-energetische und Abstrahlung in großem Winkel, welche die Variablen der Jet-Substruktur verändern. Um dieses Problem anzugehen, werden hier Jet-Grooming-Algorithmen angewendet.

In der ersten Analyse werden präzise Korrekturfaktoren für die Simulation der groomed Jetmasse von W-Jets und Top-Jets als Funktion des transversalen Impulses von 500 bis 1200 GeV des groomed Jets in Proben von semileptonischen $t\bar{t}$-Ereignissen und vollhadronischen $W(q\bar{q})$-Ereignissen gemessen. Die Korrekturfaktoren werden in einem simultanen Fit an Daten in Signal- und Kontrollregionen gemessen. Die Hauptherausforderung besteht darin, den dominierenden QCD-Multijet-Untergrund im $W(q\bar{q})$ abzuschätzen, was mit einer dedizierten Methode aus Daten der Kontrollregion erreicht wird. Die Signal- und Kontrollregionen werden unter Verwendung von zwei verschiedenen Tagging-Ansätzen konstruiert, um deren Auswirkungen auf die Jetmassenskala zu vergleichen. Darüber hinaus werden die Korrekturfaktoren untersucht, um die Korrelation zwischen der Jetmassenskala und der Jetenergieskala abzuschätzen. Die finale Messung der Korrekturfaktoren erreicht eine hohe Präzision von etwa ~ 1% und zeigt eine Restdifferenz der Jetmassenskala zwischen Daten und Simulation von nur 2% in einem weiten Bereich des transversalen Impulses bis zu 1200 GeV unter erstmaliger Verwendung des $W(q\bar{q})$-Prozesses für die Kalibration.

Die zweite Analyse stellt die erste Messung der Verteilung der groomed Jetmasse von W-Jets in Bins des transversalen Impulses auf Teilchenebene in Daten dar. Für diese Messung wird eine zweidimensionale Maximum-Likelihood-Entfaltungstechnik in vollhadronischen $W(q\bar{q})$-Ereignissen angewendet. Die entfalteten Daten und die verglichene Vorhersage aus Simulation mit LO+MLM-Genauigkeit mit NLO-Korrekturen (QCD und EWK) stimmen innerhalb der Unsicherheiten überein. Die Unsicherheit der entfalteten Datenverteilung im W-Massen-Peak-Bereich liegt zwischen 30–80% im Bereich des transversalen Impulses von 650–1200 GeV. Mit dem geplanten HL-LHC kann eine höhere Präzision erreicht werden, was jedoch eine deutliche Verringerung der systematischen Unsicherheiten bei der Modellierung der Jet-Substruktur erfordert. Diese Studie ist ein wichtiger erster Schritt auf dem

Weg zur Messung der W-Masse im Endzustand des All-Jets.

List of own contributions

Validation of the Legacy Re-Reconstruction

I performed detailed validation studies for the legacy re-reconstruction of the Run 2 data recorded by the CMS experiment, which was analyzed in the context of this thesis (see Section 6.1). For this, I performed detailed validation studies to test the impact of the re-reconstruction on jet substructure during the first year of my doctoral studies. This included the analysis and comparison of key jet kinematic and jet substructure observables in multiple data-taking periods and reconstruction scenarios. I presented regular updates in internal working group meetings and collaboration-wide general meetings. As a result of this successful validation, CMS re-reconstructed the data and simulation of the full Run 2, used by all upcoming publications of the CMS collaboration. This work was supervised by Dr. Andreas Hinzmann and Dr. Laurent Thomas.

Calibration of the jet mass scale

I was the principal analyzer of the effort to measure calibration factors for the jet mass scale using boosted W bosons and top quarks, which is detailed in Section 7. I was the primary author of a CMS Detector Performance Summary [1], which summarizes the final results of this analysis. The initial technical setup was developed in collaboration with Dr. Dennis Schwarz. The technique for data-driven QCD multijet background estimation was based on previous work [2], and was further developed with input from the original developers Dr. Nick Smith, Dr. David Yu, Prof. Dr. Phil Harris, Dr. Ka Hei Martin Kwok and Prof. Dr. Javier Duarte. The higher-order NLO QCD and EWK corrections are based on [3] and [4], and their technical implementation was provided by Dr. Nick Smith. The work I performed on the analysis included the development and study of jet merging and tagging categories (Sections 6.1.1 and 6.1.2), further development and optimization of the QCD multijet background estimate for multiple jet tagging approaches (Section 6.4), further development and optimization of the technical framework to measure the correction factors with maximum-likelihood fits (Sections 7.1 and 7.2) and the extensive study of the final correction factors (Section 7.4 – 7.6). I presented updates in working group and general meetings. This work was supervised by Dr. Andreas Hinzmann.

Measurement of the jet mass distribution of boosted W bosons

I was the principal analyzer of the effort to perform the measurement of the jet mass distribution of boosted W bosons, which included the work detailed in Section 8. The technical setup is based on the efforts for the calibration of the jet mass scale and was extended to perform the two-dimensional unfolding. The further work I performed entailed: The study of different phase space definitions for the unfolding which led to the definition in Section 8.1, the optimization of the binning scheme described in Section 8.2, the implementation, testing and validation of regularization schemes for the unfolding described in Section 8.3.1. Finally, I performed the bias and coverage tests and performed the final Unfolding as described in Section 8.3.2 and 8.4. I presented updates on this work in working group and general meetings. This work was supervised by Dr. Andreas Hinzmann

Contents

1 Introduction

The Standard Model (SM) of particle physics describes nature at the microscopic scale, including the characterization of the known elementary particles and three out of the four fundamental interactions. It has been tested to very high precision in numerous experimental setups, covering a wide range of energy scales. Despite its great success, the SM is not a complete description of all aspects of elementary particle physics. For example, in its current form, it does not account for dark matter, gravity or massive neutrinos. As an example for isolated experimental deviations from the SM, the recent result on the improved, precise measurement of the mass of the W boson published by the CDF collaboration [5] puts additional stress on the SM, as it is ~ 7 standard deviations higher than the previous global average of m_W. For these reasons, numerous efforts exist to probe standard model processes occurring in the high-energy proton-proton collisions at the Large Hadron Collider (LHC) for further deviations from the SM prediction and search for processes arising from new physics.

This thesis focuses on the measurement of SM properties. Heavy SM particles, such as the W boson and the top quark, are produced at high rates in pp collisions at the LHC, spanning a wide range of energy scales. The high center of mass energy allows for the study of the case where these particles reach very high transverse momenta exceeding their mass by a large factor. Thus they are subject to a high Lorentz-boost, and consequently, their decay products will be highly collimated. The decay products form parton showers in the detector that are reconstructed as a single large-radius jet. These merged jets contain often the complete decay of the initiating particles like W bosons or top quarks. They are the subject of many direct searches for new physics and measurements of SM properties [6, 7]. The substructure is a very helpful testbed for predictions of perturbative and non-perturbative QCD [8, 9] and a valuable instrument for the identification of jets. The invariant mass of jets is crucial in all of this, as it is highly sensitive to the mass of the jet-initiating particle, thus offering a way to identify particles and even measure their mass. Further, jets acquire mass from splittings of quarks and gluons in the parton shower. To counter effects that result in extra soft and wide-angle radiation obscuring the jet kinematics, jet grooming is applied, which removes soft and wide-angle radiation under a given threshold from the jet.

In the context of this thesis, the groomed mass of large-radius jets originating from W bosons or top quarks with large Lorentz-boost is studied in two analyses of data from pp collision at $\sqrt{s} = 13\,\text{TeV}$ recorded by the CMS experiment [10] at the LHC in the years 2016–2018 (Run 2), corresponding to an integrated luminosity of $\mathscr{L}_{\text{int}} \approx 138\,\text{fb}^{-1}$.

The uncertainties of boosted jet tagging methods to identify heavy objects are often the limiting systematic uncertainty in searches or measurements involving boosted heavy objects. To improve this, precise calibrations for the soft drop mass were previously derived for boosted W and

top jets in a semileptonic $t\bar{t}$ sample. With the increased center of mass energy in Run 2 and Run 3 of the LHC, higher Lorentz-boost is achievable and the previous method can be extended with a new $W(q\bar{q})$+jets sample with an increased range of transverse momentum. In the first analysis, simulation to data correction factors for the groomed jet mass scale of W jets and top jets are measured in semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets events. The correction factors are derived in bins of jet transverse momentum in simultaneous fits to the data in four periods of data-taking separately. These variations are propagated to shifts in the groomed jet mass. While the jet energy scale is measured precisely, the understanding of the correlation between jet mass scale and jet energy scale has to be improved to be able to perform precise measurements of the jet mass. The measurement of the correction factors is used to estimate the correlation between the jet mass scale and the jet energy scale and consequently studies the applicability of the dedicated jet energy scale corrections on the jet mass scale. Further, the correction factors are measured using two different jet tagging approaches to construct the pure samples of W and top jets, to estimate the sensitivity of the jet mass to these different approaches. One approach uses substructure variables motivated by QCD theory and the other one uses state-of-the-art deep learning algorithms to distinguish jets from heavy SM particles against jets from light quarks and gluons.
Measurements of jet mass distribution have been performed previously for light quark and gluon jets, top jets and Z jets [11–15]. The measurements of the quark jets offer a testbed for perturbative and non-perturbative QCD predictions and a proxy for the measurement of the top quark but are difficult to study due to color reconnection to the rest of the event, while the measurement of the Z jets offers important insights on the systematic uncertainties for searches involving b quark pairs in the final state. To study a simpler case, in the second analysis in this thesis, the groomed jet mass distribution of fully-hadronic $W(q\bar{q})$+jets events is measured in bins of ungroomed jet transverse momentum on particle-level in the full Run 2 dataset for the first time. For this measurement, a two-dimensional unfolding of the data is performed using the maximum-likelihood approach. The jet mass distribution on particle-level of the unfolded data can be compared to predictions from simulations and semi-analytical calculations at different levels of accuracy including perturbative and non-perturbative effects, and is a first step towards measuring the mass of the W boson in the all-jets final state. In both analyses, the sample of fully-hadronic $W(q\bar{q})$+jets events is subject to a large background coming from QCD multijet processes. This introduces one of the main challenges in both analyses and is addressed with a data-driven approach to estimate this background.

The thesis is structured in the following way. Section 2 presents an overview of the current state of the standard model. In Section 3 the experimental setup of the LHC and the CMS experiment with its individual sub-detectors is described. Section 4 outlines the theoretical aspects of simulating pp collision events and the reconstruction of high-level physical objects from detector signals in the recorded and simulated collision events using the particle-flow algorithm is described. In the following Section 5 the reconstruction and calibration of jets, as well as pileup mitigation techniques are described before an overview of jet substructure and recent measurements and searches using jet substructure is given. Section 6 introduces the

aspects of the analysis strategy common for both analyses, which includes the event selection and categorization, the summary of recorded and simulated event samples, a description of the background estimation and considered sources of systematic uncertainty. In Section 7 the measurement of the correction factors for the mass scale of groomed jets is presented and discussed. The measurement of the groomed jet mass distribution of boosted W bosons is presented and discussed in Section 8. Finally, Section 9 summarizes the result of the two analyses and gives an outlook on possible improvements that can be made.

2 Standard Model of Elementary Particle Physics

The Standard Model (SM) of elementary particle physics is the phenomenological and mathematical description of the particle content of the visible (anti-) matter and three out of the four fundamental interactions. Although it has been tested experimentally to high precision, it is not without its problems (e.g. gravity not being included). This Section covers the basic concepts of the SM and is based on [16–19].
From here on natural units will be used, where $\hbar = c = 1$, unless stated otherwise.

2.1 Symmetries and Particles

In the SM, fundamental particles and fundamental interactions are described as components of a quantum field theory (QFT).
A fundamental particle is considered to be point-like (i.e. has no spatial extent) and non-composite (i.e. cannot be broken down further into smaller particles). Figure 2.1 summarizes the 17 fundamental particles we have discovered so far. The fundamental interactions described by the SM are the *strong interaction*, the *weak interaction* and the *electromagnetic interaction*. With QFT we move classical field theory into the realm of special relativity and quantum mechanics and ultimately derive the mathematical formalism that describes particles as fermionic fields, which interact with each other via bosonic fields. With the requirement of invariance under local transformations of a gauge symmetry group in a QFT, the corresponding Lagrangian must include additional terms corresponding to interactions and the QFT becomes a quantum gauge field theory. In the case of the SM, we have the following gauge symmetry group:

$$SU(3)_C \otimes SU(2)_L \otimes U(1)_Y,$$

where $SU(3)_C$ represents rotations of the color degrees of freedom (*red*, *green*, *blue*) of the theory of Quantum Chromodynamics (QCD), while $SU(2)_L$ are rotations in space of the weak isospin I and $U(1)_Y$ rotations in the hypercharge Y charge space of the theory of electroweak (EW) unification. While the strong interaction is described by QCD, the electromagnetic interaction can be either described by pure Quantum electrodynamic (QED) or in combination with the weak interaction as the electroweak interaction in the framework of EW unification. A particle is classified by quantum numbers, of which the most important are: mass m, spin s, color charge C, (the third component of the) weak isospin I_W^3 and electric charge Q. The latter three are the ones, that determine the type of interaction a particle will undergo, so whether a particle is subject to the strong, weak or electromagnetic interaction respectively. Each particle has an anti-particle, which has the same mass but opposite sign charges.

The 17 known or so far discovered fundamental particles come in two types, fermions and bosons. In simple terms, fermions are particles with half-integer spin, while bosons have an integer spin. While the bosons (or gauge bosons in this case) are the mediator particles of their respective interactions (e.g. the $W^{\pm}$ and Z bosons are the mediators of the weak interaction), the fermions are what we call matter. The fermions are divided further into six quarks and six leptons. The six quarks, namely up- (u), down- (d), charm- (c), strange- (s), bottom- (b) and top-quark (t), are subject to all three interactions as they carry each of the three charges. While all quarks carry color charge in the same way (either red, green or blue), the up-type quarks (u, c, t) carry $Q = \frac{2}{3}e$ and the down-type quarks (d, s, b) carry $Q = -\frac{1}{3}e$, where e denotes the elementary charge $e = 1.602\,176\,634 \times 10^{-19}\ C$ [20]. The six leptons on the other hand are subject to electroweak interactions only, as they do not carry color charge. There are the electrically charged ($Q = 1e$) leptons electron (e^-), muon (μ^-) and tau (τ^-) and their neutrino ($Q = 0$) counterparts (ν_e,ν_μ,ν_τ).

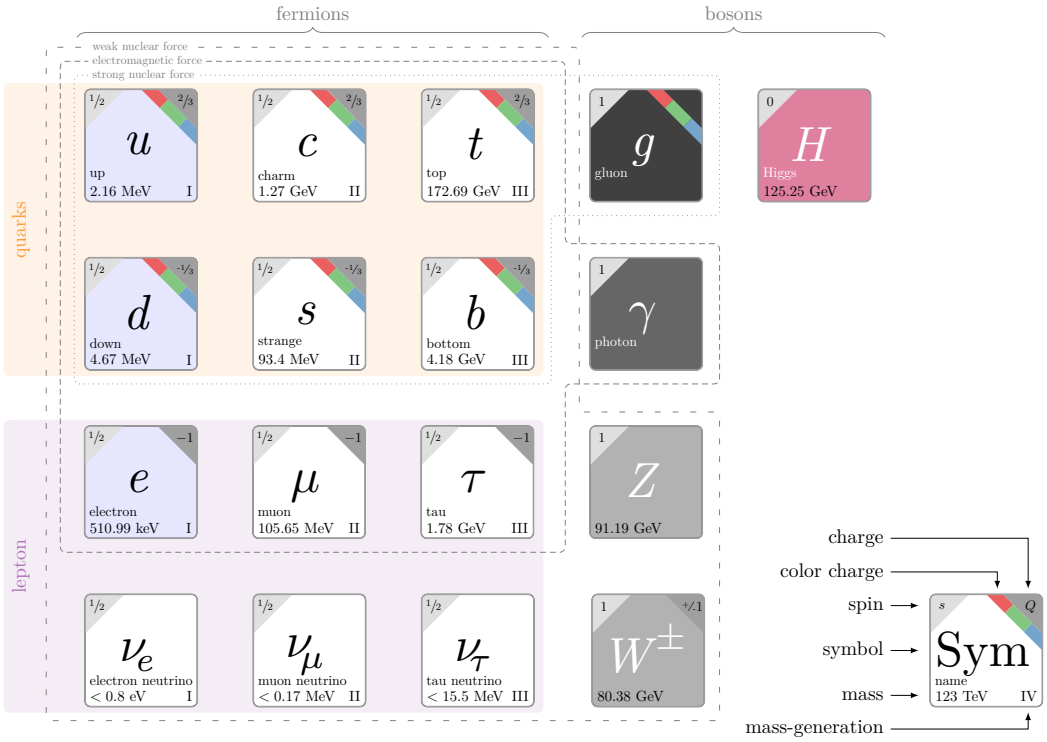


Figure 2.1: Overview of standard model particles and their properties and categorization. Adapted from [21], particle masses taken from [22], with most recent upper limit on electron neutrino mass from [23].

Only the left-handed fermions and right-handed anti-fermions have a non-zero weak isospin I_W^3 and consequently participate in the weak interaction (see 2.1.1). Reflecting this left-handed fermions appear in doublets of I_W^3:

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

where the upper row carries a weak isospin of $I_W^3 = +\frac{1}{2}$ and the lower row $I_W^3 = -\frac{1}{2}$. The right-handed counterparts appear in singlets as:

$$e^-_R, \quad \nu_{eR}, \quad \mu_R, \quad \nu_{\mu R}, \quad \tau^-_R, \quad \nu_{\tau R}, \quad u_R, \quad c_R, \quad c_R, \quad d'_R, \quad s'_R, \quad b'_R, \tag{2.2}$$

and do not carry weak isospin ($I^3_W = 0$), thus do not participate in weak interactions. It is worth noting though that right-handed neutrinos have not been observed.

Both quarks and leptons are arranged in 3 mass generations with increasingly higher particle masses. The electrically charged first-generation fermions (shaded blue in Figure 2.1) make up the stable matter in our universe, with the up- and down-quark being the primary constituents of the proton and neutron, which in turn are - together with the electron - the constituents of atoms. The second and third-generation fermions are not stable (i.e. have a short lifetime) but can be produced in high-energy collisions and thus be found in e.g. cosmic rays or collider experiments. The neutrinos are the only fermions that are subject to the weak interaction only and are thus very hard to detect. They were initially believed to be massless, but with neutrino oscillation [24, 25], there is an experimentally confirmed addition to the SM, that dictates them to be massive. Recent experiments have put an upper limit on the electron neutrino mass of $m_{\nu_e} < 0.08$ eV [23]. In the following the three interactions and their respective mediator gauge bosons are discussed in more detail.

2.1.1 *Electroweak unification*

The initial development of SM as a QFT was driven by QED, which offers a full description of the electromagnetic interaction. As mentioned above, one postulates invariance of the Lagrangian under local gauge transformations of $U(1)$, such that an additional gauge field A_μ (the photon) and its interaction with fermions carrying the charge Q arises. Similarly for electroweak unification local gauge invariance under $SU(2)_L \otimes U(1)_Y$ transformations introduces the three gauge fields W^1_μ, W^2_μ, and W^3_μ from $SU(2)_L$ and the B_μ from $U(1)_Y$. Here the symmetry group $U(1)_Y$ is not to be confused with the $U(1)$ symmetry of QED, but rather a subgroup with the charge Q replaced with the hypercharge Y, that relates the electric charge Q to the weak isospin I^3_W as $Y = 2(Q - I^3_W)$. The physical gauge bosons of the electroweak sector we observe are the $W^\pm$ bosons, the Z boson and the photon γ, which are the linear combinations of the W^1_μ, W^2_μ, W^3_μ and B_μ. The $W^\pm$ boson reads as:

$$W^\pm_\mu = \frac{1}{\sqrt{2}}(W^1_\mu \mp W^2_\mu). \tag{2.3}$$

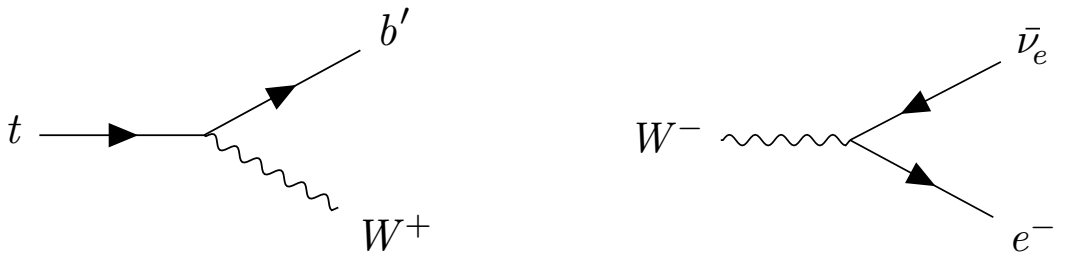
While the Z boson and the photon γ (here as A_μ) follow from:

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

with the weak mixing angle θ_W, which also relates the masses of $W^\pm$ and Z bosons via:

$$\cos\theta_W = \frac{m_W}{m_Z}. \tag{2.5}$$

The masses of $W^\pm$ and Z are not part of the formalism EW unification, but rather a consequence



(a) (b)

Figure 2.2: Feynman diagrams of the decay of a t-quark into a W^+-boson and a b'-quark and the decay of a W^--boson into a $\bar{\nu}_e$-neutrino and a e^-. Here the b' denotes the flavor eigenstate of the b-quark, to distinguish it from the mass eigenstate b.

of the addition of the mechanism of spontaneous symmetry breaking described in 2.1.2, and are measured to be $m_W = 80.377 \pm 0.012$ GeV and $m_Z = 91.1876 \pm 0.0021$ GeV [22].
As mentioned above only left-handed fermions (and right-handed anti-fermions) are subject to the weak interaction. Here handedness refers to the chiral components of a particles wave function[1]. Additionally one should note, that the doublets of I_W^3 hold the flavor eigenstates of the particles. That means, for example, the W^+ will couple to the flavor eigenstate b' in the decay of a t-quark as shown in the Feynman-diagram in Figure 2.2a. The mass eigenstates we observe are the result of the flavor mixing described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} = \begin{pmatrix} 0.974 & 0.225 & 0.004 \\ 0.224 & 0.974 & 0.042 \\ 0.009 & 0.041 & 0.999 \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{2.6}$$

with the values as reported by the PDG group [22]. Similarly, the flavor eigenstates of the neutrinos that participate in weak interactions are superpositions of the mass eigenstates given by the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix:

$$\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = \begin{pmatrix} U_{e1} & U_{e2} & U_{e3} \\ U_{\mu 1} & U_{\mu_2} & U_{\mu 3} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix}. \tag{2.7}$$

Figure 2.2b shows a decay of a W^- boson into a $\bar{\nu}_e$ neutrino and an e^-. Here it has to be an anti-neutrino, since an important quantum number for the weak interaction, the lepton number L, has to be conserved. Any lepton has $L = 1$ while any anti-lepton has $L = -1$. Additionally, there are the similar lepton flavor numbers L_e, L_μ and L_τ, for which partial conservation is broken by neutrino oscillation.

[1] The chiral components of any dirac spinor are derived from the chiral projection operators P_L and P_R. In the ultrarelativistic limit ($E >> m$) particles with positive helicity are almost completely right-handed and particles with negative helicity are almost completely left-handed, with the helicity being $h = \frac{\vec{s}\cdot\vec{p}}{p}$. Thus for left-handed particles in the ultrarelativistic limit, the direction of the momentum is opposite to that of its spin.

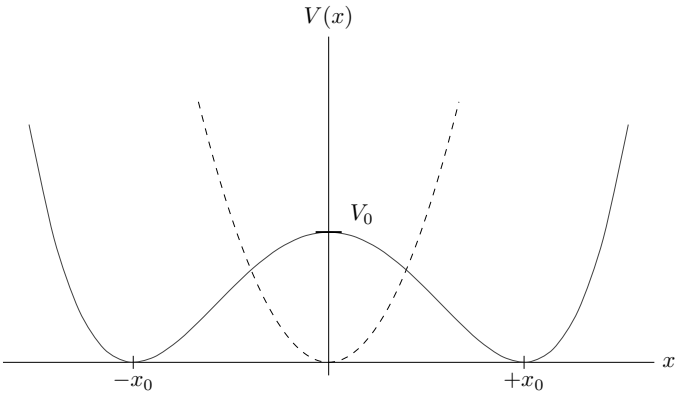


Figure 2.3: Sketch of a one-dimensional well (dashed line) $V(x) = kx^2$ as well as a double-well potential (solid line) $V(x) = \frac{V_0}{x_0^4}(x^2 - x_0^2)^2$

2.1.2 Higgs mechanism

The Lagrangian of the EW sector does not hold any terms responsible for the masses of the massive gauge bosons $W^{\pm}$ and Z bosons without breaking the gauge invariance. To address this problem François Englert, Robert Brout [26] and Peter W. Higgs [27] introduced the mechanism of spontaneous symmetry breaking in the electroweak sector of the SM. Considering a simple quadratic potential for a physical system such as $V(x) = kx^2$, as shown with the dashed line in Figure 2.3, the potential itself is symmetric, i.e. $x \to -x : V(x) = V(-x)$ and so would be the description of the system if one would perturb it around its minima at $x = 0$. For a different system with a double-well potential $V(x) = \frac{V_0}{x_0^4}(x^2 - x_0^2)^2$ as shown in Figure 2.3 with the solid line the situation changes. The potential itself is still symmetric around $x = 0$, but the description of the system after perturbation around one of the minima at $x = \pm x_0$ is not. The choice of the minima is arbitrary and the system can be in either of the two minima. This is called *spontaneous symmetry breaking*. The system is still symmetric, but the description of the system is not.
With the Higgs mechanism a doublet of $SU(2)_L$ of scalar complex fields with four degrees of freedom is introduced:

$$\vec{\Phi} = \begin{pmatrix} \Phi^+ \\ \Phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \Phi_1 + i\Phi_2 \\ \Phi_3 + i\Phi_4 \end{pmatrix}, \tag{2.8}$$

where the Φ_i are real scalar fields. This doublet transforms under the same gauge transformations of $SU(2)_L \otimes U_Y$, so with the choice for the potential of $V(\vec{\Phi})$ which reads as

$$V(\vec{\Phi}) = \lambda \left(|\vec{\Phi}|^2 - \frac{v^2}{2} \right)^2, \tag{2.9}$$

with its *vaccum expectation value* vev at $\langle |\vec{\Phi}| \rangle = \frac{v}{\sqrt{2}}$, one can show and the quartic coupling λ, that three of the four degrees of freedom of the doublet can be rotated away by a suitable transformation of the EW gauge group. The resulting doublet

$$\vec{\Phi} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}, \tag{2.10}$$

holds only the vev and the remaining degree of freedom in the form of the real scalar field $H(x)$, which is associated with the *Higgs boson*. By inserting the doublet in this form into the EW Lagrangian we get the following mass terms:

$$\mathscr{L}_{\text{mass}} = \frac{g_I^2 v^2}{4}\frac{1}{2}((W_\mu^1)^2 + (W_\mu^2)^2) + \frac{v^2}{2}(-\frac{g_I}{2}W_\mu^3 + \frac{g_Y}{2}B_\mu)^2, \tag{2.11}$$

where g_I and g_Y are the coupling constants of the $SU(2)_L$ and $U(1)_Y$ gauge groups respectively. These relate to the weak mixing angle $\theta_W = \tan^{-1}(g_Y/g_I)$, thus one can rewrite the expression of the Z_μ boson and photon A_μ in terms of the electroweak scalar fields from Equation 2.4 as:

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \frac{g_I}{\sqrt{g_I^2+g_Y^2}} & \frac{g_Y}{\sqrt{g_I^2+g_Y^2}} \\ -\frac{g_Y}{\sqrt{g_I^2+g_Y^2}} & \frac{g_I}{\sqrt{g_I^2+g_Y^2}} \end{pmatrix} \cdot \begin{pmatrix} B_\mu \\ W_\mu^3 \end{pmatrix}, \tag{2.12}$$

which in turn can be used together with Eq. 2.3 to identify the weak gauge bosons in Eq. 2.11 and rewrite it into:

$$\mathscr{L}_{\text{mass}} = \left(\frac{g_I v}{2}\right)^2 W_\mu^+ W^{-\mu} + \frac{1}{2}\frac{v^2(g_I^2+g_Y^2)}{4} Z_\mu Z^\mu,. \tag{2.13}$$

From these mass terms follow directly the masses of the massive gauge bosons:

$$m_W = \frac{g_I v}{2}, \quad m_Z = \frac{v\sqrt{g_I^2+g_Y^2}}{2}. \tag{2.14}$$

There is no mass term for the photon, thus remains massless.
Taking the doublet $\vec{\Phi}$ in the form with only the scalar Higgs field of Eq. 2.10 and inserting it into the potential Eq. 2.9 one gets the final form of the Higgs potential:

$$V(H(x)) = \frac{\lambda}{4}H(x)^4 + \lambda v H(x)^3 + \lambda v^2 H(x)^2, \tag{2.15}$$

where one can identify the quadratic term in H as the mass term of the Higgs boson, yielding the mass itself to be $m_H = \sqrt{2\lambda}v$. The masses of the fermions are strictly not a direct consequence of the Higgs mechanism, since they originate from their interaction with the scalar field $\vec{\Phi}$, thus being proportional to the Yukawa coupling y_i and the vev v:

$$m_i = \frac{y_i v}{\sqrt{2}}. \tag{2.16}$$

With the discovery of the Higgs boson at the LHC in 2012 by the CMS and ATLAS collaborations [28, 29], the mechanism of electroweak symmetry breaking was confirmed. Its mass has been measured to be $m_H = 125.15 \pm 0.17$ GeV [22].

2.1.3 Quantum chromodynamics

In the formalism of quantum chromodynamics the local gauge invariance under transformations of the $SU(3)_C$ gives rise to the strong interaction with its mediators - the gluons. These massless gauge bosons carry combinations of color and anti-color, such that there can be eight

independent superpositions of color states. Besides interaction with the only color-carrying fermions - the quarks - gluons also interact with each other. This, together with the running of the coupling of the strong interaction α_s, leads to different phenomena arising at different energy scales. At low energies (i.e. large distances) α_S becomes increasingly large, which leads to the *confinement* of the quarks. This means at larger distances and small energies there can not be free quarks, but only bound states. As the distance between quarks in a bound state increases, the gluon field between them does not decrease, but rather the energy increases, to the point where it is more favorable to form a new quark anti-quark pair resulting in two bound states. This process of *fragmentation* will lead to particle showers to be produced in high energy collisions, which are called *jets* and will be discussed in more detail in section 5. On the other hand, when one probes interactions at small distances or high energies, the coupling α_S decreases to the point of *asymptotic freedom*, where quarks can be treated as quasi-free particles.

3 Experimental Setup and Methods

The analyses presented in this thesis study data from *pp* collisions recorded with the Compact Muon Solenoid experiment (CMS) at the Large Hadron Collider (LHC) during the years 2016-2018 (Run 2). In this Section, the LHC is briefly introduced and the CMS detector is described in more detail, focusing on the detector sub-systems, that are relevant for the presented analyses.

3.1 Large Hadron Collider

The LHC [30] is the largest two-ring circular hadron accelerator and collider used for the study of *pp* and heavy-ion collisions. It is located and operated at the European Organization for Nuclear Research (Conseil Européen pour la Rechere Nucléaire - CERN) near Geneva, Switzerland. The LHC is built in a 26.7 km long tunnel, which was previously used by the Large Electron Positron (LEP) collider. The LHC is designed to collide protons at a center of mass energy $\sqrt{s} = \sqrt{4E_{p_1}E_{p_2}}$ of up to 14 TeV and was operated during 2016-2018 at $\sqrt{s} = 13$ TeV. For this, first, protons are collected from ionizing hydrogen atoms, which are accelerated using a series of pre-accelerators of the CERN accelerator complex: the protons collected from the source are accelerated in the linear accelerator LINAC 2, then in the circular accelerators Proton Synchrotron Booster (BOOSTER), Proton Synchrotron (PS) and Super Proton Synchrotron (SPS). At the end of this chain, the proton beams reach an energy of 450 GeV. The proton beams are split into $n_b = 2808$ bunches (25 ns bunch spacing) with $N_b = 1.15 \times 10^{11}$ protons per bunch, which are injected into the two beam pipes of the LHC. Here they will circulate the LHC rings in opposite directions at a revolution frequency of $f_{\text{rev}} = \frac{c}{26.7\text{ km}} \sim 11.24$ Hz. The LHC consists of eight arcs where the protons are guided on circular trajectories by superconducting dipole magnets with a nominal magnetic field of 8.33 T, and eight straight sections, in which the protons are accelerated using high-frequency radiofrequency (RF) cavities tuned at 400 MHz. The protons are accelerated successively in each revolution until both proton beams reach an energy of 6.5 TeV. The two beams are then brought to collision at the four interaction points (IP) located at the center of detectors of the four main experiments at the LHC: LHCb (**L**arge **H**adron **C**ollider **b**eauty), ALICE (**A** **L**arge **I**on **C**ollider **E**xperiment), ATLAS (**A** **T**oroidal **L**HC **A**paratu**s**) and CMS (**C**ompact **M**uon **S**olenoid). While CMS and ATLAS are detectors designed as general-purpose detectors, covering a wide range of physics, LHCb is designed to study the physics of B-mesons and ALICE is designed to study heavy-ion collisions.

The rate at which a process occurs in a collider experiment $\frac{dN}{dt} = \sigma \mathscr{L}$ is characterized by the instantaneous luminosity $\mathscr{L}$ and the cross section σ of the process. The instantaneous luminosity

for the LHC is described by:

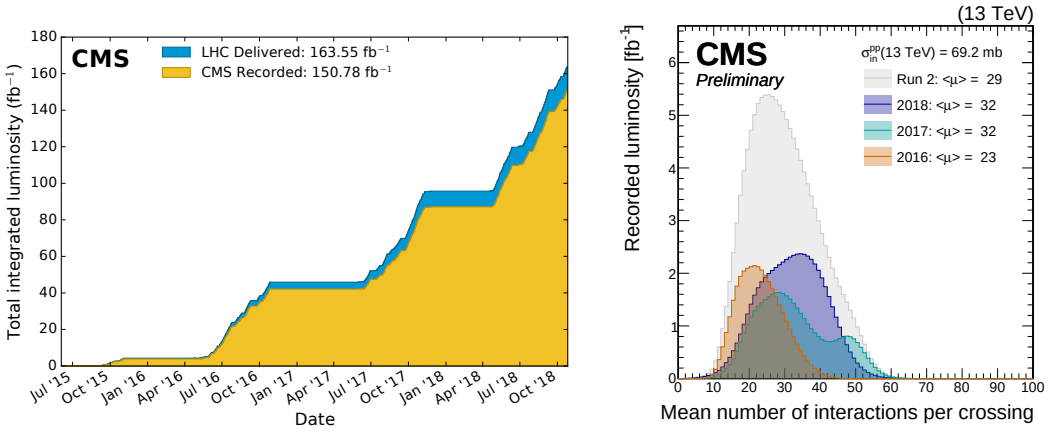
$$\mathscr{L} = \frac{n_b N_b^2 f_{\text{rev}}}{4\pi\sigma_{b,x}\sigma_{b,y}}, \tag{3.1}$$

where n_b and N_b are the number of bunches and number of protons per bunch, respectively, f_{rev} is the revolution frequency, and $\sigma_{b,x}$ and $\sigma_{b,y}$ are the transverse beam sizes in the horizontal and vertical directions, respectively. The transverse beam sizes are reduced using focusing quadrupole magnets to the order of 16 μm and below. The maximum collision rate is determined by the frequency of the RF cavities and the consequent bunch spacing of 25 ns, to 40 MHz corresponding to $n_b^{\text{max}} = \frac{1}{f_{\text{rev}} \cdot 25\text{ ns}} \sim 3560$. To allow for abort and injection gaps only 80% of the bunches are filled, yielding in $n_b = 2808$ and thus an effective bunch crossing rate of 32 MHz [31]. With this, the LHC was designed to reach a peak instantaneous luminosity of $\mathscr{L} \sim 10^{34}\ \text{cm}^{-2}\text{s}^{-1}$. The cross-sectional size of the beams and their crossing angle are the metrics that can be controlled to increase the instantaneous luminosity, everything else is fixed. During LHC Run 2 the design luminosity was reached and surpassed.

The total number of events N of a process produced in a collider experiment is characterized by the integrated luminosity recorded by the experiment $\mathscr{L}_{\text{int}}$ and the cross section σ of the process:

$$N = \sigma\mathscr{L}_{\text{int}} = \sigma\int \mathscr{L}\,\mathrm{d}t. \tag{3.2}$$

The evolution of the integrated luminosity delivered by the LHC and the one recorded by the CMS experiment is shown as blue and yellow histograms in Figure 3.1a. The data recorded during Run 2 by the CMS experiment, which passed all quality criteria correspond to an integrated luminosity of $\mathscr{L}_{\text{int}} \approx 138\,\text{fb}^{-1}$ [32–35]. Though a higher luminosity yields a higher rate of physics processes that are of interest for physics analyses, i.e. the hardest inelastic scattering of two protons, it also increases the chances to have collisions of other protons from the same or adjacent bunch crossing. These processes overlay the process of interest in each collision event and are called pileup. They are one of the challenges in the process of analyzing the data, as they can obscure many crucial observables of the process of interest. To minimize the effect the pileup has on the performance of analyses, dedicated pileup mitigation techniques are employed, which are discussed in Section 5.2.1. The distribution of the mean number of interactions per bunch crossing (pileup profile) is measured by the CMS experiment using the instantaneous luminosity and a total inelastic pp collision cross section of $\sigma_{\text{in}}^{pp} = 69.2$ mb and shown in Figure 3.1b for each year of Run 2 separately and integrated over the whole Run. The mean number of interactions per bunch crossing reached up to 70 and was on average 29 during Run 2.



(a) (b)

Figure 3.1: The left plot shows the evolution of the integrated luminosity that was delivered by the LHC (blue histogram) and recorded by the CMS experiment (yellow histogram) during Run 2 [35]. The right plot shows the distribution of the mean number of interactions per bunch crossing (pileup profile) in Run 2 pp collision estimated from the instantaneous luminosity and a total inelastic cross section $\sigma_{\text{in}}^{pp} = 69.2$ mb. The orange, cyan and blue histograms correspond to the data taken in 2016, 2017 and 2018 respectively. The grey histogram corresponds to the integrated profile of all years [36].

3.2 The Compact Muon Solenoid Detector

The Compact Muon Solenoid (CMS) detector is a general-purpose detector, located at Interaction Point 5 at the LHC ring [10, 37]. The design of the detector is driven by the primary physics goal of the LHC, to study the nature of the electroweak symmetry breaking, search for the Higgs boson and to search for physics beyond the SM at the TeV scale. The detector is structured in a cylindrical barrel part and two forward parts, with an overall length of 21.6 m, a diameter of 14.6 m and a total weight of 12500 tons. The detector is designed to be as hermetic as possible to allow for the reconstruction of all particles produced in the particle collisions, covering a solid angle of almost 4π. The detector is composed of several subdetectors, which are arranged in layers around the interaction point, as depicted in Figure 3.2. At the innermost layer is the tracking system consisting of a pixel detector and silicon strip detectors. The tracking system is surrounded by an electromagnetic calorimeter, which is followed by a hadronic calorimeter. A solenoid magnet coil is located outside of the calorimeters and is followed by the muon system, which is located in the outermost layer of the detector.

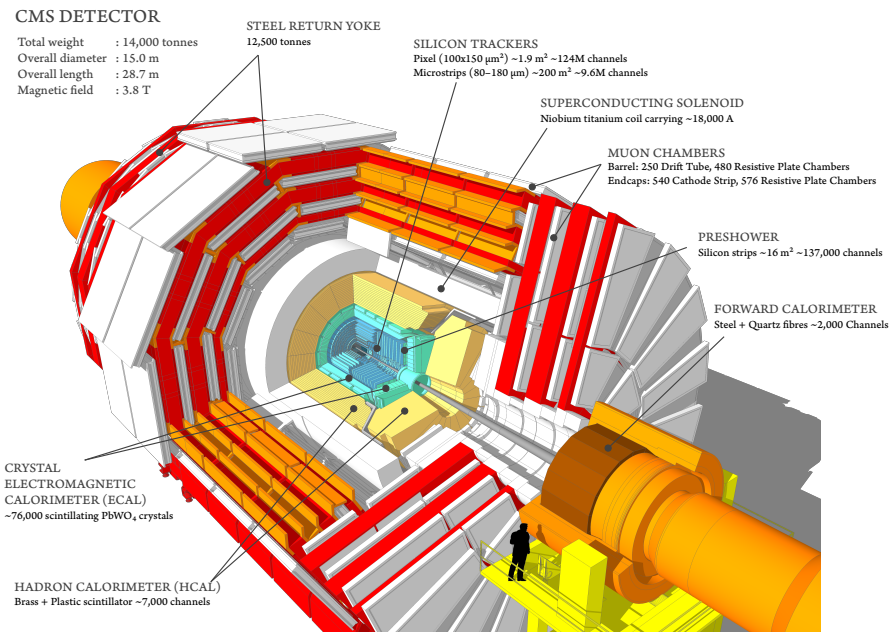


Figure 3.2: Overview of the CMS-detector after the Phase-1 upgrade [38, 39].

The magnet is a superconducting solenoid coil, that is 12.9 m and has an inner diameter of 5.8 m in order to encapsulate the majority of the subdetectors. The magnetic field strength is 3.8 T and is used to bend the trajectories of charged particles.

The CMS experiment uses a right-handed coordinate system, that has its origin in the interaction point. The x-axis points radially inwards towards the center of the LHC ring, the y-axis points upwards and the z-axis is parallel to the beam axis. Since the detector is approximately a cylinder, the cylindrical coordinates azimuthal angle ϕ, the polar angle θ and the radial distance r from the z-axis are used. The azimuthal angle ϕ is measured in the (x, y)-plane from the x-axis and the polar angle θ is measured from the z-axis. The pseudorapidity η is defined as $\eta = -\ln(\tan(\theta/2))$ and is used instead of the polar angle, since differences in η are invariant under

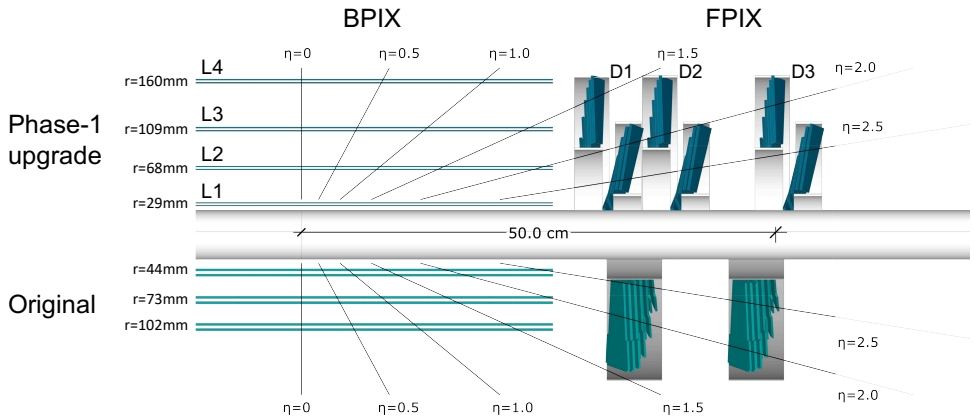


Figure 3.3: Schematic Overview of Pixel detector as part of the tracking system before and after the Phase-1 upgrade, which increased both the layer in the barrel region (BPIX) from three to four and the number of discs in the forward region (FPIX) from two to three. The design of the discs also got updated, to include inner and outer rings. Taken from [40]

Lorentz boosts along the z-axis. The transverse components of the momentum and energy are computed from the measured components in the (x, y)-plane, $p_T = \sqrt{p_x^2 + p_y^2} = p\sin(\theta)$ and $E_T = \sqrt{E_x^2 + E_y^2} = E\sin(\theta)$. Any object reconstructed in the detector can be identified by a four-vector constructed from only the transverse momentum p_T, the pseudorapidity η the azimuthal angle ϕ and the invariant mass m of the object. The distance ΔR between two objects in the (η, ϕ)-plane is defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} = \sqrt{\eta_1 - \eta_2|^2 + |\phi_1 - \phi_2|^2}$.
In the following, the individual sub-systems of the detector are discussed in more detail.

3.2.1 The Tracker

The Tracker or, tracking system is the innermost subdetector at the heart of the CMS detector. It comprises a pixel detector and a strip detector. The tracker system is the first detector, which the outgoing particles produced in the collisions traverse, hence it receives the highest particle flux and has to be made of radiation-hard material and be able to process a large number of input signals (hits). Both in the pixel and the strip detector silicon sensors are used to detect charged particles, which will ionize detector material, causing electrons to travel toward electrodes and induce a signal current. The individual hits are used to reconstruct the trajectories of the charged particles as tracks. The tracks are then used to reconstruct primary and secondary interaction vertices. The charged particles are subject to the magnetic field of the solenoid, thus the tracks are bent, and their curvature is characterized by the strength of the magnetic field B and the momentum of the particle p. Hence the tracks play a crucial role in the reconstruction of the

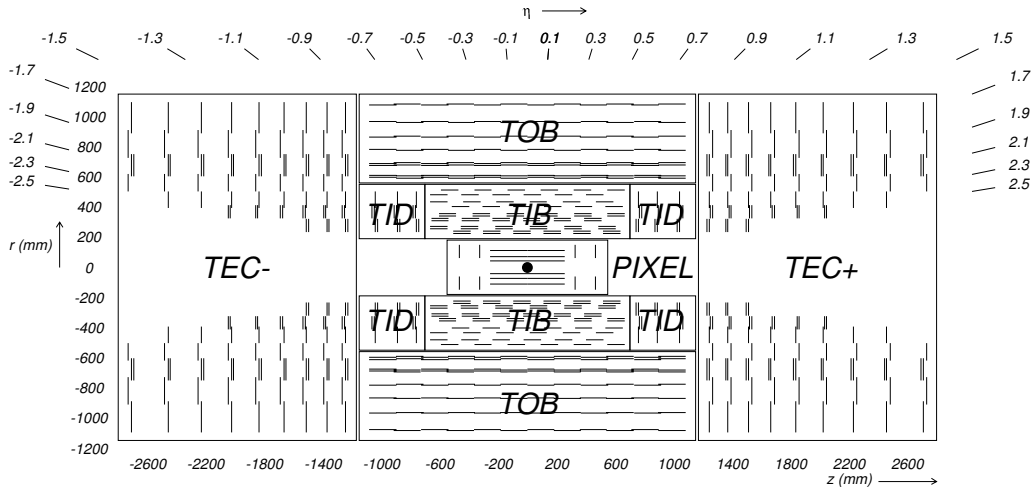


Figure 3.4: Schematic Overview of tracking system as of 2016 in the Phase-0 design, before the Phase-1 upgrade. Taken from [10]

momentum of the particles. Consequently, the momentum resolution and the vertex position resolution depend on the spatial resolution achieved in the tracking system.

The pixel detector was replaced with a new, upgraded pixel detector (Phase-I upgrade) after the data-taking period in 2016. Figure 3.3 shows a schematic overview of the design before and after the upgrade on the lower and upper half of the figure respectively. Before its upgrade, it consisted of three layers of sensors located in the barrel region around the beam pipe at different radii, the smallest being $r = 4.4$ cm, and two disks of sensors in each endcap region. In the upgrade, an additional layer was added to the barrel region, where the innermost layer moved closer to the beam pipe ($r = 2.9$ cm), and the number of discs in the endcap regions was increased from two to three [40]. The number of readout channels in the pixel detector was almost doubled from 66 million to 124 million channels. All silicon sensors of both pixel detector designs have a size of $100 \times 150\,\mu\text{m}^2$.

The silicon strip detector surrounds the pixel detector and is segmented similarly into separate barrel and endcap sections. Figure 3.4 shows the schematic overview of the tracking system with the Phase-0 pixel detector. The strip detector segments in both the barrel region and the endcap regions consist of an inner and an outer layer. For the inner layer strip sensors with a pitch between $80-120\,\mu$m and a thickness of $320\,\mu$m are used, while the outer layer uses sensors with a pitch of $120-180\,\mu$m and thickness of $500\,\mu$m. In total there are 11 layers in the barrel region and 12 discs in each endcap region. The total number of readout channels in the strip detector is 9.3 million.

Overall the tracker system covers the region $|\eta| < 2.5$ in the Phase-0 design and $|\eta| < 2.7$ in the Phase-1 design. The four-hit coverage of the Phase-1 design reaches up to $|\eta| < 2.5$. The spatial hit position resolution of the pixel detector is approximately $10\,\mu$m in the transverse and $20-40\,\mu$m in the longitudinal coordinate. For the strip detector the resolution in the (r,ϕ)-plane reaches $13-38\,\mu$m for the inner layers and $18-47\,\mu$m for the outer layers. For isolated muons (reconstruction $\sim 100\%$ efficient) this propagates to momentum resolutions of $\lesssim 1\%$ for central muons with $p_T < 20\,\text{GeV}$. For higher momenta, the resolution degrades to (e.g. $\sim 4\%$ for $p_T = 100\,\text{GeV}$) [41].

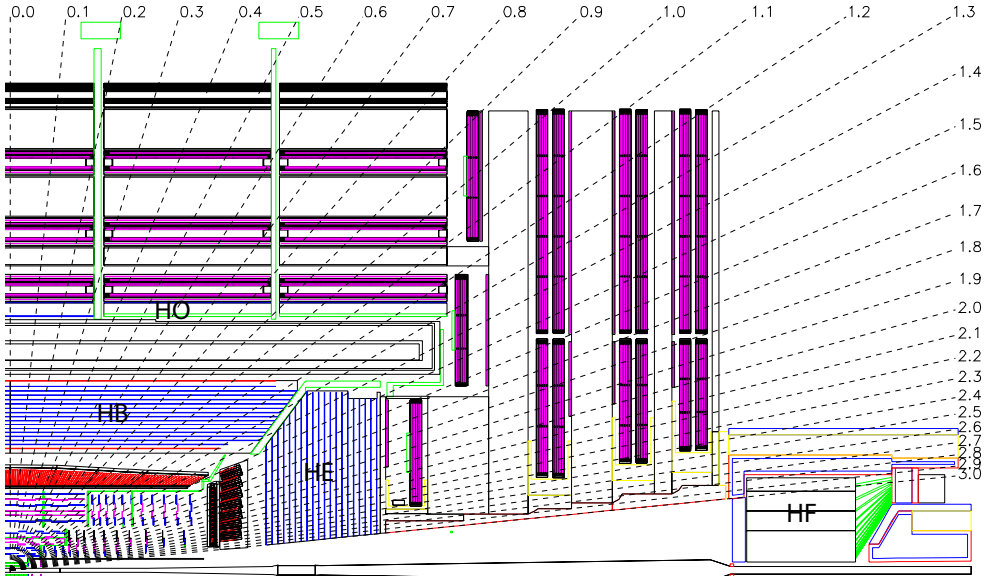


Figure 3.5: Schematic Overview of one quadrant of the electromagnetic and hadronic calorimeter. Taken from [10]

3.2.2 The Calorimeters

The CMS experiment uses calorimeters for the energy measurement of all charged and neutral particles, except for the muon and neutrinos. When particles traverse a calorimeter they interact with the calorimeter material and decay in cascades forming showers of secondary particles. The subsequent shower particles with low energy will be absorbed by the scintillating detector material, which will emit the energy in form of photons, which are detected by photodetectors. The energy deposited by the shower can be reconstructed from the cumulative intensity of the emitted photons. The energy resolution of a calorimeter can be generally parameterized as the quadratic sum of three terms:

$$\frac{\sigma_E}{E} = \frac{S}{\sqrt{E}} \oplus \frac{N}{E} \oplus C, \tag{3.3}$$

where S is the stochastic term, N is the noise term and C is the constant term. The energy resolution of calorimeters, in contrast to the resolution of the tracking system, improves at high energies but saturates when the constant term becomes dominant. The characteristic radiation length X_0 (for electrons and photons) and the interaction length λ_I (for hadrons) of the absorber material are crucial metrics in the design of calorimeters. These are the distances over which the energy of an electron, photon or hadron is reduced by a factor of $1/e$, respectively. The two different calorimeters used by the CMS experiment are shown in the schematic overview in Figure 3.5 and will be described in the following.

3.2.2.1 Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) is used to stop and measure energy of mainly photons and electrons. It is a homogenous calorimeter constructed with scintillating lead-tungstate ($PbWO_4$) crystals. It is segmented into a barrel part (EB), covering $|\eta| < 1.479$ and two endcaps

(EE), which cover $1.479 < |\eta| < 3.0$. The barrel part is constructed using 61200 tapered-shaped crystals, each with a cross section corresponding to approximately 0.0174×0.0174 in the (η, ϕ)-plane. They are 230 mm long, which corresponds to $25.8 X_0$. The light emitted by the crystals is detected by silicon avalanche photodiodes (APDs). In the endcaps, the crystals have a cross section of 28.6×28.6 mm and a length of 220 mm, which corresponds to $24.7 X_0$. In the endcaps, the light is detected by vacuum phototriodes (VPTs). The energy resolution of the ECAL was measured in test beam experiments to be [42]:

$$\frac{\sigma_E}{E} = \frac{2.8\%}{\sqrt{E}} \oplus \frac{12\%}{E} \oplus 0.3\%. \tag{3.4}$$

Additional preshower detectors (ES) are located in front of the endcaps to improve the identification of neutral pions. Each preshower detector is a sampling calorimeter and consists of two layers of lead absorbers ($2X_0$ and $3X_0$) and two planes of silicon strip detectors as active material.

3.2.2.2 Hadronic Calorimeter (HCAL)

The hadronic calorimeter (HCAL) is used to measure the energy of hadrons, that were not stopped by the ECAL. The interaction length λ_I of hadrons is typically much longer than the radiation length of electrons or photons, thus a homogenous calorimeter is not feasible with the limited space left inside the solenoid. Therefore the HCAL is designed as a sampling calorimeter, where the absorber material is used to initiate the hadronic shower and the active material is used to measure the energy of the shower. The HCAL is segmented into a barrel part (HB), covering $|\eta| < 1.3$, two endcaps (HE), covering $1.3 < |\eta| < 3.0$, two forward calorimeters (HF), covering $3.0 < |\eta| < 5.0$ and an additional outer calorimeter (HO), which located outside of the solenoid as a tail catcher and covers $|\eta| < 1.3$. As material for the absorber layers brass is used for the HB and HE, and steel is used for the HF and HO. The active material is a plastic scintillator for the HB, HE and HO and quartz fibers for the HF. The raw energy resolution of the combined ECAL+HCAL system was measured in test beam experiments to be [43]:

$$\frac{\sigma_E}{E} = \frac{111.5\%}{\sqrt{E}} \oplus 8.6\%. \tag{3.5}$$

3.2.3 The Muon System

The outermost subdetector of the CMS experiment is the muon system, which is used to identify and measure the momentum of muons. It is embedded in the return yoke of the solenoid and uses different types of gaseous detectors in different regions of η. In the central region $|\eta| < 1.2$ Drift Tubes (DT) are used, while in the two endcap regions $0.9 < |\eta| < 2.4$ Cathode Strip Chambers (CSC) are used. Additionally, Resistive Plate Chambers (RPC) are used in both the barrel and endcap regions, to improve the timing resolution of the measurements and to provide a trigger signal.

3.2.4 Trigger

With a bunch crossing rate of up to 40 Hz, the LHC delivers a *pp* collision rate of the order of $\mathcal{O}(10^9)$ per second. To save all collision events is not feasible, as the rate to save data is limited to $\mathcal{O}(100\ \text{Hz})$. In order to reduce the rate of events, without discarding events that are interesting for physics analyses, a trigger system is used to select events that are stored. The trigger system is comprised of two stages: the hardware-based Level-1 (L1) triggers and the software-based High-Level triggers (HLT). The L1 triggers are deciding within $\sim 4\mu s$ based on information from the calorimeters and the muon system, to discard an event or to pass it on to the HLT, and reduce the rate to about 100 kHz. The HLT is a software-based trigger, which uses information from all subdetectors to perform a basic reconstruction of high-level objects, to decide whether an event is saved or discarded and reduces the rate to an average of 400 Hz [44].

4 Event Simulation and Reconstruction

The precise analysis of the recorded data requires a comparison to the theoretical prediction of the processes of interest. This section provides an overview of the simulation and reconstruction of pp collision events employed in CMS data analyses. The analyses presented in this thesis study jet final states, thus the reconstruction and calibration of jets and the methods that exploit the substructure of the jets are discussed in the dedicated Section 5.

4.1 Simulation of pp collision events

The simulation of the pp collisions is performed in several steps. The proton is a composite particle, with an inner structure that includes three valence quarks (u, u, d), sea quarks of different flavors and gluons. Following the *parton model* [45], the deep inelastic scattering of the two protons is governed by the scatterings of individual partons of the protons. The scattering among partons that has the largest momentum transfer is referred to as the *hard scattering*. The inner structure of the proton can be described by *parton distribution functions* (PDFs) $f_i(x_i; Q^2)$, which gives the probability, that the parton i carrying the fraction x_i of the momentum of the proton is observed in a proton at a probed energy scale Q^2. The momentum fraction is defined as $x \equiv \frac{Q^2}{2q\cdot P}$ [19] (Bjorken scale) with the total momentum transfer of the scattering $Q^2 \equiv -q^2$ and the four-momentum of the proton P. Following the *factorization theorem* [46], the deep-inelastic pp scattering cross section with the final state X can be expressed by the convolution of the parton distribution functions $f_i(x_i; \mu_\mathrm{F}^2)$ and the partonic cross section $\hat{\sigma}_{ij\to X}(x_1 x_2 s, \mu_\mathrm{R}^2, \mu_\mathrm{F}^2)$:

$$\sigma_{pp\to X} = \sum_{i,j} \int\int f_i(x_1; \mu_\mathrm{F}^2) \hat{\sigma}_{ij\to X}(x_1 x_2 s, \mu_\mathrm{R}^2, \mu_\mathrm{F}^2) f_j(x_2; \mu_\mathrm{F}^2) dx_1 dx_2, \tag{4.1}$$

where the sum runs over all possible initial-state partons of each proton, with their respective momentum fractions $x_{1,2}$, that can give rise to the process with the final state X and the center-of-mass energy $\sqrt{\hat{s}} = \sqrt{x_1 x_2 s}$. The partonic cross section $\hat{\sigma}_{ij\to X}(x_1 x_2 s, \mu_\mathrm{R}^2, \mu_\mathrm{F}^2)$ can be calculated perturbatively in the strong coupling constant α_s (pQCD). It will depend on the renormalization scale μ_R^2 and factorization scale μ_F^2 if the expansion is truncated, to avoid ultraviolet (UV) and infrared (IR) divergencies, respectively. The PDFs, especially at low energy scales, can not be described by pQCD, but are rather the subject of non-perturbative QCD. Thus they are parameterized by fits to data at given momentum fractions x and factorization scale μ_F^2. In the context of this thesis the NNPDF3.1 [47] PDFs are used in simulations of pp collisions. Figure 4.1 shows the PDFs for different parton flavors scaled by their momentum fraction x, derived for $\mu_F^2 = 10\,\mathrm{GeV}^2$ in the left plot and for $\mu_F^2 = 10^4\,\mathrm{GeV}^2$ on the right plot. The evolution of the PDFs to the scale Q is derived using the DGLAP equations [48–51].

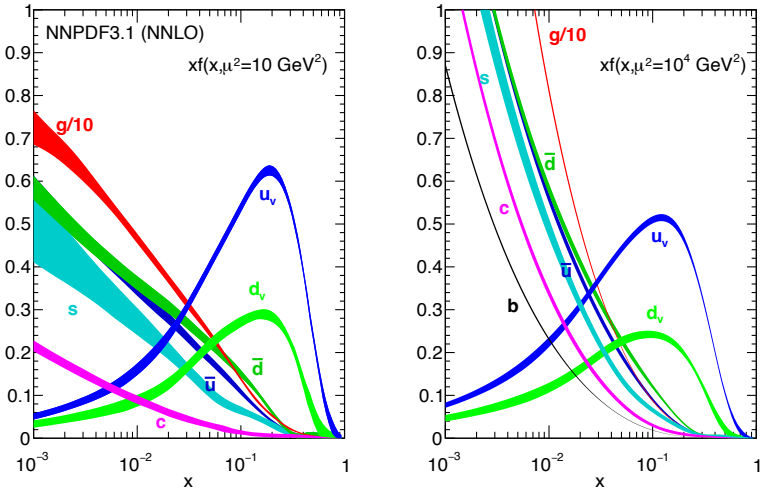


Figure 4.1: Overview of the NNPDF3.1 NNLO parton distribution functions $f_i(x_i;Q^2)$ for different partons i of different flavors, scaled by their momentum fraction x. Evaluated at low momentum transfer $Q^2 = 10\,\text{GeV}^2$ on the left and high momentum transfer $Q^2 = 10^4\,\text{GeV}^2$ on the right as a function of Bjorken x. Taken from [47]

The first step in the MC simulation chain is the generation of the hard scattering. For this, the expansion in α_s for the calculation of $\hat{\sigma}_{ij\to X}$ is typically performed at leading order (LO) or next-to-leading order (NLO), so summing first and second order terms in the perturbation series. Higher order corrections at NLO or next-to-next-to-leading order (NNLO) can be derived using fixed-order prediction calculations both in QCD and EW and used as a correction factor to the lower order cross section prediction. The hard scattering is then simulated numerically using MC matrix-element generators, by populating the phase space specific to the process $pp \to X$ with events. In the context of this thesis the simulation of the hard scattering is performed by either of the following: The matrix-element generation for the samples used in the analyses is handled by one or more of the following generators: MADGRAPH5_aMC@NLO [52] v2.6.5 (v2.6.1), POWHEG 2 [53–55] or PYTHIA 8.244 [56].

The generation of the hard scattering is followed by the simulation of *parton showers* and *hadronization*, which is performed by PYTHIA. The parton showering is used to approximate the orders of the expansion in the momentum transfer for the splitting processes of the initial and final state particles, the initial state radiation (ISR) and the final state radiation (FSR) respectively [57]. The evolution of the parton shower is stopped when the energy scale of the showered particles reaches energies out of reach of the perturbative expansion. For event samples, which include extra partons in the simulation of the hard scattering, at this stage final state partons from the parton shower and any additional final state partons added to the hard scattering by the matrix-element generator, are matched to one another to avoid double counting using the MLM scheme [58] for samples generated at LO and using the FXFX scheme [59] for samples generated at NLO. After this, in the hadronization process, the final state particles that carry color charge will form color-neutral hadrons, which is approximated using the Lund-String model [60]. The remnant partons, that do not participate in the hard scattering, form the underlying event (UE), which is modeled also by PYTHIA. The events are

then overlayed with additional *pp* collisions to simulate the effect of pileup.
Finally, the events are subjected to a simulation of the response of the CMS detector, using GEANT4 [61], which is interfaced with a detailed description of the material and geometry of the CMS detector, as well as the status of the detector components during given periods of data-taking.

4.2 Object reconstruction and particle identification

The *pp*-collision events in the recorded data and the simulation undergo the same procedure of event reconstruction. The event reconstruction consists of multiple steps and yields collections of high-level objects, representing particles, jets and other physics objects. In the CMS experiment the particle-flow (PF) algorithm [42] is used to reconstruct final state particles (PF particles or PF candidates) by linking information from the individual sub-detectors. The following Section 4.2.1 provides an overview of the PF algorithm. The reconstruction of jets is covered in the dedicated Section 5.

4.2.1 Particle-Flow

The particle-flow approach aims to take advantage of the strengths of the individual sub-detectors in different energy regions, to maximize the reconstruction efficiency and resolution. Additionally, particles traversing the detector will leave characteristic signatures in the different parts of the detector as depicted in Figure 4.2. Photons and neutral hadrons will be detected primarily in the ECAL and HCAL, respectively. Electrons and charged hadrons will produce hits in the inner tracking system and showers primarily in the ECAL and HCAL, respectively. Muons are traversing the detector with minimal interactions and energy loss and will produce hits in the inner tracker and the muon system. Thus the combination of the sub-detectors will also help in the identification of particles.
The PF algorithm used in CMS is segmented into the local reconstruction of tracks and energy clusters and the linking of these elements into PF blocks. The PF blocks are then used to identify stable final-state particles as PF candidates. The individual steps of the PF algorithm are described in the following.
For the local reconstruction of tracks the combinatorial track finding (CTF) [41] algorithm based on an extension of Kalman Filtering (KF) [62] is used. The CTF algorithm is performed iteratively, where each iteration uses information from both the inner tracking system and the muon system to seed and fit tracks from hits in the individual tracking systems. After each iteration, the hits corresponding to tracks that satisfy certain quality criteria are removed from the collection of hits. The algorithm starts with strict quality criteria to first find the most trivial tracks of high p_{T} objects and then loosens criteria after each iteration. In the local reconstruction of energy clusters, the hits in each calorimeter subsystem are clustered with the cells with local energy maxima serving as seeds. Adjacent cells that observe energies above a noise level threshold are merged to form topological clusters.
In the second step, the link algorithm will attempt to link tracks and energy clusters from the local reconstruction and form PF blocks. The search for links is restricted to neighboring elements in

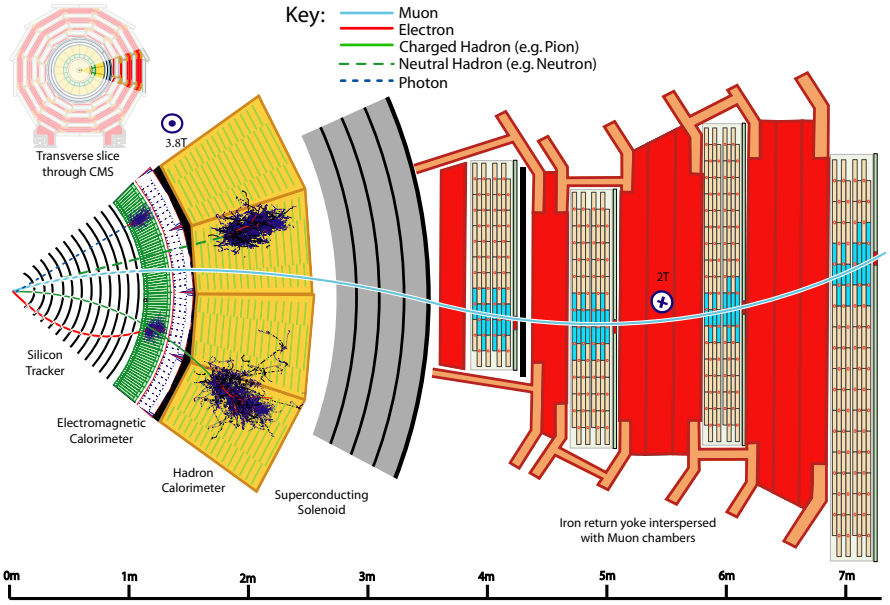


Figure 4.2: Schematic overview of a transverse slice through the CMS detector. Taken from [42]

the (η, ϕ) plane, to prevent the computing time from growing too fast. Links between tracks and energy clusters are established if the extrapolated track position is compatible with the position of the energy cluster. Furthermore, links between energy clusters in different calorimeters are established if the cluster position in the more granular calorimeter agrees with the envelope of the cluster in the less granular calorimeter. Finally, links between tracks in the inner tracker and the muon system are established. The identification of PF candidates from these PF blocks is performed in the last step of the PF algorithm and is described in the following.

4.2.1.1 Muons

First, muon candidates are identified from PF blocks that contain a track in the inner tracker, in the muon system or both. PF blocks that link tracks in both the inner tracker as well as the muon system are considered *global muons*. PF blocks that contain only a track in the muon system are considered *standalone muons* and PF blocks that link a track in the inner tracker with a single muon segment in the muon system are considered *tracker muons*. Inside the coverage of the muon system ($|\eta| < 2.4$) about 99% of the muons are reconstructed as global or tracker muons. All tracks that are identified as muon candidates are removed from the considered collection of tracks for the next step in the PF candidate identification. The candidate muons are required to fulfill an additional identification (ID) criteria [63]. In the context of this thesis, the *loose muon ID* and the *tight muon ID* criteria are used. A *loose* muon is required to be either a tracker or a global muon. A *tight* muon is required to fulfill the following criteria:

- the muon must *loose* muon reconstructed with a track that has hits in at least six layers of the inner tracker and at least one hit in the pixel detector,
- it must be reconstructed both as a tracker muon and a global muon,
- where the tracker muon must have segment matching in at least two muon stations,

- the fit of the global muon track must have $\chi^2/\text{ndof} < 10$ and include at least one hit in the muon system,

- the muon must be compatible with the primary vertex, with a transverse impact parameter $|dXY| < 0.2\,\text{cm}$ and a longitudinal impact parameter $|dz| < 0.5\,\text{cm}$.

4.2.1.2 Electrons and photons

Candidates for electron and isolated photons are identified in the same step. Electrons are identified from PF blocks linking tracks with energy clusters in the ECAL. Candidates for isolated photons are reconstructed from the remaining ECAL clusters that are not linked to any tracks. The electron candidates are required to fulfill identification criteria. The identification criteria include isolation criteria, shower shape criteria and cuts on variables like the difference between the energies reconstructed from the tracks and energy clusters. The electron IDs used in the context of this thesis are the *loose electron ID* and the *medium electron ID*.

4.2.1.3 Hadrons

The remaining particles to be reconstructed are charged and neutral hadrons, while neutral hadrons can also be picked up as non-isolated photons. Charged hadrons are identified from PF blocks linking tracks with energy clusters in the HCAL. All remaining ECAL clusters without a link to a track are considered non-isolated photons, while all remaining HCAL clusters without a link to a track are considered neutral hadrons.

4.2.2 Missing transverse momentum

Particles that do not interact with any of the detectors, for example, neutrinos, can not be directly detected. Instead, they can be inferred by the imbalance of the momentum in the transverse plane. The *pp* interaction has almost no transverse component in the momentum transfer. Any deviation from zero in the vectorial sum of the transverse momenta of all particles in an event can be attributed to the presence of particles that were not detected, or mismeasurements. The missing transverse momentum vector is defined as the negative of the vectorial sum of the transverse momenta of all PF candidates, balancing out the event:

$$\vec{p}_{T,\text{PF}}^{\,\text{miss}} = -\sum_{i}^{N_{\text{PFcand.}}} \vec{p}_{T,i}. \tag{4.2}$$

This uncorrected version is considered the raw PF missing transverse momentum, while any corrections applied to the PF candidates (pileup mitigation, jet energy corrections) are propagated to the missing transverse momentum by recomputing the vectorial sum. The absolute magnitude $|p_T^{\text{miss}}|$ is used as a measure for the undetected neutrinos in W decays in the context of this analysis.

4.2.3 *Primary vertices*

The position of the individual pp collision vertices is reconstructed from tracks, which are reconstructed using the CTF algorithm as described above. For this, tracks that fulfill strict quality criteria are first clustered using the deterministic annealing (DA) algorithm [64], to identify candidate vertices. The adaptive vertex fitter [65] is then used to estimate vertex parameters for candidate vertices, which consist of at least two tracks. The vertex with the highest sum of squared transverse momenta $\sum_{i \in \text{jets}} p_{T,i}^2$ of jets, which are clustered from tracks only using the anti-k_T algorithm [66], assigned to it is considered the leading primary vertex (PV).

5 Jet reconstruction and identification

The *pp* collisions studied at the LHC and in this thesis produce final states with quarks and gluons in abundance. They decay in cascades and form collimated streams of particles in the detector, known as jets. The reconstructed and calibrated jets are vital for the analysis of the *pp* collision, as they serve as a proxy of the jet-initiating particles. The reconstruction of jets is a complex task, as the detector response is non-linear and the jets are affected by pileup. The reconstruction of jets using jet clustering algorithms is described in Section 5.1, while the calibration of jets is described in 5.2. Section 5.3 introduces methods and techniques exploiting and describing jet substructure. The application of these methods in measurements and searches is described in 5.4. More detailed discussions of these topics can be found in [7–9].

5.1 Jet clustering algorithm

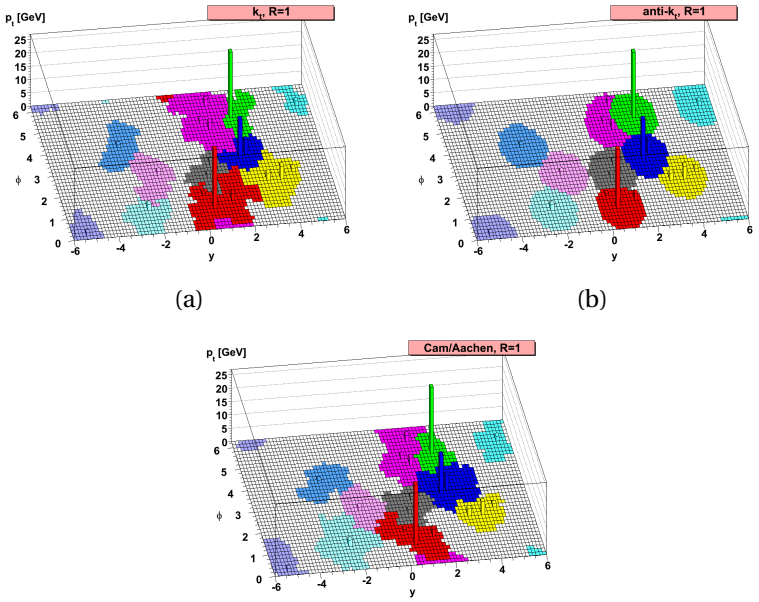
In modern particle physics a variety of jet clustering algorithms are used, that are implemented using the FastJet framework [67]. The reconstruction of the jets in CMS involves most often sequential jet clustering algorithms like the k_T algorithm [68], the anti-k_T algorithm [66] and the Cambridge/Aachen (CA) [69]. Additional jet clustering algorithms adapted in the CMS experiment include clustering algorithms with variable jet radius like HOTVR [70] and algorithms with a fixed number of derived jets like XCone [71, 72]. In the context of this analysis, only sequential jet clustering algorithms are used, which are described in the following. All of the three aforementioned sequential clustering algorithms use common metrics and the same procedure to cluster the PF candidates reconstructed by the PF algorithm as described in Section 4.2.1. The individual PF candidates are considered *pseudojets* and are compared using the distance metrics d_{ij} between two pseudojets i and j and $d_{i\mathrm{B}}$ between a pseudojet i and the beam axis are given by:

$$d_{ij} = \min(p_{T,i}^{2k}, p_{T,j}^{2k}) \frac{\Delta R_{ij}^2}{R^2}, \tag{5.1}$$

$$d_{i\mathrm{B}} = p_{T,i}^{2k}, \tag{5.2}$$

with the transverse momenta $p_{T,i}$ of particle i, the characteristic radius parameter R of the final jet and the geometrical distance between the particles i and j defined as $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_2)^2$, where y_i and ϕ_i is the rapidity and azimuthal angle of the particle i, respectively. The clustering algorithms start with a list of pseudojets and proceed with the following steps:

1. The distance metric d_{ij} are calculated for all combinations of pseudojets i and j and the



(a) (b)

(c)

Figure 5.1: Rendering of three dimensional histogram of $\eta - \varphi$-plane filled with jet-constituents. The different colors correspond to jets clustered using the different algorithms of the family of generalized k_T algorithms. The k_t, anti-k_T and the CA algorithm are shown on the top left, top right and bottom plot, respectively. Taken from [66].

metric $d_{i\mathrm{B}}$ is calculated for all pseudojets i,

2. if the smallest distance is a d_{ij} merge the pseudojets i and j, add the merged pseudojet to the list of pseudojets and remove i and j from the list. If the smallest distance is a $d_{i\mathrm{B}}$, the pseudojet i is considered a jet and removed from the list of pseudojets.

3. Steps 1 and 2 are repeated until the list of pseudojets is empty.

The exponent k characterizes the three algorithms of the family of generalized k_T algorithms and determines the behavior of the clustering procedure. The k_T algorithm uses $k = 1$, thus the clustering prioritizes pseudojets, which have a low p_{T} (soft) or are collinear. This yields irregular-shaped jets and is more sensitive to effects from pileup and the underlying event (UE) [8]. The anti-k_T, which sets $k = -1$ favors the pseudojets with large p_{T} (hard) in the clustering, which yields cone-shaped jets. The anti-k_T algorithm is more resilient against pileup and UE effects and is overall least sensitive to background particles overlaying the event [7]. The CA algorithm sets $k = 0$, so particles are solely compared by their proximity in (y, ϕ)-plane while ignoring their transverse momenta. This makes the CA algorithm less sensitive to soft radiation than the k_T algorithm and results in jets shaped less irregularly than when using k_T, but not as cone-shaped as when using the anti-k_T algorithm. The different clustering algorithms are illustrated in Figure 5.1, where each plot shows the three-dimensional distribution of the jet constituents in $(y, \phi, p_{\mathrm{T}})$. The different colors correspond to the individual jets the respective algorithm clustered from the same list of input pseudojets.
Due to its resilience to soft radiation, the anti-k_T algorithm is widely used in the studies of pp collisions at the LHC. In CMS, thus in the context of this thesis, it was used to reconstruct

jets with a radius parameter of $R = 0.8$ and $R = 0.4$. They are referred to as AK8 and AK4 jets respectively and are used in different contexts. While AK4 jets are used to reconstruct jets that originate from single objects like b-quarks from a leptonic decay of a top quark, the large radius AK8 jets are used to reconstruct the complete hadronic decays of objects like top quarks or W bosons. The pp collisions at the LHC produce heavy particles such as the top quark and W boson often with transverse momenta, that exceed their mass by a lot ($p_T \gg m$). In these cases the particle itself and subsequently their decay products are subject to large Lorentz-boosts, resulting in the decay products being highly collimated. The opening angle ΔR can be approximated in this case by $\Delta R \approx \frac{2m}{p_T}$ [7]. Thus the choice of $R = 0.8$ for large-radius jets allows to expect, that reconstructed jets from W bosons and top quarks will encapsulate most of the decay for transverse momenta starting at $p_T \gtrsim \frac{2 \cdot m_W}{0.8} \approx 200\,\text{GeV}$ and $p_T \gtrsim \frac{2 \cdot m_t}{0.8} \approx 440\,\text{GeV}$ respectively. The underlying jet substructure of the large-radius jets is of special interest since it offers valuable information about the origin of the jet, which is discussed in more detail in Section 5.3. For the study of jet substructure, technically the anti-k_T algorithm is not useful as the last contributions added are soft or collinear. With its prioritization of angular relations, the CA algorithm is more useful in this context, as the last clustering step can be used to identify multiple hard cores of the jet.

5.2 Jet calibration

Before calibration, the jets reconstructed by the clustering algorithms do not reflect the true kinematics of the jet-initiating particle. In the following the methods used to correct for effects from the non-linear detector response and pileup are described. The respective corrections are applied to the jets or underlying jet constituents in data, simulation or both.

5.2.1 Pileup mitigation

Additional pp interactions in the same bunch crossing are considered *pileup* (PU), which produce additional particles that overlie the rest of the event and potentially obscure the kinematics of the reconstructed physics objects (most importantly jets). To mitigate the effect of the PU, different methods are used on particle-level. In CMS and in the context of this thesis the two PU mitigation techniques *charged hadron subtraction* (CHS) [73] and *pileup per particle identification* (PUPPI) [74] are used and will be described in the following.

The CHS technique treats the charged component of PU by removing charged particles from the jet clustering, whose tracks are associated with PU vertices. This does not remove all charged particles from PU, because the track-vertex association is not 100% efficient and because the tracker has no full $|\eta|$-coverage. Neutral particles from PU are not handled by CHS at all. To mitigate the remaining PU after applying CHS are treated by applying additional jet-area based PU offset corrections to the four-momenta of jets [75]. These are part of the jet energy calibration which is described in Section 5.2.2. Besides spurious PU particles in analyzed jets, there are also additional spurious jets reconstructed purely from PU particles. They are identified with a dedicated PU jet identification (PU jet ID) [76], which is based on a multivariate analysis using boosted decision trees (BDT) trained on track and observables related to jet-shapes, to reject jets

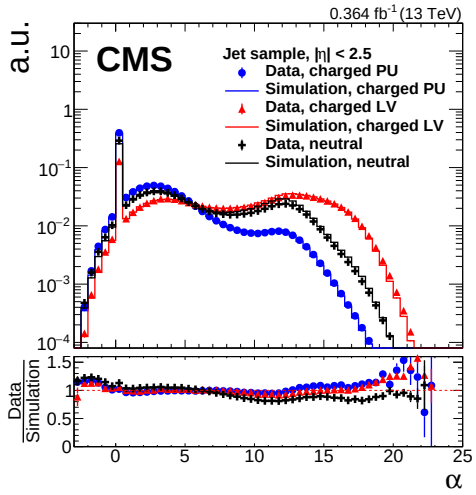


Figure 5.2: The distribution of α for jets with $|\eta| < 2.5$ in a jet sample for neutral particles in black and for charged particles from the LV and PU vertices in red and blue, respectively. The distributions are shown both for jets from simulation as lines and from data as makers. Taken from [36].

both in the central and forward region. The CHS method and PU jet ID are used complementary in tandem but are both sensitive to tracking inefficiencies.

For better resilience to tracking inefficiencies, to be able to recover jet substructure and jet-shapes closer to the truth, that are more stable with respect to an increasing amount of pileup, and to treat the neutral component in a more direct way, PUPPI was introduced. In this approach weights w_i between 0 and 1 are assigned to each PF particle, scaling their four-momenta based on the probability that they originate from a PU vertex or the PV respectively. For a charged particle i the weight is set based on the track-vertex association and the distance $|d_z|$ of the particle to the leading vertex (LV) in z according to the following cases. If the particle was used in the LV fit, the weight is set to 1. If the particle is used in any PU vertex fit the weight is set to either set to 1, if the associated PU vertex is the first or second and if $|d_z| < 0.2$ cm, or to 0 otherwise. If the particle is not associated with any vertex there are four cases that are distinguished: If the particle has $p_T > 20\,\text{GeV}$ the weight is set to 1. If the particle has $|\eta| > 2.4$ and $p_T < 20\,\text{GeV}$, the weight is set to 1 if $|d_z| < 0.3$ cm and to 0 if $|d_z| > 0.3$ cm. Lastly, if the particle has $|\eta| < 2.4$ and $p_T < 20\,\text{GeV}$ it is treated as a neutral particle in the proceeding procedure. The determination of the weights for the neutral particles is based on tracking information and starts by calculating the discriminator α_i for each particle i as:

$$\alpha_i = \log \sum_{j \neq i, \Delta R_{ij} < R_0} \left(\frac{p_{T,j}}{\Delta R_{ij}} \right)^2 \begin{cases} \text{for } |\eta_i| < 2.5, & \text{j are charged particles from LV,} \\ \text{for } |\eta_i| > 2.5, & \text{j are all kinds of reconstructed particles,} \end{cases} \tag{5.3}$$

where the sum runs over all other particles j of the respective category according to $|\eta_i|$, that are in a cone of radius $R_0 = 0.4$ around the probed particle i. The sum of the transverse momenta $p_{T,j}$ of particles j over its angular distance ΔR_{ij} to the probed particle i makes the discriminator a measure of the isolation of particle i. This offers good separation between particles from PU and the LV as demonstrated in Figure 5.2.

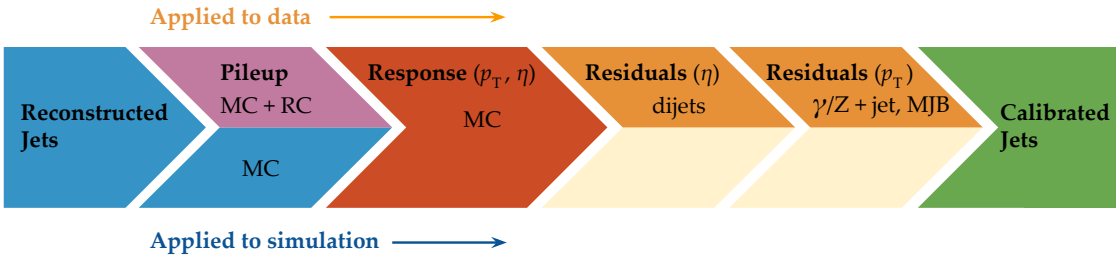


Figure 5.3: Schematic depiction of the workflow to derive and apply the jet energy corrections in the factorized approach. The top half of the arrows indicate which corrections are applied to data, while the bottom half indicates which corrections are applied to simulation. Taken from [78].

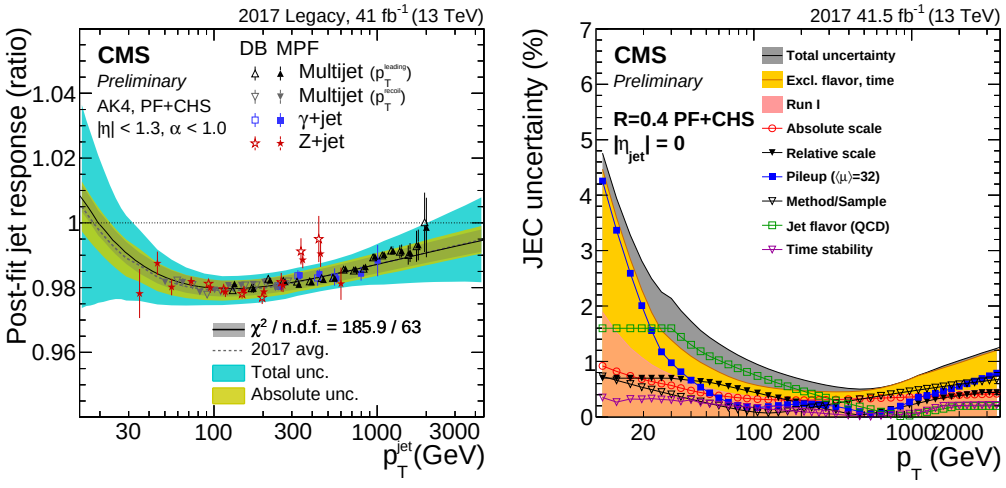
Next, a measure for the probability that the particle i is originating from pileup is estimated under the assumption, that the distribution of α_i is similar for neutral and charged particles. The probability is derived by comparing the α_i of the neutral particle i with the expected median $\bar{\alpha}_{\mathrm{PU}}$ and root-mean-square (RMS) $\alpha_{\mathrm{PU}}^{\mathrm{RMS}}$ of the α distribution of charged particles from PU vertices in the event in terms of a signed χ^2:

$$\mathrm{signed}\chi_i^2 = \frac{(\alpha_i - \bar{\alpha}_{\mathrm{PU}})|\alpha_i - \bar{\alpha}_{\mathrm{PU}}|}{(\alpha_{\mathrm{PU}}^{\mathrm{RMS}})^2}. \tag{5.4}$$

The weights for the neutral particles are then derived as $w_i = F_{\chi^2,\mathrm{ndf}=1}(\mathrm{signed}\chi_i^2)$, where $F_{\chi^2,\mathrm{ndf}=1}$ is the cumulative distribution function of a χ^2 distribution with one degree of freedom.
The weights are used to scale the four-momenta of PF candidates before jet clustering. This was optimized and studied in terms of its performance in Run 2 data [77]. CHS and PUPPI are compared in terms of the efficiency and purity of LV jets. Both PUPPI and CHS perform well in terms of efficiency, where both achieve stable efficiencies close to 100% in the tracker coverage $|\eta| < 2.4$ and efficiencies $> 95\%$ in the forward regions $|\eta| > 3.0$. The purity is similarly high and stable in the central region for PUPPI, while for CHS it degrades more than for PUPPI with increasing pileup. Outside of the tracker coverage it reaches close to 90% in $2.4 < |\eta| < 3.0$ and $\sim 70\%$ in $|\eta| > 3.0$ at low pileup for both CHS and PUPPI, but degrades fast with increasing pileup [77]. Applying different working points of the PU jet ID in tandem with CHS recovers the purity to some degree while degrading the efficiency [36]. The results of these studies with respect to jet energy resolution are discussed in section 5.2.2. In the context substructure PUPPI yields more stable results with respect to increasing pileup, which is shown in Section 5.3. This is one of the main reasons, why in the context of this thesis CHS is used for AK4 jets, while for AK8 jets PUPPI is used to mitigate pileup.

5.2.2 Jet energy corrections

After the pileup mitigation, the jet energy calibration is performed in a factorized manner [79], to achieve the best possible precision corrections for different effects. The total calibration chain is depicted schematically in Figure 5.3, showing which correction is applied to data (top row) and simulation (bottom row), where each correction is a multiplicative correction factor, that is applied to the four-momentum of the jets. As a first step, the jets are corrected for the

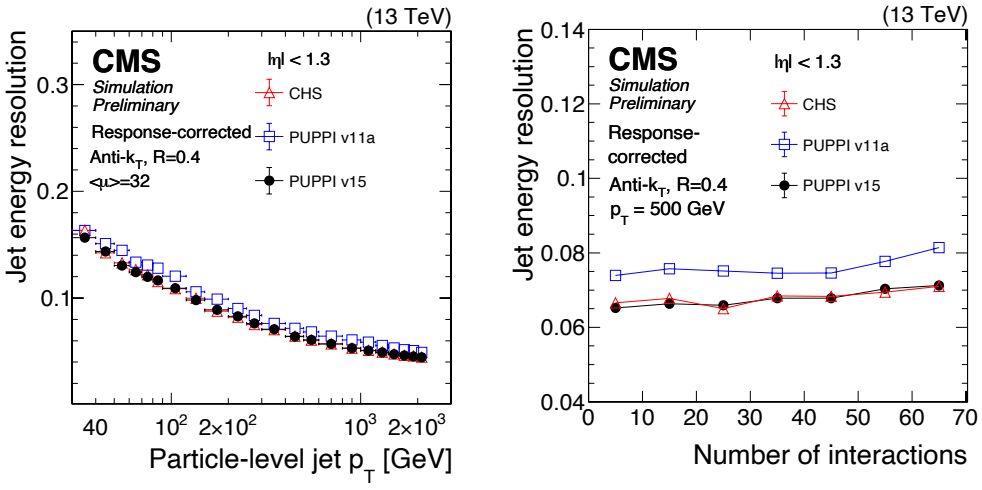


(a) (b)

Figure 5.4: The left plot shows the last level of the residual jet energy corrections (L3Res) derived from a global fit of all measurements using 2017 legacy dataset. [78]. The plot on the right shows the breakdown of the systematic uncertainties of the jet energy corrections as a function of p_T for the 2017 dataset [80].

remnant PU in the jets, which consists of the jet-area-based pileup offset corrections mentioned in Section 5.2.1. Dedicated corrections using the random cone (RC) method are derived, to account for residual difference between data and simulation. The PU corrections are applied only on CHS jets. They are not necessary for PUPPI, since the pileup offset is already low when using PUPPI. The next step is handling the bulk of the correction of the jet energy scale (JES), by correcting the response of jets to unity on average. For this, the response is defined as $R = \frac{\langle p_T^{reco} \rangle}{\langle p_T^{ptcl} \rangle}$, with the transverse momentum of the reconstruction-level jet p_T^{reco} and the particle-level jet p_T^{ptcl}. The response R is measured in bins of p_T and η in simulation by comparing detector-level jets with matched particle-level jets. The corrections are applied to the jets, that have been corrected for pileup, in data and simulation. After the simulated response corrections, the simulated jets are calibrated and the response closes with unity within 0.1% for AK4 CHS jets with $30 < p_T < 2000\,\text{GeV}$ and $|\eta| < 2.5$, and within 1% elsewhere.

The remaining residual differences between data and simulation are corrected using two consecutive corrections, which are only applied to the data. First, a η-dependent corrections factor that corrects the response of jets relative to the response of jets in the barrel region ($|\eta| < 1.3$), and then a p_T-dependent correction measured in the barrel region is applied. The η-dependent correction is measured in dijet events using the missing transverse momentum projection fraction (MPF) method [79] and is within 1% different from unity for barrel jets, within 5% for $|\eta| < 2.5$ and reaches up to 18% above unity in the transition regions between the sub-detectors. The final correction of the JES is the p_T-dependent correction, which is measured in a combination of Z/γ + jet, QCD multijet, and $W(q\bar{q})$ from $t\bar{t}$ samples using both the MPF and direct balance (DB) methods [79]. The resulting correction factor is within 2% different from unity across the measured range of p_T, while the uncertainty is < 1% in a large phase-space as shown in Figure5.4a [78]. The total uncertainty of the jet energy corrections reaches down to a sub-percent level as shown in the uncertainty breakdown in Figure 5.4b.



(a) (b)

Figure 5.5: Distribution of the jet energy resolution, defined as the Gaussian width of the response $p_{\mathrm{T}}^{\mathrm{ptcl}}/p_{\mathrm{T}}^{\mathrm{reco}}$ as a function of the transverse momentum $p_{\mathrm{T}}^{\mathrm{ptcl}}$ of the particle-level jet on the left and as a function of the number of interactions on the right. Both compare CHS (red triangles) to the newest PUPPI tune (v15 - black points) as well as to the previous PUPPI tune (v11 - blue squares). Taken from [77].

After both residual corrections are applied the JES is fully calibrated. The last step is the correction of the jet energy resolution (JER), which is typically smaller in simulation than the resolution observed in data. For this reason, JER scale factors are derived in bins of η, that are used to smear the resolution of simulated jets, in order to match the resolution of jets in data. The measurement involves parameterizing the reference resolution obtained from simulation as a function of $p_{\mathrm{T}}^{\mathrm{ptcl}}$ and mean number of interactions μ [79]. The JER scale factors are determined in dijet events and are between 1.0 and 1.2 and larger in the transition region of the ECAL endcap and the forward HCAL [78].

Figure 5.5 compares the different PU mitigation techniques CHS and PUPPI in their impact on the JER. The left plot shows the JER of barrel jets as a function of p_{T} for $\langle\mu\rangle = 32$, while the right shows the JER of barrel jets with $p_{\mathrm{T}} = 500\,\mathrm{GeV}$ as a function of the number of interactions μ. Both CHS and PUPPI perform similarly well, reaching JER $\lesssim 10\%$ for jets with $p_{\mathrm{T}}^{\mathrm{ptcl}} > 100\,\mathrm{GeV}$ at $\langle\mu\rangle = 32$, while slowly degrading with increasing pileup.

The fraction of energy carried by the individual PF candidate flavors is measured in data and simulation [80] and shown in Figure 5.6. On average the measured contribution to the total jet energy of charged hadrons is $\approx 65\%$, photons $\approx 30\%$ and neutral hadrons $\approx 5\%$ [79]. Due to different reconstruction efficiencies and detector responses for the individual flavors, this is 1–2% different from the simulated composition [80], thus measurements of observables based on jet constituents reconstructed with the particle-flow algorithm are subject to only small corrections.

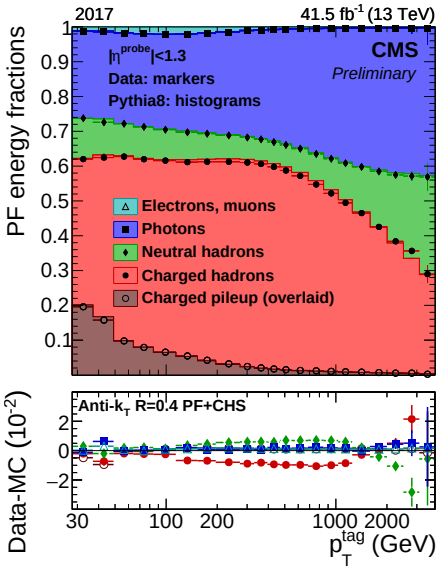


Figure 5.6: Overview of the fraction of energy of anti-k_T jets carried by each PF candidate flavor in bins of p_{T}. The considered PF candidate flavors are: charged hadrons (red), neutral hadrons (green), photons (blue), electrons & muons (cyan). Additionally, the fraction of energy from charged hadrons, that were removed from the jets using charged hadron subtraction is overlayed in brown. The composition was measured using 2017 data (markers) and simulation (histograms). Taken from [80].

5.3 *Jet substructure*

The study of hadronic final states of processes with boosted heavy particles, like W, Z, H bosons and top quarks, relies on jet substructure. It is crucial in the identification and classification, constituting an entire field of dedicated research. This section will discuss the main observables and challenges in jet substructure relevant in the context of this thesis, before giving an overview of jet substructure measurements and applications of jet substructure in measurements and searches.

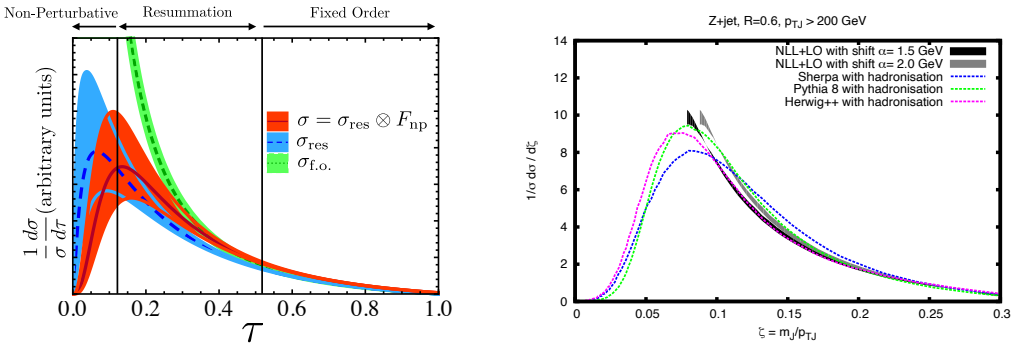
In the following general aspects of calculations with substructure observables are discussed with the example of the invariant jet mass. A more detailed discussion of these topics can be found in [7–9].

The invariant jet mass, which is an important observable at the center of many analyses and studies of jet substructure, is defined as:

$$m_{\mathrm{jet}}^2 = P_\mu P^\mu = \left(\sum_{i \in \mathrm{jet}} p_i \right)^2 \tag{5.5}$$

with P_μ the four-momentum of the jet, constructed from the summed four-vectors p_i of all jet constituents. For jets initiated by boosted heavy particles the jet mass is coming primarily from the mass going into the energy of the constituents, while jets initiated by quarks and gluons acquire mass primarily from the collinear splitting of the partons. Assuming partons itself to be massless, e.g. The jet mass of a jet originating from a massless parton a that undergoes a collinear splitting into partons b and c can be approximated as $m_{\mathrm{jet}} \approx p_{T,b} p_{T,c} R_{bc}$ [7, 81].

In the collinear or soft limit (i.e. $R_{bc} \to 0$ or either $p_{T,i} \to 0$) the fixed-oder calculations in

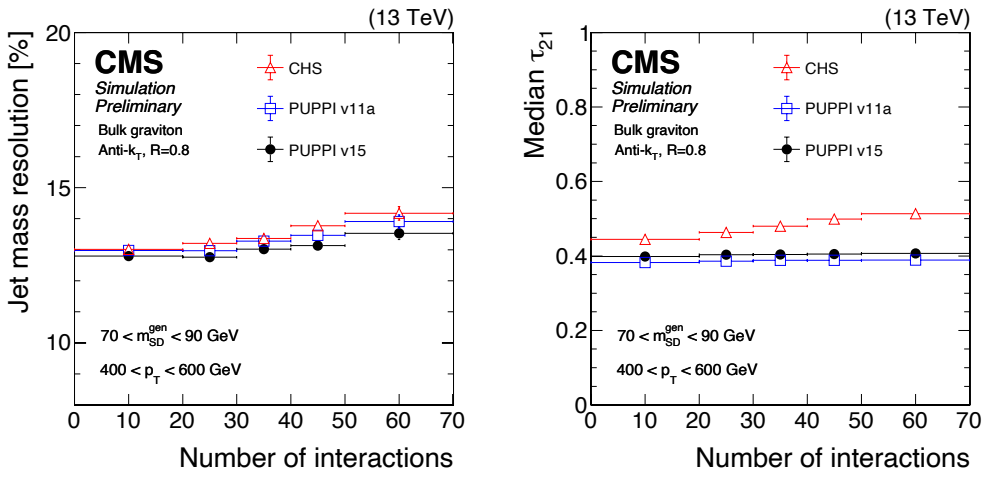


(a) (b)

Figure 5.7: Example of (semi-)analytical calculations of substructure observables. The left shows the differential cross section for a general substructure variable τ. The green line corresponds to the purely perturbative calculation, the blue line includes resummation of the leading logarithms and the red line includes convolution with a shape function accounting for non-perturbative effects [9]. The right plot shows the predicted differential cross section of the scaled jet mass $\zeta = m_J/p_{TJ}$ from different MC generators with hadronization as dashed lines as well as NLL calculations with a shift corresponding to the power order corrections accounting for the non-perturbative hadronization effects.Taken from [83].

perturbative QCD (pQCD), e.g. of the production cross section $d\sigma/dm_{\rm jet}$, do not adequately describe the problem, as large logarithmically enhanced terms of the disparate energy scales (e.g. $\log^n(p_{\rm T}/m_{\rm jet})$) are introduced to the cross section of the process by the emissions, and at each order, it diverges like $1/m_{\rm jet}$. This is overcome by *resumming* the leading logarithms over all orders in the computation of the cross section while introducing exponentially suppressing *Sudakov formfactors* [19, 82]. The level of accuracy of these calculations is given by the terms considered in the approximation, e.g. leading logarithm (LL) calculations only consider the first, and next-to-leading logarithm (NLL) also considers the second-order terms. While jet mass descriptions from the parton shower in general-purpose MC event generators such as PYTHIA [56] typically reach precisions to LL, recent semi-analytical calculations reach NLL or NNLL accuracy. The predictions from semi-analytical calculations can be matched to the prediction using pQCD from MC generators at different precision (LO, NLO) as well as predictions using non-perturbative models to derive distributions of jet substructure observables, describing the effects from the respective regions at best possible accuracy (e.g. LO+NLL, NLO+NLL+NP, etc.). Figure 5.7a shows an example of a calculation of a general substructure observable τ, which can be taken as the jet mass [9], where the green line corresponds to the prediction from fixed-order perturbation which diverges for $\tau \to 0$, while the blue line includes resummation of the logarithmically enhances terms. The latter features a peak known as the *Sudakov peak*, where in the case of the jet mass the peak position scales linearly with $p_{\rm T}$.

In addition to considering these perturbative effects, non-perturbative effects such as hadronization and the underlying event are additional limiting factors when performing calculations involving jet substructure observables. For example, corrections derived for the squared jet mass $m^2_{\rm jet}$ in approximation as power corrections for hadronization effects scale with $p_{\rm T}R$ and for the effects from the underlying event with $p_{\rm T}R^4$ [7, 81]. Figure 5.7b shows an example of this non-perturbative hadronization correction applied as a shift to the scaled jet mass $\zeta = m_{\rm jet}/p_T$,



(a) (b)

Figure 5.8: Distribution of the soft drop jet mass resolution (JMR) and N-Subjettiness ratio τ_2/τ_1 as a function of the number of interactions. Both compare CHS (red triangles) to the newest PUPPI tune (v15 - black points) as well as to the previous PUPPI tune (v11 - blue squares). Taken from [77].

compared to distributions derived from parton shower generators with hadronization models. The approximated power correction is only valid right from the Sudakov peak [7, 83]. Figure 5.7a shows also the effect of including non-perturbative effects in the calculation of the general substructure observable τ as red line. Here the perturbative calculation is convoluted with the non-perturbative shape function F_{np} [9]. Due to the sensitivity of the jet mass to the mass of the initiating particle, the jet mass is often used to discriminate signal (heavy object decay) from background (light quark and gluon jets). For the high-p_{T} jets, the Sudakov peak can be an experimental challenge, as jets initiated from quarks and gluons will have a jet mass peaking at much higher values than what might be naively expected, where the distribution would be steeply falling with the maximum close to zero. As a result, these jets can appear similar to the signal jets, therefore complicating the differentiation.

5.3.1 Soft drop grooming

To mitigate the effects coming from pileup and to limit the range where non-perturbative effects play a role, jets are often treated with grooming techniques, which aim to remove soft and wide-angle radiation from jets. The CMS experiment uses the soft drop algorithm [84] to groom jets and derive groomed substructure variables, most importantly the soft drop groomed jet mass or soft drop mass m_{SD}. The soft drop algorithm is a generalized version of the modified mass drop tagger [85]. To perform soft drop grooming on an anti-k_T jet with the radius parameter R_0, the jet constituents are reclustered using the CA algorithm, which converts the momentum-ordered structure of the anti-k_T jet constituents into an angular-ordered structure. The last clustering step is reversed resulting in the two subjets j_1 and j_2, which are tested for the soft drop condition:

$$\frac{\min(p_{T,1}, p_{T,2})}{p_{T,1} + p_{T,2}} > z_{\text{cut}} \left(\frac{\Delta R_{12}}{R_0} \right)^{\beta}. \tag{5.6}$$

Here $p_{T,i}$ are the transverse momenta of the subjets and ΔR_{12} is their angular separation. The soft drop threshold z_{cut} controls the strength of the procedure and the exponent β the angular dependence. If the condition is met by the two subjets j_1 and j_2, the jet j is taken to be the groomed jet. If the condition does not hold, the procedure is repeated with the subjet with the larger transverse momentum p_{T}. The CMS experiment adopts the soft drop algorithm with the parameters $z_{\text{cut}} = 0.1$ and $\beta = 0$, which yields the same definition as the modified mass drop tagger (mMDT) [85, 86]. In the following, the soft drop mass is denoted as m_{SD}. Figure 5.9 shows examples of the soft drop mass in different representations. Figure 5.9a shows the calculation at NNLL precision of the scaled soft drop mass $e_2^{(2)} = \frac{m_{\text{SD}}^2}{E_{\text{SD}}^2}$ [87] in green, compared to predictions from simulation using PYTHIA with hadronization effects in blue and without hadronization effects in red. The bulk of the non-perturbative contributions are limited to regions at the low end of the spectrum, the intermediate part is dominated by the resummation, while the high end of the spectrum is described by perturbation theory. Figure 5.9b shows a comparison of the jet mass before and after applying soft drop grooming (solid and dashed lines respectively) to q/g jets (red lines) and W jets (blue lines). In both cases the ungroomed mass peaks at high masses (Sudakov peak), where the q/g jets peak close to the region of the W boson mass. After the grooming is applied the jet mass peaks near the mass of the jet-initiating particles. In Figure 5.9c the reconstructed soft drop mass of AK8 jets from CMS simulation is shown for different heavy particles in the final state (W, Z, H boson) in colors and for light quark and gluon jets from QCD multijet backgrounds in black. This plot shows the characteristic peaks of the heavy particles, which are often used to discriminate signal from background processes. Figure 5.8a shows the resolution of the soft drop jet mass of boosted bosons with $400 < p_{\text{T}} < 600\,\text{GeV}$ for CHS (red triangles) and PUPPI (black points) as a function of the number of interactions in an event. While both pileup mitigation techniques achieve similar resolutions of 13 – 14%, PUPPI is more stable and reaches smaller resolutions even in high pileup scenarios [77].
In the context of this thesis dedicated simulation to data scale factors were derived for the calibration of the soft drop mass, which is described in Section 7. Further, the distribution of the soft drop mass is measured in a two-dimensional unfolding described in Section 8.

5.3.2 Idenfication of jet flavor and origin

Besides the soft drop mass, there are various other substructure variables or discriminators using jet substructure, which are used to identify jet signatures (jet tagging). In this section, the substructure variables and tagging approaches are described, which are used in the context of this thesis to discriminate jets originating from W bosons and top quarks jets from QCD multijet and other background processes. All substructure variables are in some way correlated to the jet mass, which is mostly undesirable for the purpose of jet tagging. In Section 5.3.2.4 techniques to decorrelate the substructure variables from the jet mass are described.

5.3.2.1 N-Subjettiness

As a measure of the probability of a jet consisting of N or fewer subjets, the N-Subjettiness [90] is widely used in the tagging of N-prong final states. The N-Subjettiness τ_N is defined as:

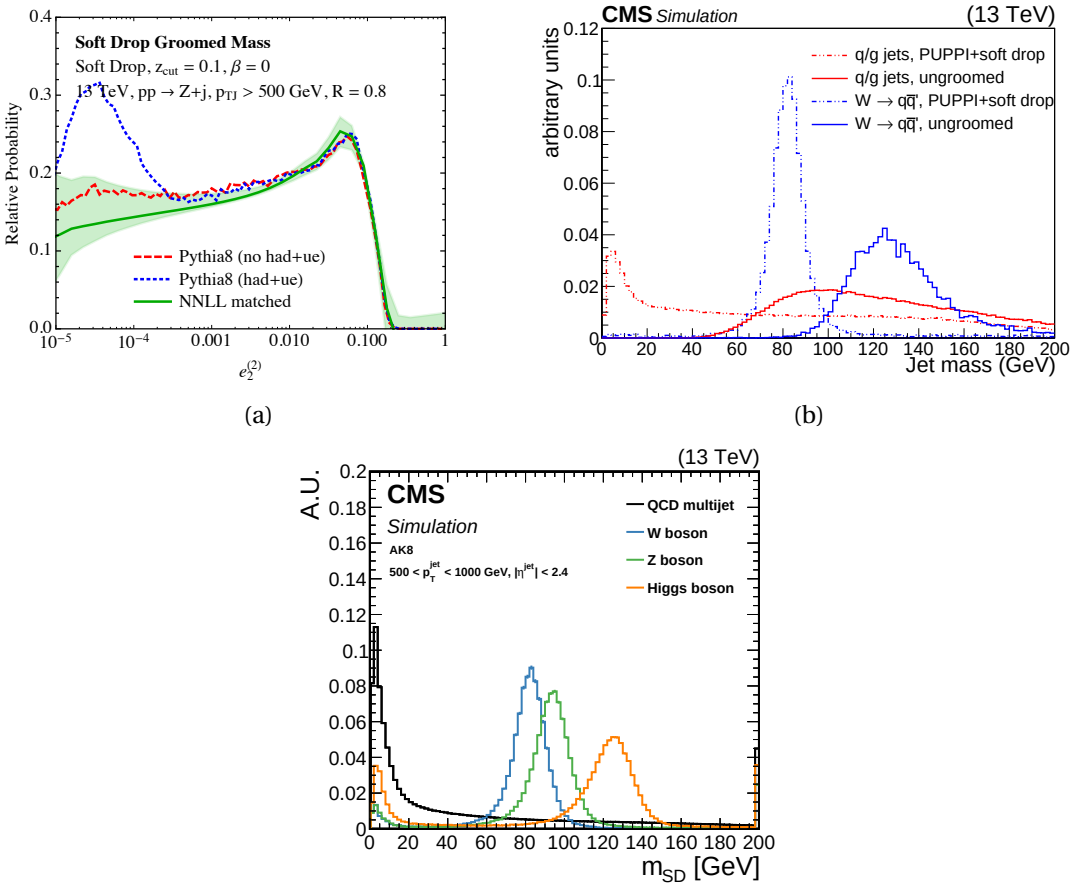


Figure 5.9: The top left plot shows the relative probability of finding a jet with a scaled jet mass $e_2^{(2)} = m_{SD}^2/E_{SD}^2$, with a calculation at NNLL precision matched to NLO pQCD predictions in green compared to predictions derived using PYTHIA with and without hadronization effects in blue and red, respectively [87]. The top right plot shows the reconstructed jet mass with and without soft drop grooming for q/g jets in red and for W jets in blue. Taken from [88]. The bottom plot shows the reconstructed soft drop mass m_{SD} of AK8 jets with $500 < p_T < 1000\,\text{GeV}$ and $|\eta| < 2.4$ from simulations of different processes. While the black line shows the prediction for QCD multijet processes, the blue, green and orange lines show the prediction for processes involving W, Z and H bosons. Taken from [89].

$$\tau_N = \frac{1}{\sum_k p_{T,k} R_0} \sum_k p_{T,k} \min\{\Delta R_{1,k}, \Delta R_{2,k}, ..., \Delta R_{N,k}\}, \tag{5.7}$$

where k runs over all jet constituents, $p_{T,k}$ is the transverse momentum of the constituent k, $\Delta R_{i,k}$ is the angular separation in (η, ϕ) between the constituent k and a candidate subjet axis i and R_0 is the radius parameter of the original jet. The candidate axes are derived by re-clustering the jet with the exclusive k_t-algorithm with the *winner-takes-all* (WTA) recombination scheme [91], which defines the four-vector of pairwise recombination to be massless, i.e. $p = (E_r, \hat{n}_r E_r)$, with the recombined energy and $E_r = E_1 + E_2$. The recombined momentum is pointing in the direction of the harder particle:

$$\hat{n} = \begin{cases} \frac{\vec{p}_1}{|\vec{p}_1|} & E_1 > E_2, \\ \frac{\vec{p}_2}{|\vec{p}_2|} & E_1 < E_2. \end{cases} \tag{5.8}$$

The N jets clustered from this are then taken to be the candidate axes. For $\tau_N \approx 0$ most constituents are localized around the N subjet axes, while for $\tau_N \gg 0$ they are more uniformly distributed with respect to the N candidate subjet axes. The discrimination power between N prong and $N-1$ prong structures is increased when looking at ratios of N-Subjettiness variables τ_N / τ_{N-1}. Figure 5.8b shows the median N-Subjettiness ratio τ_2/τ_1 for AK8 jets from boosted bosons with $400 < p_T < 600\,\text{GeV}$ for CHS (red triangles) and PUPPI (black points) as a function of the number of interactions in an event, which shows that the N-Subjettiness variable is more stable against increasing pileup when using PUPPI [77].
In the context of this thesis, the ratio τ_2/τ_1 is used to identify W jets, while τ_3/τ_2 is used to discriminate top jets from W jets and background processes. Figure 5.10b shows the resulting soft drop mass distribution when tagging top jets with $\tau_3/\tau_2 < 0.5$ in a sample of jets with $p_T > 200\,\text{GeV}$ of semileptonic decaying $t\bar{t}$ events as solid black line and when tagging W jets from the same sample with $\tau_2/\tau_1 < 0.45 \wedge \tau_3/\tau_2 > 0.5$. The tagged top jets have a considerable contribution from jets where only the decay products of the W boson are merged into the AK8 jet. The tagged W jets (black and green dashed line) show the peak at the mass of the W boson, similar to when no tagging requirement is used (red dotted line).

5.3.2.2 Energy correlation functions

The energy correlation functions (ECF) [92, 93] are a generalization of the N-Subjettiness variables, which do not depend on the identification of candidate subjet axes. There are different series of ECF variables, which are designed to be sensitive to jet substructure in different limits. They are based on the generalized energy correlation functions, which are defined as:

$${}_{\nu}e_i^{(\beta)} = \sum_{1 \le i_1 \le i_2 < ... < i_n \le n_J} \left(\prod_{a=1}^{n} z_{i_a} \right) \prod_{m=1}^{\nu} \min_{s<t \in i_1, i_2, ..., i_n}^{(m)} \left(\Delta R_{st} \right)^{\beta}, \tag{5.9}$$

with the energy fractions $z_i = \frac{p_{T,i}}{p_{T,J}}$, n_J being the number of constituents in the jet, n denoting the number of particles to correlate, ν the number of angular pairwise angles entering the product and the angular exponent β. The N_i series of ECF variables are designed to mimic N-Subjettiness

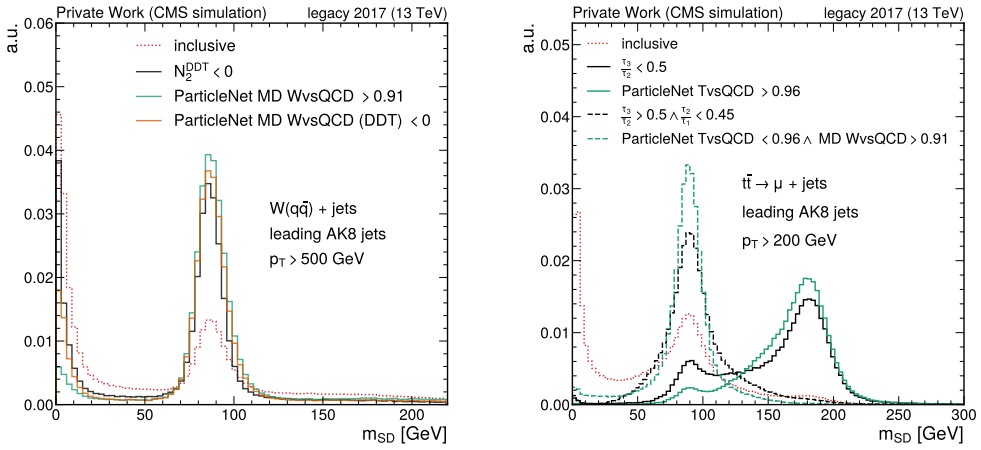
in the limit of resolved substructure. They are defined by:

$$N_i^{(\beta)} = \frac{{}_2e_{i+1}^{(\beta)}}{\left({}_1e_i^{(\beta)}\right)^2} \tag{5.10}$$

In the context of this thesis, the variable $N_2^{\beta=1}$ is used for the identification of W jets. Using the definition of generalized ECFs in Equation 5.9 with $\beta = 1$, the variable $N_2^{\beta=1}$ is defined using the following two-point and three-point correlation functions:

$${}_2e_3^{\beta=1} = \sum_{1\leq i\leq j\leq k\leq n_J} z_i z_j z_k \min\left\{\Delta R_{ij}\Delta R_{ik}, \Delta R_{ij}\Delta R_{jk}, \Delta R_{ik}\Delta R_{jk}\right\} \tag{5.11}$$

$${}_1e_2^{\beta=1} = \sum_{1\leq i\leq j\leq n_J} z_i z_j \Delta R_{ij}. \tag{5.12}$$



(a) (b)

Figure 5.10: Comparison of the soft drop mass, when using the different W and top jet tagger, that are used in the context of this thesis. The left plot shows the tagger used to select W jets in fully hadronic $W(q\bar{q})$+jets events, while the right plot shows the tagger used to select top and W jets in the semileptonic $t\bar{t}$ sample. The black lines show the distributions when taggers based on the substructure variables N-Subjettiness and energy correlation function ratio $N_2^{\beta=1}$ are used. The green and orange lines show the distributions when ParticleNet discriminators with and without further treatment using DDT are used. On the right plot, the dashed lines show the distributions of tagged W jets and the solid lines show tagged top jets from semileptonic decaying $t\bar{t}$ events. The red dotted line in both plots shows the inclusive distribution of m_{SD} without requiring any tagging criteria to be met.

5.3.2.3 *Machine learning techniques*

In modern particle physics multivariate machine learning (ML) methods are widely used. In the context of jet substructure, they are especially useful to combine the large amount of information contained in the jet constituents into high-level discriminative variables, improving the background rejection with limited signal efficiency degradation [89]. In the context of this

thesis, the ML-based heavy object tagger ParticleNet [94] is used in comparison with the non-ML substructure variable tagger approaches described above.

ParticleNet is a graph neural network (GNN) algorithm based on Dynamic Graph Convolutional Neural Network (DGCNN) [95], designed to discriminate hadronic decays of highly-boosted W, Z and H bosons and top quarks from QCD multijet background. While many other ML approaches treat the jet as some ordered structure [89], the ParticleNet algorithm treats jets as particle-clouds, i.e. an unordered, permutation invariant set of particles. ParticleNet treats these point-clouds of input data with edge convolutions (*EdgeConv* [95]), which are graph convolutional operations that conserve the number of points by creating a new graph in a higher dimensional latent space, where the vertices are the points itself and the edges are constructed for the k nearest neighbors. This makes EdgeConv operations stackable and useable in the framework of a DGCNN. ParticleNet implements consecutive EdgeConv operation blocks, where each one compares $k = 16$ nearest neighbors and constructs the edges and aggregates them using the mean. To find the nearest neighbors, the first EdgeConv block uses the spatial distance in (η, ϕ), while the other blocks use the distance of the feature vectors in the point-cloud in latent space. Following a global average pooling is a fully connected layer with ReLu activation, and to prevent overfitting a dropout layer with a probability of 10%. A second fully connected layer with a softmax activation function is used in the binary classification task.

Finally, the GNN yields probability scores (in the following called P^{GNN}) for each class of process used in the training. Each class corresponds to a decay mode of the considered particle-category (e.g. $t \to bc$ or $t \to bq\bar{q}$, where $q \in [u, d, s]$). ParticleNet is trained with particle-clouds consisting of the jet constituents (PF particles) of a jet and point-clouds consisting of secondary vertices associated with the jet. The latter is included to improve the performance of the tagging of heavy-flavor jets, such as jets from b quarks. These include B-hadrons, which give rise to secondary vertices because of their long lifetime. Each point or particle in the clouds carries a feature vector, which includes for the particles: its kinematic properties (p_{T}, η, ϕ, energy), its charge and differences between the kinematic properties of the particles and of the jet. The properties used for the secondary vertices include additionally the displacement and quality criteria. In the context of this thesis the TvsQCD tagger is used to identify top jets in comparison with the N-Subjettiness $\frac{\tau_3}{\tau_2}$. TvsQCD is with the respective GNN probability scores for the top decay modes and the QCD multijet background modes:

$$\text{TvsQCD} = \frac{P^{\text{GNN}}_{t\to bc\bar{q}} + P^{\text{GNN}}_{t\to bq\bar{q}}}{P^{\text{GNN}}_{t\to bc\bar{q}} + P^{\text{GNN}}_{t\to bq\bar{q}} + \sum P^{\text{GNN}}_{\text{QCD}}}, \tag{5.13}$$

where the sum over the QCD multijet background modes includes all decay modes considered by ParticleNet[2]. The ParticleNet discriminators are highly correlated with the jet substructure observables, especially the correlation with the jet mass is typically introduces problems in many methods applied in analyses. In the context of this thesis for example the data-driven technique to estimate the QCD multijet background described in Section 6.4 relies on a mass-decorrelated

[2] with $\sum P^{\text{GNN}}_{\text{QCD}}$ and $\sum P^{\text{GNN-MD}}_{\text{QCD}}$ as the sum of the respective scores over all possible QCD classes: QCD $\to b\bar{b}$, QCD $\to c\bar{c}$, QCD $\to b$, QCD $\to c$, QCD $\to$ others.

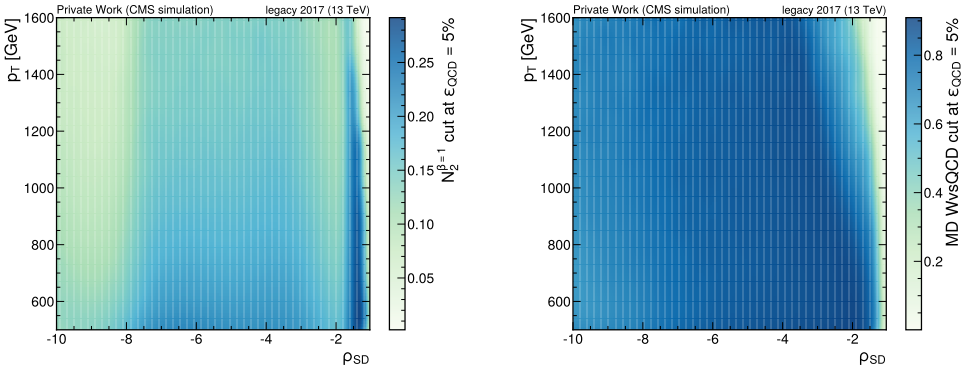
variable for the construction of control and signal region. To construct a mass-decorrelated version of ParticleNet discriminants, the GNN architecture is trained using a signal sample, where a generic spin-0 particle X with a flat mass spectrum decays into two highly Lorentz-boosted particles. Additionally, the jets used in the training are reweighted both in the signal and in the QCD background sample to have a flat jet p_{T} and m_{SD} spectrum [96]. The resulting mass-decorrelated MDWvsQCD discriminator is used in the context of this thesis to identify W jets in comparison with the ECF $N_2^{\beta=1,\mathrm{DDT}}$ and N-Subjettiness $\frac{\tau_2}{\tau_1}$. The MDWvsQCD discriminator is defined using the probability scores of the mass-decorrelated version of the GNN corresponding to the:

$$\mathrm{MDWvsQCD} = \frac{P_{X\to c\bar{c}}^{\mathrm{GNN-MD}} + P_{X\to q\bar{q}}^{\mathrm{GNN-MD}}}{P_{X\to c\bar{c}}^{\mathrm{GNN-MD}} + P_{X\to q\bar{q}}^{\mathrm{GNN-MD}} + \sum P_{\mathrm{QCD}}^{\mathrm{GNN-MD}}} \tag{5.14}$$

In Figure 5.10 the green lines show the resulting soft drop mass distribution when using the MDWvsQCD and TvsQCD discriminators to tag W and top jets. Figure 5.10a shows that ParticleNet improves the signal purity especially, at low masses, where less q/g jets are wrongly identified as W jet. Figure 5.10b shows the distribution when using the mass-decorrelated W tagger MDWvsQCD as green dashed line, and the distribution when using the top tagger TvsQCD as solid green line. The ParticleNet taggers show again improved signal purity. Especially the contribution of the W jets is smaller in the top tagged sample of events when using ParticleNet, rather than N-Subjettiness variables.
Additionally, another deep-learning approach is used in the context of this thesis to identify heavy-flavor jets. The DeepJet [97] algorithm is used to identify b-jets originating from the decay of t quarks. DeepJet is a multiclass deep-learning model used for jet flavor classification. It uses sequences of convolutional layers to process the approximately 650 input features and classifies jets into b, bb, leptonic b, c or uds and g jets. The input features are lists of properties of jet constituents (charged and neutral hadrons) as well as properties of secondary vertices associated with the jet. To identify AK4 b-jets from semileptonic $t\bar{t}$ decays in the analysis presented in this thesis (see Section 7), a medium working point is used, with a misidentification rate of 1% and a signal efficiency in $t\bar{t}$ events of $\approx 70\%$ [98].

5.3.2.4 Mass decorrelated taggers

Data in ranges of soft drop mass are used to predict the QCD multijet background in regions where boosted objects peak, under the assumption of a smooth soft drop mass distribution without any peak-like structure, which could be induced by the selection of substructure observables with the jet mass. Thus, the mass-decorrelation of the ECF variable $N_2^{\beta=1}$ is vital for the application of the taggers in the context of the data-driven estimation of the QCD multijet background. Additionally, the mass-decorrelated version of the ParticleNet tagger does introduce residual mass sculpting, which is still significant in the context of this thesis. Thus mass-decorrelated versions of the $N_2^{\beta=1}$ variable and the ParticleNet tagger MDWvsQCD are constructed and used in the analyses presented in this thesis, as specified in Section 6. Here an adaption of the Design Decorrelated Tagger (DDT) [99] approach, which was introduced in [2] was used. This modified version of DDT brute-force decorrelates the tagging variable by



(a) (b)

Figure 5.11: Two-dimensional distribution as a function of jet p_T and ρ_{SD} of the 5-th QCD multijet selection efficiency percentile of the tagger $N_2^{\beta=1}$ on the left and ParticleNet MDWvsQCD on the right.

enforcing a constant selection efficiency for QCD jets across the probed phase space. The DDT approach introduces the variable ρ as:

$$\rho = \log\left(\frac{m_{\text{jet}}^2}{p_T^2}\right). \tag{5.15}$$

which makes sure the phase space is scaled appropriately for QCD jets, while in the context of this thesis mostly the soft drop mass instead of the plain jet mass is used, which defines $\rho_{SD} = 2\log\left(\frac{m_{SD}}{p_T}\right)$. Here, the approach pioneered by [2] is adopted, thus the decorrelation is flattening the (ρ_{SD}, p_T)-dependence of the tagging variable. For this a constant selection efficiency X% for QCD jets across (ρ_{SD}, p_T) is enforced when cutting on $N_2^{\beta=1}$. The steps necessary to achieve this are described in the following. First, a three-dimensional histogram is constructed with axes in $(\rho_{SD}, p_T, N_2^{\beta=1})$ and filled with the respective properties of the leading AK8 jet from simulated QCD multijet events (see Section 6.2 for details on sample generation). The histogram is then projected onto the (ρ_{SD}, p_T)-plane by calculating the Xth-percentile of the $N_2^{\beta=1}$ distribution in each (ρ_{SD}, p_T)-bin. The resulting two-dimensional map is shown in Figure 5.11 for the $N_2^{\beta=1}$ tagger on the left and the ParticleNet tagger MDWvsQCD on the right. The map filled with $N_2^{\beta=1}$ percentiles shows a significant dependence of $N_2^{\beta=1}$ on ρ_{SD} in the QCD multijet sample, while the right shows that the mass-decorrelation methodology in ParticleNet is yielding a flat dependence in most of the phase space, however at high ρ_{SD} ParticleNet is not well decorrelated. Important to note here is, that while $N_2^{\beta=1}$ peaks at low values for signal (e.g. W jets) and high values for background (e.g. QCD jets), the ParticleNet tagger peaks at high values for signal and low values for background, thus the z-axis of the map has to be interpreted in the opposite way. The maps are then used to define the new variables $N_2^{\beta=1,\text{DDT}}$ and $\text{MDWvsQCD}^{\text{DDT}}$ as:

$$N_2^{\beta=1,\text{DDT}} = N_2^{\beta=1} - P^X_{N_2^{\beta=1}}(\rho_{SD}, p_T) \tag{5.16}$$

$$\text{MDWvsQCD}^{\text{DDT}} = P^X_{\text{MDWvsQCD}}(\rho_{SD}, p_T) - \text{MDWvsQCD}, \tag{5.17}$$

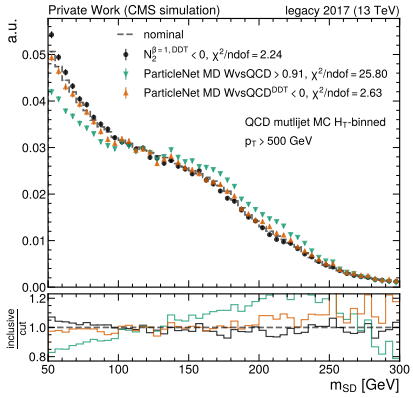


Figure 5.12: Distribution of the soft drop mass of the leading AK8 jet in QCD multijet events with $p_T > 500\,\mathrm{GeV}$. The inclusive distribution is shown as the dark grey dashed line, while the markers show the distribution after a cut on the different W taggers. The black points correspond to $N_2^{\beta=1,\mathrm{DDT}} < 0$, the green triangles correspond to MDWvsQCD > 0.91 and the orange triangles correspond to MDWvsQCD$^{\mathrm{DDT}} < 0$.

where P_i^X denotes the X-th QCD selection efficiency percentile of the tagger i, which is taken from the bin corresponding to $(\rho_{\mathrm{SD}}, p_T)$ in the respective map. The tagger are then used by requiring the new variables to be negative, i.e. $N_2^{\beta=1,\mathrm{DDT}} < 0$ and MDWvsQCD$^{\mathrm{DDT}} < 0$. To check for closure of this method and estimate the performance compared to the plain ParticleNet mass decorrelation Figure 5.12 shows the soft drop mass distribution of the leading AK8 jet with $p_T > 500\,\mathrm{GeV}$ in the QCD multijet simulated events. The black dashed line shows the distribution without any cut on any tagger. The black circles show the resulting distribution when requiring $N_2^{\beta=1,\mathrm{DDT}} < 0$, the green triangles show the distribution after cutting on the ParticleNet discriminant without further brute-force decorrelation (MDWvsQCD > 0.91) and the orange triangles show the distribution of the events passing MDWvsQCD$^{\mathrm{DDT}} < 0$. Comparing the green and orange triangles the benefit of the additional decorrelation of the ParticleNet tagger becomes apparent.
Both the decorrelated variables MDWvsQCD$^{\mathrm{DDT}}$ as well as $N_2^{\beta=1,\mathrm{DDT}}$, introduce minimal sculpting in the soft drop jet mass, quantified by the χ^2/ndof of the comparison of the inclusive distribution with the distributions after the cut, which is around 2.63 and 2.24 for the variables respectively. The χ^2 was derived by comparing the weighted histograms[3]. The maps and closure tests for the remaining data-taking years and also in bins of p_T can be found in appendix C.1.

[3] When comparing two weighted histograms 1 and 2, the χ^2 can be estimated with $\chi^2 = \sum_i^n \frac{\left((\sum_i^n w_{1i}) w_{2i} - (\sum_i^n w_{2i}) w_{1i}\right)^2}{(\sum_i^n w_{1i})^2 s_{2i}^2 + (\sum_i^n w_{2i})^2 s_{1i}^2}$, with the sum of weights $w_{1,2i}$ in each bin i and the estimator for the variances $s_{1,2i}^2$ of bin i. The resulting distribution follows approximately a χ^2_{n-1} distribution. [100]

5.4 *Jet substructure in measurements and searches*

Jet substructure is at the center of many LHC data analyses, either it is the subject of a measurement itself, or it is used as a tool in measurements of SM properties or searches for new physics. There are already numerous measurements of jet substructure performed by the four large LHC experiments ALICE and LHCb, ATLAS and CMS, including measurements of jet shapes and soft drop observables [101–103], fragmentation properties [104–109] and the Lund plane [110, 111]. In the following, recent examples related to the measurements of this thesis are presented, to introduce the current status of research and point out open questions.

Since jet mass plays a vital role in many analyses involving jets, it is important to understand the jet mass and its uncertainties as well as possible. For light quark and gluon jets the jet mass was measured for groomed and ungroomed jets both by CMS and ATLAS shown in Figure 5.13. Figure 5.13a shows the measurement of the differential cross section of inclusive jet production,

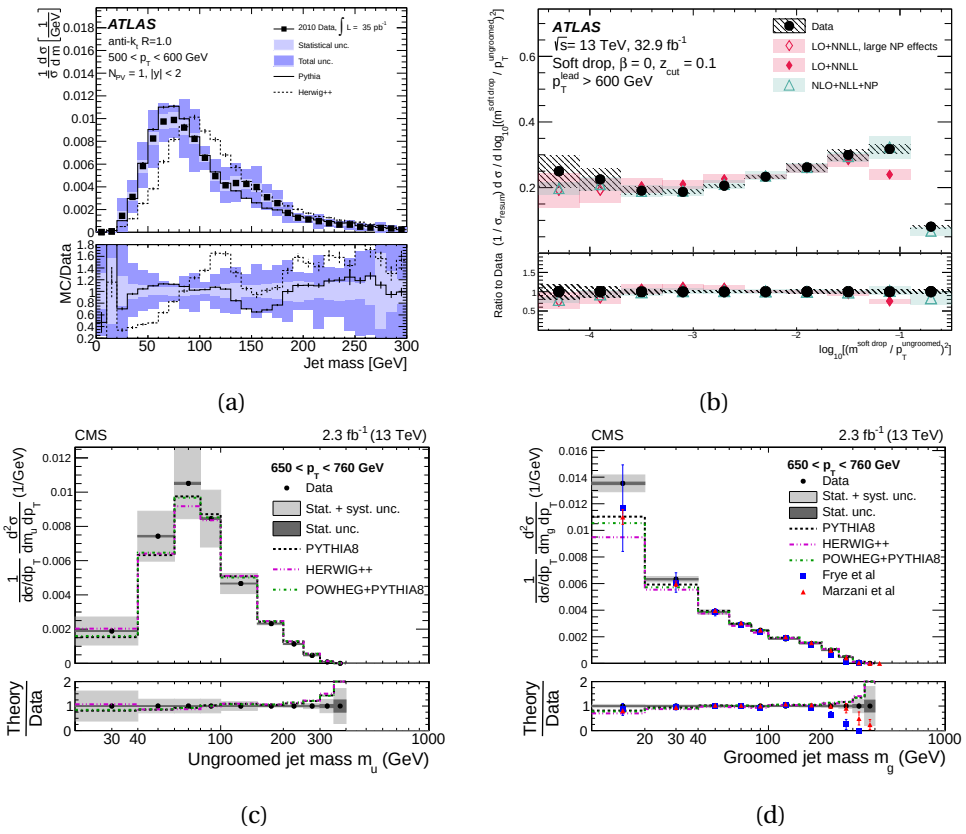
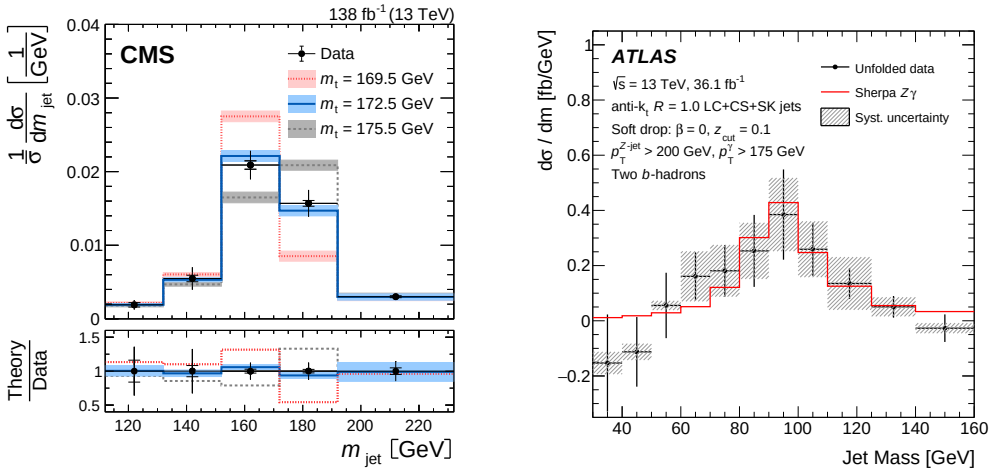


Figure 5.13: Measurement of the normalized differential cross section of light jet production of the plain ungroomed jet mass (left column) and the groom soft drop mass (right column) from different studies performed in bins of p_T. Black markers show the unfolded data, which are compared to different predictions from simulation or (semi-) analytical calculations at different levels of precision. The top left shows the differential inclusive jet cross section measured in $\sqrt{s} = 7\,\text{TeV}$ data by ATLAS [11] compared to simulations using different MC generators. The top right shows the measurement normalized differential dijet cross section of the scaled groomed soft drop mass $\rho = \log(m_{SD}^2/p_T^2)$ in $\sqrt{s} = 13\,\text{TeV}$ data by ATLAS [12]. The bottom row shows the measurement of the normalized differential cross section of the ungroomed and groomed jet mass in dijet events, performed at $\sqrt{13}\,\text{TeV}$ by CMS [13].



(a) (b)

Figure 5.14: Unfolded distribution of the soft drop jet mass measured for jets initiated by hadronically decaying top quarks from the measurement of the differential $t\bar{t}$ cross section and subsequent m_t measurement performed by CMS [14] on the left and for jets initiated by $Z \to b\bar{b}$ from a measurement by ATLAS [15] on the right.

which was measured in dependence of the ungroomed jet mass and bins of jet p_T in $\sqrt{s} = 7\,\text{TeV}$ data by ATLAS [11]. The measurement is compared to the prediction of the LO MC generators PYTHIA and HERWIG++ and demonstrates that with the Sudakov peak the jet mass of light jets peaks at masses around 50–100 GeV for the bin $500 < p_T < 600\,\text{GeV}$ shown in the figure. In this measurement, which requires events to have a single PV (thus reducing pileup effects to an absolute minimum), the unfolded data agree with the simulation from both generators within uncertainties. However, it typically lies between the prediction from PYTHIA and HERWIG++. This systematic difference is often a common source of uncertainty in substructure measurements and makes it important to compare unfolded data with multiple MG generators or theoretical models. However, this effect can be reduced when using grooming techniques [7].

The measurement of the groomed jet mass at $\sqrt{s} = 13\,\text{TeV}$ in dijet events was performed by both CMS [13] shown in Figure 5.13d in terms of the normalized differential dijet cross section in dependence of the groomed jet mass and by ATLAS [12] shown in Figure5.13b in terms of the normalized differential dijet cross section in dependence of the scaled soft drop jet mass $\rho = \log(m_{SD}^2/p_T^2)$. Both Figures show the result using soft drop grooming with $z_{cut} = 0.1$ and $\beta = 0$. The measurement by CMS was also performed for the ungroomed jet mass, shown in Figure 5.13c. The latter shows good agreement with the simulation using PYTHIA, HERWIG++ and POWHEG +PYTHIA, while the uncertainty is systematically dominated. The unfolded data of the groomed jet mass are compared to semi-analytical calculations including resummation at NLL or NNLL accuracy. Both results show good agreement in the intermediate scaled mass range $0.03 < m_{SD}/p_T < 0.25$ where the resummation is expected to be most accurate. At high m_{SD}, the perturbative calculations become more important, thus in Figure 5.13b the data agree best with the NLO+NLL+NP prediction. At low masses in the non-perturbative regime, the uncertainties become larger and the data are not as well described, even when including the non-perturbative corrections in the theory calculations in both measurements. These

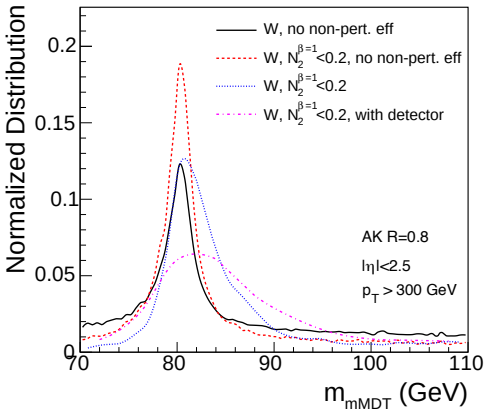
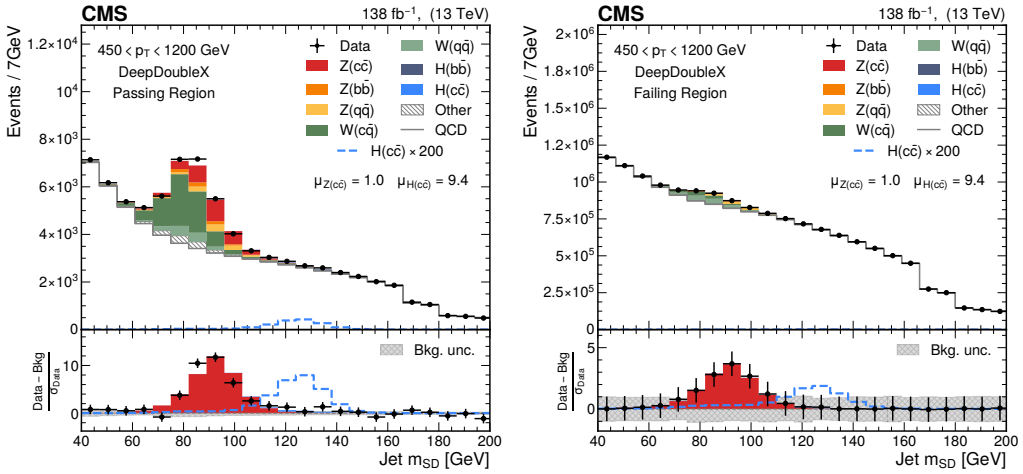


Figure 5.15: Jet mass distribution of AK8 W jets with $p_T > 300\,\text{GeV}$ and $|\eta| < 2.5$ on different level and in different scenarios. The black line shows the particle-level distribution without including non-perturbative effects during the simulation. The red line shows the same but with an additional selection using $N_2^{\beta=1} < 0.2$. The blue line shows the distribution on particle-level with non-perturbative effects included in the simulation and the selection using $N_2^{\beta=1} < 0.2$. Finally, the purple line shows the same as the blue line, but on the detector-level, so with consequent simulation of the detector response. Taken from [113].

measurements are a helpful testbed for the theory calculations including resummation at NLL, NNLL accuracy as well as different levels of non-perturbative corrections. For a deeper understanding of how well the MC generators model the jet substructure, CMS measured the generalized jet angularities $\lambda_\beta^\kappa = \sum_{i\in\text{jet}} z_i^\kappa \left(\frac{\Delta R_i}{R}\right)^\beta$ in 2016 data corresponding to a integrated luminosity of $\mathscr{L}_{\text{int}} \approx 35.9\,\text{fb}^{-1}$ [112]. Here z_i is the transverse momentum fraction and ΔR_i is the displacement of the i-th jet constituent with respect to the jet. The measurement is performed separately in a quark-enriched Z+jets sample and a gluon-enriched dijet sample with and without soft drop grooming. The resulting unfolded data distributions are again located in between predictions of MADGRAPH +PYTHIA and HERWIG++. Further, the difference in prediction for quark and gluons is within large uncertainties, while no tested generator is able to describe this perfectly. Both could change with new and improved tuning of the generators.

In other studies, the jet mass can be used as a direct proxy to measure the properties of the initiating particle. The measurement of the differential $t\bar{t}$ cross section for example shows great sensitivity to the top quark mass as shown in Figure 5.14a. Different mass hypotheses are compared to the unfolded data, in order to measure the top quark mass to $m_t = 173.06 \pm 0.24(\text{stat.}) \pm 0.61(\text{exp.}) \pm 0.47(\text{model}) \pm 0.23(\text{theo.})\,\text{GeV}$ [14]. The leading uncertainties are the ones connected to the experimental sources (JER, JES, etc.) followed by uncertainties related to the signal modeling (Choice of m_t, color reconnection scheme). The largest theoretical uncertainties are connected to non-perturbative effects (UE, hadronization) where the largest ones contribute of the order of 10% to the total uncertainty. A study by ATLAS measured the unfolded groomed soft drop mass of jets from $Z \to b\bar{b}$ decays in $Z\gamma$ events at $\sqrt{s} = 13\,\text{TeV}$ [15], as shown in Figure 5.14b. This aims to help better understand important systematic uncertainties and identification techniques for measurements involving decays to heavy-flavor quark pairs. The unfolded data are in good agreement with the LO prediction from SHERPA within uncertainty. The uncertainty

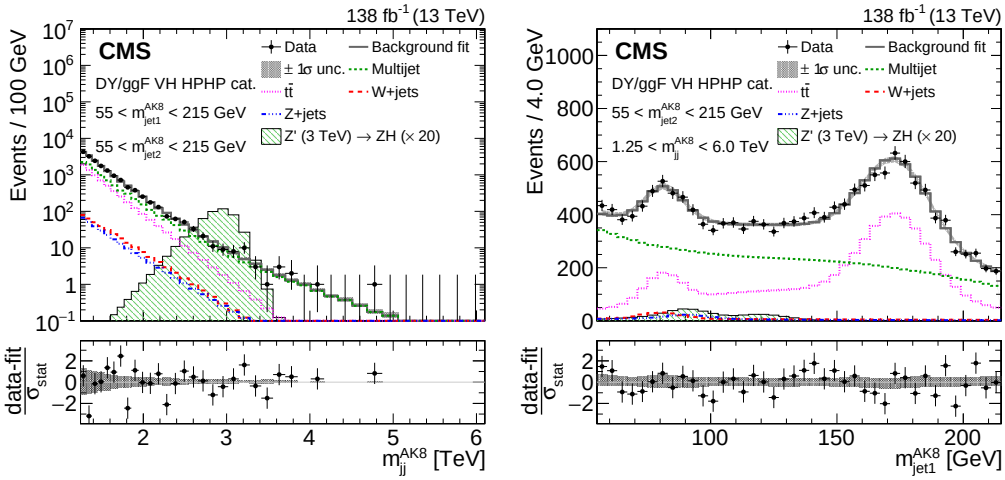
on the integrated fiducial cross section is about 37 – 46% depending on whether soft drop grooming or trimming was used to treat the jet and reduce the impact of non-perturbative effects. Among the largest uncertainties in this measurement are the experimental uncertainty on the jet energy and jet mass calibration (JES 7.2.1 – 7.4% and JMR 5.1 – 6.0%), uncertainties related to b-jet tagging and subsequently the background estimate (4.0 – 7.5%) and the uncertainty coming from the signal modeling and overall non-closure of the unfolding(5.8–15%).
So far, W jets are often used as standard candles in measurements involving jet substructure, e.g. for the calibration of the jet mass scale (JMS), where the determination of the JMS in separate fits to simulation and data is limited primarily in statistics and reaches resolutions of below 10% [114]. To complement and ultimately complete the testbed of the Standard Model using jet substructure, the measurement of the jet mass distribution of W jets from hadronically decaying $W(q\bar{q})$+jets events is going to be essential. The study of these offers a theoretically simpler configuration, as the W boson form jets without color reconnection to the rest of the event, unlike W bosons from $t\bar{t}$ decays. Additionally, a measurement of the mass of the W



(a) (b)

Figure 5.16: Distribution of the soft drop jet mass for AK8 jets with $p_T > 450\,\text{GeV}$ after a fit to Run 2 data recorded by the CMS experiment ($\mathscr{L}_{\text{int}} \approx 138\,\text{fb}^{-1}$) [115]. The left plot shows the signal region, where events are required to pass the full list of selection criteria imposed in the analysis, which includes requirements on the flavor-tagger DEEPDOUBLEX. The left plot shows the same, but for the control region, with events failing said selection criteria. The QCD multijet background, shown as a white histogram, is estimated from data. Taken from [115]

boson m_W in this final state could be of great interest, with the recent measurements of the CFD collaboration [5] putting stress on the SM. The global average of measurements excluding the refined CDF measurement put the W mass to $m_W = 80377 \pm 12\,\text{MeV}$ [22], while the CDF measurement yielded $m_W = 80433.5 \pm 9.4\,\text{MeV}$. The discrepancy between these most precise estimates remains subject to further investigation. So a possible complementary approach, which would also benefit the understanding of jet substructure, is to measure m_W in the all-jets final state similar to the measurement of the top quark mass [14]. Here the main challenges are the massive QCD multijet background and the dominant experimental uncertainties coming from the jet energy scale and resolution and the non-perturbative contributions to the invariant jet mass. The prospects of feasibility to reach an adequate level of uncertainties in such a



(a) (b)

Figure 5.17: Distribution of the invariant mass of the dijet system on the left and the soft drop mass of the leading AK8 jet on the right in Drell-Yan and gluon-gluon fusion events in the VH final state. The distributions are shown after a fit to full Run 2 data recorded by CMS ($\mathscr{L}_{\text{int}} \approx 138\,\text{fb}^{-1}$). Taken from [116]

measurement were studied with respect to the expected integrated luminosity of the high-luminosity pp program at the LHC (HL-LHC) [113]. With an expected integrated luminosity of $\mathscr{L}_{\text{int}} \approx 3000\,\text{fb}^{-1}$ the statistical uncertainty on a m_W measurement in the all-jets final state could reach 30 MeV when exploiting the Z boson as a standard candle and performing the measurement in an indirect measurement by measuring the mass difference $m_W - m_Z$. By measuring the difference, important common systematic uncertainties related to the calibration of the jet energy and jet mass scale cancel out since the relative jet mass scale is measured. Additionally, for this estimate the study assumes that with advances in the high-level trigger system, the statistical uncertainty can be reduced, as similar efficiency for online trigger jets as for offline jets can be achieved, allowing for a lowering of the p_{T} threshold. The final statistical uncertainty estimated on the mass difference measurement with the HL-LHC dataset is 32 MeV. The main limiting factor in measuring the W mass m_W or the mass difference $m_W - m_Z$ is the current precision achieved for predictions of the non-perturbative contributions to the invariant jet mass. With the current modeling of effects due to hadronization and the underlying event, the systematic uncertainties arising from these effects are estimated to be of the level of 1100 MeV in the direct m_W measurement and 100 MeV in the measurement of mass difference $m_W - m_Z$. The magnitude of the effect was estimated by comparing MC prediction with and without the modeling of the effects of hadronization and underlying event, with a subsequent selection using the ECF ratio $N_2^{\beta=1} < 0.2$. Figure 5.15 shows the resulting jet mass distribution for AK8 W jets with $p_{\text{T}} > 300\,\text{GeV}$ and $|\eta| < 2.5$, that are treated with the mMDT technique with $z_{\text{cut}} = 0.1$ and $\beta = 0$ (equivalent to soft drop) in different scenarios. The jet mass distribution on particle-level is shown in the scenarios with and without including non-perturbative effects in the simulation and after the $N_2^{\beta=1}$ cut as blue and red lines respectively.

Besides probing the jet substructure itself to gain a deeper understanding of the Standard model, jet substructure is a very powerful tool, that is used in many analyses probing the properties

of the Standard Model and beyond. A recent example is the search for $H \to c\bar{c}$ and consequent observation of $Z \to c\bar{c}$ using $\sqrt{s} = 13\,\text{TeV}$ data from Run 2 ($\mathscr{L}_{\text{int}} \approx 138\,\text{fb}^{-1}$ at $\sqrt{s} = 13\,\text{TeV}$) from CMS [115]. Here events with jets with a two-prong structure are selected using the $N_2^{\beta=1}$ variable, which is decorrelated following the same prescription as described in Section 5.3.2.4 with a fixed QCD multijet selection efficiency of 26%. The events are further split into signal and control regions (shown in Figure 5.16) using the dedicated DEEPDOUBLEX DNN algorithm [117], which is loosely inspired by the DeepJet model and enriches the signal region in $b\bar{b}$ and $c\bar{c}$ decay modes. With a simultaneous fit of signal and control regions with an advanced data-driven QCD multijet background estimation the study is able to present the first observation of $Z \to c\bar{c}$ in association with jets at a hadron collider and places an upper expected and observed limit on the Higgs boson production cross section times the branching fraction to $c\bar{c}$ of 39 and 47 times the SM expectation at 95% confidence level. The background estimation method used to estimate the massive QCD multijet background shown in white in Figure 5.16 is also adopted in the context of this thesis and described in detail in Section 6.4 The uncertainty in this measurement is statistically dominated, while the largest systematic effects are connected to the background estimation.

Another recent example of a direct search for physics beyond the standard model using jet substructure is the search for new heavy resonances, which decay to heavy boson pairs (WW, WZ, ZZ, WH or ZH) into the all-jets final state using full Run 2 data recorded by CMS at $\sqrt{s} = 13\,\text{TeV}$. The search employs ML-based jet tagging algorithms (DEEPAK8) and many other techniques discussed in this section to identify AK8 fat jets to reconstruct the heavy boson final state and place limits on the production cross section of multiple heavy resonances, which are predicted by different standard model extending theories. The key feature of this analysis is the simultaneous fit to data in three dimensions: the soft drop jet mass of the two AK8 jets, as well as the invariant mass of the consequent dijet system. Figure 5.17 shows the invariant mass of the dijet system on the left and the soft drop mass of the leading AK8 jet on the right. The analysis covers multiple production channels (ggF, VBF and DY) and splits the probed phase space into categories of different purity to maximize the sensitivity of the search. The dominant systematic uncertainties for the signal processes are connected to the reconstruction and identification of the heavy H, W and Z bosons.

In Summary, the modeling of jet substructure is crucial to search for physics beyond the SM and to probe rare SM decays like $H \to c\bar{c}$ at the LHC in boosted topologies. For an SM precision measurement like the measurement of the mass of the top quark or W boson and with an outlook to the HL-LHC systematic uncertainties need to be reduced by much. While multiple jet substructure measurements exist, including measurements of the mass of quark, gluon, Z and top jets, that help to improve the modeling of jet substructure, a straightforward standard candle, the mass of W jets has so far not been measured at the LHC.

6 Analysis Strategy

This section describes the strategy of the two analyses presented in this thesis. The main goal is to study the substructure of jets initiated by hadronically decaying W bosons and top jets. In the first analysis (Section 7) a correction factor for the jet mass scale of W and top jets is measured and the second analysis (Section 8) is the measurement of the jet mass distribution of W jets. In Section 6.1 the selection and categorization of the semileptonic $t\bar{t}$ events and fully-hadronic $W(q\bar{q})$+jets events is described. In Section 6.2 an overview of the simulated event samples is given, followed by a description of the methodology of the statistical data analysis. The data-driven estimation of the dominant QCD multijet background in the fully-hadronic $W(q\bar{q})$+jets is described in Section 6.4. Finally, in Section 6.5 the list of considered systematic uncertainties is outlined.

6.1 Event Selection and Categorization

The events referred to in this section are reconstructed with the PF algorithm described in Section 4.2. After a full reconstruction promptly after data-taking a series of re-reconstructions were performed to include improvements, e.g. calibrations of sub-detectors or new techniques. A final legacy re-reconstruction in 2020, which improved for example the tracker-alignment [118] and included the new PUPPI [77]. In the context of this thesis, detailed validation studies were performed to test the impact of the new reconstruction on jet substructure.

The event selection is performed in multiple stages: First, a suitable collection of HLT-Trigger paths is chosen. While we explicitly require them to be fired in simulation as part of the preselection, the data are already recorded with collections of triggers. The choice of trigger collection defines what is called a primary dataset (PD). For the semileptonic $t\bar{t}$ sample, the *SingleMuon* PD is used, and further in the offline selection a trigger is required to have fired, where the online muon had a transverse momentum $p_{\mathrm{T}} > 50\,\mathrm{GeV}$ (HLT_Mu50). For 2016 data, a logical OR of the two triggers HLT_Mu50 and HLT_TkMu50 is used for the event selection.

For the fully-hadronic $W(q\bar{q})$+jets sample the only objects that are of interest are jets. To select jets in the boosted regime, i.e. jets with high transverse momentum, the *JetHT* PD is used, and in the offline selection specifically one of two triggers with a single large-radius ParticleFlow jet with a minimum p_{T} are required to have fired online. In events, where the candidate jet has a $p_{T,AK8} < 650\,\mathrm{GeV}$ the trigger HLT_AK8PFJet450 is used, while for events where $p_{T,AK8} \geq 650\,\mathrm{GeV}$ the trigger HLT_AK8PFJet500 is used. While HLT_AK8PFJet450 is prescaled starting in 2017, HLT_AK8PFJet500 is not prescaled during the whole data-taking. However, in early 2016 data, both were introduced after the start of data-taking and thus only partially available. To account for differences in trigger efficiency in data and the modeled efficiency in the simulation, data

to simulation scale factors for both triggers are measured in each era separately, to make sure the efficiency in simulation matches with that in data. The trigger efficiency is measured in an orthogonal sample of *SingleMuon* events, selected using the trigger HLT_IsoMu27. Additionally, there has to be at least one muon in the event, with a $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$ and passing the requirements of the *tight muon ID* (Section 4.2.1.1). Furthermore, the candidate jet has to pass the tight working point of the jet ID criteria as described in [75]. and should not overlap with the muon. The efficiency is then measured as the ratio of events that pass the probed trigger, and the total number of events in their orthogonal sample. For example for the trigger HLT_AK8PFJet450 the efficiency is measured as:

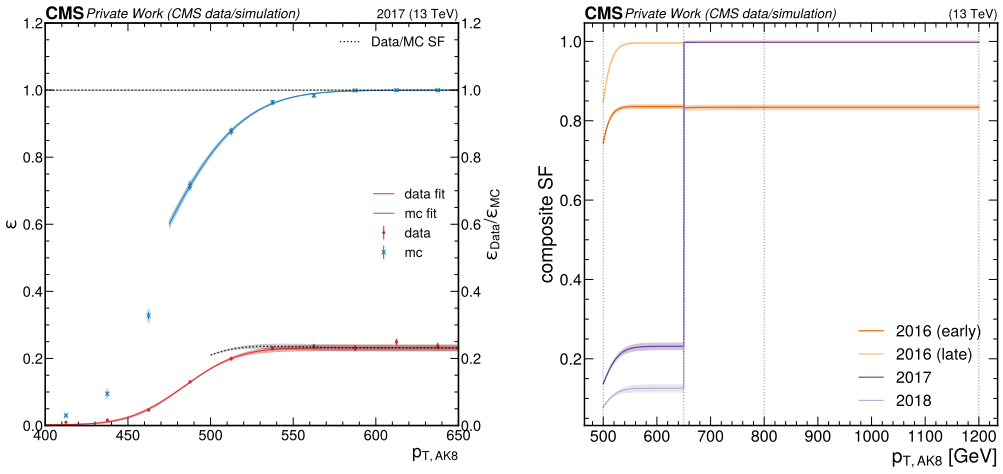
$$\epsilon = \frac{N_{\text{pass AK8PFJet450 \&\& pass reference selection}}}{N_{\text{pass reference selection}}}. \tag{6.1}$$

The measurement is performed in bins of offline jet p_T and then fitted with an error function of the form:

$$\epsilon(p_T) = a\left[b + \frac{1-b}{2}\cdot\left(1 - \text{erf}\left(\frac{p_T - c}{d}\right)\right)\right]. \tag{6.2}$$

The efficiencies for both triggers are measured in data and simulation as a function of p_T of the offline reconstructed jet. Figure 6.1a demonstrates the efficiency measurement for the lower p_T threshold trigger in 2017. In the plot the measured efficiency (points) and the fit of the error function (line) both in data and simulation in red and blue, respectively, are shown in the range $400 < p_T \leq 650\,\text{GeV}$. The plots corresponding to all measurements can be found in Appendix D for each era separately. The scale factors are then calculated as the ratio of the efficiencies in data and simulation. The scale factor is shown as a black dotted line in Figure 6.1a. The composite scale factors, which combine the individual scale factors for different triggers, are depicted in Figure 6.1b for all years. The scale factors cover the range of $500 < p_T < 1200\,\text{GeV}$ and reach different plateaus in the different p_T regions depending on the trigger used in that region. For $p_T < 650\,\text{GeV}$ the scale factors are taken from the measurement of the trigger HLT_AK8PFJet450, which is prescaled in 2017 and 2018 and not present during the whole year of 2016. Consequently, the scale factors are reaching different plateaus, which correspond to the effective luminosity, where the trigger was 100% efficient. For 2017 and 2018, the plateau of the scale factor is at 0.23 and 0.13 respectively. For 2016 the trigger was not present during the whole year and thus the plateau is at 0.83 for early 2016, but close to unity for late 2016. For $p_T > 650\,\text{GeV}$, the scale factors are taken from the measurement of the trigger HLT_AK8PFJet500 which is not prescaled during the whole data-taking, but was also not present during the whole year of 2016. Consequently, the scale factors are reaching a plateau at 1.0, except in early 2016, where it again reaches 0.83. The scale factors are applied as event weights in the final selection to the simulation.

The events selected by the trigger will then be used in the full offline selection. As aforementioned this is split into the pre-selection and the final selection, for solely technical reasons. The pre-selection is performed on a customized `ROOT`-tree format that is based on the CMS internal `MiniAOD` format. The individual steps of the pre-selection that succeed the requirement of the trigger to have fired are described in the following for the semileptonic $t\bar{t}$ and the fully-hadronic $W(q\bar{q})$+jets sample separately. Additionally to the selection of events, the candidate AK8 jets are identified among these steps as the AK8 jet leading in p_T.



(a) (b)

Figure 6.1: The measurement of trigger efficiency ϵ of the HLT_AK8PFJet450 trigger in 2017 data and simulation is shown in the left plot. Additionally, the plot shows the resulting trigger scale factor taken as the ratio of efficiency in data over simulation. The right plot shows the composite trigger scale factor for the two triggers used in the fully-hadronic $W(q\bar{q})$+jets sample. The p_T bins used in the analysis are shown as vertical dotted lines. The scale factor is calculated for the lower threshold jet trigger in the first bin, whereas for the remaining bins, the higher threshold jet trigger is utilized.

The pre-selection consists of the following consecutive requirements for the semileptonic $t\bar{t}$ sample:

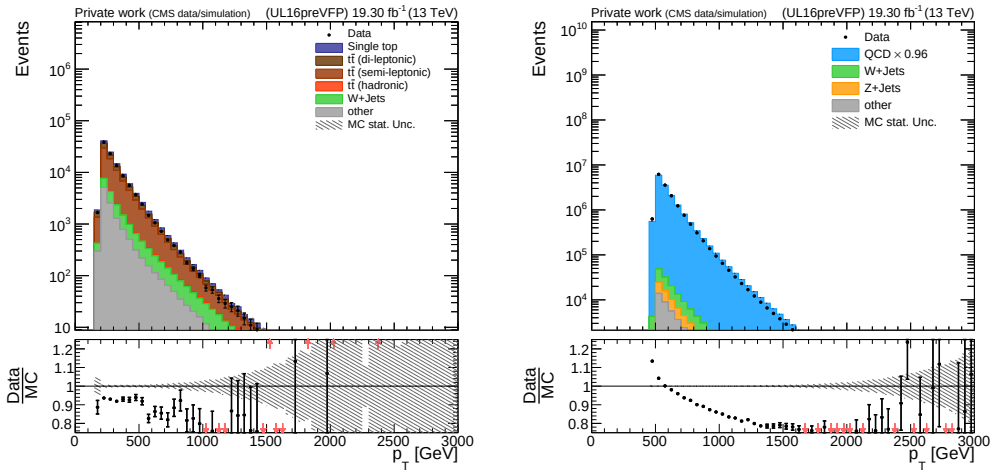
- exactly one muon and no electron ($N_\mu = 1$ and $N_e = 0$) with $p_T > 55\,\text{GeV}$ and $|\eta| < 2.4$ and passing the *tight muon ID* or *medium electron ID* criteria, respectively (Sections 4.2.1.1, 4.2.1.2),
- at least one AK8 jet with $p_T > 200\,\text{GeV}$ and $|\eta| < 2.4$,
- missing transverse momentum $p_T^{\text{miss}} > 50\,\text{GeV}$,
- at least one b-tagged (medium working point) AK4 jet with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$,
- 2D-cut on $\Delta R(\mu, \text{jets}) > 0.4$ or $p_{T,\text{rel}}(\mu, \text{jets}) > 25\,\text{GeV}$.

For the fully-hadronic $W(q\bar{q})$+jets sample the requirements for the pre-selection are as follows:

- lepton veto ($N_\mu = 0$ and $N_e = 0$) with $p_T > 10\,\text{GeV}$ and $|\eta| < 2.4$ and passing *loose muon ID* or *loose electron ID* criteria, respectively (see Section 4.2.1.1 and Section 4.2.1.2),
- at least one AK8 jet with $p_T > 500\,\text{GeV}$ and $|\eta| < 2.4$,
- additionally the candidate AK8 jet is required to have $\rho_{\text{SD}} = 2\log\frac{m_{\text{SD}}}{p_T} < -2.1$ to reject events with high soft drop mass but low transverse momentum, which are not well modeled by simulation.

In 2018 two sectors of the HCAL endcap in the region $-1.57 < \phi < -0.87$ and $-3.2 < \eta < -1.3$ were not functional for the last ≈ 65% of the data-taking (HEM15/16 issue). In data recorded during this time, events are vetoed if the leading jet is reconstructed in this region. Events in simulation are reweighted with the weight $(1 - 0.65)$ if the leading jet is reconstructed in this region to reflect the resulting effective luminosity.

After the pre-selection, the data to MC agreement is within 10 %. This is demonstrated in Figure 6.2, which shows the transverse momentum p_T of the leading AK8 jet in events that pass the pre-selection for both the semileptonic $t\bar{t}$ sample on the left and the fully-hadronic $W(q\bar{q})$+jets sample on the left. More control plots can be found in Appendix B. The data to MC disagreement in the $W(q\bar{q})$+jets sample stems primarily from the massive contribution of QCD multijet events, which are very difficult to model properly in simulation. Due to this, the QCD multijet simulation is not used directly in the analysis, but rather a data-driven method is used to estimate the contribution of QCD multijet events in the signal region. This method is described in Section 6.4. The disagreement in the $t\bar{t}$ sample is mainly because the simulation does not model the $t\bar{t}$ p_T spectrum correctly, which is accounted for in the analysis by applying a scale factor depending on the p_T of the generator top-quark as described in Section 6.2.



(a) (b)

Figure 6.2: Control plots of the transverse momentum p_T of the leading AK8 jet from events in the semileptonic $t\bar{t}$ sample and the fully-hadronic $W(q\bar{q})$+jets sample on the left and right side respectively. Both plots show early 2016 data and simulation corresponding to an integrated luminosity $\mathcal{L}_{\text{int}} \approx 19.3\,\text{fb}^{-1}$

The events selected by the pre-selection criteria based on the properties of the jets selected are then further split into categories that enter the simultaneous fit. First, the events are binned in the p_T of the candidate jet. The events are divided into 4 p_T-bins ranging from 200 GeV to 650 GeV in the semileptonic $t\bar{t}$ sample, whereas for the fully-hadronic $W(q\bar{q})$+jets events 3 p_T-bins between 500 GeV and 1.2 TeV are used. Table 1 summarizes the p_T-edges of the used bins for each sample.

jet p_T-edges	[200,300)	[300,400)	[400,500)	[500,650)	[650,800)	[800,1200)
semileptonic $t\bar{t}$	✓	✓	✓	✓		
fully-hadronic $W(q\bar{q})$+jets				✓	✓	✓

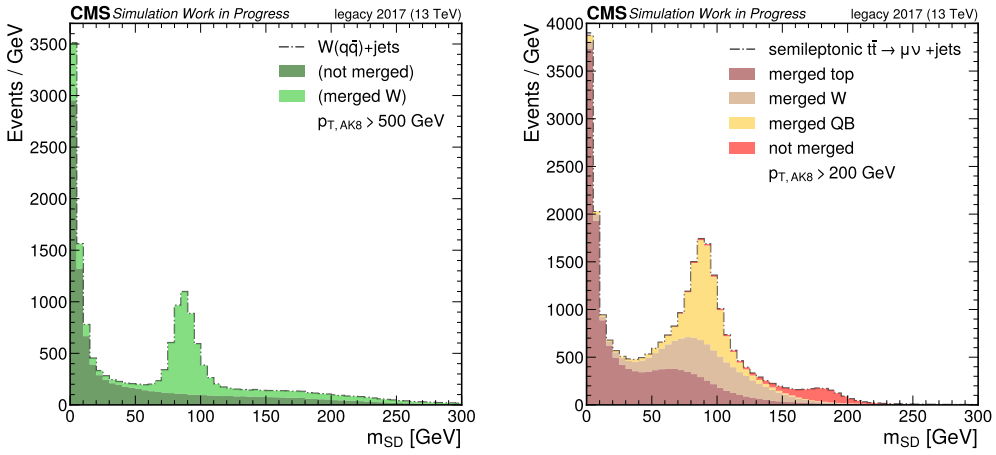
Table 1: Summary of jet p_T-bins used for the semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets samples used in the jet mass calibration. Checkmarks mark the bins a sample is split into.

6.1.1 *Merging categories*

The two signal processes semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets are categorized as described in the following. This categorization is done by comparing the decay products of the jet-initiating particle on generator-level with the reconstructed AK8 jet. For the full-hadronic $W(q\bar{q})$+jets, events are categorized as *merged W* and *not merged*. The W boson decays into two quarks q_1, q_2, if both of these decay products of the generator W boson are matched to the AK8 jet with $\Delta R(\text{AK8}, q_i) < 0.8$ the event is categorized as *merged W*, otherwise it is labeled as *not merged*. In the semileptonic $t\bar{t}$ decays, the hadronically decaying t quark of interest decays into a b quark and a W boson. The W boson decays further into two quarks q_1, q_2. The semileptonic $t\bar{t}$ events are categorized into four categories based on which subset of decay products are ΔR matched to the AK8 jet:

- *merged top*: the b quark and both decay products (q_i) of the W boson,
- *merged W*: both decay products (q_i) of the W boson,
- *merged QB*: the b and one of the decay products (q_i) of the W boson,
- *not merged*: none of the above is fulfilled.

The application of merging categorization of the simulated signal events after the pre-selection is demonstrated in Figure 6.3 for both the fully-hadronic $W(q\bar{q})$+jets and the semileptonic



(a) (b)

Figure 6.3: Demonstration of merging categories in fully-hadronically decaying $W(q\bar{q})$+jets events (left) and semileptonically decaying $t\bar{t}$ events simulated with 2017 detector conditions. The dashed line corresponds to the total inclusive sample, i.e. without merging categorization.

$t\bar{t}$ sample. For the $W(q\bar{q})$+jets sample, Figure 6.3a shows the merged W component, which peaks around the W mass, in light green and the not merged component in darker green. The latter is a steeply falling distribution with the majority of events at very low values for m_{SD}. The merged component has a second peak at low masses, which likely consists of events, where the candidate jet (leading jet in p_{T}) is not originating from the W boson but rather from the recoiling quark/gluon jet.
For the semileptonic $t\bar{t}$, sample Figure 6.3b shows the merged top and merged W components in red and yellow. They both peak around the mass of the respective parent particle, while the merged QB and not merged component, shown in light brown and burgundy, peak at very low masses and values around $50-80\,\mathrm{GeV}$ indicating the remnant radiation of the W boson and top quark into these jets.

6.1.2 W and t tagging categories

The events in each p_{T} bin are further categorized into signal and control regions. For this, either a set of jet substructure variables or state-of-the-art machine learning jet flavor/content discriminators (ParticleNet) are employed as jet tagging variables, both described in Section 5.3.2.
For the semileptonic $t\bar{t}$ sample, there are two signal regions and one control region. The two signal regions are designed to be enriched with jets, that contain all decay products of either a hadronically decaying top-quark (*pass* region) or a W boson originating from the top-decay (*pass-W* region). Consequently, the control (*fail*) region should ideally be depleted of these types of signal and contain jets, which are neither initiated by the top quark nor contain all decay products of the top or W.
The regions are defined by requirements on either the ratios of N-Subjettiness variables $\frac{\tau_2}{\tau_1}$ and $\frac{\tau_3}{\tau_2}$ or the ParticleNet discriminators TvsQCD and the mass-decorrelated version of WvsQCD. The tagger regions are summarized in Table 2 showing the requirements and Table 3 shows the corresponding efficiencies for selecting signal (W and top) and background jets. The efficiencies are calculated using events from simulated signal samples, which pass the pre-selection. The efficiency to select top and W jets is the fraction of events in the merged top and merged W category that are selected in the respective region. The background efficiency is the fraction of background events that are selected in the respective region. The efficiencies are averaged over the years of data-taking.
With a top jet selection efficiency of 52% and 79% when using substructure variables or ParticleNet discriminators, the pass top region is enriched in top jets, compared to a background efficiency of less than 5% in both approaches. The same holds for W jets in the pass-W region with efficiencies of 80% and 37%. The fail region is enriched with background with respective efficiencies of 30% and 78% when using substructure variables or ParticleNet discriminators. This is also reflected in the soft drop mass distribution of the jets in the individual regions shown in Figure 6.4 for the substructure tagging approach with 2017 data and simulation. From left to right the plots show the pass-top, pass-W and fail region. The distribution in the pass-top region is peaking around the masses of the top quark and the W boson. The peak at the mass of the W boson is less pronounced and originates from the lower end of the probed p_{T} range, where

the top jet is not fully merged yet. This yields subdominant contributions not only from the merged W component but also from the merged QB and not merged component. In the pass-W region the distribution peaks at the mass of the W boson with considerable contributions from the merged QB and not merged component. The fail region consists mostly of the not merged component of the $t\bar{t}$ events, peaking at low masses. The individual regions offer overall pure samples of the respective signal and background processes.
The region definition with ParticleNet discriminators is less efficient in selecting W jets in the pass-W region but still yields a more pure sample of top jets in the pass-top region and a more pure sample of W jets in the pass-W region than when using substructure variables. The data to MC agreement has improved with respect to the one shown in Figure 6.2a. This is due to the fact, that the scale factors for the top-p_{T} reweighting have been applied in the histograms shown in Figure 6.4. The remaining differences visible in the pass-top and pass-W regions mainly stem from the efficiency of the jet tagging. This will be accounted for by using dedicated scale factors for the jet tagging efficiency in the final fit as described in Section 6.5.

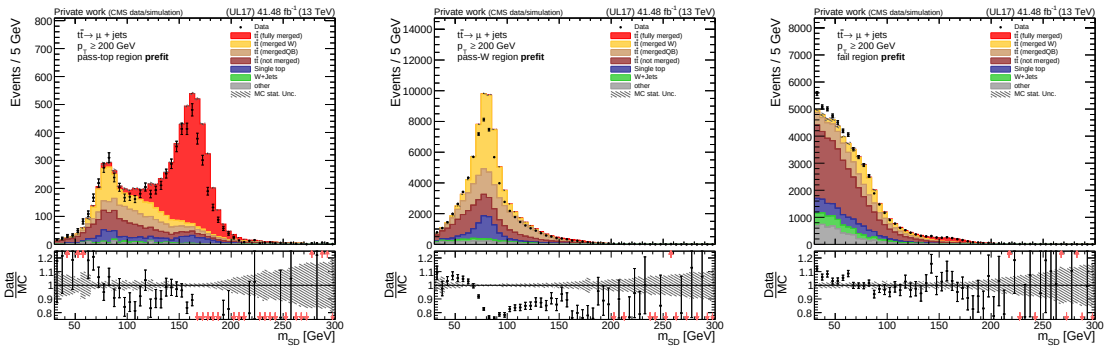
	Substructure	ParticleNet
pass-top	$\frac{\tau_3}{\tau_2} < 0.5$	TvsQCD > 0.96
pass-W	$\frac{\tau_3}{\tau_2} > 0.5 \wedge \frac{\tau_2}{\tau_1} < 0.45$	TvsQCD ≤ 0.96 ∧ MDWvsQCD > 0.91
fail	$\frac{\tau_3}{\tau_2} > 0.5 \wedge \frac{\tau_2}{\tau_1} > 0.45$	TvsQCD ≤ 0.96 ∧ MDWvsQCD ≤ 0.91

Table 2: Summary of signal and control regions (passing and failing tagger criteria respectively) for approaches using substructure variables or ParticleNet discriminators for the semileptonic $t\bar{t}$ selection.

	Substructure			ParticleNet		
	$\varepsilon_{W\text{ jets}}$	$\varepsilon_{\text{topjets}}$	$\varepsilon_{\text{backg.}}$	$\varepsilon_{W\text{ jets}}$	$\varepsilon_{\text{topjets}}$	$\varepsilon_{\text{backg.}}$
pass-top	3.63%	52.22%	4.73%	1.09%	78.75%	4.95%
pass-W	79.94%	23.12%	65.13%	37.40%	2.56%	16.33%
fail	16.43%	24.66%	30.13%	61.51%	18.69%	78.72%

Table 3: Summary of signal and background efficiencies in the different signal and control regions for the substructure and ParticleNet-based taggers in the semileptonic $t\bar{t}$ selection. The efficiencies are estimated from simulation and averaged over the years of data-taking.

For the fully-hadronic $W(q\bar{q})$+jets sample, the signal region (*pass*) - enriched in W jets - and the control region (*fail*) are constructed using three similar tagging approaches. The first one is based on the energy correlation function $N_2^{\beta=1}$, which is decorrelated with respect to the jet soft drop mass m_{SD} and jet transverse momentum p_{T} using the designed decorrelate tagger (DDT) technique as described in Section 5.3.2.4. The second and third are based on the mass decorrelated version of the WvsQCD ParticleNet discriminator. In the third approach (in the following called ParticleNet$^{\mathrm{DDT}}$), the ParticleNet discriminator is further decorrelated in the same way as the $N_2^{\beta=1}$ tagger, in order to reduce the dependence on the jet p_{T} and m_{SD} and to reduce the complexity of the background estimate.



(a) (b) (c)

Figure 6.4: Summary of tagger regions using the substructure tagger variables for the example of 2017 data and MC in the semileptonic $t\bar{t}$ sample. From left to right the *pass-top, pass-W* and *fail* region is shown. The distributions are for jets with $p_T > 200$ GeV.

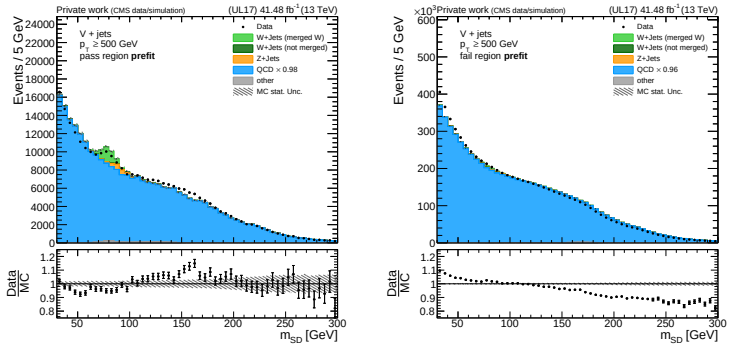
The regions are summarized in Table 4 showing the requirements and Table 5 shows the selection efficiencies for signal (W and top jets) and background. The efficiencies are calculated in the same way as the ones for the $t\bar{t}$ sample. With a W jet selection efficiency of 18%, 45% and 38% when using substructure variables, ParticleNet discriminators or ParticleNet$^{\text{DDT}}$, the pass region is enriched in W jets, when compared to the background efficiency of less than 5% in all tagging approaches. The fail region is consequently enriched with background jets with background efficiencies of around 95% for all tagging approaches.
Figure 6.5 shows the soft drop mass distribution of the jets in the pass region in the left plot and the fail region in the right plot for the substructure tagging approach with 2017 data and simulation. The dominant QCD multijet background shown as blue histograms in the left and right plots is scaled by a factor of 0.98 and 0.96 respectively to match the normalization observed in data. In the pass region the signal peaks at the mass of the W boson, while there is still a massive amount of QCD multijet background events. In the fail region the QCD multijet background dominates the distribution. The data to MC disagreement in both regions is mainly caused by the imperfect modeling of the QCD multijet background and is avoided by estimating the QCD multijet background from data as mentioned above.
For reference, the distribution of all other years and taggers can be found in Appendix B.2.

	Substructure	ParticleNet	ParticleNet$^{\text{DDT}}$
pass	$N_2^{\beta=1,\text{DDT}} < 0$	MDWvsQCD > 0.91	MDWvsQCD$^{\text{DDT}} < 0$
fail	$N_2^{\beta=1,\text{DDT}} > 0$	MDWvsQCD ≤ 0.91	MDWvsQCD$^{\text{DDT}} > 0$

Table 4: Summary of signal and control regions (passing and failing tagger criteria respectively) for approaches using substructure variables or ParticleNet discriminators for the fully-hadronic $W(q\bar{q})$+jets selection.

	Substructure		ParticleNet		ParticleNet$^{\text{DDT}}$	
	$\varepsilon_{W\text{ jets}}$	$\varepsilon_{\text{backg.}}$	$\varepsilon_{W\text{ jets}}$	$\varepsilon_{\text{backg.}}$	$\varepsilon_{W\text{ jets}}$	$\varepsilon_{\text{backg.}}$
pass	18.54%	4.14%	44.48%	4.71%	38.20%	4.20%
fail	81.46%	95.86%	55.52%	95.29%	61.80%	95.80%

Table 5: Summary of signal and background efficiencies in the different signal and control regions for the substructure and ParticleNet-based taggers in the fully-hadronic $W(q\bar{q})$+jets selection. The efficiencies are estimated from simulation and averaged over the years of data-taking.



(a) (b)

Figure 6.5: Summary of tagger regions using the substructure tagger variables for the example of 2017 data and MC in the fully-hadronic $W(q\bar{q})$+jets sample. The left plot shows the *pass*, while the right shows the *fail* region. The distributions are for jets with $p_T > 500\,\text{GeV}$.

6.2 Simulated event samples

The simulation of the pp collision events is performed using various software packages in multiple steps as summarized in Section 4.1. The matrix-element generation for the samples used in the analyses is handled by one or more of the following generators: MADGRAPH5_aMC@NLO [52] v2.6.5 (v2.6.1), POWHEG 2 [53–55] or PYTHIA 8.244 [56]. For samples, where PYTHIA is not used to generate the hard scattering, the respective generator is further interfaced with PYTHIA for the parton showering and hadronization. In the following, the generators used for the different samples are summarized. POWHEG is used for the simulation of the NLO matrix-elements of the processes $t\bar{t}$ [119], single top in association with a W boson [120] as well as single top in the t-channel [121]. The latter is additionally interfaced with the MADSPIN [122] package, which handles the decay of the top quark and consequent W bosons.

Processes with vector-bosons + jets, in the variations $W(\rightarrow qq)$ + jets (0-3), $W(\rightarrow \ell\nu)$ + jets (0-4), $Z(\rightarrow qq)$ + jets (0-4) and $Z(\rightarrow \ell\ell)$ + jets (0-4), are generated at LO using MADGRAPH5_aMC@NLO with additional jets added to the final state. The $W(\rightarrow qq)$ + jets sample is generated with up to three jets in the final state, while the others are generated with up to four jets in the final state. To avoid double counting of events with additional jet radiations during the matrix-element generation in MADGRAPH5_aMC@NLO and the parton showering in PYTHIA, jets from MADGRAPH5_aMC@NLO and jets added during the parton shower in PYTHIA are matched and events with double counting are removed. The matching is done with the MLM scheme for LO matrix-element generators [58] within the MADGRAPH5_aMC@NLO package. Further MADGRAPH5_aMC@NLO is used to generate matrix-elements at NLO for single-top processes in the s-channel, while MADSPIN handles the decay of the top quark and W boson similar to the t-channel sample. For QCD multijet processes two different generators are used. MADGRAPH5_aMC@NLO is used to generate MC-samples of QCD multijet events binned in H_T to use in the fully-hadronic $W(q\bar{q})$+jets sample, while PYTHIA is used to simulate MC-samples of QCD multijet events binned in p_T and enriched with muons (at least one μ with $p_T > 5\,\text{GeV}$ must be present in an event) to use in the semileptonic $t\bar{t}$ sample. The QCD H_T-binned sample is generated at LO with up to four jets in the final state and then treated for double counting using the MLM scheme similar to the V +jets samples.

All samples were generated with the PYTHIA settings according to the CP5 tune [123] and using the NNPDF3.1 [47] PDF sets for the modeling of the parton distribution functions. The generated events are then passed through the GEANT4 [61] package to simulate the detector response before they are treated with the same reconstruction algorithms as the real data.

The simulated events are corrected for various effects, which are not well modeled in the simulation. These correction factors are applied to the simulated events as event weights. The events are generated with a profile of the number of pileup interactions per bunch crossing, which is not the same in data. To correct for that all events are reweighted to match the profile of that in the respective data-taking period. The next-to-leading order POWHEG + PYTHIA $t\bar{t}$ samples used in the analyses show deviations in the p_T spectrum of top quarks when compared to data [124, 125]. Events containing a top quark pair are reweighted based on the measurements of the top p_T spectrum. The W+jets and Z+jets LO+MLM samples are corrected to higher order precision using the same dedicated correction factors as they were used in [4]. These

corrections consist of multiplicative correction factors to the differential cross section of the V +jets processes in dependence of the p_{T} of the vector boson. They are correcting for higher-order QCD and electroweak effects. The QCD corrections were derived from NLO V+jets samples, that were generated using MADGRAPH5_aMC@NLO, while the electroweak corrections are based on higher-order calculations at NLO precision [3].

6.3 Statistical analysis

For the statistical analysis of the data, maximum likelihood template fits are performed using histograms of the soft drop jet mass m_{SD}. In a maximum likelihood fit, a set of parameters for a model, which aims to accurately describe observations within the data, is determined. In general, the model or hypothesis can be described by a probability density function $f(x_i|\boldsymbol{\theta})$, where x_i is the random variate and $\boldsymbol{\theta}$ is the set of parameters of the model. The likelihood function $\mathscr{L}$ describes the probability of the observed data x_i given the model parameters $\boldsymbol{\theta}$:

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}). \tag{6.3}$$

For large amounts of data, as typical for datasets of pp collision events from experiments in high-energy physics, it is more convenient to work with the data in the form of histograms. The data are then described by a set of N bins with $\mathbf{n} = (n_1, ..., n_N)$ entries (events). With the expectation value $\nu_i(\boldsymbol{\theta}) = n_{\mathrm{tot}} \int_{\mathrm{bin\ i}} f(x|\boldsymbol{\theta})\mathrm{d}x$ for the number of events in bin i, the likelihood function can be simplified by constructing a joint probability density function in the form of Poisson distributions to the negative log-likelihood:

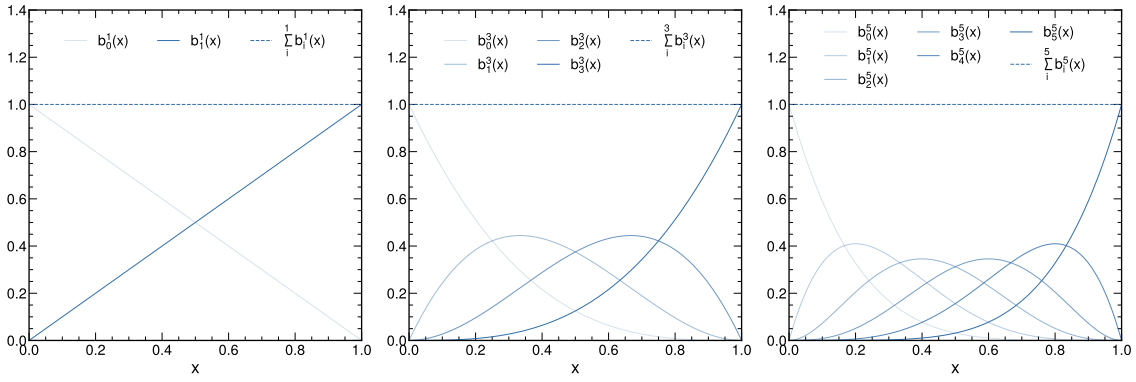
$$-\log\mathscr{L}(\nu_{\mathrm{tot}}, \boldsymbol{\theta}) = -\log\prod_{i=1}^{N} \frac{\nu_i^{n_i}(\boldsymbol{\theta})}{n_i!} e^{-\nu_i(\boldsymbol{\theta})} = -\nu_{\mathrm{tot}} + \sum_{i=1}^{N} n_i \log\nu_i(\boldsymbol{\theta}) - \log n_i!\,. \tag{6.4}$$

Here the terms that do not depend on the parameters $\boldsymbol{\theta}$ can be dropped, as they do not affect the minimization of the negative log-likelihood. As per convention the negative log-likelihood $-2\log\mathscr{L}$ is used, which is then minimized to obtain the maximum-likelihood estimators of the parameters $\hat{\boldsymbol{\theta}}$. A more detailed description of this procedure can be found in [126]. The construction of the likelihood functions and the minimization of the negative log-likelihood is performed using the COMBINE toolkit [127], which is an advanced interface of the ROOFIT [128] and ROOSTATS [129] packages.

6.4 Background Estimation

All background processes in the semileptonic $t\bar{t}$ sample are estimated from simulation. The fully-hadronic $W(q\bar{q})$+jets sample is dominated by QCD multijet events, which is not modeled well enough in simulation as demonstrated before in Figure 6.2b. Therefore a data-driven approach is used to estimate the QCD multijet background, while the remaining processes are estimated from simulation.

The data-driven approach for the QCD multijet background estimation is based on an established differential alphabet method as used in several CMS analyses [2, 4]. The method is based



(a) (b) (c)

Figure 6.6: A series of demo plots of Bernstein polynomials $b_i^n(x)$ of different orders n ($n \in [1,3,5]$) and the sums with respect to i.

on the assumption, that the shape of the m_{SD} distribution of the QCD multijet events is similar in the pass and fail region. Consequently, one can estimate the shape of the QCD background in the signal-depleted fail region from data, and use a transfer factor to extrapolate the shape to the signal-enriched pass region. In case of a perfect description of the data by MC simulation, the transfer factor would be a constant value, the tagging efficiency in a QCD multijet enriched data-sample ε_{QCD}. With current simulations, the jet tagging efficiency is not constant but depends on the jet kinematics, primarily on the jet p_{T}. Additionally, the jet tagging variables are not entirely decorrelated to the jet mass or m_{SD}, despite efforts to explicitly decorrelate them (as described in Section 5.3.2). Thus a two-dimensional transfer function in dependence of the p_{T} and ρ_{SD} of the jet is used, where ρ_{SD} is the dimensionless scaling variable $\rho_{\mathrm{SD}} = 2\log(m_{\mathrm{SD}}/p_{\mathrm{T}})$, which was already used in the decorrelation procedure of $N_2^{\beta=1}$ described in Section 5.3.2.

As a basis for the two-dimensional transfer function the Bernstein basis polynomials [130] are used. The i-th Bernstein basis polynomial of n-th degree ($i \in [0,\cdots,n]$) is defined for $x \in [0,1]$ as:

$$b_i^n(x) = \binom{n}{i} x^i (1-x)^{n-i}. \tag{6.5}$$

Figure 6.6 demonstrates the basis polynomials for arbitrary choice $n \in [1,3,5]$ as well as their sum $\sum_i^n b_i^n(x)$, which is equal to one for all $x \in [0,1]$. These basis polynomials were first introduced by S. Bernstein to formulate a proof of the Weierstrass approximation theorem, stating that one can approximate any continuous function $f(x)$ on the interval $[0,1]$ with arbitrary precision, using linear combinations of polynomials. The linear combinations are called Bernstein polynomials and read as follows:

$$B_n(x) = a_i \sum_{i=0}^{n} b_i^n(x), \tag{6.6}$$

where a_i are the coefficients of the linear combination. To use Bernstein polynomials for the transfer function the dimensions of the input variables (p_{T} and ρ_{SD}) need to be scaled to the

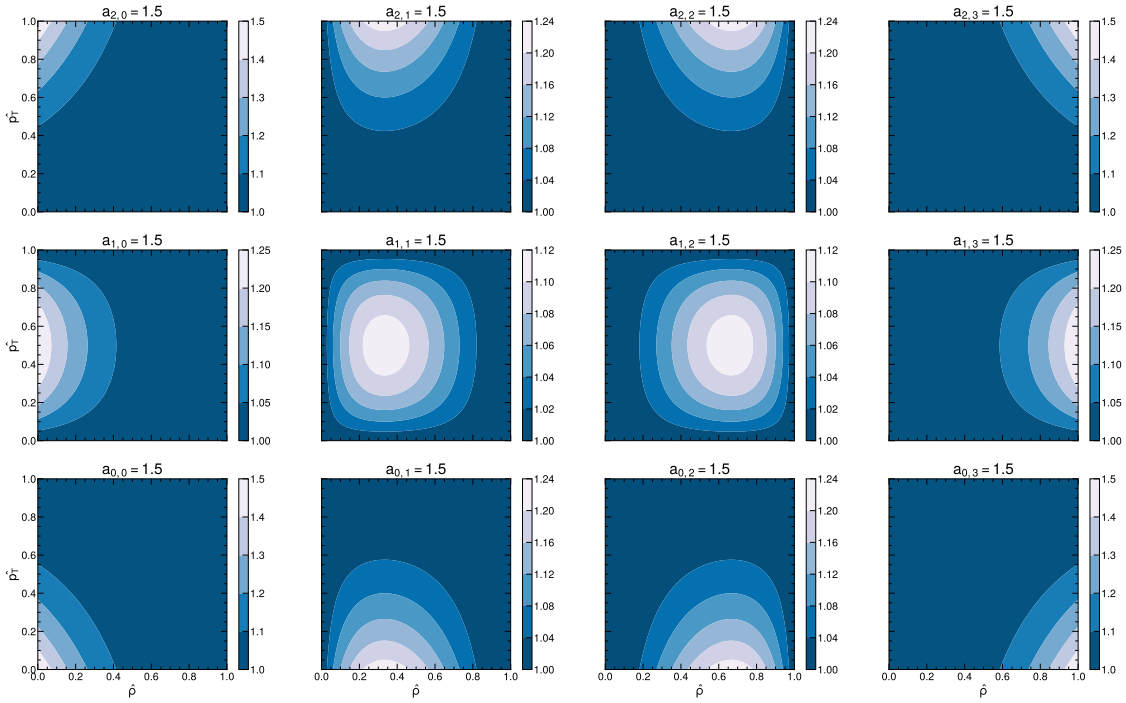


Figure 6.7: Demo of sums of products of Bernstein polynomials $b_i^{n_{\rho_{\mathrm{SD}}}}(\hat{\rho}_{\mathrm{SD}})$ and $b_j^{n_{p_{\mathrm{T}}}}(\hat{p}_T)$ for $n_{p_{\mathrm{T}}} = 2$ and $n_{\rho_{\mathrm{SD}}} = 3$ with setting each scaling parameter $a_{i,j}$ to non-zero value separately.

interval $[0,1]$. This is done by the following transformation:

$$\hat{x} = \frac{x - x_{\mathrm{min}}}{x_{\mathrm{max}} - x_{\mathrm{min}}}, \tag{6.7}$$

where x denotes the respective variable, while x_{min} and x_{max} are the minimum and maximum values in the binning of that variable used in the fit. For ρ_{SD} the latter refers technically to the binning of both p_{T} and m_{SD}, since these are the input variables for the fit, so $\hat{\rho}_{\mathrm{SD}}$ is calculated for each bin in the $(\hat{p}_T, \hat{m}_{\mathrm{SD}})$.

The scaled variables are then used to calculate the transfer function $R_{p/f}(\hat{\rho}_{\mathrm{SD}}, \hat{p}_T)$:

$$R_{p/f}(\hat{\rho}_{\mathrm{SD}}, \hat{p}_T) = \sum_i^{n_{\rho_{\mathrm{SD}}}} \sum_j^{n_{p_{\mathrm{T}}}} a_{ij} b_i^{n_{\rho_{\mathrm{SD}}}}(\hat{\rho}_{\mathrm{SD}}) b_j^{n_{p_{\mathrm{T}}}}(\hat{p}_T), \tag{6.8}$$

where $n_{\rho_{\mathrm{SD}}}$ and $n_{p_{\mathrm{T}}}$ are the orders of the Bernstein polynomials in $\hat{\rho}_{\mathrm{SD}}$ and $\hat{p}_T$, respectively, such that $i \in [0, \cdots, n_{\rho_{\mathrm{SD}}}]$ and $j \in [0, \cdots, n_{p_{\mathrm{T}}}]$. The coefficients a_{ij} of these linear combinations are used in the fit to data as free parameters. Figure 6.7 demonstrates for a two-dimensional Bernstein polynomial of order $n_{\rho_{\mathrm{SD}}} = 3$ and $n_{p_{\mathrm{T}}} = 2$ how the parameters a_{ij} can be used to shape the transfer function in distinct regions on the $(\hat{p}_T, \hat{\rho}_{\mathrm{SD}})$-plane. The final transfer factor has also to account for the tagging efficiency ε_{QCD}, so the final transfer factor for a given dataset corresponding to the year of data-taking reads as:

$$\mathrm{TF}(\hat{\rho}_{\mathrm{SD}}, \hat{p}_T) = \varepsilon_{QCD}(\hat{p}_T) \cdot R_{p/f}(\hat{\rho}_{\mathrm{SD}}, \hat{p}_T), \tag{6.9}$$

where the efficiency ε_{QCD} is evaluated in dependence of p_{T} from data. The p_{T}-dependent tagger

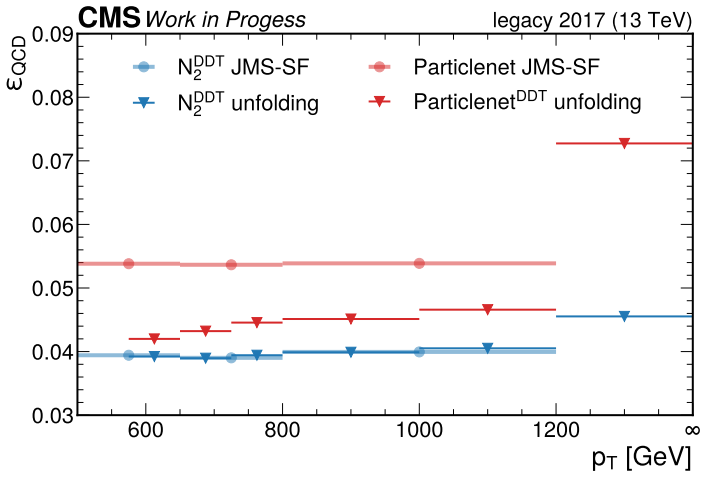


Figure 6.8: Summary of the selection efficiency for QCD multijet background measured in 2017 data when using the different tagging approaches in the data-driven background estimation.

efficiency measured in 2017 data is shown in Figure 6.8 for the different tagging approaches and different p_T-binning schemes used in the analyses. The blue and red circles correspond to QCD mistag rates when using $N_2^{\beta=1,\mathrm{DDT}}$ and ParticleNet discriminators respectively as they are used for the jet mass calibration discussed in Section 7. Here the efficiency is measured in three p_T bins with the edges $[500, 650, 800, 1200]$. The blue and red triangles correspond to the QCD mistag rates when using the $N_2^{\beta=1,\mathrm{DDT}}$ and ParticleNet$^{\mathrm{DDT}}$ respectively. These are used in the measurement of the jet mass distribution discussed in Section 8. Here the efficiency is measured in six p_T bins with the edges $[575, 650, 725, 800, 1000, 1200, \infty]$. The same plots for all periods of data-taking can be found in Appendix C.2

6.4.1 *Goodness of Fit and Fisher's F-test*

To determine the optimal order of the Bernstein polynomials $n_{\rho_{SD}}$ and n_{p_T} the goodness of fit using the saturated model approach is evaluated for a series of orders. The resulting test statistics are then compared in F-tests and the final orders of the polynomials are chosen as described in the following.

With the goodness of fit one tests how well the null hypothesis H_0 fits the data, without the need for an alternative hypothesis. There is no best definition for a test statistic for the goodness of fit, but often likelihood ratios are used, as opposed to the absolute likelihood of the null hypothesis, since the latter typically depends highly on the parameter space [131]. Since the $-2\log$ of likelihood-ratios will asymptotically follow a χ^2-distribution under certain conditions (Wilk's Theorem [132]), one can choose the test statistic for the goodness of fit as the ratio of likelihoods. A suitable choice for the denominator of the ratio is a hypothesis that introduces one parameter per observation, thus fitting the data perfectly. Hypotheses like this are called *saturated* models [133] and are maximally complex (number of degrees of freedom ndf = 0), thus include the null hypothesis, which is mandatory for Wilk's theorem. With this choice, the likelihood ratio for independent poisson distributed binned data is given by:

$$\lambda = \frac{\mathscr{L}(H_0)}{\mathscr{L}(\text{saturated})} = \frac{\prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}}{\prod_{i=1}^{N} \frac{n_i^{n_i}}{n_i!} e^{-n_i}} = \prod_{i=1}^{N} \left(\frac{\nu_i}{n_i}\right)^{n_i} e^{n_i - \nu_i}, \tag{6.10}$$

and the more commonly used negative log-likelihood ratio follows:

$$-2\log\lambda = 2\sum_{i=1}^{N}\left[\nu_i - n_i + n_i \log\frac{n_i}{\nu_i}\right], \tag{6.11}$$

which will have the same minimum as the absolute $-2\log\mathscr{L}(H_0)$, which will follow a χ^2-distribution, when certain regularity conditions are met and the estimators ν_i are sufficiently large. [22]

In an F-test, the variances of two random variates are tested for equality. Given the two independent random variables X_1 and X_2 that are distributed according to a scaled χ^2-distribution with the degrees of freedom d_1 and d_2, then the ratio defined as the F-value:

$$F = \frac{X_1/d_1}{X_2/d_2} \tag{6.12}$$

is F-distributed with the degrees of freedom d_1 and d_2. The probability density function $f(x)$ of the F-distribution for $x > 0$ is given by:

$$F \sim f(x) = \frac{\Gamma\left[\frac{d_1+d_2}{2}\right] d_1^{d_1/2} d_2^{d_2/2}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \cdot \frac{x^{d_1/2-1}}{(d_2 + d_1 x)^{(d_1+d_2)/2}} \tag{6.13}$$

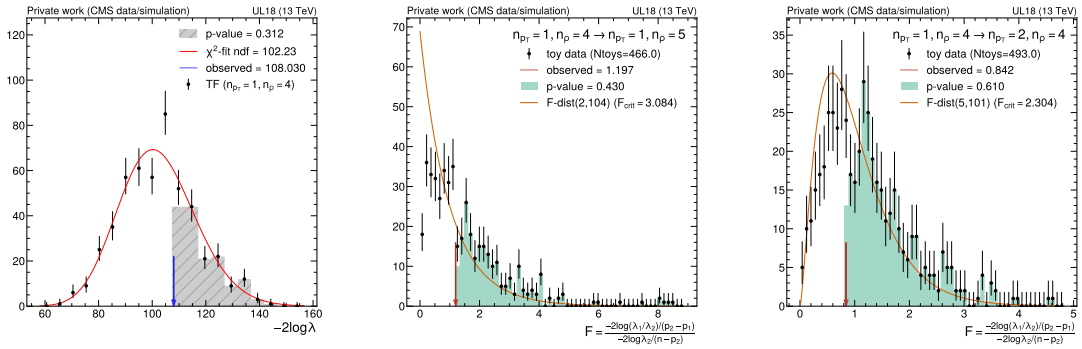
and $f(x) = 0$ for $x \leq 0$ [134, 135].

When comparing two models in a fit to data, that have different numbers of parameters p_1 and p_2 ($p_1 < p_2$) and model 1 is nested in model 2, the F-test can be used to test the null hypothesis that the less complex model (i.e. the one with the lower number of parameters) is sufficient to describe the data. In this case, the variances to compare are the residuals of the two models and the difference in degrees of freedom. Thus for the null hypothesis, we assume the following should hold:

$$\frac{(-2\log\lambda_1) - (-2\log\lambda_2)}{(-2\log\lambda_2)} \approx \frac{(n_{\text{bins}} - p_1) - (n_{\text{bins}} - p_2)}{n_{\text{bins}} - p_2}, \tag{6.14}$$

where $-2\log\lambda_i$ is the goodness of fit test statistic (e.g. using the *saturated* model) for the model i, p_i is the number of parameters of the model i and n_{bins} is the number of bins holding independent observations. If the null hypothesis is accepted an increase of parameters should not improve the fit and the residual on the left-hand side of Eq. 6.14 would be small. In this case, the F value would be close to 1. If the null hypothesis is rejected, the F value would be larger than 1. Consequently, the F value used in the following is defined as the ratio of both sides of Eq. 6.14 :

$$F = \frac{(-2\log\lambda_1) - (-2\log\lambda_2)}{(n_{\text{bins}} - p_1) - (n_{\text{bins}} - p_2)} \cdot \frac{n_{\text{bins}} - p_2}{(-2\log\lambda_2)}, \tag{6.15}$$



(a) (b) (c)

Figure 6.9: Summary of the last iteration of F-Tests to determine the maximum order in degrees of Bernstein-polynomials to be used for the transfer factor when using substructure variables for the jet tagging in the fit to 2018 data. The plots show the relevant distributions of the F-Test, which tests if the base model with the order combination $(n_{p_T} = 1, n_{\rho_{SD}} = 4)$ describes the data sufficiently well. The left plot shows the goodness of fit calculated for the 500 toys generated with the base model and the data using the saturated model approach. The p-value= 0.312 indicates a well-behaved fit. The plot in the middle and on the right shows the F-distribution derived from Equation 6.16 for the toys (shown as black markers) and for the data (shown as a red arrow) and the expected F-distribution following Equation 6.13. In both, the toys describe the expected F-distribution well and the p-values are both above $\alpha = 0.05$, thus the F-test iteration can be used to reject both higher order models and accepting the base model.

which can be simplified to:

$$F = \frac{\left(\frac{-2\log\lambda_1 + 2\log\lambda_2}{p_2 - p_1}\right)}{\left(\frac{-2\log\lambda_2}{n_{\text{bins}} - p_2}\right)} \tag{6.16}$$

Under the null hypothesis, the F value is then F-distributed with the degrees of freedom $p_2 - p_1$ and $n_{\text{bins}} - p_2$ [136].

As mentioned before the F-test is then performed for a series of $(n_{p_T}, n_{\rho_{SD}})$-combinations, comparing a reference model $(n_{p_T}, n_{\rho_{SD}})$ with alternative models where either n_{p_T} or $n_{\rho_{SD}}$ is incremented by one. For each reference model, the full fit model is constructed and fit to data. Next toy MC datasets are generated using the best-fit parameters of this model before the goodness of fit using the saturated model approach is calculated by comparing both the toys and the data distribution to the reference model $(-2\log\lambda_1)$ and one of the alternative models $(-2\log\lambda_2)$. Then one can compute the F-value for each toy and the data. Figure 6.9a shows both the goodness of fit of the reference model for both the toys as a histogram and for data as an arrow. Figure 6.9b and figure 6.9c show the full F-Test information for the comparisons to both alternative models in a similar manner. If the toys correctly follow an F-distribution with the degrees of freedom $p_2 - p_1$ and $n_{\text{bins}} - p_2$ (orange line) the assumption 6.14 holds. In case the observed F-value falls under the critical value F_{crit} for which $\int_{F_{\text{crit}}}^{\infty} f(d_1, d_2, x)\,dx = 0.05$ the less complex model describes the data better than the alternative, and we choose it over the alternative model. In some cases, the goodness of fit for the alternative model is not changing or worsening (i.e. higher values of $-2\log\lambda_2$) with respect to the reference model, yielding values close to or below zero, respectively. In these cases, the reference model is also chosen over

	Substructure		ParticleNet		ParticleNet$^{\text{DDT}}$	
	n_{p_T}	n_{ρ_SD}	n_{p_T}	n_{ρ_SD}	n_{p_T}	n_{ρ_SD}
2016 (early)	1	4	2	5	2	3
2016 (late)	1	2	2	5	2	3
2017	2	3	2	6	2	3
2018	1	4	2	6	2	3

Table 6: Chosen orders for the Bernstein polynomials for the different years of data-taking and tagging strategies.

the alternative model. The scan over reference models is performed for each tagging strategy (Substructure, ParticleNet and ParticleNet$^{\text{DDT}}$) and all years of data-taking (2016 is split into early and late 2016) separately. The final chosen order combinations are summarized in Table 6 and the distributions corresponding to the last step in the F-Test for all years and tagging approaches can be found for reference in Appendix C.3 for each tagging approach. The tagger based in the ECF ratio $N_2^{\beta=1}$ with further decorrelation with respect to the jet p_T and ρ_SD as discussed in Section 5.3.2.4 requires orders $n_{p_\text{T}} = 1$ for most of the years, except for 2017, where $n_{p_\text{T}} = 2$ is chosen with the F-Tests. In the dimension of the scaled jet mass, ρ_SD orders in the range 2–4 are required. Both ParticleNet-based taggers generally require higher orders both in p_T and ρ_SD to account for the larger residual differences in tagger responses in data and simulation, especially when the ParticleNet tagger MDWvsQCD without further decorrelation with respect to the jet p_T and ρ_SD. This is expected from the observed non-closure discussed in Section 5.3.2.4, where the sculpting introduced by the tagger was probed for the different taggers. For MDWvsQCD the order combination ($n_{p_\text{T}} = 2, n_{\rho_\text{SD}} = 5$) is chosen for early and late 2016 and the combination ($n_{p_\text{T}} = 2, n_{\rho_\text{SD}} = 6$) is chosen for 2017 and 2018. With further decorrelation using the adapted DDT method the orders in ρ_SD required for a good fit result can be reduced to $n_{\rho_\text{SD}} = 3$ for all years.

6.5 Systematic Uncertainties

The systematic uncertainties considered in the analyses can affect the jet soft drop mass in different ways. They can cause differences in shape, normalization or both. Several sources of systematic uncertainties are accounted for in the fit as nuisance parameters θ in the likelihood function. In the following, the treatment of the individual systematic uncertainties is described.

- **Jet energy correction** To incorporate both shape uncertainties and normalization, the total uncertainties of the jet energy correction are propagated to the soft drop jet mass. While measuring the jet mass distribution, a nuisance parameter is employed for each year to account for the uncertainties. However, when performing fits for the jet mass calibration, no nuisance parameters are used, except when explicitly mentioned otherwise.

- **Parton shower initial and final state radiation** The uncertainty from initial state radiation (ISR) and final state radiation (FSR) is estimated by varying the respective renormalization scale up and down by a factor of two. It is treated as a shape uncertainty by propagating it to the soft drop mass by reweighting the simulated events based on the varied scales.

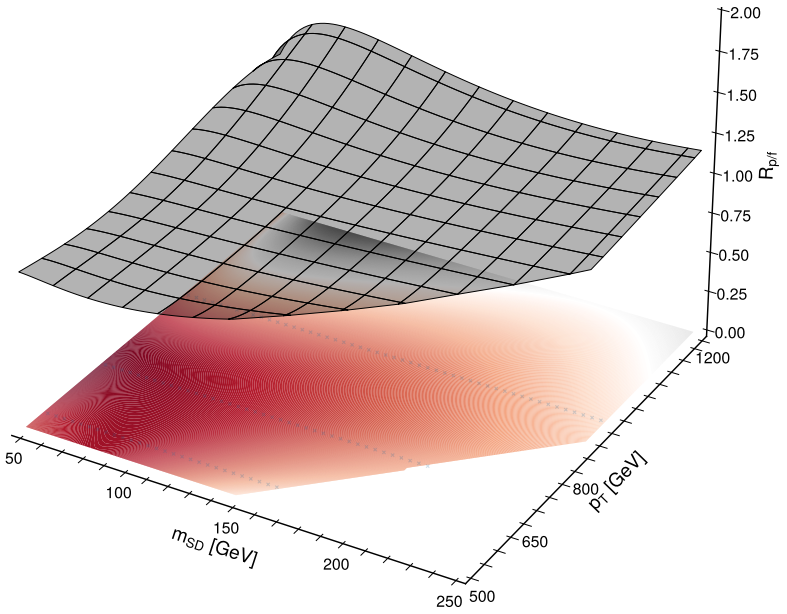


Figure 6.10: Postfit 2D transfer factor for QCD estimate using 2017 data as surface plot (grey) and contour plot (red-grey gradient). The grey crosses mark the (m_{SD}, p_T)-points, at which the polynomials were evaluated for the fit. The region at high m_{SD} and low p_T is cut off, due to the selection criteria $\rho_{SD} = 2\log\left(\frac{m_{SD}}{p_T}\right) < -2.1$.

- **Luminosity measurement** The uncertainty on the measurement of the full Run II luminosity $\mathscr{L}_{int} \approx 137.2\,fb^{-1}$ is estimated to be 1.6% [34, 137]. This uncertainty is accounted for in the fits as normalization nuisance parameters for signal and background processes. It is treated as correlated across the years of data-taking, where applicable.

- **Pileup reweighting** The number of interactions per bunch crossing is estimated from the measurements of the luminosity and the inclusive inelastic cross section of proton-proton collisions. The simulation is reweighted to match the distribution of the number of interactions per bunch crossing in data. The inclusive inelastic cross section is varied up and down by 4.6% and the reweighting is repeated. With the weight corresponding to up and down variation the uncertainty is propagated to the soft drop mass and treated as a shape uncertainty in the fit for both signal and background processes.

- **Top quark p_T reweighting** To account for any uncertainty in the top quark p_T reweighting a shape nuisance parameter is used. The up and down variations are constructed by comparing the soft drop mass distribution with and without applying the reweighting. The difference is then used to reweight the simulated $t\bar{t}$ events.

- **Trigger scale factor** The trigger scale factors are estimated as the ratio of the fit of the trigger efficiency in simulation over the fit in data as described in 6.1. The uncertainty is estimated by varying the numerator and denominator histograms used in the trigger efficiency measurement up and down by the statistical uncertainty and repeating the respective fit. The variations are propagated to the soft drop mass and used as a shape uncertainty in the fit for signal and background processes.

- **V+jets NLO correction** For the higher order correction factors correcting the $W(q\bar{q})$+jets and $Z(q\bar{q})$+jets simulation to the NLO cross section predictions, the uncertainty is estimated from the envelope of the QCD and EWK systematic uncertainties developed in [3]. The QCD and EWK uncertainties are treated as separate shape uncertainties in the fit correlated across years of data-taking, where applicable.

There are no nuisance parameters accounting for jet tagging efficiency uncertainties. Instead dedicated tagging correction nuisance parameters are used for the signal processes allowing for different efficiencies in simulation and data. The correction factors adjust the normalization of the respective process in each signal and control region. For the fully-hadronic $W(q\bar{q})$+jets sample the correction factor is implemented as a nuisance parameter in the fit allowing for anti-correlated changes of the normalization in the pass and fail region while conserving the total number of events

$$N_{\text{total}} = N_{\text{pass,postfit}} + N_{\text{fail,postfit}} \tag{6.17}$$

$$= \hat{\theta} \cdot N_{\text{pass}} + [(1-\hat{\theta}) \frac{N_{\text{pass}}}{N_{\text{fail}}} + 1] \cdot N_{\text{fail}}. \tag{6.18}$$

For the semileptonic $t\bar{t}$ sample, there are two correction factors, one for the efficiency of the W jet tagging, which enriches the pass-W region with W jets, and one for the top jet tagging, which enriches the pass-top region with top jets. Both correction factors are implemented similarly as for the $W(q\bar{q})$+jets sample, each conserving the total number of events across both signal regions (pass-W and pass-top region) and the control region (fail region):

$$N_{\text{total}} = \hat{\theta}_{\text{W}} \cdot N_{\text{pass-W}} + [(1-\hat{\theta}_{\text{W}}) \cdot \frac{N_{\text{pass-W}}}{N_{\text{fail}} + N_{\text{pass-top}}} + 1] \cdot (N_{\text{pass-top}} + N_{\text{fail}}) \tag{6.19}$$

$$N_{\text{total}} = \hat{\theta}_{\text{top}} \cdot N_{\text{pass-top}} + [(1-\hat{\theta}_{\text{top}}) \cdot \frac{N_{\text{pass-top}}}{N_{\text{fail}} + N_{\text{pass-W}}} + 1] \cdot (N_{\text{pass-W}} + N_{\text{fail}}). \tag{6.20}$$

Finally, MC normalization uncertainties are included as nuisance parameters with log-normal constraint with different uncertainties for the processes: 20% for $t\bar{t}$, 100% for QCD (only in semileptonic $t\bar{t}$), 100% for $Z(\rightarrow \ell\ell)$ +jets, 23% for single top, 19% for $W(\rightarrow \ell\nu)$ + jets, 20% for $W(\rightarrow qq)$ + jets and 20% for $Z(\rightarrow qq)$ + jets.

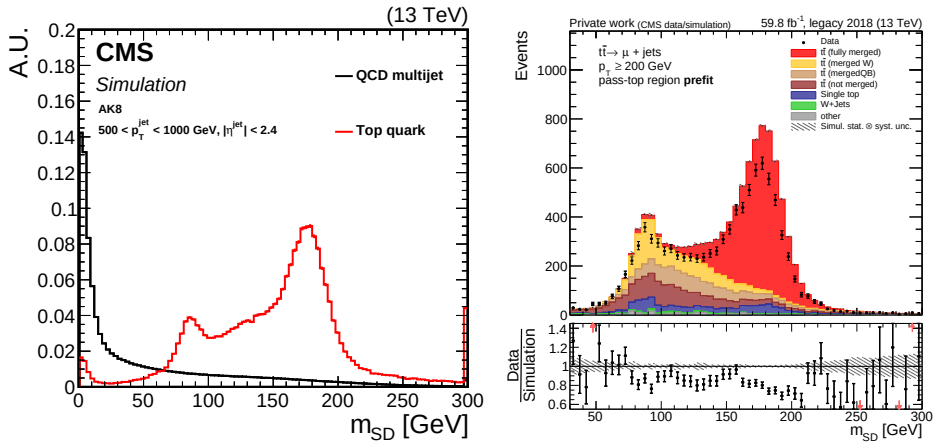
7 Calibration of the jet mass scale

The first main analysis discussed in this thesis is the calibration of the jet mass scale using jets originating from hadronically decaying W bosons and t quarks in the boosted regime. The jet mass scale is a crucial component in many physics analyses, from measurements of standard model physics to searches for physics beyond the standard model. When analyses employ jets with large transverse momentum (e.g. $p_{\mathrm{T}} > 200\,\mathrm{GeV}$), they probe the boosted regime as discussed in Section 5. In the boosted regime the jet-initiating particles are subject to a high Lorentz-boost and result in highly collimated particle showers in the detector. Due to this, it becomes more likely with increasing transverse momentum to cluster the decay products of for example t quarks or W bosons in a single large-radius jet. The mass of these jets and more specifically the soft drop mass m_{SD} is consequently a useful proxy to identify the jet-initiating particle. With most of the soft and wide-angle radiation removed from the jet, the soft drop mass typically has a very distinct peak around the mass of the jet initiating particle. The left plot in Figure 7.1 shows for instance the simulated soft drop mass of high p_{T} jets initiated by top quarks and by quarks and gluons as red and black lines respectively. The complicated nature of jet substructure is not perfectly modeled in simulation as shown in the data to simulation comparison on the right plot in Figure 7.1. Here imperfect modeling of the soft drop mass of top jets becomes apparent when one compares the distribution of the merged top jets (red histogram) with the distribution in data. The analysis presented in this chapter aims to derive correction factors, that correct for this mismodeling of the jet mass scale of soft drop jets. The results of this analysis are published in [1].

The jet mass scale correction factors are measured for high p_{T} jets initiated by both W bosons as well as t quarks in maximum-likelihood fits to data. The fits are performed separately to data recorded in the four periods of data-taking (early 2016, late 2016, 2017 and 2018). The measurement is performed in the semileptonic $t\bar{t}$ sample, as well as in the fully hadronic $W(q\bar{q})$+jets sample as they are described in Section 6. The jets are corrected by applying the full set of dedicated jet energy corrections, which are described in Section 5.2. The energy scale and mass scale of jets are strongly correlated, therefore the jet energy correction should apply also to the jet mass scale. Any further residual deviations of the jet mass scale from unity have to be covered by the jet mass scale correction factors. The jet mass scale correction factors are derived in dependence of p_{T} by categorizing the events into the p_{T} bins shown in Table 1.

7.1 Proxy for the jet mass scale

In order to measure the jet mass scale correction factors in data, the energy scale of the jet constituents is used as a proxy. The jet constituents are the particle-flow particles that are



(a) (b)

Figure 7.1: Distribution of the soft drop jet mass of high p_T jets. The left plot (taken from [89]) shows the soft drop mass of simulated jets initiated by quarks and gluons (black line) and top quarks (red line). The right plot shows the soft drop mass of jets from the *merged top* category and with $\frac{\tau_3}{\tau_2} < 0.5$ in 2018 data and simulation.

clustered into the jet as discussed in Section 5.1. The four momenta of all jet constituents are varied up and down by an arbitrary multiplicative factor $1 \pm \lambda_{\text{input}}$ and the jet mass is recalculated from the sum of the varied four momenta of the jet constituents. In this way, the variation on the per-particle level is propagated to shifts in the soft drop jet mass m_{SD} as demonstrated in Figure 7.2.

The variations in m_{SD} are then used to morph the nominal soft drop mass template in p_T bin j in the maximum-likelihood fit, by employing parameters λ_j that have these variations as $\pm 1\sigma$ effects. In this way, a maximum-likelihood estimator of the parameter in p_T bin j of $\hat{\lambda}_j = 1.0$ would indicate that a shift in m_{SD} was measured in data and that this shift corresponds to $+\lambda_{\text{input}}$. In the following $\lambda_{\text{input}} = 0.005$ is used, such that a shift in parameter λ_j by $\pm 1\sigma$ corresponds to a shift in jet constituent energy scale of 0.5%. In the context of this analysis, it was studied, whether one can employ dedicated parameters for each flavor of particle-flow particles (i.e. charged hadrons, neutral hadrons, photons, etc.), but it was observed, that the parameters dedicated to the charged hadrons would dominate the fit in terms of sensitivity and the other parameters could not be constrained, due to the majority of the jet consisting of charged hadrons as discussed in section 5.2. Therefore, only one common parameter is used to scale the four momenta of all jet constituents up and down. The jet mass scale parameters λ_j measure m_{SD} shifts in data in units of λ_{input}. From this, the simulation to data correction factors $c_{\text{JMS}}(p_T^j)$ for a given p_T bin j is derived from:

$$c_{\text{JMS}}(p_T^j) = 1 + \hat{\lambda}_j \cdot \lambda_{\text{input}}. \tag{7.1}$$

7.2 Maximum-likelihood fit

The fits to data in the context of this analysis are maximum-likelihood fits, performed using the procedure described in section 6.3 for each period of data-taking separately. The parameters λ_j

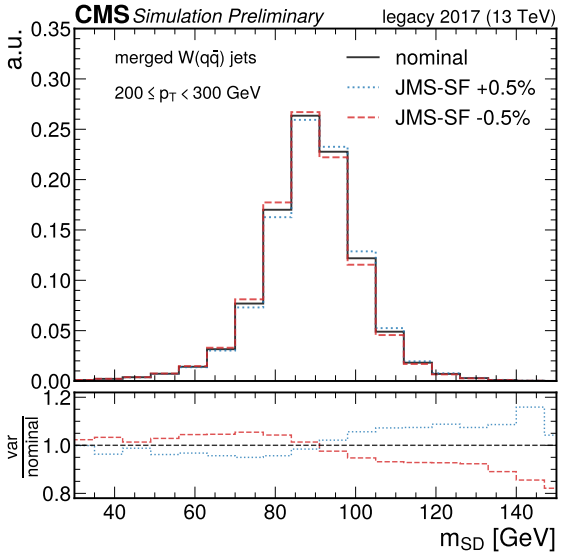


Figure 7.2: The distribution of soft drop mass m_{SD} of the AK8 jet with the highest transverse momentum p_{T} in the *pass-W* tagging category, *merged W* merging category and the transverse momentum bin $200 < p_{\mathrm{T}} \leq 300\,\mathrm{GeV}$. The black line shows the nominal distribution of the jets, whose p_{T} and m_{SD} have been corrected using the jet energy corrections. The blue and red lines correspond to the distribution where the four momenta of the jet constituents have been scaled up and down by 0.5% respectively. This demonstrates how the shift in jet constituent energy scale propagates to the soft drop mass m_{SD}.

described above are used as parameters of interest and will adjust the number of signal events expected by the model in m_{SD} bins of the templates in each p_{T} bin. The processes considered as signals in the fit are the *merged W* component of the simulated $W(q\bar{q})$+jets events in the fully-hadronic $W(q\bar{q})$+jets regions and the *merged W* and *merged top* component of semileptonic $t\bar{t}$ events in the semileptonic $t\bar{t}$ regions. The fits are performed both separately in the semileptonic $t\bar{t}$ and the fully-hadronic $W(q\bar{q})$+jets sample but also combined in both simultaneously. All fits are performed simultaneously in all tagging regions, hence the likelihood function is always the product of the individual likelihood functions of all tagging regions. The likelihood function for a fit of one of the samples is given by

$$\mathscr{L}(\vec{\mu},\vec{\theta}) = \prod_{k} \prod_{j}^{N_{p_{\mathrm{T}}}} \prod_{i}^{N_{m_{\mathrm{SD}}}} P(n_{i,j,k} | \mathbf{s}(\lambda_j, i, j, k, \vec{\theta}) + \mathbf{b}(i, j, k, \vec{\theta})) \cdot \Pi(\vec{\theta}), \tag{7.2}$$

where k runs over all tagging regions, so over [fail, pass-W, pass-top] in the case of the semileptonic $t\bar{t}$ sample and over [fail, pass] in the case of the fully-hadronic $W(q\bar{q})$+jets sample. The indices i and j correspond to the bins in m_{SD} and p_{T} respectively. The number of events in a given (p_{T}, m_{SD}) bin and in tagger region k observed in data is denoted as $n_{i,j,k}$. The number of signal and background events in the same region and bin predicted by the fit model is denoted as $\mathbf{s}(\lambda_j, i, j, k, \vec{\theta})$ and $\mathbf{b}(i, j, k, \vec{\theta})$ respectively. The number of signal events depends on the jet mass scale parameter λ_j corresponding to the respective p_{T} bin j. In the case of the semileptonic $t\bar{t}$ sample either one common jet mass scale parameter λ_j for both W jets and top jets is used per p_{T} bin, or one dedicated parameter λ_j^W varying the jet mass scale of W jets and one parameter λ_j^t are used per p_{T} bin in the fit. For the second scenario the parameters λ_j^W vary only the *merged*

W component of the signal and the parameters λ_j^t vary only the *merged top* component of the signal. The number of background events is not depending on the jet mass scale parameters λ_j. Both the number of signal events and the number of background events depends on the vector of nuisance parameters $\vec{\theta}$, which is constrained in the fit by the Gaussian constraint terms $\Pi(\vec{\theta})$. The set of nuisance parameters consists of one parameter per source of systematic uncertainty. The considered systematic uncertainty sources and the estimation of the size of the uncertainty connected were described in Section 6.5.
For demonstration purposes, the post-fit distributions of the soft drop mass m_{SD} per p_{T} bin and tagging region resulting from a combined fit of both the fully-hadronic $W(q\bar{q})$+jets sample and the semileptonic $t\bar{t}$ to 2017 data are shown in Figures 7.3 and 7.4. For the jet tagging the ParticleNet discriminators listed in Table 4 and 2 are used. The dominant QCD multijet background in the fully-hadronic $W(q\bar{q})$+jets sample is shown as a blue histogram in Figure 7.3 and is estimated using the data-driven method described in Section 6.4. The remaining background processes are estimated from the simulated event samples described in Section 6.2.

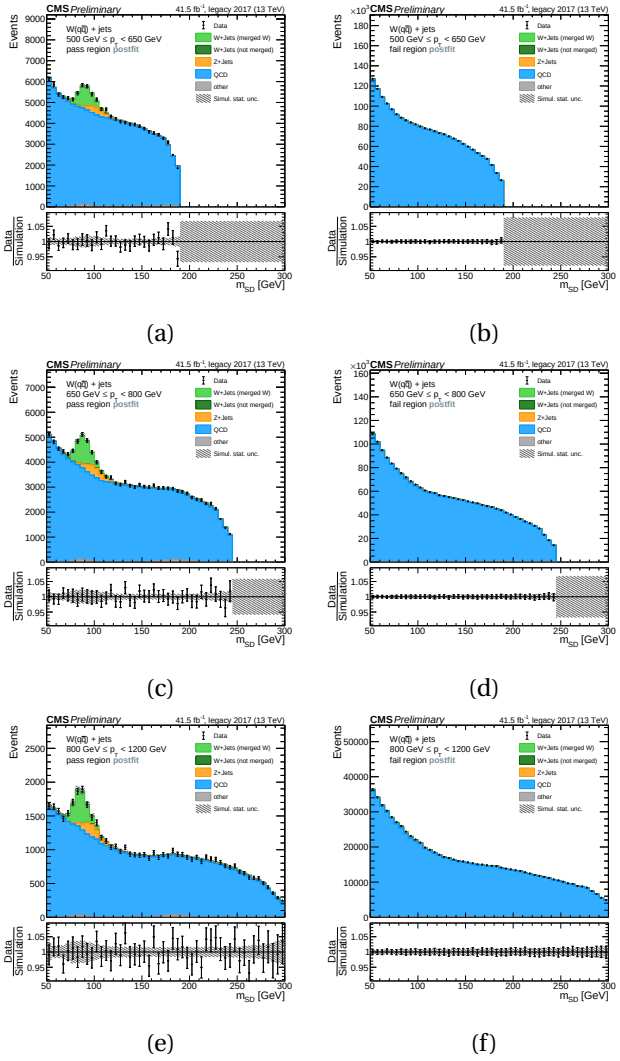


Figure 7.3: Post-fit distributions of the soft drop mass m_{SD} of the leading AK8 jet in the fully-hadronic $W(q\bar{q})$+jets sample of a fit to 2017 data. The rows correspond to the p_{T} bins and the left and right column corresponds to the pass region and fail region respectively. The fit was performed simultaneously in the semileptonic $t\bar{t}$ sample and the fully-hadronic $W(q\bar{q})$+jets sample to derive the common jet mass scale correction factor c_{JMS} for W and top jets.

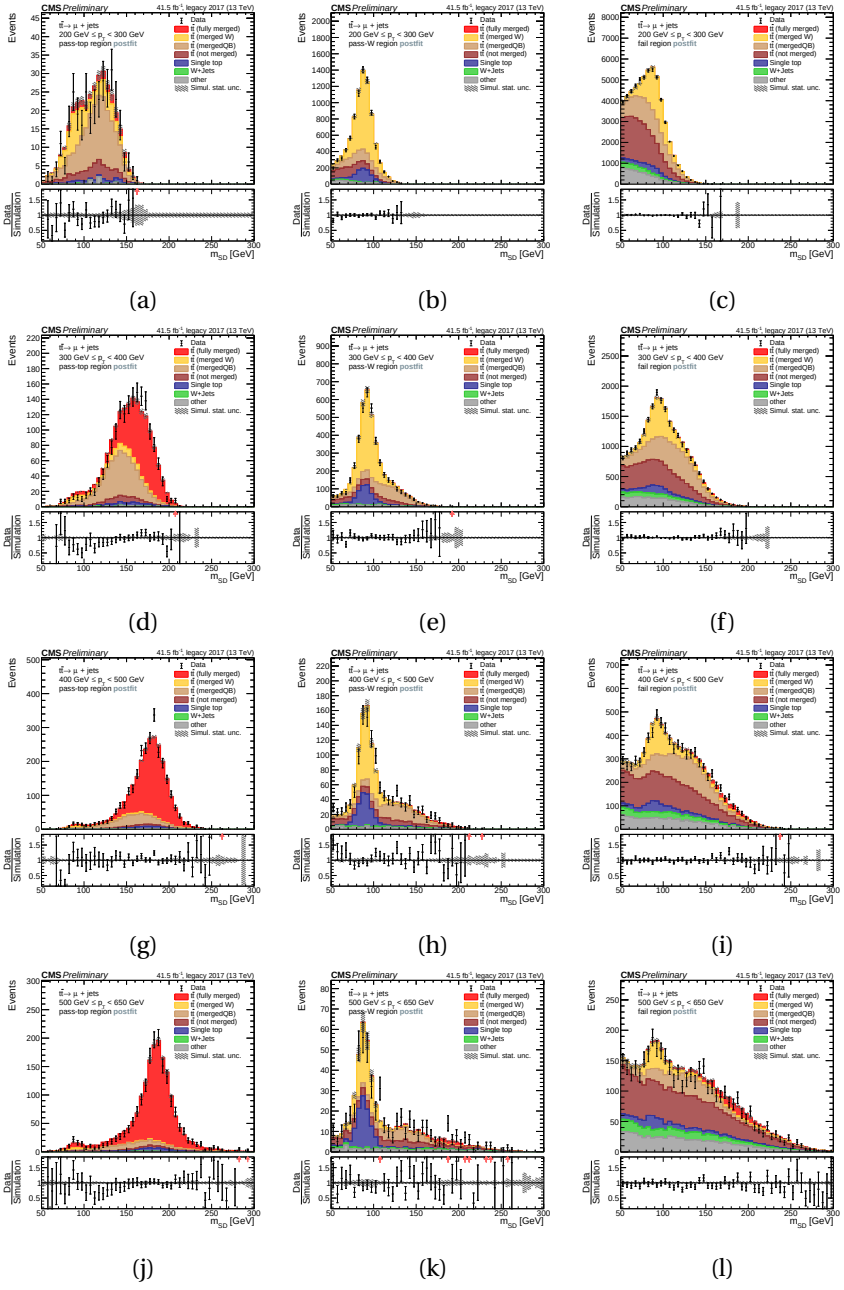


Figure 7.4: Post-fit distributions of the soft drop mass m_{SD} of the leading AK8 jet in the semileptonic $t\bar{t}$ sample of a fit to 2017 data. The rows correspond to the p_{T} bins and the columns correspond from left to right to the pass-top, pass-W and fail region. The fit was performed simultaneously in the semileptonic $t\bar{t}$ sample and the fully-hadronic $W(q\bar{q})$+jets sample to derive the common jet mass scale correction factor c_{JMS} for W and top jets.

7.3 Analysis strategy

The measurement and the consequent analysis of the p_{T} dependent jet mass scale correction factors c_{JMS} consist of deriving the correction factors in three different scenarios and for the two different tagging approaches using either substructure variables or ParticleNet discriminators for the tagging of top and W jets. The signal regions (passing tagger criteria, i.e. pass, pass-W and pass-top) are designed to be enriched in either W jets (pass and pass-W) or top jets (pass-top). This was already demonstrated in Figures 6.4 and 6.5 and offers a good handle to study the jet mass scale of W jets and top jets separately. Additionally, both of the samples use the p_{T} bin $500 < p_{\mathrm{T}} \leq 650\,\mathrm{GeV}$, which allows for the study of the jet mass scale correction factors in an overlapping region and testing of the samples for consistency. The three scenarios are described in the following.

First, the consistency of the studied samples is tested, by measuring the jet mass scale of W jets in the semileptonic $t\bar{t}$ sample and the fully-hadronic $W(q\bar{q})$+jets sample separately. For this two separate jet mass scale parameters λ_j^W and λ_j^t for W jets and top jets respectively are used in the fit in the semileptonic $t\bar{t}$ sample. In this way, the correction factors c_{JMS} measured in the fully-hadronic $W(q\bar{q})$+jets sample can be compared to the correction factors $c_{\mathrm{JMS},W}$ measured in the semileptonic $t\bar{t}$ sample since both measure the correction factors for the jet mass scale of W initiated jets. This comparison will show, if any deviations are visible, that could arise from the larger amount of final state particles or color-reconnection effects in the semileptonic $t\bar{t}$ sample.

The second scenario aims to check the jet mass scale of top and W jets for consistency. The larger amount of b quarks in the top jets could lead to different detector responses for W jets and top jets since the b quark is massive compared to light quarks and leads to secondary vertices. The comparison in the second scenario is done by measuring the one individual jet mass scale correction factors $c_{\mathrm{JMS},W}$ for W initiated jets and one correction factors $c_{\mathrm{JMS},t}$ for top initiated jets in one simultaneous fit in both the semileptonic $t\bar{t}$ sample and the fully-hadronic $W(q\bar{q})$+jets sample. This way the jet mass scale of W jets and top jets can be compared to each other.

The final correction factors c_{JMS} for the jet mass scale of both top and W initiated jets are measured in the third scenario. Here one common jet mass scale parameter λ_j both for top and W jets is used in a simultaneous fit in both samples. The correction factors derived in this scenario are further studied to better understand the correlation between the jet energy scale and the jet mass scale.

7.4 Correction factor consistency across samples

The resulting p_{T} dependent jet mass scale correction factors in the first scenario with dedicated scale parameters λ_j^t and λ_j^W for top and W jets are shown in Figures 7.5 and 7.6 corresponding to the fits in signal and control regions constructed using substructure variables or ParticleNet discriminators as jet taggers respectively.

The red triangles show the p_{T} dependent correction factor $c_{\mathrm{JMS},W}$ measured in semileptonic $t\bar{t}$ events and the blue squares show the p_{T} dependent correction factor c_{JMS} measured in fully-

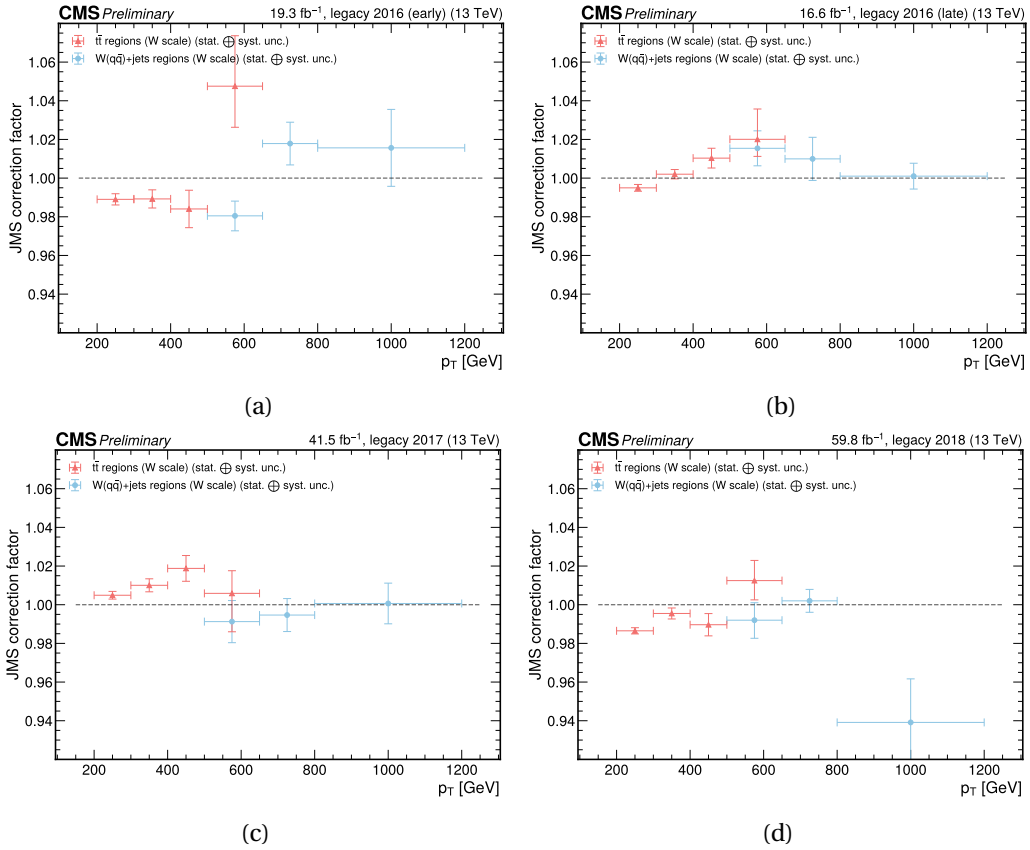


Figure 7.5: Comparison of p_T dependent jet mass scale correction factors for the scale of W jets when using substructure variables for the jet tagging. The correction factors are derived separately in the semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets sample and with two separate mass scale parameters for W jets and t jets per p_T bin. From the top left to the bottom right the plots show the separate results of each period of data-taking: early 2016, late 2016, 2017 and 2018. The red triangles show the resulting correction factors $c_{\mathrm{JMS},W}$ for the W jet mass scale derived from jets in the semileptonic $t\bar{t}$ sample, while the blue circles show the correction factors c_{JMS} derived from jets in the fully-hadronic $W(q\bar{q})$+jets sample.

hadronic $W(q\bar{q})$+jets events. The error bars correspond to the full statistical and systematic uncertainty after the fit to data. The correction factors are measured using the *merged W* component of the events in each sample. The correction factors are consistent across samples within the uncertainties for both tagging approaches and all years, except for early 2016 in the substructure tagging approach.

Here the correction factor $c_{\mathrm{JMS},W}$ measured in the overlap p_T bin $500 < p_T \leq 650\,\mathrm{GeV}$ disagrees between the two samples. The correction factor measured in the $t\bar{t}$ sample in the overlap p_T bin is more than 2 standard deviations larger than the correction factor measured in the $W(q\bar{q})$+jets sample. The correction factor measured in $W(q\bar{q})$+jets agrees with the trend of the c_{JMS} measured in the $t\bar{t}$ sample for $p_T < 500\,\mathrm{GeV}$, which indicates that the deviation in this bin could be an outlier caused by the limited statistics in the early 2016 data. Another possible explanation for this deviation could be the different jet composition in the $t\bar{t}$ sample compared to the $W(q\bar{q})$+jets sample, but this hypothesis is disfavored as the effect is not observed in the other data-taking periods. The correction factor $c_{\mathrm{JMS},W}$ measured in $t\bar{t}$ is most sensitive to how purely the pass-W region consists of jets in the *merged W* component. With increasing

p_T the pass-W region is getting further enriched in jets from the *merged QB* and *merged top* since the decay products of the top quark will be stronger collimated due to the larger p_T. The jet mass scale parameter in the highest p_T bin is measured in a less pure sample of W jets, which could explain the deviation in this bin. Overall the correction factors are consistent across samples within the uncertainties, which indicates that a common W jet mass scale correction factor $c_{JMS,W}$ can be measured in the fully-hadronic $W(q\bar{q})$+jets sample and the semileptonic $t\bar{t}$ sample simultaneously. The uncertainties of the correction factors are smaller than 0.5% for the lowest p_T-bin and reach up to 1 – 2% in the highest p_T-bin in the $t\bar{t}$ sample. In the $W(q\bar{q})$+jets sample the uncertainty is around 1 – 2% across p_T.
The correction factors are mostly within 2% different from unity, except for some outliers which reach up to 6%. The outliers are all in the high p_T region of the respective sample and are likely caused by the limited statistics in these regions. The largest deviation from unity and from the observed trend in the other p_T bins is the correction factor measured in the largest p_T bin in 2018 $W(q\bar{q})$+jets data when using substructure variables as the W jet tagger. Here the correction factor is measured to be close to 6% lower than unity, which is more than two standard deviations. This large shift in jet mass scale can already be seen when comparing the data distribution in the pass region in the last two p_T bins to one another, which can be found in the appendix in Figures B.11c and B.11e. This indicates, that this is not a problem in the fit itself, but rather is a feature introduced or uncovered by the $N_2^{\beta=1,\mathrm{DDT}}$ tagger. This is also supported by the fact, that this deviation is not present when using the ParticleNet discriminators as jet W tagger.

7.5 Correction factor consistency between top and W jets

The resulting jet mass scale factors following the procedure for the second and third scenarios as described above are shown in Figure 7.7 and 7.8 for the substructure and ParticleNet tagging approach, respectively. The blue circles show the p_T dependent jet mass scale correction factors $c_{JMS,W}$ for the W jets, while the p_T dependent correction factors $c_{JMS,t}$ for top jets are shown as green triangles. Both are measured simultaneously in the semileptonic $t\bar{t}$ sample and the fully-hadronic $W(q\bar{q})$+jets sample. The purple triangles correspond to the third scenario described above and show the p_T dependent jet mass scale correction factors c_{JMS} measured in both samples simultaneously and with one common jet mass scale parameter for top and W jets. The error bars correspond similarly as before to the total statistical and systematic uncertainty. The correction factors $c_{JMS,W}$ and $c_{JMS,t}$ for W and top jets are consistent with each other within their uncertainties.
The fit with separate jet mass scale parameters for top and W jets to late 2016 data using the tagger approach with ParticleNet discriminators failed to converge. Since the fits with one common jet mass scale parameter for W and top jets in this region and the fits in the remaining tagger approach and data-taking period combinations did converge, this is not followed up further.
In the fits that did converge the resulting separate jet mass scale correction factors for W jets $c_{JMS,W}$ (blue circles) and the ones for top jets $c_{JMS,t}$ (green triangles) agree within their uncertainties. This indicates that the jet mass scale correction factors can be measured simultaneously for

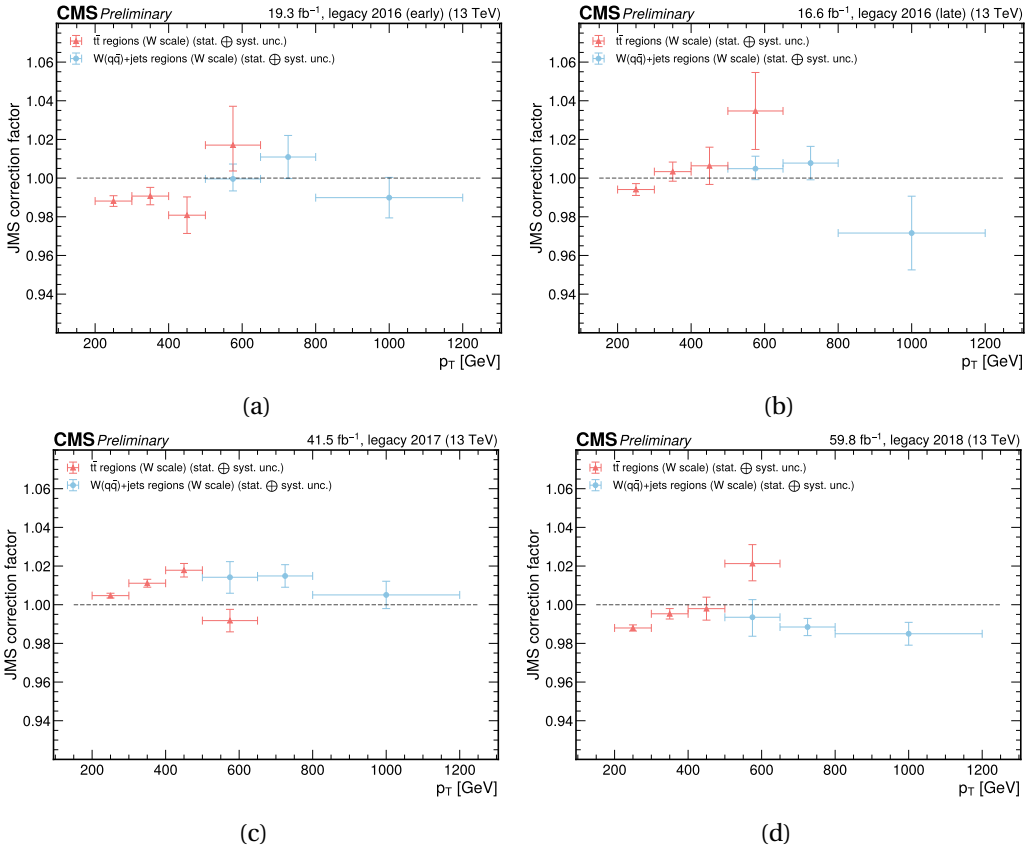


Figure 7.6: Comparison of p_T dependent jet mass scale correction factors for the scale of W jets when using ParticleNet discriminators for the jet tagging. The correction factors are derived separately in the semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets sample and with two separate mass scale parameters for W jets and t jets per p_T bin. From the top left to the bottom right the plots show the separate results of each period of data-taking: early 2016, late 2016, 2017 and 2018. The red triangles show the resulting correction factors $c_{\mathrm{JMS},W}$ for the W jet mass scale derived from jets in the semileptonic $t\bar{t}$ sample, while the blue circles show the correction factors c_{JMS} derived from jets in the fully-hadronic $W(q\bar{q})$+jets sample.

W and top jets with one common jet mass scale parameter λ_j.

The common jet mass scale correction factor for top and W jets c_{JMS} measured in both samples simultaneously (purple triangles) is stable across p_T except for the outlier in the highest p_T bin when using the substructure variables for the jet tagging in the measurement using 2018 data, which was already discussed above. They all agree with the correction factors $c_{\mathrm{JMS},W}$ and $c_{\mathrm{JMS},t}$ within their uncertainties.

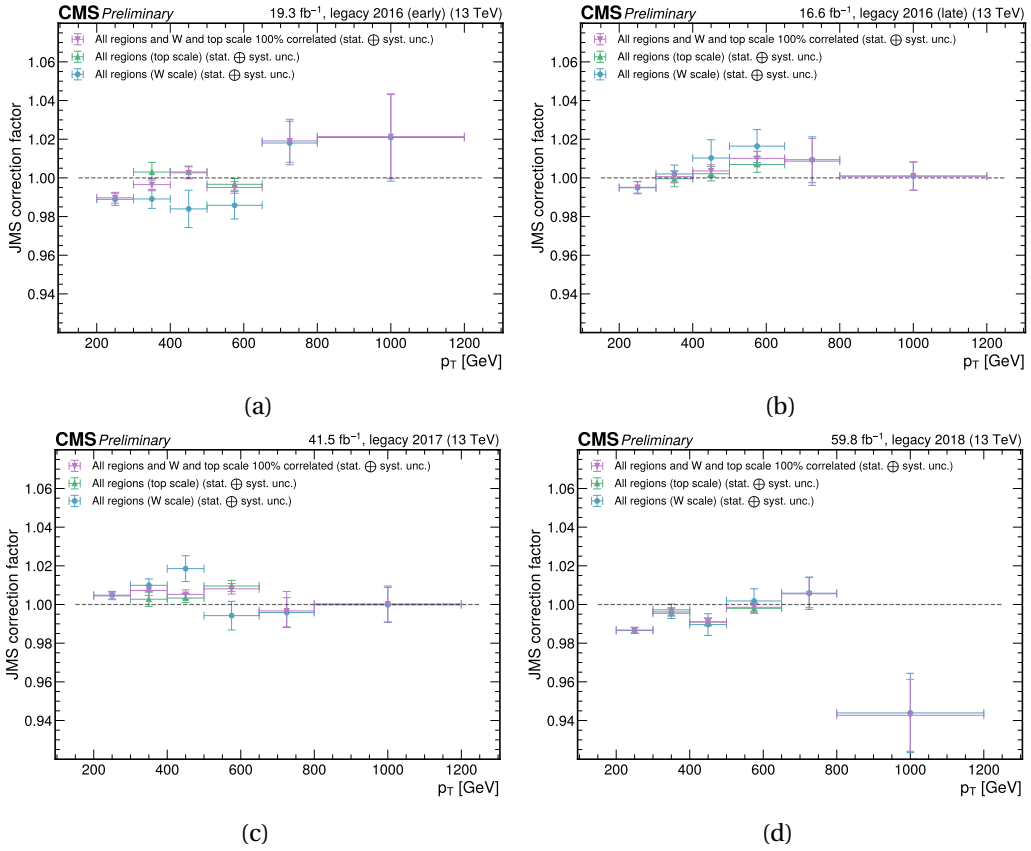


Figure 7.7: Comparison of p_{T} dependent jet mass scale correction factors for the scale of top jets and W jets when using substructure variables for the jet tagging. The correction factors are derived simultaneously in the semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets sample and with two separate mass scale parameters for W jets and t jets per p_{T} bin. From the top left to the bottom right the plots show the separate results of each period of data-taking: early 2016, late 2016, 2017 and 2018. The green triangles show the resulting correction factors $c_{\mathrm{JMS},t}$ for top jets and the blue circles show the correction factors $c_{\mathrm{JMS},W}$ for W jets. The purple triangles show the combined correction factors c_{JMS} for top and W jets.

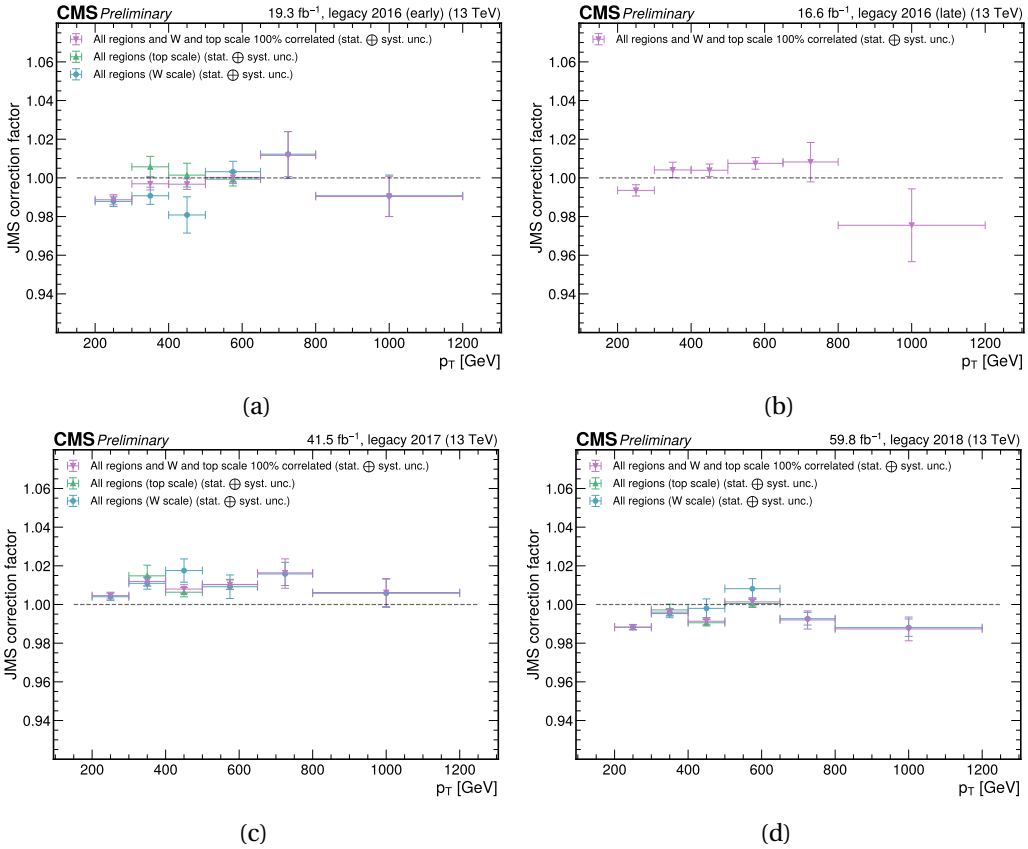


Figure 7.8: Comparison of p_{T} dependent jet mass scale correction factors for the scale of top jets and W jets when using ParticleNet discriminators for the jet tagging. The correction factors are derived simultaneously in the semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets sample and with two separate mass scale parameters for W jets and t jets per p_{T} bin. From the top left to the bottom right the plots show the separate results of each period of data-taking: early 2016, late 2016, 2017 and 2018. The green triangles show the resulting correction factors $c_{\mathrm{JMS},t}$ for top jets and the blue circles show the correction factors $c_{\mathrm{JMS},W}$ for W jets. The purple triangles show the combined correction factors c_{JMS} for top and W jets.

7.6 Final correction factors for top and W jets

The final jet mass scale correction factors for top and W jets are shown again as purple triangles in Figure 7.9 and 7.10 from fits using substructure variables and ParticleNet discriminators respectively. Here the nominal JEC are used to correct both the p_{T} and m_{SD} of the jets, while the blue squares show the correction factors, where only the p_{T} is corrected. Comparing the two indicates that the jet energy corrections apply also to the jet mass scale to some extent, bringing the jet mass scale closer to unity. The residual differences corrected by the correction factor for m_{SD} derived with JEC applied on m_{SD} is at the percent level and reaches up to $\approx 2\%$, except for the outlier in 2018 discussed above, when using substructure variables. The deviations from unity show that the jet energy corrections do not fully apply to the jet mass scale, indicating that the jet mass scale and jet energy scale are not fully correlated.
The purple band around the purple triangles shows an estimate of the systematic uncertainty arising from the uncertainties of jet energy corrections. The band is derived by measuring the correction factors with the templates of the simulated processes for the fit constructed using not the nominal JEC but rather $\pm 1\sigma$ of the total uncertainty of the JEC, i.e. propagating the JEC uncertainty to the soft drop mass m_{SD}. The differences between the resulting correction factors and the nominal correction factors (purple triangles) are added in quadrature to the total uncertainty of the nominal correction factors. This estimates the effect the JEC uncertainty has on the measurement if one assumes a 100% correlation between the jet mass scale and the jet energy scale. The differences in the final correction factors from unity are covered by this uncertainty band, except for the outlier in 2018 discussed above, showing that the deviation from unity can be covered by the jet energy correction uncertainties.
This is also shown by the green triangles, which show the correction factors derived with the nominal JEC applied and using an additional nuisance parameter accounting for the JEC uncertainty. The resulting correction factor is close to unity demonstrating that the residual differences between data and simulation after applying the jet energy corrections to the jet mass scale can be covered by the jet energy correction uncertainties under the assumption of 100% correlation between jet mass scale and jet energy scale.
The correction factor using the nominal JEC and no nuisance parameter for the JEC uncertainty (purple triangles) is more stable and precise with the tagging approach using ParticleNet discriminants. Here, the correction factors are within 1% different from unity with uncertainties reaching down to < 0.25–0.5% at low p_{T} and < 1–2% at high p_{T}, depending on the data-taking period and consequent sample size.

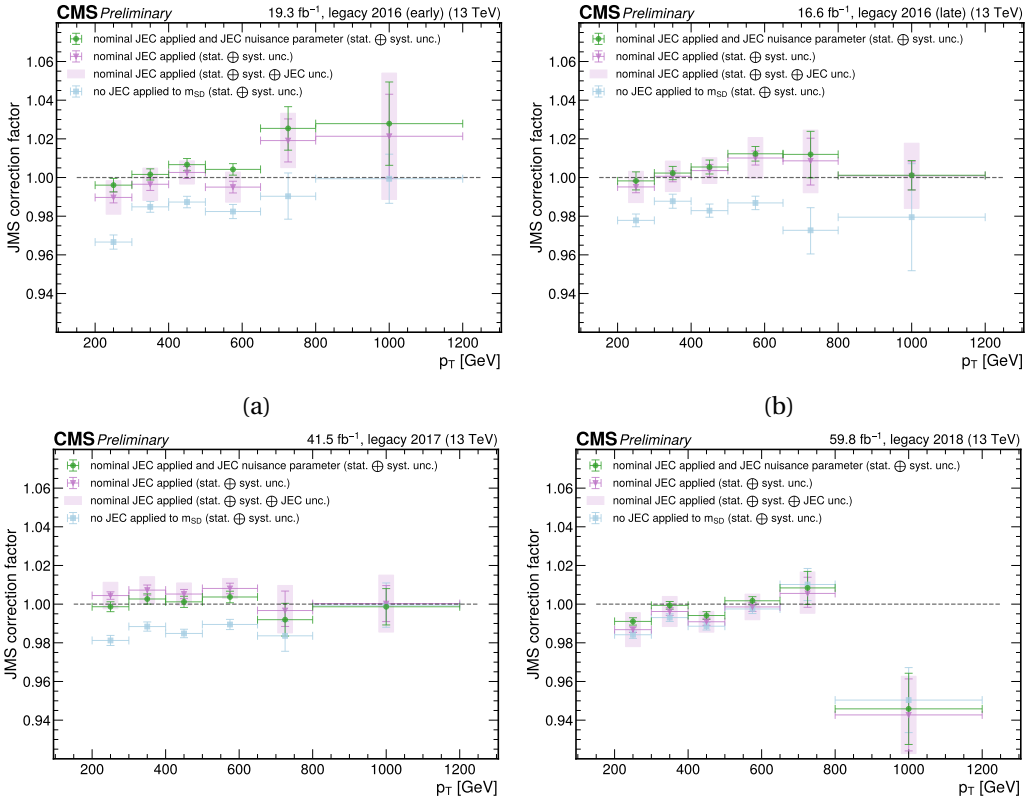


Figure 7.9: Comparison of p_T dependent jet mass scale correction factors for the scale of top jets and W jets when using substructure variables for the jet tagging. The correction factors are derived simultaneously in the semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets sample and with one common mass scale parameter for W jets and t jets per p_T bin. From the top left to the bottom right the plots show the separate results of each period of data-taking: early 2016, late 2016, 2017 and 2018. The purple triangles show the common correction factor c_{JMS} measured in the combined fit of both samples, same as the purple triangles in Figure 7.7. The purple band shows the result of the fits, where the pre-fit templates have been constructed using $\pm 1\sigma$ of the JEC uncertainty instead of the nominal JEC, added in quadrature to the purple triangles. The green circles show the resulting jet mass scale correction derived with the fit that includes a dedicated nuisance parameter accounting for the JEC uncertainty. The blue squares show the result from fits, where the jet p_T has been corrected with the JEC, but the soft drop mass m_{SD} has not.

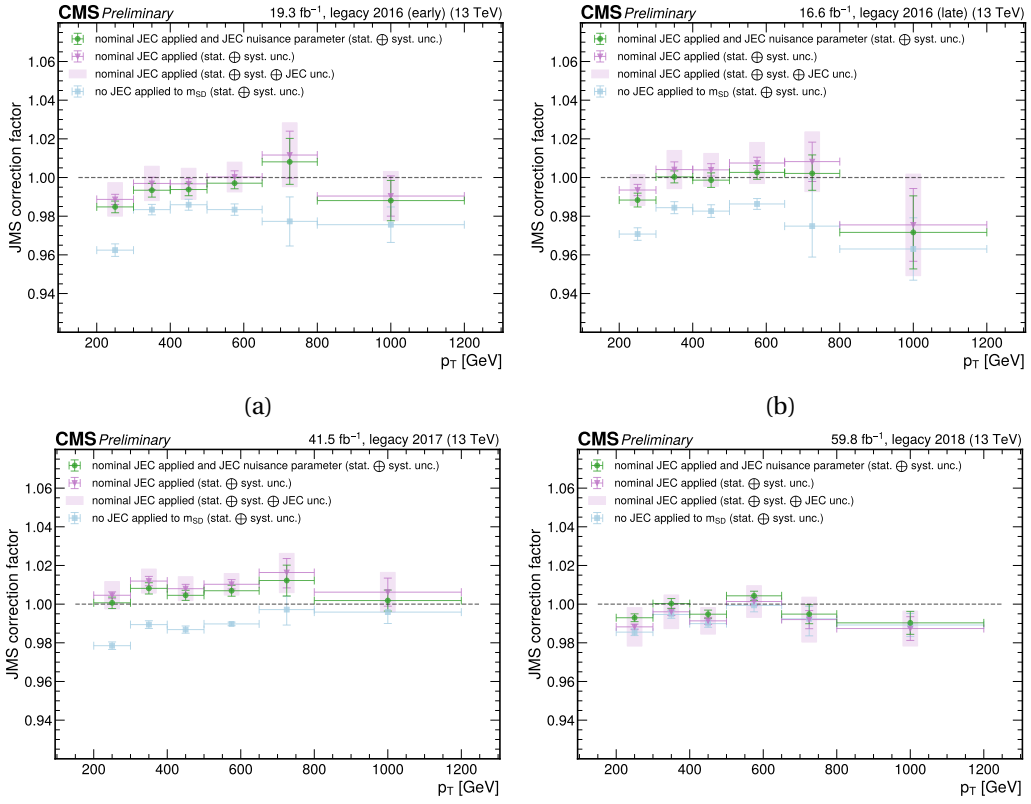


Figure 7.10: Comparison of p_{T} dependent jet mass scale correction factors for the scale of top jets and W jets when using ParticleNet discriminators for the jet tagging. The correction factors are derived simultaneously in the semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets sample and with one common mass scale parameter for W jets and t jets per p_{T} bin. From the top left to the bottom right the plots show the separate results of each period of data-taking: early 2016, late 2016, 2017 and 2018. The purple triangles show the common correction factor c_{JMS} measured in the combined fit of both samples, same as the purple triangles in Figure 7.8. The purple band shows the result of the fits, where the pre-fit templates have been constructed using $\pm 1\sigma$ of the JEC uncertainty instead of the nominal JEC, added in quadrature to the purple triangles. The green circles show the resulting jet mass scale correction derived with the fit that includes a dedicated nuisance parameter accounting for the JEC uncertainty. The blue squares show the result from fits, where the jet p_{T} has been corrected with the JEC, but the soft drop mass m_{SD} has not.

8 Measurement of the jet mass distribution of boosted W bosons

The second analysis presented in this thesis is the measurement of the distribution of the jet mass of hadronically decaying W bosons on particle-level in data. For this, a two-dimensional maximum-likelihood unfolding of the jet transverse momentum p_{T} and soft drop mass m_{SD} is performed in a fit to the full Run II data. In the following first the definition of the measurement phase-space is described, then the unfolding procedure and finally the results are presented.

8.1 Phase space definition

The measurement is performed in the phase space defined by selection criteria on detector-level and particle-level. The detector-level selection is the same as the one for the fully-hadronic $W(q\bar{q})$+jets sample described in 6. The criterion on the transverse momentum of the reconstructed jet is raised to $p_{\mathrm{T}}^{\mathrm{reco}} > 575\,\mathrm{GeV}$. The tagging criteria using $N_2^{\beta=1,\mathrm{DDT}}$ and the ParticleNet$^{\mathrm{DDT}}$ tagger is only used for the estimation of the QCD multijet background as described in Section 6.4. Both the pass-region and the fail-region contribute to the measurement phase space.
The particle-level selection consists of two criteria: there must be at least one AK8 jet and it must have a transverse momentum $p_{\mathrm{T}}^{\mathrm{ptcl}}$ of at least 500 GeV.

8.2 Binning definition

The binning used for the transverse momentum on particle-level $p_{\mathrm{T}}^{\mathrm{ptcl}}$ is chosen to be similar to the one used for the $p_{\mathrm{T}}^{\mathrm{reco}}$ on detector-level in the jet mass calibration as defined in Section 6. The bin edges are $[500, 650, 800, 1200, \infty]$. On detector-level the binning of $p_{\mathrm{T}}^{\mathrm{reco}}$ is chosen to have half the bin width of the binning on particle-level, but start at 575 GeV, thus using the following bin edges: $[575, 650, 725, 800, 1000, 1200, \infty]$. With this choice for the p_{T} binning on particle-level and detector-level, the bin width is well above the experimental resolution of the order 5–10% and the number of bins is small enough to have good statistics in all bins. Figure 8.1a shows the p_{T} distribution in 2018 simulation of $W(q\bar{q})$+jets on particle-level and detector-level with equidistant binning with a bin width of 30 GeV, while Figure 8.1b shows the same distributions with the binning used for the measurement. In both the full unfolding selection is applied and all scale factors are applied. The peak at $p_{\mathrm{T}}^{\mathrm{reco}}$ around 600 GeV is caused by the trigger scale factor.

8.2.1 Detector-level jet mass correction and $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ binning

Before the binning definitions for p_{T} and m_{SD} are optimized the agreement of soft drop mass on detector-level $m_{\mathrm{SD}}^{\mathrm{reco}}$ and on particle-level $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ is studied by constructing the two-dimensional

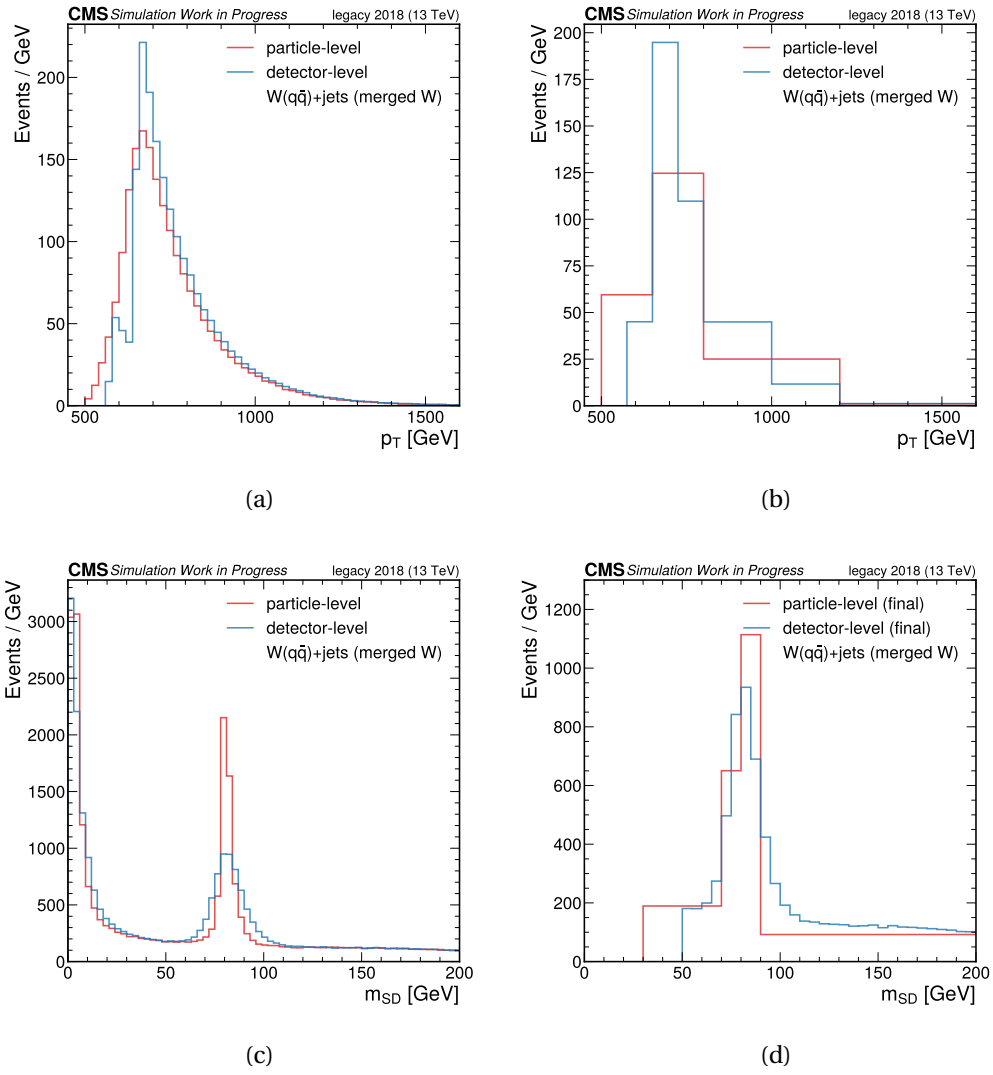
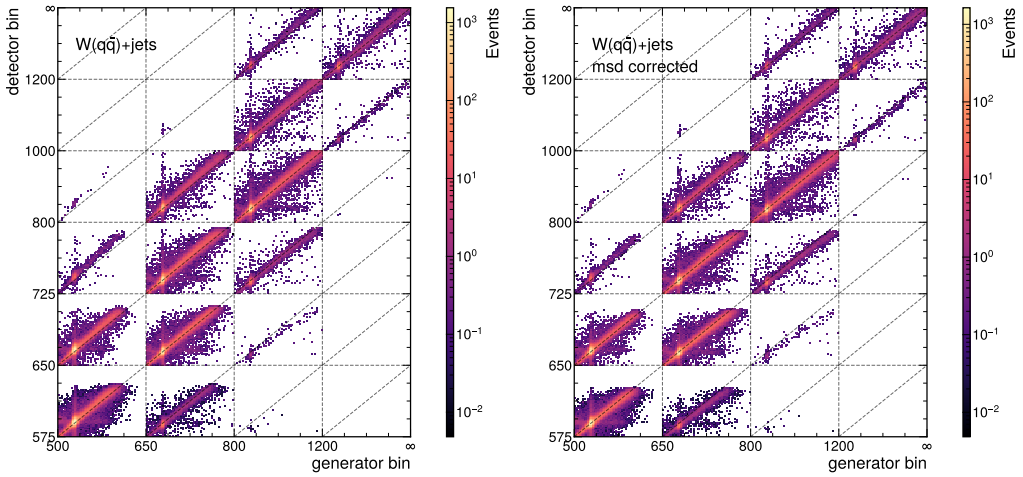


Figure 8.1: Distribution of p_{T} on the top row and m_{SD} on the bottom row for the jets on particle-level as red lines and for the jets on detector-level as blue lines. The distributions are derived from the fully-hadronic $W(q\bar{q})$+jets simulation using 2018 detector conditions. Both the particle-level and the detector-level selection criteria including the matching between detector-level jet and the generator W decay products are required. The left column shows the distribution with fine equidistant binning, while the right plot shows the distribution using the final binning, which is used in the unfolding. For both histograms of $m_{\mathrm{SD}}^{\mathrm{reco}}$ the correction factor explained in Section 8.2.1 is applied.



(a) (b)

Figure 8.2: Two-dimensional response matrix $\mathbf{R}(m_{\mathrm{SD}}^{\mathrm{ptcl}}, m_{\mathrm{SD}}^{\mathrm{reco}}, p_{\mathrm{T}}^{\mathrm{ptcl}}, p_{\mathrm{T}}^{\mathrm{reco}})$ filled with $W(q\bar{q})$+jets simulated events of all data-taking years corresponding to the full Run II luminosity of $\mathscr{L}_{\mathrm{int}} \approx 137\,\mathrm{fb}^{-1}$, that fulfill both the particle-level criteria and detector-level criteria, as well as the matching criteria described in Section 8.1. The right plot shows the response matrix with the additional jet mass correction factor applied on $m_{\mathrm{SD}}^{\mathrm{reco}}$, and the left plot shows the same, but without the correction factor applied. For both plots, the x-axis shows the bins on particle-level (generator bins), where the major ticks mark with label mark the $p_{\mathrm{T}}^{\mathrm{ptcl}}$ edges and the minor edges show the repeating $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins. The y-axis shows the same but on detector-level (detector bin).For each ($p_{\mathrm{T}}^{\mathrm{ptcl}}$,$p_{\mathrm{T}}^{\mathrm{reco}}$) bin the dashed line shows the diagonal in the ($m_{\mathrm{SD}}^{\mathrm{ptcl}}$,$m_{\mathrm{SD}}^{\mathrm{ptcl}}$)-plane as a visual aid, to show how diagonal the respective response matrix is.

response matrix $R(m_{\mathrm{SD}}^{\mathrm{reco}}, m_{\mathrm{SD}}^{\mathrm{ptcl}}, p_{\mathrm{T}}^{\mathrm{ptcl}}, p_{\mathrm{T}}^{\mathrm{reco}})$ with the final p_{T} binning definition, but with fine m_{SD} binning with a bin width of 5 GeV both on detector-level, as well as on particle-level. This matrix is shown as a two-dimensional histogram in Figure 8.2a. Here the x-axis corresponds to the particle-level bins (also called generator bins), where the major ticks with labels mark the $p_{\mathrm{T}}^{\mathrm{ptcl}}$ edges and the minor ticks mark the repeating $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins. The y-axis corresponds to the same but on detector-level. The grey dashed lines in each ($p_{\mathrm{T}}^{\mathrm{ptcl}}$,$p_{\mathrm{T}}^{\mathrm{reco}}$) bin show the diagonal in the ($m_{\mathrm{SD}}^{\mathrm{ptcl}}$,$m_{\mathrm{SD}}^{\mathrm{reco}}$)-plane, and demonstrate, that the response matrix is not perfectly diagonal. This is expected since the detector-level soft drop mass is not calibrated perfectly. To counter this in order to achieve a more stable unfolding fit, a p_{T} dependent correction factor for the detector-level soft drop mass $m_{\mathrm{SD}}^{\mathrm{reco}}$ is measured in the simulated $W(q\bar{q})$+jets events and applied to $m_{\mathrm{SD}}^{\mathrm{reco}}$ both in simulation and in data. The correction factor is measured as the inverse of the mean value of the response $\langle m_{\mathrm{SD}}^{\mathrm{reco}} / m_{\mathrm{SD}}^{\mathrm{ptcl}} \rangle$ in each $p_{\mathrm{T}}^{\mathrm{reco}}$ bin. The correction factor is measured for each year of data-taking separately. The final correction factors are shown in Figure 8.3. The resulting response matrix is shown in Figure 8.2b. Here the response matrix is more diagonal, and the unfolding fit is expected to be more stable.

The binning of m_{SD} on particle-level is chosen such that the stability and purity are approximately above 50% for all bins. The stability is the fraction of events in a given bin on particle-level, that are reconstructed in the same bin on detector-level. Similarly, purity is the fraction of events that are reconstructed in a given bin on detector-level and originate from the same

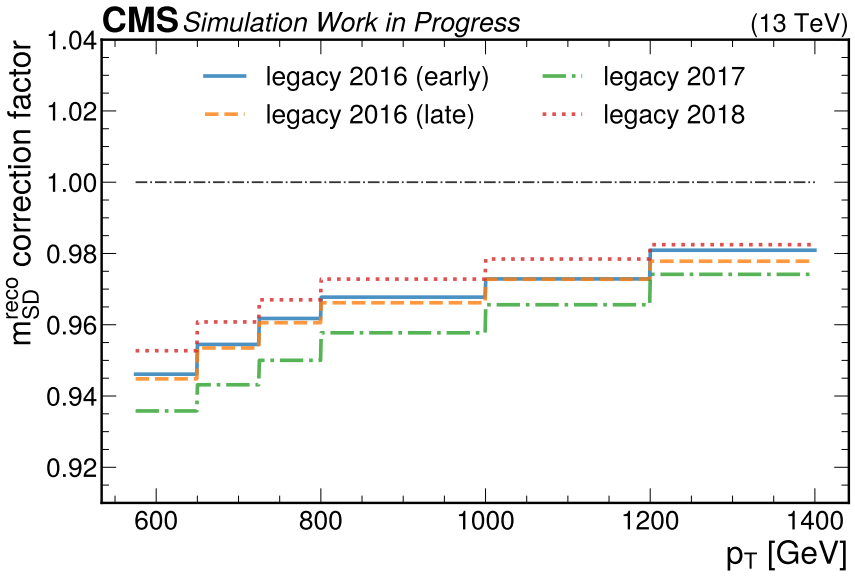


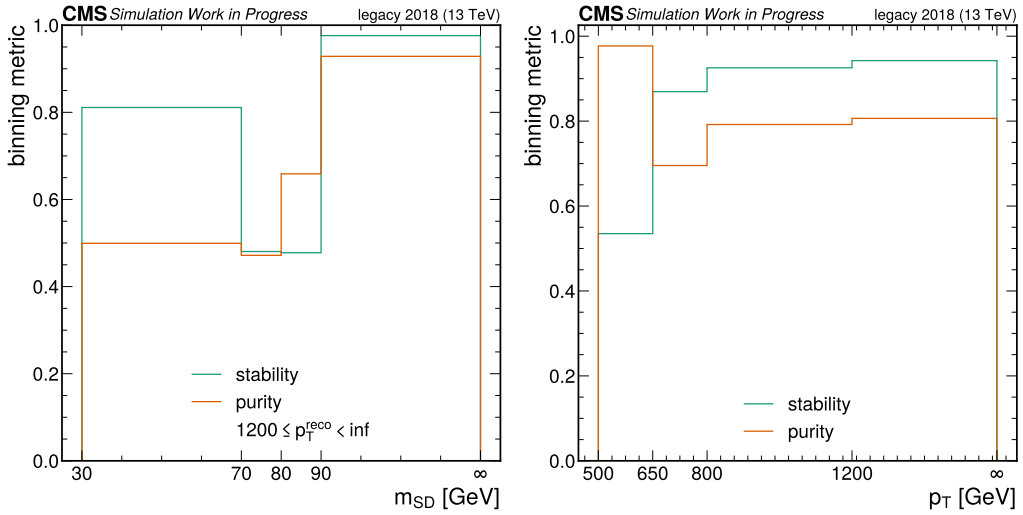
Figure 8.3: Correction factors for the soft drop mass on detector-level $m_{\mathrm{SD}}^{\mathrm{reco}}$ as a function of the transverse momentum on generator-level $p_{\mathrm{T}}^{\mathrm{ptcl}}$. The correction factors are measured in simulated $W(q\bar{q})$+jets events with 2016 (early), 2016 (late), 2017 and 2018 detector conditions. The correction factors are applied to $m_{\mathrm{SD}}^{\mathrm{reco}}$ both in simulation and in data.

bin on particle-level. Starting with a fine binning of 0.1 GeV a bi-directional scan starting from $m_{\mathrm{SD}} = 80\,\mathrm{GeV}$ is performed along the m_{SD} axis, merging bins until stability and purity above 50% or the end of the scanned m_{SD} range ($10 - 300\,\mathrm{GeV}$) is reached. This scan was repeated for each bin of $p_{\mathrm{T}}^{\mathrm{reco}}$ and for each year of data-taking separately. The final bin edges for $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ were chosen to be: $[30, 70, 80, 90, \infty]$. This is consistent with most scan results while maintaining sufficient statistics in each bin, and having at least two bins to measure the peak region around the W mass. The binning of $m_{\mathrm{SD}}^{\mathrm{reco}}$ is chosen as equidistant bins in the range $50 - 300\,\mathrm{GeV}$, with a bin width of 5 GeV, which is half the bin width of the binning on particle-level. This is the same binning as used in the jet mass calibration as described in Section 7, thus the same order Bernstein polynomials for the background estimate can be used.

Figure 8.1c shows the m_{SD} distribution in 2018 simulation of $W(q\bar{q})$+jets on particle-level and detector-level with equidistant binning with a bin width of 3 GeV, while Figure 8.1d shows the same distributions with the binning used for the measurement. In both the full unfolding selection is applied and all scale factors including the $m_{\mathrm{SD}}^{\mathrm{reco}}$ correction described above are applied. The choice of the lowest bin edge for the $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ binning at 30 GeV removes most of the events from the phase space where the correct jet was well reconstructed on detector-level but not on particle-level. These events are considered *fakes* and are treated as background.

The stability and purity of the final $p_{\mathrm{T}}^{\mathrm{ptcl}}$ and $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ binning are shown in Figure 8.4. For the calculation of both metrics, the binning chosen for the particle-level was used for the detector-level as well. Both stability and purity for the $p_{\mathrm{T}}^{\mathrm{ptcl}}$ binning using 2018 simulation shown in Figure 8.4b are above 50% across the whole p_{T} range. This is shown for all other years as well and can be found in Figure E.1 in the appendix. Figure 8.4a shows the purity and stability for the final $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ binning for the bin $p_{\mathrm{T}}^{\mathrm{reco}} > 1200\,\mathrm{GeV}$ using 2018 simulation. Both purity and stability

are above 40% for all bins. For most of the other $p_{\mathrm{T}}^{\mathrm{reco}}$ bins and the other years, the purity and stability are above 50% and can be found in Figures E.2 and E.3 in the appendix.



(a) (b)

Figure 8.4: Distribution of stability and purity of the final chosen particle-level binning for $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ (left) and $p_{\mathrm{T}}^{\mathrm{ptcl}}$ (right). While the right plot shows the $p_{\mathrm{T}}^{\mathrm{ptcl}}$ metrics for the whole range of m_{SD}, the left plot shows the $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ metrics only for one of the $p_{\mathrm{T}}^{\mathrm{reco}}$ bins

8.3 Unfolding

In an unfolding procedure, the measured distribution on detector-level is corrected for detector effects to obtain the distribution on particle-level in data. This is done by inverting the detector response matrix **R**, which is obtained from simulation, and applying the inverted matrix to the measured distribution in data, after subtracting the background contribution [138]. In the maximum likelihood approach the background subtraction and matrix inversion is done in one step in a fit to data. The response matrix holds the information about the probabilities that an event in a given bin j on particle-level is reconstructed in a bin i on detector-level. The total number of expected events in bin i on detector-level is given by:

$$y_i = \sum_{j=1}^{M} \tilde{R}_{ij} \cdot x_j + b_i, \tag{8.1}$$

with the probability response matrix $\tilde{R}_{ij}$, the expectation values of the true number of events in the bins j on particle-level x_j and the total number of expected background events in the i-th bin on detector-level b_i [126]. In a different representation using signal strength modifiers μ_j, which are the ratios of the expected number of events on particle-level predicted by simulation over the number of events predicted by the fit, the total number of expected events in bin i on detector-level is then given by:

$$y_i = \sum_{j=1}^{M} R_{ij} \cdot \mu_j \cdot x_j + b_i. \tag{8.2}$$

Here R_{ij} is the response matrix holding the number of events that fall in a bin i on detector-level given that they were in a bin j on particle-level. If the response matrix can be inverted the solution for the expectation values of the true number of events on particle-level x_j is given by:

$$\mu_j = \sum_{i=1}^{N} R_{ji}^{-1} \cdot (y_i - b_i). \tag{8.3}$$

This solution is obtained by minimizing the negative log-likelihood function:

$$-2\ln \mathscr{L}(\vec{\mu}) = -2\sum_{i}^{N} \ln P(n_i, y_i), \tag{8.4}$$

where n_i is the total number of data events in the i-th detector-level bin and $P(n_i, y_i)$ refers to the Poisson distribution with .

The response matrix R_{ij} is obtained from simulation. It is constructed by filling the number of events in the i-th bin on detector-level given that they were in the j-th bin on particle-level. Only events that pass the detector-level selection and the particle-level selection as described in Section 8.1 are considered.

Furthermore, events are only filled in the response matrix if the matching criteria between the W daughter particles on generator-level and the AK8 jet on detector-level (*merged W*, see Section 6.1.1) is fulfilled. If there is a jet both on detector-level and particle-level that fulfill the respective selection requirements an additional requirement must be met to ensure that the jet on particle-level is matched to the jet on detector-level. This is done by requiring that the

angular distance $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ between the two jets is smaller than 0.4.
The response matrix constructed with the binning and the $m_{\mathrm{SD}}^{\mathrm{reco}}$ correction factor as described in Section 8.2 is shown in Figure 8.5, using the full Run II set of simulated $W(q\bar{q})$+jets events. The performance of the unfolding and more specifically the inversion of the response matrix is limited by how well posed the system of equations following from Equation 8.3 is. This is quantified by the condition number of the response matrix, which is defined as the ratio $\mathrm{cond}(\mathbf{R}) = \frac{\sigma_1}{\sigma_n}$, where σ_1 is the largest and σ_n is the smallest singular value of the matrix. The singular values are obtained by singular value decomposition (SVD) of the response matrix $\mathbf{R} \in \mathbb{R}^{m\times n}$:

$$\mathbf{R} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} = \sum_{j=1}^{n} \sigma_j \mathbf{u}_j \mathbf{v}_j^{\mathrm{T}}, \tag{8.5}$$

where $\mathbf{U} \in \mathbb{R}^{m\times n}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are matrices with orthonormal rows and Σ is diagnoal matrix holding the singular values $\sigma_1, ..., \sigma_n$ [139]. The dimension m corresponds to the detector-level bins and n corresponds to the particle-level bins. The condition number of the response matrix shown in Figure 8.5 is 10151.16. The large condition number shows that the problem is ill-posed and regularisation is needed in the unfolding procedure. The choice of regularisation scheme is described in Section 8.3.1. One possible reason for this large condition number could be the significant off-diagonal features. These are, at least partially, arising from events, in which the recoiling quark/gluon jet is selected rather than the W jet. The selection of the W is not trivial, since the kinematic properties of the W jet and the recoiling quark/gluon jet are very similar at Born-level.

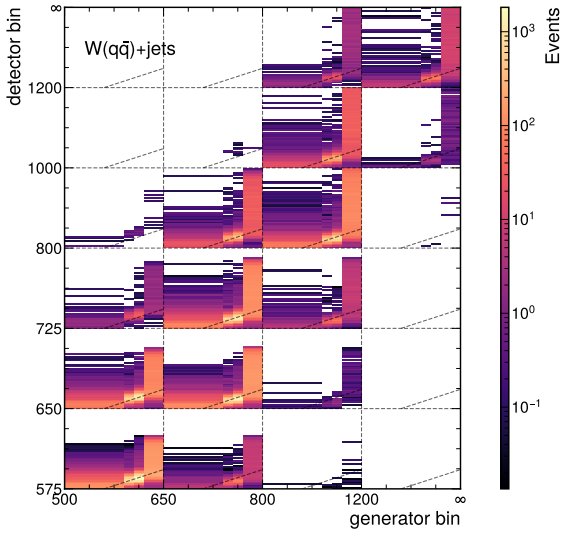


Figure 8.5: Final response matrix filled with events in the measurement phase space using the final binning and with the $m_{\mathrm{SD}}^{\mathrm{reco}}$ corrections applied, described in Section 8.2. The matrix holds simulated events from all Run II data-taking years, corresponding to an integrated luminosity of $\mathscr{L}_{\mathrm{int}} \approx 137\,\mathrm{fb}^{-1}$.

From the above, it follows that in the maximum-likelihood fit approach of the unfolding the response matrix enters in the form of one-dimensional histograms of the fit discriminant, which in this case is the soft drop mass on detector-level $m_{\mathrm{SD}}^{\mathrm{reco}}$. The $p_{\mathrm{T}}^{\mathrm{reco}}$ axis is accounted for by

combining the likelihood functions of the different $p_{\mathrm{T}}^{\mathrm{reco}}$ bins and tagger regions, which means in terms of Equation 8.4, that i runs over every ($m_{\mathrm{SD}}^{\mathrm{reco}}$, $p_{\mathrm{T}}^{\mathrm{reco}}$, [pass, fail]) combinations. That way the particle-level axes can be encoded in the one-dimensional $m_{\mathrm{SD}}^{\mathrm{reco}}$ histograms, having one histogram per particle-level bin. Thus the total likelihood function for a given dataset follows from the product of the likelihood functions of the individual $p_{\mathrm{T}}^{\mathrm{reco}}$ bins and tagger regions:

$$\mathscr{L}(\vec{\mu},\vec{\theta}) = \prod_{k\in[\mathrm{pass,fail}]} \prod_{i_{p_{\mathrm{T}}^{\mathrm{reco}}}}^{N_{p_{\mathrm{T}}^{\mathrm{reco}}}} \prod_{i_{m_{\mathrm{SD}}^{\mathrm{reco}}}}^{N_{m_{\mathrm{SD}}^{\mathrm{reco}}}} P(y_{i_{p_{\mathrm{T}}^{\mathrm{reco}}},i_{m_{\mathrm{SD}}^{\mathrm{reco}}},k} | \sum_{j=1}^{16} \mu_j \cdot \mathbf{s}^j(i_{p_{\mathrm{T}}^{\mathrm{reco}}}, i_{m_{\mathrm{SD}}^{\mathrm{reco}}}, k, \vec{\theta}) + \mathbf{b}(i_{p_{\mathrm{T}}^{\mathrm{reco}}}, i_{m_{\mathrm{SD}}^{\mathrm{reco}}}, k, \vec{\theta})) \cdot \Pi(\vec{\theta}), \tag{8.6}$$

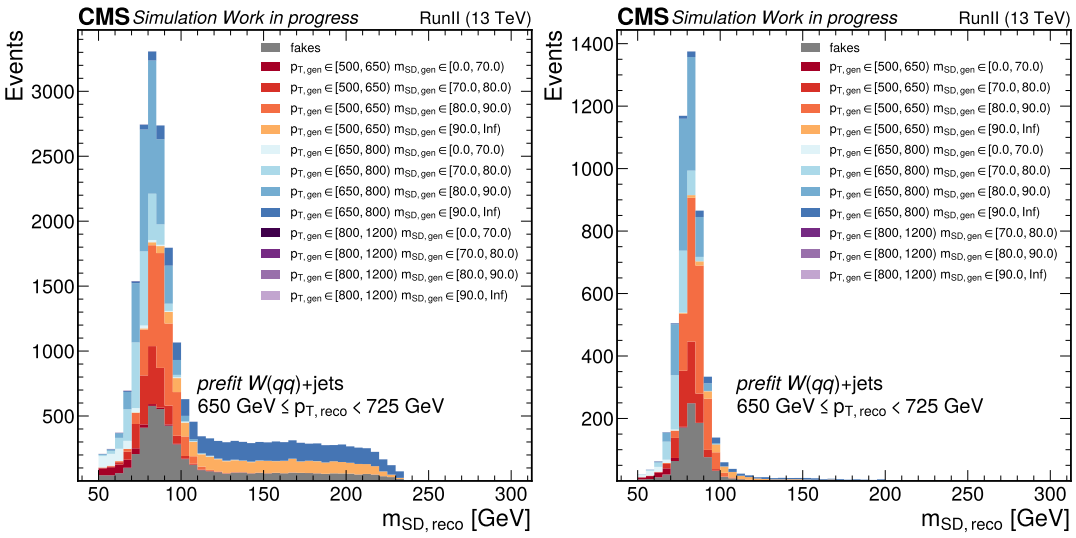
where j is running over the 16 particle-level bins, k is running over the tagger regions denoted as *pass* and *fail*, $i_{p_{\mathrm{T}}^{\mathrm{reco}}}$ is running over the $p_{\mathrm{T}}^{\mathrm{reco}}$ bins, $i_{m_{\mathrm{SD}}^{\mathrm{reco}}}$ is running over the $m_{\mathrm{SD}}^{\mathrm{reco}}$ bins and P is the poissonian probability to observe $y_{i_{p_{\mathrm{T}}^{\mathrm{reco}}},i_{m_{\mathrm{SD}}^{\mathrm{reco}}},k}$ events in data in the detector-level bin $(i_{p_{\mathrm{T}}^{\mathrm{reco}}}, i_{m_{\mathrm{SD}}^{\mathrm{reco}}})$ and the tagger region k given the expected number of events of a specific model configuration. The expected number of signal events in a given particle-level bin, detector-level bin and tagger region combination is given by $\mathbf{s}^j(i_{p_{\mathrm{T}}^{\mathrm{reco}}}, i_{m_{\mathrm{SD}}^{\mathrm{reco}}}, k, \vec{\theta})$, where $\vec{\theta}$ denotes the vector of nuisance parameters in the final unfolding fit. Similarly $\mathbf{b}(i_{p_{\mathrm{T}}^{\mathrm{reco}}}, i_{m_{\mathrm{SD}}^{\mathrm{reco}}}, k, \vec{\theta})$ denotes the expected number of background events in a detector-level bin and tagger region combination. The nuisance parameters affect the expected number of signal and background events and are constrained in the likelihood via Gaussian constraints $\Pi(\vec{\theta})$. The fit is performed in all four eras of data-taking (early 2016, late 2016, 2017 and 2018) simultaneously by using the product of the individual likelihoods as a combined likelihood, with one common set of parameters of interest $\vec{\mu}$. The nuisance parameters are treated either fully correlated as well, or fully uncorrelated, as described in Section 6.5.

The fit uses one signal strength modifier μ_j per particle-level bin as a parameter of interest, scaling the contribution of each particle-level bin simultaneously in all detector-level bins. Figure 8.6 shows the templates of the *merged W* signal processes in the first $p_{\mathrm{T}}^{\mathrm{reco}}$ bin $650 \leq p_{\mathrm{T}}^{\mathrm{reco}} < 725\,\mathrm{GeV}$, constructed from $W(q\bar{q})$+jets simulation of all years of data-taking scaled to full Run II luminosity ($\mathscr{L}_{\mathrm{int}} \approx 137\,\mathrm{fb}^{-1}$). The templates are shown for both the *fail* and *pass* region of the $N_2^{\beta=1,\mathrm{DDT}}$ tagger in Figure 8.6a and Figure 8.6b respectively.

Acceptance and efficiency effects originating from events exiting the fiducial phase space by failing either the detector-level, particle-level criteria or the matching are not included in the response matrix R_{ij}. They are instead handled separately in a factorized approach.

The events that both pass the detector-level criteria and are categorized as *merged W*, but do not have a matched jet on particle-level that fulfills the particle-level criteria, are considered *fakes* and are treated as background in the fit. In Figure 8.6 the *fakes* are shown as the grey part of the stacked histograms.

Events that fail the detector-level selection criteria but pass the particle-level selection criteria are accounted for by dividing the final unfolded distribution by the acceptance. The acceptance $\mathscr{A}$ is defined with the ratio of the number of events passing the particle-level selection but failing the detector-level selection which is measured in each particle-level bin in simulation:



(a) (b)

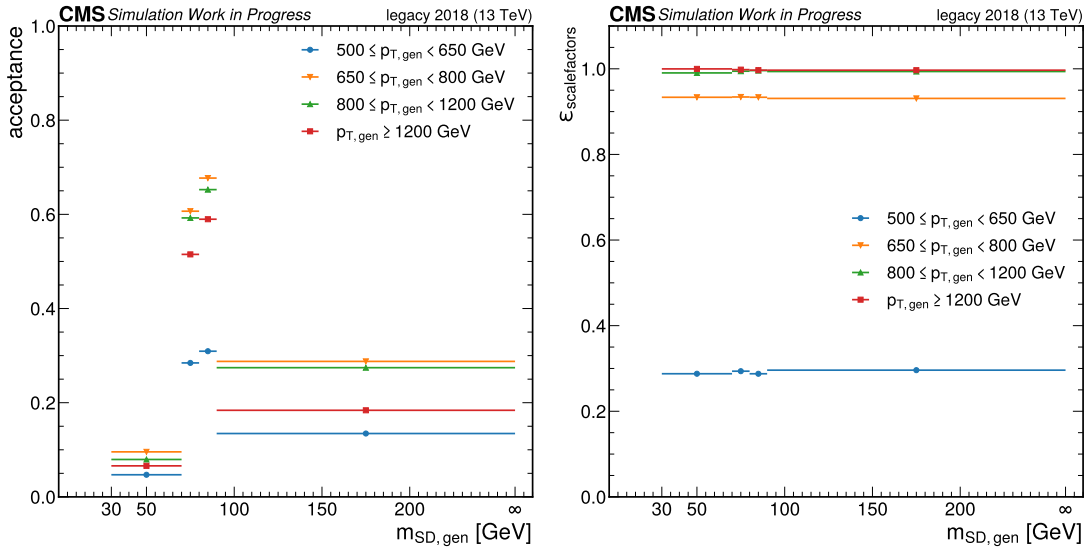
Figure 8.6: Templates of the *merged W* signal processes in the first p_T^{reco} bin $650 \leq p_T^{reco} < 725\,GeV$, constructed from $W(q\bar{q})$+jets simulation normalized to full Run II luminosity. The events are required to pass or fail the $N_2^{\beta=1,DDT}$ tagger in the right or left plot respectively. The contribution from each $(p_T^{ptcl}, m_{SD}^{ptcl})$ particle-level bin is represented by a different color, as denoted in the legend. The events that pass the detector-level criteria, but fail any particle-level criteria are considered *fakes* and shown as the grey part of the histogram.

$$\mathscr{A} = 1 - \frac{N(\text{pass particle} - \text{level} \wedge \text{fail detector} - \text{level})}{N(\text{pass particle} - \text{level})}. \tag{8.7}$$

The measured acceptance is shown for 2018 simulation in Figure 8.7a, using the final binning and the m_{SD}^{reco} correction factor. The first and last bin in m_{SD}^{ptcl} has low acceptance of 5 – 20%, especially in the first and last p_T^{ptcl} bin. The central m_{SD}^{ptcl} bins in the W mass peak region have a much larger acceptance of 30 – 70%. This shows how much phase-space extrapolation is involved when deriving the final unfolded jet mass distribution. The acceptance correction does not account for any scale factor or the treatment of the HEM15/16 issue described in 6. To account for all per-event scale factors, the acceptance is multiplied by the scale factor efficiency $\varepsilon_{scalefactors}$, which is defined as the ratio of the number of events passing the particle-level selection, the matching and the detector-level selection, with and without all scale factors applied to the events. The scale factor efficiency is measured in simulation in each particle-level $(p_T^{ptcl}, m_{SD}^{ptcl})$ bin for each year of data-taking. Figure 8.7b shows the measured efficiency in 2018 simulation. In the two highest p_T^{ptcl} bins the efficiency is almost 100%, while for the lowest and second to lowest p_T^{ptcl} bin efficiency is around 30% and 94% respectively. The majority of the difference stems from the trigger scale factors applied to the events as described in 6.1.

The final unfolded distribution is obtained by dividing the unfolded distribution by the product of the acceptance and the scale factor efficiency:

By maximizing the likelihood function, the maximum-likelihood estimator of the signal strength modifier μ_j is determined for each particle-level bin as $\hat{\mu}_j$. The unfolded distribution in data in the fiducial phase space only with criteria on particle-level ($m_{SD}^{ptcl} > 30\,GeV$ and $p_T^{ptcl} > 500\,GeV$) is then obtained by multiplying each column of the response matrix as it is shown in Figure 8.5



(a) (b)

Figure 8.7: Distribution of the acceptance $\mathscr{A}$ and scale factor efficiency $\varepsilon_{\text{scalefactors}}$ measured in the $W(q\bar{q})$+jets signal simulation sample using 2018 detector conditions. Both are measured in the same bins of $m_{\text{SD}}^{\text{ptcl}}$ and $p_{\text{T}}^{\text{ptcl}}$.

with the corresponding $\hat{\mu}_j$ and applying the acceptance and efficiency corrections discussed above by dividing each measured bin by the corresponding acceptance and efficiency:

$$N(m_{\text{SD}}^{\text{unfold}}, j) = \frac{\hat{\mu}_j}{\mathscr{A}^j \cdot \varepsilon^j_{\text{scalefactors}}} \cdot \sum_i \cdot R_{ij}. \tag{8.8}$$

To derive the unfolded jet mass distribution in terms of a differential cross section in the fiducial particle-level phase space the number of events is divided further by the integrated luminosity $\mathscr{L}_{\text{int}} \approx 137\,\text{fb}^{-1}$ and the bin width $(\Delta m_{\text{SD}})_j$:

$$\left(\frac{d\sigma}{dm_{\text{SD}}}\right)_j = \frac{1}{\mathscr{L}_{\text{int}} \cdot (\Delta m_{\text{SD}})_j} \cdot N(m_{\text{SD}}^{\text{unfold}}, j). \tag{8.9}$$

8.3.1 Regularisation of multidimensional distribution

As discussed above, the condition number response matrix is large, thus the inversion problem is ill-posed. For this reason, the use of regularisation is considered. Multiple regularisation schemes constraining the solution to be smooth along the m_{SD} and p_{T} axes are tested. The regularisation of the multidimensional distribution is performed using *Tikhonov* regularisation [140]. For this penalty terms $P(\vec{\mu})$ are added to the likelihood function:

$$-2\ln\mathscr{L} = -2\ln\mathscr{L}(\vec{\mu}) + \tau P(\vec{\mu}), \tag{8.10}$$

where $\vec{\mu}$ is the vector of the signal strength modifiers of the particle-level bins and τ is the regularisation strength. For the form of the penalty terms the Singular Value Decomposition (SVD) approach [141] is used, which will dampen fluctuations in the curvature of the signal strength modifiers of neighboring particle-level bins, by using derivatives. With 4 equidistant

bins in one dimension on particle-level the penalty terms P read as:

$$P = \vec{\mu} \cdot C, \tag{8.11}$$

with the curvature matrix C given by:

$$C = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \tag{8.12}$$

Here the first and last lines correspond to the first-order derivatives

$$(x_{j_2} - x_{j_1}) = 1 \cdot x_{j_2} + (-1) \cdot x_{j_1}, \tag{8.13}$$

while the second and third correspond to the second-order derivatives

$$(x_{j_3} - x_{j_2}) - (x_{j_1} - x_{j_2}) = (1) \cdot x_{j_1} + (-2) \cdot x_{j_2} + (1) \cdot x_{j_3}. \tag{8.14}$$

Since the binning of the distributions on particle-level is non-uniform it can be more accurate to include information on the bin widths in the penalty term. This is achieved with the procedure adapted from [138] and explained in more detail in the appendix E.1. To find the optimal regularisation scheme and regularisation strength τ the procedure is repeated for different values of τ with each scheme and compared by the average global correlation coefficient $\bar{\rho} = \sum_i \rho_i / N$, where ρ_i is the correlation coefficient of the signal strength modifier μ_i of the i-th bin on particle-level and N is the number of bins. The correlation coefficient is defined as:

$$\rho_i = \sqrt{1 - \frac{1}{(\mathbf{V_{xx}}^{-1})_{ii} (\mathbf{V_{xx}})_{ii}}}, \tag{8.15}$$

with $\mathbf{V_{xx}}$ being the covariance matrix of the signal strength modifiers of the particle-level bins estimated from the fit. For each scheme, the regularisation strength τ is chosen such that the average global correlation coefficient is minimized.

The regularisation scheme for both tagging approaches is chosen to be the SVD regularisation with derivatives accounting for non-equidistant binning as described in E.1, constraining fluctuations in the curvature across μ_j along both the m_{SD} and p_{T} axes. This scheme was chosen over constraining only along the p_{T} axis, as it introduces less bias in the unfolded distribution, especially for the lowest bin in $m_{\mathrm{SD}}^{\mathrm{ptcl}}$. The procedure of testing for bias is described in Section 8.3.2 The results of the regularisation strength scan for the chosen scheme are shown in Figure 8.8. The minimum average global correlation coefficient is reached at $\delta = 1.45$ for the $N_2^{\beta=1,\mathrm{DDT}}$ tagger and at $\delta = 1.3$ for the ParticleNet$^{\mathrm{DDT}}$ tagger. The regularisation strength τ corresponding to the minimum average global correlation coefficient is then used for the final unfolding.

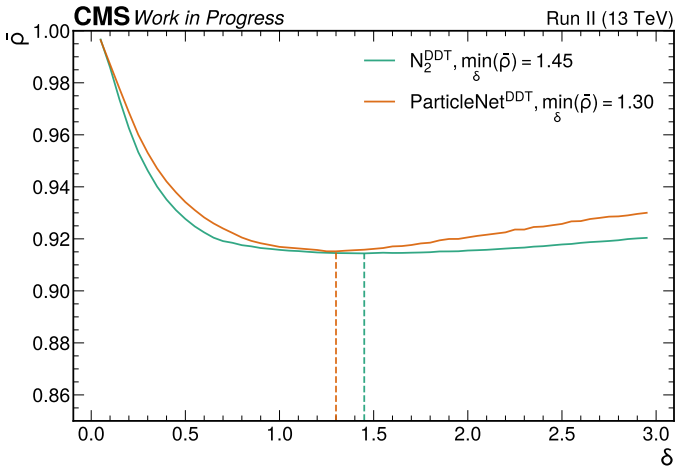


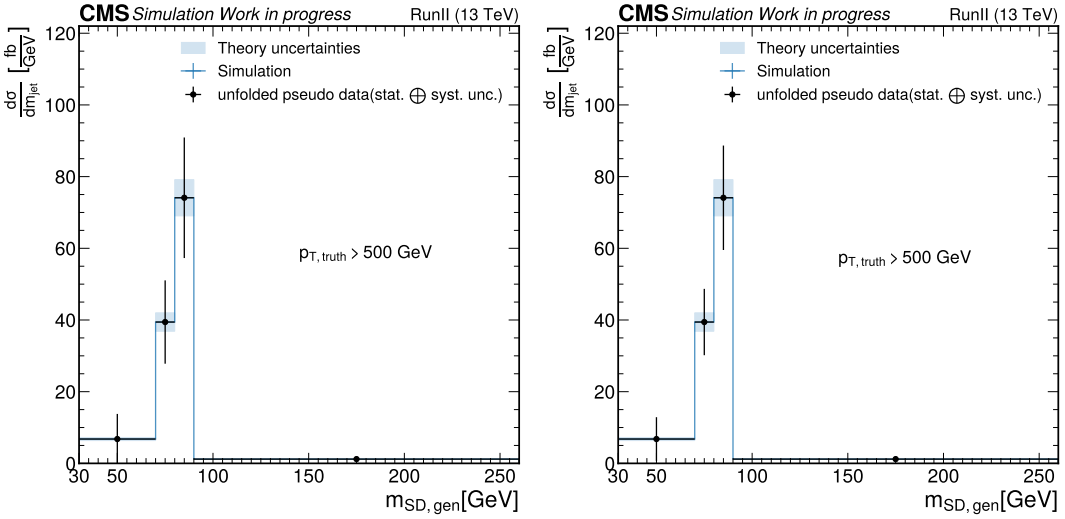
Figure 8.8: Average global correlation coefficient $\bar{\rho}$ computed for a scan of the regularisation strength parameter $\delta = \frac{1}{\sqrt{\tau}}$. The green line corresponds to fits using $N_2^{\beta=1,\mathrm{DDT}}$ as the W tagger for the QCD multijet background estimation. The minimum average global correlation coefficient is reached at $\delta = 1.45$. The orange line similarly corresponds to fits using the ParticleNet$^{\mathrm{DDT}}$, where the minimal global correlation coefficient is found to be at $\delta = 1.3$. The SVD regularisation with derivatives accounting for non-equidistant binning as described in E.1 is used for both tagging approaches.

8.3.2 Bias and Coverage test

To estimate the bias and over- or under-coverage introduced by the regularisation a series of tests are performed using pseudo-data in the simultaneous fit instead of real data. The pseudo-data are constructed from simulation in all regions except for the initial QCD multijet estimate in the *fail* tagger regions. In these control regions, the pseudo-data is constructed from Data with the simulated predictions of minor backgrounds and signal processes subtracted. Two different pseudo-data are constructed for each tagging approach ($N_2^{\beta=1,\mathrm{DDT}}$ and ParticleNet$^{\mathrm{DDT}}$): first using 100% of signal MC statistics in both the construction of pseudo-data and in the fit (in the following called *Asimov data*) and a *statistically independent* pseudo-data, where 60% of signal MC statistics is used in the construction of the pseudo-data and 40% is used in the fit. In the statistically independent setup, each sub-histogram is scaled up by 1/0.6 or 1/0.4 respectively, so each corresponds to $\mathscr{L}_{\mathrm{int}} \approx 137\,\mathrm{fb}^{-1}$. For the first test, the unfolding is performed with the full set of nuisances, the regularisation and the Asimov data. In this test, the pre-fit model is exactly equal to the pseudo-data, thus the extracted maximum-likelihood estimators of the signal strength modifiers $\hat{\mu}_j$ should yield unity. Any deviations could indicate systematic biases in the model (e.g. in the background estimation or nuisances). Figure 8.9 shows the resulting jet mass distribution in terms of the fiducial differential cross section $\frac{d\sigma}{dm_{\mathrm{SD}}}$ inclusive in $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins, derived using Equation 8.9. The unfolded Asimov data is shown as black markers and the prediction from simulation as the blue curve. The resulting unfolded distribution in pseudo-data closes perfectly with the simulated expectation.

The figures include theory uncertainties as the blue shaded band around the prediction from simulation. This uncertainty is estimated from the same sources connected to the V+jets higher-order corrections [3] used in the fit as nuisance parameters as described in Section 6.5. Hence it

is important to note, that the uncertainty source is also already accounted for in the uncertainty on the unfolded result (shown as the black uncertainty bars).



(a) (b)

Figure 8.9: Jet mass distribution of the sum of all p_T^{ptcl} bins resulting from the unfolding using $N_2^{\beta=1,DDT}$ (left) and ParticleNet tagger (right) in the background estimation using Asimov data. The unfolded Asimov data is shown as black markers, the blue line shows the true distribution from simulation with theory uncertainties as a shaded blue band.

To test the model against statistical fluctuation in the pseudo-data, the Asimov data are perturbed by generating 500 toy datasets and computing the pull distribution of the maximum-likelihood estimators $\hat{\mu}_j$ of the signal strength modifiers. The pull distribution is the histogram filled with the pull:

$$\text{pull}_i = \frac{\hat{\mu}_j^i - \mu}{\sigma_{\hat{\mu}_j}^i}, \tag{8.16}$$

with the expected signal strength per bin μ (here $\mu = 1$) the maximum-likelihood estimator $\hat{\mu}_j^i$ and its uncertainty $\sigma_{\hat{\mu}_j}^i$ for each toy dataset i. If there is no bias and perfect coverage the pull distribution should be a Gaussian with mean zero and standard deviation one. Any bias would shift the mean of the pull distribution, while a width significantly larger or smaller than one would indicate over-coverage or under-coverage respectively. Figure 8.10 shows the resulting pull distributions for the signal strength modifiers corresponding to the four m_{SD}^{ptcl} bins in the first p_T^{ptcl} bin $\hat{\mu}_{p_T^0,\, m_{SD}^0}$,$\hat{\mu}_{p_T^0,\, m_{SD}^1}$,$\hat{\mu}_{p_T^0,\, m_{SD}^2}$,$\hat{\mu}_{p_T^0,\, m_{SD}^3}$ the $N_2^{\beta=1,DDT}$ and ParticleNet^{DDT} taggers. The pulls are not 100% Gaussian distributed, but rather have a more pronounced tail towards negative pulls and are not centered perfectly at zero, but the mean of the distribution is consistent with zero within one standard deviation for most bins. The pull distributions for the other p_T^{ptcl} bins can be found in the appendix in Figures E.6 and E.7.

To further test for biases introduced by the regularisation the same tests are performed using the statistically independent pseudo-data. The unfolded distribution of the pseudo-data is shown in Figure 8.11 as black markers, which closes within uncertainties with the expected distribution on particle-level in simulation, which is shown as the blue curve. Figure 8.12 shows the pull

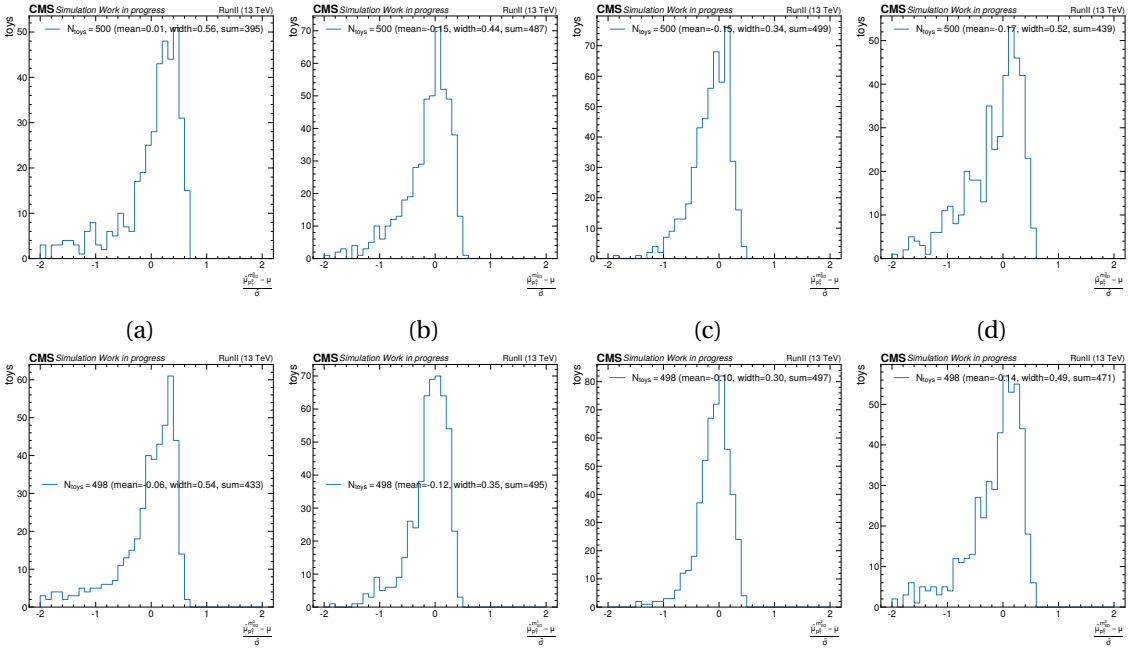
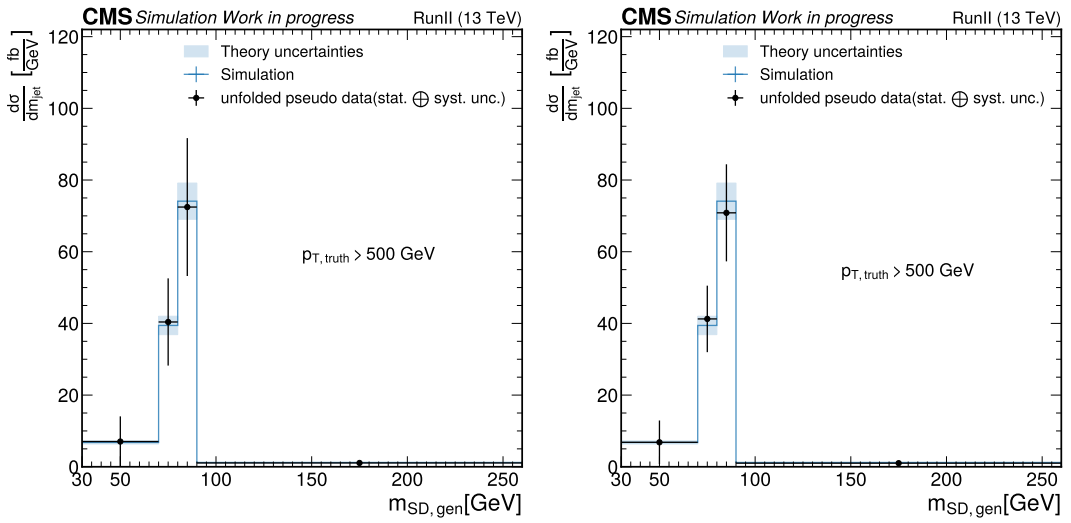


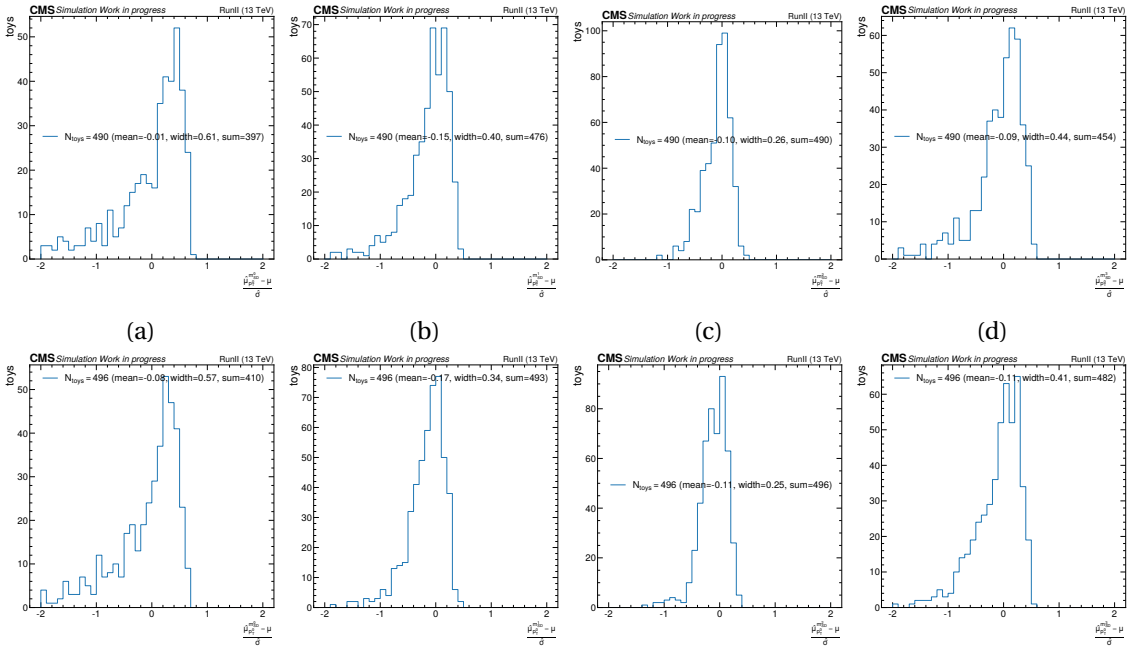
Figure 8.10: Pull distribution of the signal strength modifiers of the lowest $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin: $\hat{\mu}_{p_T^0,\, m_{\mathrm{SD}}^0}$,$\hat{\mu}_{p_T^0,\, m_{\mathrm{SD}}^1}$,$\hat{\mu}_{p_T^0,\, m_{\mathrm{SD}}^2}$,$\hat{\mu}_{p_T^0,\, m_{\mathrm{SD}}^3}$ from the left to the right using toys derived from Asimov data. The upper row shows pulls from fits using $N_2^{\beta=1,\mathrm{DDT}}$ as tagger for the background estimation and the lower shows the same but with using ParticleNet$^{\mathrm{DDT}}$ as tagger. The metrics in the legend are derived from either the total histogram (N_{toys}) or only from the visible range ([-2,2]) (mean, width and sum).

distributions of the signal strength modifiers of the first $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin from toy datasets derived from the statistically independent pseudo-data. Similarly to the pulls from the Asimov data, the pulls from the statistically independent pseudo-data are not perfectly centered around zero, but with the mean within one standard deviation around zero.
The toys for both the Asimov data as well as for the pseudo-data indicate that in the lowest $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bin there is a small bias towards higher values of $\hat{\mu}_{p_T^i,\, m_{\mathrm{SD}}^0}$, which is however not reflected in the fit to the nominal pseudo-dataset.



(a) (b)

Figure 8.11: Jet mass distribution of the sum of all p_T^{ptcl} bins resulting from the unfolding using $N_2^{\beta=1,DDT}$ (left) and ParticleNet tagger (right) in the background estimation using pseudo-data. The unfolded pseudo-data is shown as black markers, the blue line shows the true distribution from simulation with theory uncertainties as a shaded blue band.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 8.12: Pull distribution of the signal strength modifiers of the lowest p_T^{ptcl} bin: $\hat{\mu}_{p_T^0,\, m_{SD}^0}$, $\hat{\mu}_{p_T^0,\, m_{SD}^1}$, $\hat{\mu}_{p_T^0,\, m_{SD}^2}$, $\hat{\mu}_{p_T^0,\, m_{SD}^3}$ from the left to the right using toys derived from pseudo-data. The upper row shows pulls from fits using $N_2^{\beta=1,DDT}$ as tagger for the background estimation and the lower shows the same but with using ParticleNet^{DDT} as tagger. The metrics in the legend are derived from either the total histogram (N_{toys}) or only from the visible range ([-2,2]) (mean, width and sum).

8.4 Results

The final simultaneous fit to real Run II data is performed using the SVD regularisation scheme accounting for non-equidistant binning with derivative terms including distances between the particle-level bin centers. The regularisation strength parameter is chosen to be $\delta = 1.45$ when using $N_2^{\beta=1,\mathrm{DDT}}$ as W tagger in the background estimate and $\delta = 1.3$ when using ParticleNet$^{\mathrm{DDT}}$. Figure 8.13 shows the post-fit templates for both tagging approaches and for the detector-level transverse momentum bin $650 < p_{\mathrm{T}}^{\mathrm{reco}} \leq 725\,\mathrm{GeV}$, where the QCD multijet background is estimated using the data-driven approach described in 6.4. The signal efficiency is larger for the ParticleNet-based tagger, which reflects in larger signal yields in the templates in the pass region. The QCD multijet background efficiency is the same for both taggers since the DDT taggers are designed to select a constant rate of 5% for the QCD multijet background. The data to prediction agreement in the fail region is by construction very good since the QCD background is derived from the distribution in fail region data with the minor backgrounds and the signal templates subtracted. The agreement of the data and simulation in the pass region is not as good, but within the uncertainties. In the fit one transfer factor $R_{\mathrm{p/f}}$ per data-taking period is used to account for differences in tagging response depending on $p_{\mathrm{T}}^{\mathrm{reco}}$ and ρ_{SD} and the year of data-taking.

The post-fit transfer factors for 2018 simulation are shown in Figure 8.14 for both tagging strategies. On the left the transfer factor from the fit using $N_2^{\beta=1,\mathrm{DDT}}$ as tagger is shown and on the right the transfer factor from the fit using ParticleNet$^{\mathrm{DDT}}$. The definition of the transfer factors follows the description in Section 6.4. The maximum order of Bernstein-polynomial degrees is different for the two tagging approaches, which is reflected in the different resulting shapes. The maximum magnitude of differences the transfer factors have to cover is similar. The full set of post-fit templates, as well as the transfer factors for all data-taking periods and both tagging approaches, can be found in Appendix E.4.

The final extracted jet mass distribution is shown for each tagging approach separately in Figure 8.15 in terms of the double differential fiducial cross section $d\sigma/dm_{\mathrm{SD}}$ per $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin. The unfolded data is shown as black markers with the total uncertainty (statistical $\otimes$ systematic uncertainty) drawn as black bars. The blue curve shows the prediction from the simulation with an additional uncertainty band drawn as a shaded blue band. This theory uncertainty corresponds to the uncertainty sources connected to the V+jets higher order correction factors, exactly as described above for the plots showing the results using Asimov data and pseudo-data. The unfolded differential fiducial cross section in the W mass region ranges from 65 fb to 0.3 fb with uncertainties between $\approx 30\%$ to $\approx 100\%$ across the $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins in the approach using $N_2^{\beta=1,\mathrm{DDT}}$, and from 70 fb to 0.5 fb with uncertainties ranging from $\approx 30\%$ to $\approx 50\%$ when using ParticleNet$^{\mathrm{DDT}}$. In the edge $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins the uncertainty is mostly close to or larger than 100%. The unfolded and predicted jet mass distributions are also shown in the same way for $p_{\mathrm{T}}^{\mathrm{ptcl}} > 500\,\mathrm{GeV}$ in Figure 8.16 for both tagging approaches, where the individual distributions of each $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin are added together. The unfolded data for both tagging approaches agree within the total uncertainty, both in the double differential representation per $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin, as well as in the representation with the sum of $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins. The first $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bin shows the largest systematic difference when comparing the two tagging approaches across $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins. When using $N_2^{\beta=1,\mathrm{DDT}}$

the first $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins are measured to be larger than the simulation, while the measurement with ParticleNet$^{\mathrm{DDT}}$ as tagger the bins are measured to be smaller than the simulation, while both agree with the prediction by the simulation within statistical and systematic uncertainty (black bars). Thus this difference is covered by the systematic uncertainties already included in the fit. The distribution in unfolded data in the first two $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins is generally in more agreement with the simulation concerning both the normalization as well as the shape. In the last two $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins there is a shift in the peak position visible in the unfolded data compared to the simulation. This shift is covered well within the total uncertainty and can also be caused by larger anticorrelation between the $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins, which can be seen in the correlation matrices in Figure 8.17. The correlation matrices show the correlation between the signal strength modifiers μ_j of the individual particle-level bins. The $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins in the different $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins are shown unrolled, where the black dashed lines indicate the $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins. The highest correlations are found to be among neighboring bins both along the $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ and $p_{\mathrm{T}}^{\mathrm{ptcl}}$ axis. The latter ones are similar for all $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins, but smaller than correlations between neighboring bins along the $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ axis overall. The highest anti-correlations are found between the first two and last two $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins in the third and fourth $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin, In general, the fit shows higher anti-correlations between the bins when using $N_2^{\beta=1,\mathrm{DDT}}$ as the W tagger.

The impacts of the most influential groups of fit parameters on the uncertainty of the signal strength modifiers corresponding to the a given $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin are summarized in Table 7 for the substructure tagger approach and in Table 8 for the ParticleNet$^{\mathrm{DDT}}$ tagger approach. For this the maximum value among the highest impact in a given parameter group on the signal strength modifiers of all $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins in a given $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin is taken as the upper estimate, and similarly the minimum value for the lower estimate. The largest contribution to the total uncertainty is the normalization uncertainty on the $W(q\bar{q})$+jets and $Z(q\bar{q})$+jets processes, reaching up to 18% for the substructure tagger approach and up to 29% in the ParticleNet tagger approach. The second and third largest contributions stem from the QCD multijet background estimate and jet energy corrections. The uncertainty related to ISR, FSR, pileup and theoretical V+jets QCD correction k-factors, all reaching the order of 3–17%. Generally, the uncertainties and the individual impacts from the individual sources are larger in the first and last $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin.

Finally, as a prospect on the inclusion of a $N_2^{\beta=1}$ cut in the particle-level definition, the unfolded result is further corrected with the acceptance correction for the phase-space definition with the selection $N_2^{\beta=1} < 0.2$. Figure 8.18 shows the distributions for $p_{\mathrm{T}}^{\mathrm{ptcl}} > 500\,\mathrm{GeV}$ from particle-level simulation obtained with the $N_2^{\beta=1}$-cut as red lines and the adjusted unfolded data obtained with the unfolding without $N_2^{\beta=1}$-cut in the particle-level definition and additional acceptance correction including the $N_2^{\beta=1}$-cut as black markers for the substructure tagger approach on the left and for the ParticleNet tagger approach on the right. With the additional $N_2^{\beta=1}$-cut the separation between the peak region and the tails at low and high masses is improved when compared to the nominal result shown in Figure 8.16. This indicates that the inclusion of the $N_2^{\beta=1}$-cut could improve the unfolding in terms of stability in the edge bins. The upper and lower end of the spectrum contain events, where not the W but the recoiling quark/gluon jet is measured. Thus, in the first $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bin the predicted yield depends on the modeling of non-perturbative effects, which is highly model dependent as discussed in 5.3. The presented results

are so far only compared to the prediction derived with the parton shower model in PYTHIA. A comparison of the result to more than one model is crucial and should be performed in the progression of this analysis using a different parton shower model, e.g. using HERWIG++. The last bin is limited by the precision of the perturbative QCD, i.e. the fact, that the signal MC sample is generated at LO+MLM matching accuracy. Here signal modeling using simulation at a higher precision, e.g. NLO or NLO+FxFx matching could reduce the uncertainties in the highest $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ and $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin. By introducing the $N_2^{\beta=1}$-cut the sensitivity of this analysis to the modeling of the perturbative and non-perturbative effects related to the recoiling quark/gluon jet could be reduced, judging by the increased purity of the peak at the W boson mass.

Uncertainty	**Impact in $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin**			
	$p_{\mathrm{T}}^{\mathrm{ptcl}} \in [500, 650)$	$p_{\mathrm{T}}^{\mathrm{ptcl}} \in [650, 800)$	$p_{\mathrm{T}}^{\mathrm{ptcl}} \in [800, 1200)$	$p_{\mathrm{T}}^{\mathrm{ptcl}} \in [1200, \infty)$
W+jets normalization	6 – 15 %	11 – 18 %	6 – 16 %	< 0.1 – 15 %
Z+jets normalization	8 – 18 %	4 – 11 %	0.5 – 2 %	< 0.1 – 13 %
QCD multijet estimate	< 0.1 – 21 %	< 0.1 – 10 %	< 0.1 – 6 %	< 0.1 – 7 %
Jet energy correction	1 – 13 %	0.2 – 13 %	< 0.1 – 4 %	< 0.1 – 17 %
Final state radiation	2 – 14 %	5 – 8 %	3 – 8 %	< 0.1 – 12 %
Initial state radiation	3 %	2 – 4 %	0.9 – 4 %	< 0.1 – 4 %
Pileup reweighting	0.3 – 1 %	0.1 – 0.6 %	0.4 – 0.7 %	< 0.1 – 2 %
V+jets QCD NNLO	1 – 8 %	5 – 8 %	4 – 10 %	< 0.1 – 15 %
W+jets EW fixed order	0.9 – 2 %	1 – 2 %	0.6 – 1.0 %	< 0.1 – 0.4 %
Z+jets EW fixed order	0.3 – 0.7 %	0.1 – 0.4 %	< 0.1 – 0.1 %	< 0.1 – 0.3 %
Luminosity	1 – 2 %	0.7 – 1 %	0.6 – 2 %	< 0.1 – 3 %
Top p_T reweighting	0.4 – 4 %	0.2 – 2 %	0.2 – 1 %	< 0.1 – 2 %
Trigger scale factor	< 0.1 – 1 %	< 0.1 – 0.9 %	< 0.1 – 1 %	< 0.1 – 2 %

Table 7: Summary systematic uncertainty sources and their impact on the measurement using $N_2^{\beta=1,\mathrm{DDT}}$ as the tagger in the QCD multijet background estimate in each $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin.

Uncertainty	**Impact in $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin**			
	$p_{\mathrm{T}}^{\mathrm{ptcl}} \in [500, 650)$	$p_{\mathrm{T}}^{\mathrm{ptcl}} \in [650, 800)$	$p_{\mathrm{T}}^{\mathrm{ptcl}} \in [800, 1200)$	$p_{\mathrm{T}}^{\mathrm{ptcl}} \in [1200, \infty)$
W+jets normalization	< 0.1 – 29 %	13 – 17 %	13 – 19 %	11 – 24 %
Z+jets normalization	< 0.1 – 18 %	8 %	4 – 5 %	< 0.1 – 4 %
QCD multijet estimate	< 0.1 – 13 %	< 0.1 – 6 %	< 0.1 – 5 %	< 0.1 – 9 %
Jet energy correction	< 0.1 – 14 %	0.5 – 9 %	0.2 – 3 %	< 0.1 – 12 %
Final state radiation	< 0.1 – 12 %	1 – 3 %	2 – 7 %	< 0.1 – 13 %
Initial state radiation	< 0.1 – 10 %	2 – 3 %	5 – 7 %	7 – 10 %
Pileup reweighting	< 0.1 – 6 %	0.7 – 3 %	1 – 6 %	3 – 16 %
V+jets QCD NNLO	< 0.1 – 13 %	5 – 9 %	9 – 11 %	10 – 16 %
W+jets EW fixed order	< 0.1 – 3 %	0.8 – 2 %	0.6 – 1 %	< 0.1 – 0.6 %
Z+jets EW fixed order	< 0.1 – 2 %	< 0.1 – 0.3 %	< 0.1 – 0.1 %	< 0.1 – 0.6 %
Luminosity	< 0.1 – 2 %	0.8 – 0.9 %	0.5 – 1 %	< 0.1 – 1 %
Top p_T reweighting	< 0.1 – 3 %	0.2 – 1 %	< 0.1 – 0.5 %	< 0.1 – 1 %
Trigger scale factor	< 0.1 – 3 %	< 0.1 – 0.7 %	< 0.1 – 1 %	< 0.1 – 2 %

Table 8: Summary systematic uncertainty sources and their impact on the measurement using ParticleNet$^{\mathrm{DDT}}$ as tagger in the QCD multijet background estimate in each $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin.

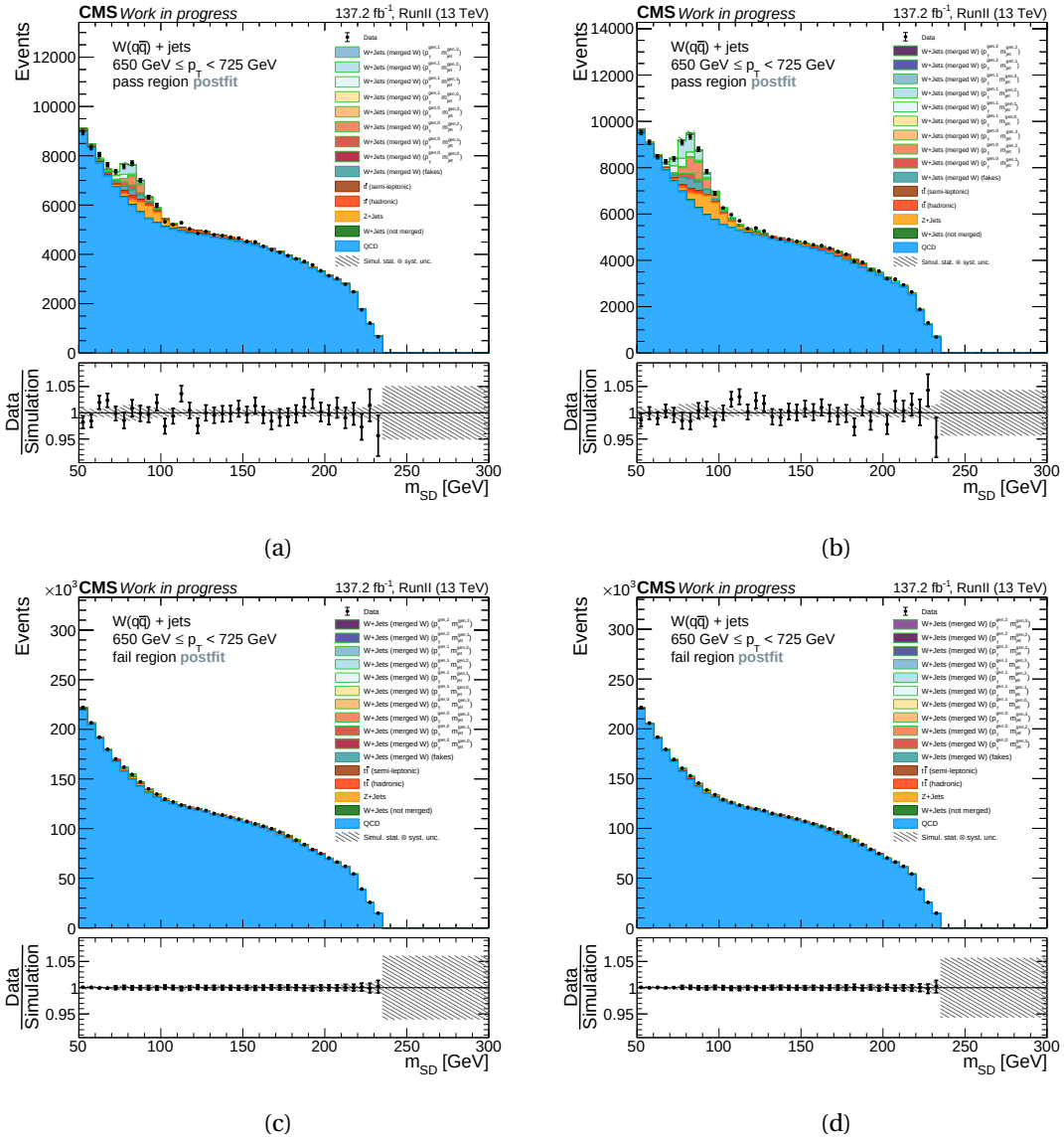
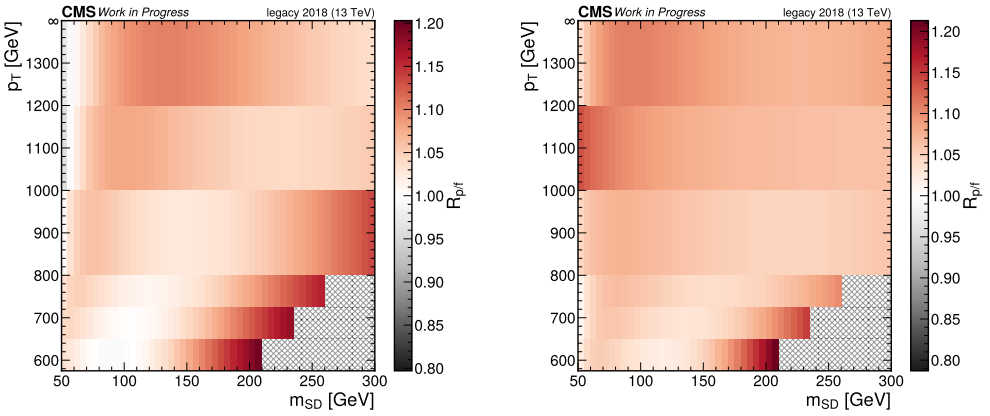
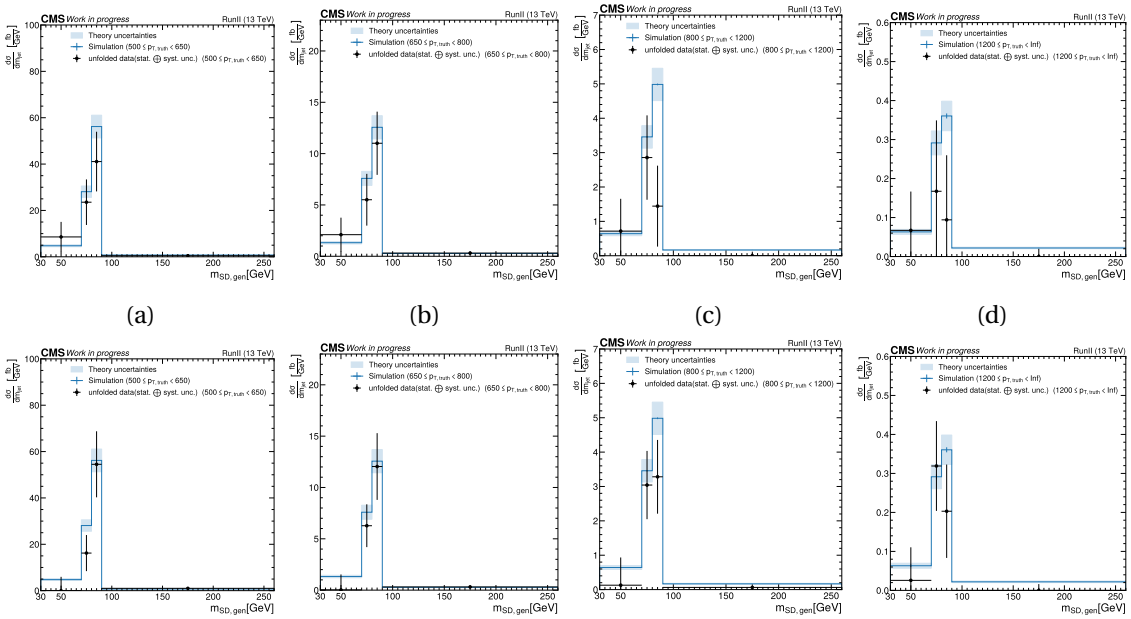


Figure 8.13: Post-fit templates in the second $p_{\mathrm{T}}^{\mathrm{reco}}$ bin $650 < p_{\mathrm{T}}^{\mathrm{reco}} \leq 725\,\mathrm{GeV}$ for events pass (top row) and failing (bottom row) the $N_2^{\beta=1,\mathrm{DDT}}$ or the ParticleNet$^{\mathrm{DDT}}$ W tagger on the left and right side respectively. All data-taking periods (early 2016, late 2016, 2017 and 2018) are summed post-fit to reflect the full Run II distribution with the integrated luminosity $\mathscr{L}_{\mathrm{int}} \approx 137\,\mathrm{fb}^{-1}$.



(a) (b)

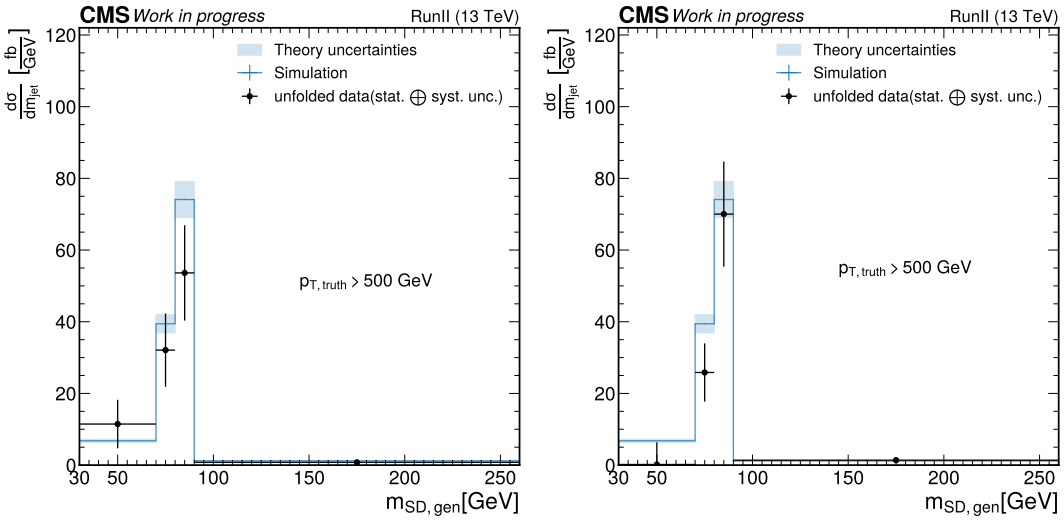
Figure 8.14: Post-fit transfer factor of the data-driven QCD multijet background estimation, when using $N_2^{\beta=1,\mathrm{DDT}}$ and ParticleNet$^{\mathrm{DDT}}$ as tagger on the left and right respectively.



(a) (b) (c) (d)

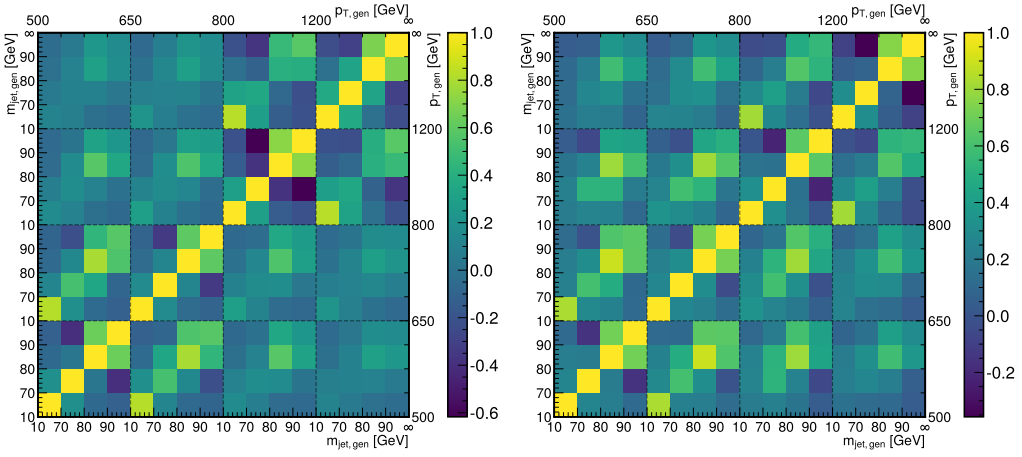
(e) (f) (g) (h)

Figure 8.15: Jet mass distribution for each $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bin resulting from the unfolding using $N_2^{\beta=1,\mathrm{DDT}}$ (top row) or ParticleNet$^{\mathrm{DDT}}$ (bottom row) in the background estimation. The unfolded data is shown as black markers, the blue line shows the true distribution from the simulation with theory uncertainties added as the shaded blue band.



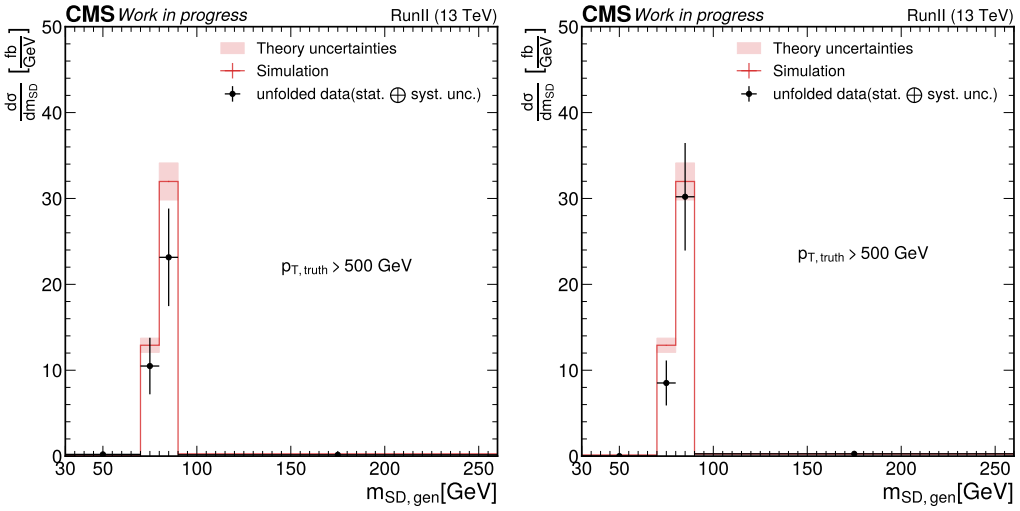
(a) (b)

Figure 8.16: Jet mass distribution of the sum of all p_T^{ptcl} bins resulting from the unfolding using $N_2^{\beta=1,DDT}$ (left) and ParticleNet tagger (right) in the background estimation. The unfolded data is shown as black markers, the blue line shows the true distribution from the simulation with theory uncertainties added as the shaded blue band.



(a) (b)

Figure 8.17: Correlation matrix of the maximum likelihood estimators of the signal strength modifiers $\hat{\mu}_{p_T,\, m_{SD}}$. The left and right plot shows the matrix from the fit to data using $N_2^{\beta=1,DDT}$ and ParcileNet^{DDT} respectively. The grey dashed lines indicate the individual p_T^{ptcl} bins.



(a) (b)

Figure 8.18: Jet mass distribution of the sum of all p_T^{ptcl} bins resulting from the unfolding using $N_2^{\beta=1,DDT}$ (left) and ParticleNet tagger (right) in the background estimation. An additional correction factor is applied to the unfolded data corresponding to the acceptance of the phase space with the additional selection $N_2^{\beta=1} < 0.2$ on particle-level. The unfolded data is shown as black markers, the red line shows the true distribution with $N_2^{\beta=1} < 0.2$ from the simulation with theory uncertainties added as the shaded red band.

9 Conclusion and Outlook

In this thesis, the decays of hadronically decaying top quarks and W bosons with high transverse momenta p_{T} were investigated in data recorded by the CMS experiment at a center of mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$ during 2016–2018 (Run 2). The recorded data correspond to a total integrated luminosity of $\mathscr{L}_{\mathrm{int}} \approx 138\,\mathrm{fb}^{-1}$. The decay products of these particles with large Lorentz-boost are highly collimated and consequently reconstructed as single large-radius jets. The substructure of these jets and specifically the invariant jet mass was studied. The jet mass is one of the most important observables related to jet substructure, as it is a very useful tool to identify jet-initiating particles (e.g. in searches) and a helpful testbed for precise theoretical predictions.

First, simulation to data correction factors for the soft drop jet mass scale of large-radius jets originating from hadronic decays of W bosons and top quarks were measured simultaneously semileptonic $t\bar{t}$ and fully-hadronic $W(q\bar{q})$+jets events. To achieve the highest purity samples in the selection of boosted W jets and top jets two different jet-tagging approaches were investigated. One based on substructure variables motivated by QCD theory, and another one based on state-of-the-art deep learning algorithms. The main challenge was the estimation of the dominant QCD multijet background in the fully-hadronic $W(q\bar{q})$+jets sample. To avoid the modeling of perturbative QCD, which is not adequately modeled in the probed phase space, a data-driven estimation technique was used. The technique was based on an established differential alphabet method using a two-dimensional transfer factor and was further developed and optimized. For this, the substructure tagging approach was decorrelated to the jet mass, to avoid sculpting and ensure similar shapes in the soft drop mass distribution in the control region and signal region. The two-dimensional transfer factor was modeled with linear combinations of Bernstein polynomials at different orders. The optimal choice of orders was determined in F-Tests employing toy datasets. The soft drop jet mass scale correction factors were derived by introducing parameters to the fit corresponding to variations of the energy scale of jet constituents, which were propagated to the soft drop jet mass. The correction factors were first measured in the two samples individually and with separate parameters acting on the mass scale of top jets and W jets separately. This showed that one common correction factor for top jets and W jets could be measured in both samples simultaneously. Next, a study on the correlation of the jet energy scale and the jet mass scale showed that the correlation is not at 100%, but very close, thus the dedicated corrections for the jet energy scale apply to some extent to the jet mass scale.

The final correction factor accounting for any further residual differences, after the jet energy corrections are applied to the jet mass, was measured in each period of data-taking separately and is mostly within 2% different from unity for both tagging approaches in the range 500 <

$p_{\mathrm{T}} < 1200\,\mathrm{GeV}$, while the tagging approach using the deep learning algorithm yielded generally more stable results. Additionally, it was demonstrated that with the uncertainties of the jet energy corrections propagated to the soft drop jet mass correction factors, closure with unity can be achieved. The measurement of the correction factors could be further improved by measuring the correction factors for each flavor of particle-flow candidates separately, which could give more insights into the performance of the method due to the different reconstruction efficiencies and fraction of jet energy of the individual flavor of jet constituents. Additionally, the measurement of correction factors using different tagging approaches and jet mass definitions could give insight into the detector response on jet substructure.

The second part of this thesis was the first W jet mass measurement at the LHC. Specifically, the measurement of the soft drop jet mass distribution of boosted W bosons on particle-level in the full Run 2 dataset. The measurement is performed by unfolding data in the same sample of boosted W jets from fully-hadronic $W(q\bar{q})$+jets events, that was used in the measurement of the soft drop jet mass correction factors. For the estimation of the dominant QCD multijet background, the same data-driven method was used. Here the approach using the mass-decorrelated substructure tagger and the approach based on the deep learning algorithm were used to construct control regions used in the estimation of the QCD background. The tagger using the deep learning algorithm was further mass-decorrelated, to achieve a more stable unfolding. To measure the jet mass distribution in bins of jet transverse momentum on particle-level, a two-dimensional unfolding of the data was performed using maximum likelihood unfolding. Here the main challenge was introduced by the fact, that in W+jets events the light QCD-like jet has similar kinematic properties as the W jet, which makes the selection of a pure sample of W jets non-trivial. As a result, the migration matrix was found to be ill-conditioned, which showed the necessity of regularization of the unfolding. Multiple schemes of binning and regularization were studied and the one introducing the smallest bias and correlation among particle-level bin estimators was chosen to be used in the final unfolding. Within uncertainties, the final unfolded data are in agreement with the prediction from the simulation at LO+MLM accuracy with further corrections derived from NLO QCD simulation and NLO EWK fixed-order calculations. As a prospect of including a substructure selection on particle-level, the final result was adjusted according to changes in acceptance induced by including the cut $N_2^{\beta=1} < 0.2$ in the particle-level definition. This showed that including such a requirement could reduce the contribution in the tails of the jet mass tails, where non-perturbative and perturbative effects related to the light QCD-like jets play a role. To study this further the unfolding has to be optimized with this new particle-level definition.

To improve the confidence in the validity of this measurement the procedure should be repeated with at least one other simulation based on a different modeling of the parton shower, to be able to compare differences in the modeling of non-perturbative QCD. Additionally, in this standard candle testbed, the comparison to semi-analytical calculations at LL or even NLL' accuracy can yield further insight into jet substructure. Further study of the possible regularization and binning schemes could reduce introduced biases in the edge bins and the overall sensitivity to the W mass could potentially be improved.

With the HL-LHC this measurement will benefit from the larger dataset expected, as the uncer-

tainty of the measurement is statistically dominated in some of the bins, and could provide a first step towards the measurement of the mass of the W boson or the mass difference $m_W - m_Z$.

References

[1] CMS Collaboration, "Calibration of the Jet Mass Scale using boosted W bosons and top quarks", 2023, Technical Report CMS-DP-2023-044.

[2] CMS Collaboration, "Search for low mass vector resonances decaying into quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. D* **100** (2019), no. 11, 112007, doi:10.1103/PhysRevD.100.112007, arXiv:1909.04114.

[3] J. M. Lindert et al., "Precise predictions for $V+$ jets dark matter backgrounds", *Eur. Phys. J. C* **77** (2017), no. 12, 829, doi:10.1140/epjc/s10052-017-5389-1, arXiv:1705.04664.

[4] CMS Collaboration Collaboration, "Search for higgs boson and observation of z boson through their decay into a charm quark-antiquark pair in boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. Lett.* **131** (Jul, 2023) 041801, doi:10.1103/PhysRevLett.131.041801.

[5] CDF Collaboration, "High-precision measurement of the W boson mass with the CDF II detector", *Science* **376** (2022), no. 6589, 170–176, doi:10.1126/science.abk1781.

[6] R. Kogler et al., "Jet Substructure at the Large Hadron Collider: Experimental Review", *Rev. Mod. Phys.* **91** (2019), no. 4, 045003, doi:10.1103/RevModPhys.91.045003, arXiv:1803.06991.

[7] R. Kogler, "Advances in Jet Substructure at the LHC: Algorithms, Measurements and Searches for New Physical Phenomena", volume 284. Springer, 5, 2021. doi:10.1007/978-3-030-72858-8, ISBN 978-3-030-72857-1, 978-3-030-72858-8.

[8] S. Marzani, G. Soyez, and M. Spannowsky, "Looking inside jets: an introduction to jet substructure and boosted-object phenomenology", volume 958. Springer, 2019. doi:10.1007/978-3-030-15709-8.

[9] 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", *Phys. Rept.* **841** (2020) 1–63, doi:10.1016/j.physrep.2019.11.001, arXiv:1709.04464.

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

[11] ATLAS Collaboration, "Jet mass and substructure of inclusive jets in $\sqrt{s} = 7$ TeV pp collisions with the ATLAS experiment", *JHEP* **05** (2012) 128, doi:10.1007/JHEP05(2012)128, arXiv:1203.4606.

[12] ATLAS Collaboration, "Measurement of the Soft-Drop Jet Mass in pp Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector", *Phys. Rev. Lett.* **121** (2018), no. 9, 092001, doi:10.1103/PhysRevLett.121.092001, arXiv:1711.08341.

[13] 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.

[14] CMS Collaboration, "Measurement of the differential $\mathrm{t\bar{t}}$ production cross section as a function of the jet mass and extraction of the top quark mass in hadronic decays of boosted top quarks", *Eur. Phys. J. C* **83** (2023), no. 7, 560, doi:10.1140/epjc/s10052-023-11587-8, arXiv:2211.01456.

[15] ATLAS Collaboration, "Measurement of the jet mass in high transverse momentum $Z(\rightarrow b\bar{b})\gamma$ production at $\sqrt{s} = 13$ TeV using the ATLAS detector", *Phys. Lett. B* **812** (2021) 135991, doi:10.1016/j.physletb.2020.135991, arXiv:1907.07093.

[16] F. Halzen and A. D. Martin, "Quarks and Leptons: An Introductory Course in Modern Particle Physics". Wiley, 1984. ISBN 0471887412, 9780471887416.

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

[18] M. Thomson, "Modern Particle Physics". Cambridge University Press, 9, 2013. doi:10.1017/cbo9781139525367, ISBN 9781107034266.

[19] A. J. Larkoski, "Elementary Particle Physics: An Intuitive Introduction". Cambridge University Press, 6, 2019. doi:10.1017/9781108633758, ISBN 978-1-108-49698-8, 978-1-108-57940-7.

[20] General Conference on Weights and Measures, "Resolution 1 – On the revision of the International System of Units (SI)". Available at https://www.bipm.org/en/committees/cg/cgpm/26-2018/resolution-1, 2018. Accessed on April 21, 2023.

[21] C. Burgard and D. Galbraith, "Standard model of physics". Available at https://texample.net/tikz/examples/model-physics/, Accessed on April 1, 2023.

[22] Particle Data Group Collaboration, "Review of Particle Physics", *PTEP* **2022** (08, 2022) 083C01, doi:10.1093/ptep/ptac097.

[23] The KATRIN Collaboration, "Direct neutrino-mass measurement with sub-electronvolt sensitivity", *Nature Physics* **18** (2022), no. 2, 160–166, doi:10.1038/s41567-021-01463-1.

[24] B. Pontecorvo, "Inverse beta process", Chalk River Laboratory, 1946, Report PD-205. see "Bruno Pontecorvo selected scientfic works: recollections" p.21–26.

[25] Super-Kamiokande Collaboration, "Evidence for oscillation of atmospheric neutrinos", *Phys. Rev. Lett.* **81** (1998) 1562–1567, doi:10.1103/PhysRevLett.81.1562, arXiv:hep-ex/9807003.

[26] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons", *Phys. Rev. Lett.* **13** (Aug, 1964) 321–323, doi:10.1103/PhysRevLett.13.321.

[27] P. W. Higgs, "Broken symmetries and the masses of gauge bosons", *Phys. Rev. Lett.* **13** (Oct, 1964) 508–509, doi:10.1103/PhysRevLett.13.508.

[28] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Physics Letters B* **716** (2012), no. 1, 30–61, doi:https://doi.org/10.1016/j.physletb.2012.08.021.

[29] ATLAS Collaboration, "Observation of a new particle in the search for the standard model higgs boson with the ATLAS detector at the LHC", *Physics Letters B* **716** (2012), no. 1, 1–29, doi:https://doi.org/10.1016/j.physletb.2012.08.020.

[30] L. Evans and P. Bryant, "LHC machine", *Journal of Instrumentation* **3** (aug, 2008) S08001, doi:10.1088/1748-0221/3/08/S08001.

[31] M. Spiropulu and S. Stapnes, "LHC's ATLAS and CMS detectors", *Int. J. Mod. Phys. A* **23** (2008), no. 15, 2917–2935, doi:10.1142/S0217751X08042341.

[32] CMS Collaboration, "CMS Luminosity Measurements for the 2016 Data Taking Period", CERN, Geneva, 2017, Technical Report CMS-PAS-LUM-17-001.

[33] CMS Collaboration, "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s}$ = 13 TeV", Geneva, 2018, Technical Report CMS-PAS-LUM-17-004.

[34] CMS Collaboration, "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s}$ = 13 TeV", CERN, Geneva, 2019, Technical Report CMS-PAS-LUM-18-002.

[35] CMS Collaboration, "CMS Luminosity - Public Results". https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults. Accessed on: August 2023.

[36] CMS Collaboration, "Pileup mitigation at CMS in 13 TeV data", *JINST* **15** (2020), no. 09, P09018, doi:10.1088/1748-0221/15/09/P09018, arXiv:2003.00503.

[37] CMS Collaboration, "CMS Physics: Technical Design Report Volume 1: Detector Performance and Software", 2006, Technical Report CERN-LHCC-2006-001, CMS-TDR-8-1, CERN-LHCC-2006-001, CMS-TDR-8-1.

[38] CMS Collaboration, T. Sakuma, "Cutaway diagrams of CMS detector", 2019. http://cds.cern.ch/record/2665537.

[39] T. Sakuma and T. McCauley, "Detector and Event Visualization with SketchUp at the CMS Experiment", *J. Phys. Conf. Ser.* **513** (2014) 022032, doi:10.1088/1742-6596/513/2/022032, arXiv:1311.4942.

[40] CMS Tracker Group Collaboration, "The CMS Phase-1 Pixel Detector Upgrade", *JINST* **16** (2021), no. 02, P02027, doi:10.1088/1748-0221/16/02/P02027, arXiv:2012.14304.

[41] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014), no. 10, P10009, doi:10.1088/1748-0221/9/10/P10009, arXiv:1405.6569.

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

[43] E. Yazgan and the CMS ECAL/HCAL Collaborations, "The CMS barrel calorimeter response to particle beams from 2 to 350 GeV/c", *Journal of Physics: Conference Series* **160** (April, 2009) 012056, doi:10.1088/1742-6596/160/1/012056.

[44] CMS Collaboration, "The CMS trigger system", *JINST* **12** (2017), no. 01, P01020, doi:10.1088/1748-0221/12/01/P01020, arXiv:1609.02366.

[45] R. P. Feynman, "The behavior of hadron collisions at extreme energies", *Conf. Proc. C* **690905** (1969) 237–258.

[46] J. M. Butterworth, G. Dissertori, and G. P. Salam, "Hard Processes in Proton-Proton Collisions at the Large Hadron Collider", *Ann. Rev. Nucl. Part. Sci.* **62** (2012) 387–405, doi:10.1146/annurev-nucl-102711-094913, arXiv:1202.0583.

[47] NNPDF Collaboration, "Parton distributions from high-precision collider data", *Eur. Phys. J. C* **77** (2017), no. 10, 663, doi:10.1140/epjc/s10052-017-5199-5, arXiv:1706.00428.

[48] V. N. Gribov and L. N. Lipatov, "Deep inelastic *ep* scattering in perturbation theory", *Sov. J. Nucl. Phys.* **15** (1972) 438–450.

[49] L. N. Lipatov, "The parton model and perturbation theory", *Yad. Fiz.* **20** (1974) 181–198.

[50] Y. L. Dokshitzer, "Calculation of the Structure Functions for Deep Inelastic Scattering and e^+e^- Annihilation by Perturbation Theory in Quantum Chromodynamics.", *Sov. Phys. JETP* **46** (1977) 641–653.

[51] G. Altarelli and G. Parisi, "Asymptotic Freedom in Parton Language", *Nucl. Phys. B* **126** (1977) 298–318, doi:10.1016/0550-3213(77)90384-4.

[52] J. Alwall et al., "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.

[53] 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.

[54] 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.

[55] 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.

[56] T. Sjöstrand et al., "An introduction to PYTHIA 8.2", *Comput. Phys. Commun.* **191** (2015) 159–177, doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012.

[57] A. Buckley et al., "General-purpose event generators for LHC physics", *Physics Reports* **504** (2011), no. 5, 145–233, doi:https://doi.org/10.1016/j.physrep.2011.03.005.

[58] J. Alwall et al., "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions", *Eur. Phys. J. C* **53** (2008) 473–500, doi:10.1140/epjc/s10052-007-0490-5, arXiv:0706.2569.

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

[60] B. Andersson, S. Mohanty, and F. Soderberg, "Recent developments in the Lund model", in *36th Annual Winter School on Nuclear and Particle Physics (PINP 2002) and 8th St. Petersburg School on Theoretical Physics*. 12, 2002. arXiv:hep-ph/0212122.

[61] GEANT4 Collaboration, "GEANT4–a simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003) 250–303, doi:10.1016/S0168-9002(03)01368-8.

[62] R. Fruhwirth, "Application of Kalman filtering to track and vertex fitting", *Nucl. Instrum. Meth. A* **262** (1987) 444–450, doi:10.1016/0168-9002(87)90887-4.

[63] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **13** (2018) P06015, doi:10.1088/1748-0221/13/06/P06015, arXiv:1804.04528.

[64] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems", *Proceedings of the IEEE* **86** (1998), no. 11, 2210–2239, doi:10.1109/5.726788.

[65] R. Frühwirth, W. Waltenberger, and P. Vanlaer, "Adaptive Vertex Fitting", Geneva, Technical Report CMS-NOTE-2007-008.

[66] 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.

[67] 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.

[68] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, "Longitudinally invariant K_t clustering algorithms for hadron hadron collisions", *Nucl. Phys. B* **406** (1993) 187–224, doi:10.1016/0550-3213(93)90166-M.

[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] T. Lapsien, R. Kogler, and J. Haller, "A new tagger for hadronically decaying heavy particles at the LHC", *Eur. Phys. J. C* **76** (2016), no. 11, 600, doi:10.1140/epjc/s10052-016-4443-8, arXiv:1606.04961.

[71] I. W. Stewart et al., "XCone: N-jettiness as an Exclusive Cone Jet Algorithm", *JHEP* **11** (2015) 072, doi:10.1007/JHEP11(2015)072, arXiv:1508.01516.

[72] J. Thaler and T. F. Wilkason, "Resolving Boosted Jets with XCone", *JHEP* **12** (2015) 051, doi:10.1007/JHEP12(2015)051, arXiv:1508.01518.

[73] CMS Collaboration, "Pileup Removal Algorithms", 2014, Technical Report CMS-PAS-JME-14-001.

[74] 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.

[75] CMS Collaboration, "Jet algorithms performance in 13 TeV data", 2017, Technical Report CMS-PAS-JME-16-003.

[76] CMS Collaboration, "Pileup Jet Identification", 2013, Technical Report CMS-PAS-JME-13-005.

[77] CMS Collaboration, "Pileup-per-particle identification: optimisation for Run 2 Legacy and beyond", 2021, Technical Report CMS-DP-2021-001.

[78] CMS Collaboration, "Jet energy scale and resolution measurement with Run 2 Legacy Data Collected by CMS at 13 TeV", 2021, technical report.

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

[80] CMS Collaboration, "Jet energy scale and resolution performance with 13 TeV data collected by CMS in 2016-2018", 2020, Technical Report CMS-DP-2020-019.

[81] G. P. Salam, "Towards Jetography", *Eur. Phys. J. C* **67** (2010) 637–686, doi:10.1140/epjc/s10052-010-1314-6, arXiv:0906.1833.

[82] V. V. Sudakov, "Vertex parts at very high-energies in quantum electrodynamics", *Sov. Phys. JETP* **3** (1956) 65–71.

[83] M. Dasgupta, K. Khelifa-Kerfa, S. Marzani, and M. Spannowsky, "On jet mass distributions in Z+jet and dijet processes at the LHC", *JHEP* **10** (2012) 126, doi:10.1007/JHEP10(2012)126, arXiv:1207.1640.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] CMS Collaboration, "Search for massive resonances decaying into WW, WZ, ZZ, qW, and qZ with dijet final states at $\sqrt{s} = 13$ TeV", *Phys. Rev. D* **97** (2018), no. 7, 072006, doi:10.1103/PhysRevD.97.072006, arXiv:1708.05379.

[89] CMS Collaboration, "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques", *JINST* **15** (2020), no. 06, P06005, doi:10.1088/1748-0221/15/06/P06005, arXiv:2004.08262.

[90] 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.

[91] A. J. Larkoski, D. Neill, and J. Thaler, "Jet Shapes with the Broadening Axis", *JHEP* **04** (2014) 017, doi:10.1007/JHEP04(2014)017, arXiv:1401.2158.

[92] A. J. Larkoski, G. P. Salam, and J. Thaler, "Energy Correlation Functions for Jet Substructure", *JHEP* **06** (2013) 108, doi:10.1007/JHEP06(2013)108, arXiv:1305.0007.

[93] I. Moult, L. Necib, and J. Thaler, "New Angles on Energy Correlation Functions", *JHEP* **12** (2016) 153, doi:10.1007/JHEP12(2016)153, arXiv:1609.07483.

[94] H. Qu and L. Gouskos, "ParticleNet: Jet Tagging via Particle Clouds", *Phys. Rev. D* **101** (2020), no. 5, 056019, doi:10.1103/PhysRevD.101.056019, arXiv:1902.08570.

[95] Y. Wang et al., "Dynamic graph cnn for learning on point clouds", *ACM Trans. Graph.* **38** (oct, 2019) doi:10.1145/3326362.

[96] CMS Collaboration, "Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques", 2020, Technical Report CMS-DP-2020-002.

[97] E. Bols et al., "Jet Flavour Classification Using DeepJet", *JINST* **15** (2020), no. 12, P12012, doi:10.1088/1748-0221/15/12/P12012, arXiv:2008.10519.

[98] CMS Collaboration, "Performance summary of AK4 jet b tagging with data from proton-proton collisions at 13 TeV with the CMS detector", 2023, Technical Report CMS-DP-2023-005.

[99] J. Dolen et al., "Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure", *JHEP* **05** (2016) 156, doi:10.1007/JHEP05(2016)156, arXiv:1603.00027.

[100] N. Gagunashvili, "Pearson's Test Modifications for Comparison of Unweighted and Weighted Histograms and Two Weighted Histograms", *PoS* **ACAT** (2009) 060, doi:10.22323/1.050.0060.

[101] ATLAS Collaboration, "Measurement of jet shapes in top-quark pair events at $\sqrt{s} = 7$ TeV using the ATLAS detector", *Eur. Phys. J. C* **73** (2013), no. 12, 2676, doi:10.1140/epjc/s10052-013-2676-3, arXiv:1307.5749.

[102] CMS Collaboration, "Measurement of jet substructure observables in $\mathrm{t\bar{t}}$ events from proton-proton collisions at $\sqrt{s} = 13$TeV", *Phys. Rev. D* **98** (2018), no. 9, 092014, doi:10.1103/PhysRevD.98.092014, arXiv:1808.07340.

[103] ATLAS Collaboration, "Measurement of soft-drop jet observables in pp collisions with the ATLAS detector at $\sqrt{s}$ =13 TeV", *Phys. Rev. D* **101** (2020), no. 5, 052007, doi:10.1103/PhysRevD.101.052007, arXiv:1912.09837.

[104] CMS Collaboration, "Measurement of Jet Fragmentation in PbPb and pp Collisions at $\sqrt{s_{NN}} = 2.76$ TeV", *Phys. Rev. C* **90** (2014), no. 2, 024908, doi:10.1103/PhysRevC.90.024908, arXiv:1406.0932.

[105] CMS Collaboration, "Observation of Medium-Induced Modifications of Jet Fragmentation in Pb-Pb Collisions at $\sqrt{s_{NN}} = 5.02$ TeV Using Isolated Photon-Tagged Jets", *Phys. Rev. Lett.* **121** (2018), no. 24, 242301, doi:10.1103/PhysRevLett.121.242301, arXiv:1801.04895.

[106] ALICE Collaboration, "Charged jet cross section and fragmentation in proton-proton collisions at $\sqrt{s} = 7$ TeV", *Phys. Rev. D* **99** (2019), no. 1, 012016, doi:10.1103/PhysRevD.99.012016, arXiv:1809.03232.

[107] LHCb Collaboration, "Measurement of charged hadron production in Z-tagged jets in proton-proton collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. Lett.* **123** (2019), no. 23, 232001, doi:10.1103/PhysRevLett.123.232001, arXiv:1904.08878.

[108] ATLAS Collaboration, "Properties of jet fragmentation using charged particles measured with the ATLAS detector in pp collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. D* **100** (2019), no. 5, 052011, doi:10.1103/PhysRevD.100.052011, arXiv:1906.09254.

[109] ATLAS Collaboration, "Measurement of b-quark fragmentation properties in jets using the decay $B^{\pm} \to J/\psi K^{\pm}$ in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector", *JHEP* **12** (2021) 131, doi:10.1007/JHEP12(2021)131, arXiv:2108.11650.

[110] ATLAS Collaboration, "Measurement of the Lund Jet Plane Using Charged Particles in 13 TeV Proton-Proton Collisions with the ATLAS Detector", *Phys. Rev. Lett.* **124** (2020), no. 22, 222002, doi:10.1103/PhysRevLett.124.222002, arXiv:2004.03540.

[111] CMS Collaboration, "Measurement of the primary Lund jet plane density in proton-proton collisions at $\sqrt{s}$ = 13 TeV", CERN, Geneva, 2023, Technical Report CMS-PAS-SMP-22-007.

[112] CMS Collaboration, "Study of quark and gluon jet substructure in Z+jet and dijet events from pp collisions", *JHEP* **01** (2022) 188, doi:10.1007/JHEP01(2022)188, arXiv:2109.03340.

[113] M. Freytsis et al., "Prospects for a Measurement of the W Boson Mass in the All-Jets Final State at Hadron Colliders", *JHEP* **02** (2019) 003, doi:10.1007/JHEP02(2019)003, arXiv:1807.07454.

[114] CMS Collaboration, "A multi-dimensional search for new heavy resonances decaying to boosted WW, WZ, or ZZ boson pairs in the dijet final state at 13 TeV", *Eur. Phys. J. C* **80** (2020), no. 3, 237, doi:10.1140/epjc/s10052-020-7773-5, arXiv:1906.05977.

[115] CMS Collaboration, "Search for boosted Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s}$ = 13 TeV", arXiv:2211.14181.

[116] CMS Collaboration, "Search for new heavy resonances decaying to WW, WZ, ZZ, WH, or ZH boson pairs in the all-jets final state in proton-proton collisions at $\sqrt{s}$ = 13 TeV", *Phys. Lett. B* **844** (2023) 137813, doi:10.1016/j.physletb.2023.137813, arXiv:2210.00043.

[117] CMS Collaboration, "Performance of the mass-decorrelated DeepDoubleX classifier for double-b and double-c large-radius jets with the CMS detector", 2022, Technical Report CMS-DP-2022-041.

[118] "Strategies and performance of the CMS silicon tracker alignment during LHC run 2", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1037** (2022) 166795, doi:https://doi.org/10.1016/j.nima.2022.166795.

[119] S. Frixione, P. Nason, and G. Ridolfi, "A Positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction", *JHEP* **09** (2007) 126, doi:10.1088/1126-6708/2007/09/126, arXiv:0707.3088.

[120] 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.

[121] R. Frederix, E. Re, and P. Torrielli, "Single-top t-channel hadroproduction in the four-flavour scheme with POWHEG and aMC@NLO", *JHEP* **09** (2012) 130, doi:10.1007/JHEP09(2012)130, arXiv:1207.5391.

[122] 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.

[123] CMS Collaboration, "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements", *Eur. Phys. J. C* **80** (2020), no. 1, 4, doi:10.1140/epjc/s10052-019-7499-4, arXiv:1903.12179.

[124] CMS Collaboration, "Measurement of normalized differential $t\bar{t}$ cross sections in the dilepton channel from pp collisions at $\sqrt{s} = 13$ TeV", *Journal of High Energy Physics* **2018** (2018), no. 4, 60, doi:10.1007/JHEP04(2018)060.

[125] CMS Collaboration, "Measurement of differential cross sections for top quark pair production using the lepton + jets final state in proton-proton collisions at 13 TeV", *Phys. Rev. D* **95** (May, 2017) 092001, doi:10.1103/PhysRevD.95.092001.

[126] G. Cowan, "Statistical data analysis". 1998. ISBN 978-0-19-850156-5.

[127] CMS Collaboration, "HiggsAnalysis Combine toolkit". http://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/. Accessed on: August 2023.

[128] W. Verkerke and D. P. Kirkby, "The RooFit toolkit for data modeling", *eConf* **C0303241** (2003) MOLT007, arXiv:physics/0306116.

[129] L. Moneta et al., "The RooStats Project", *PoS* **ACAT2010** (2010) 057, doi:10.22323/1.093.0057, arXiv:1009.1003.

[130] S. Bernstein, "Démonstration du théorème de weierstrass fondée sur le calcul des probabilités (proof of the theorem of weierstrass based on the calculus of probabilities)", *Communications of the Kharkov Mathematical Society* **13** (1912), no. 1, 1–2.

[131] R. D. Cousins, "Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models, with Application to Histograms". https://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf, 2013.

[132] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses", *The Annals of Mathematical Statistics* **9** (1938), no. 1, 60–62.

[133] J. K. Lindsey, "Parametric statistical inference". Clarendon Press, Oxford, England, February, 1996.

[134] M. H. DeGroot and M. J. Schervish, "Probability and statistics". Pearson custom library. Pearson Education Limited, 4th edition, 2014.

[135] V. Blobel and E. Lohrmann, "Statistische und numerische Methoden der Datenanalyse". V. Blobel, 2nd edition, 2012.

[136] W. C. Hamilton, "Significance tests on the crystallographic *R* factor", *Acta Crystallographica* **18** (Mar, 1965) 502–510, doi:10.1107/S0365110X65001081.

[137] CMS Collaboration, "Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS", *Eur. Phys. J. C* **81** (2021), no. 9, 800, doi:10.1140/epjc/s10052-021-09538-2, arXiv:2104.01927.

[138] S. Schmitt, "Tunfold, an algorithm for correcting migration effects in high energy physics", *Journal of Instrumentation* **7** (oct, 2012) T10003, doi:10.1088/1748-0221/7/10/T10003.

[139] V. Blobel, "Data analysis in high energy physics - a practical guide to statistical methods", ch. 6, pp. 187–226. John Wiley & Sons, New York, 2013.

[140] A. N. Tikhonov *Soviet Math. Dokl.* **4** (1963) 1035.

[141] A. Höcker and V. Kartvelishvili, "SVD approach to data unfolding", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **372** (1996), no. 3, 469–481, doi:https://doi.org/10.1016/0168-9002(95)01478-0.

Control plots

B

In the following figures *UL* refers to the legacy reconstruction of the CMS data. For example, the label *UL17* indicates the legacy reconstruction of the 2017 data was used. Further, *UL16preVFP* and *UL16postVFP* refer to early 2016 and late 2016 legacy data, respectively.

B.1 Control plots of distributions after pre-selection

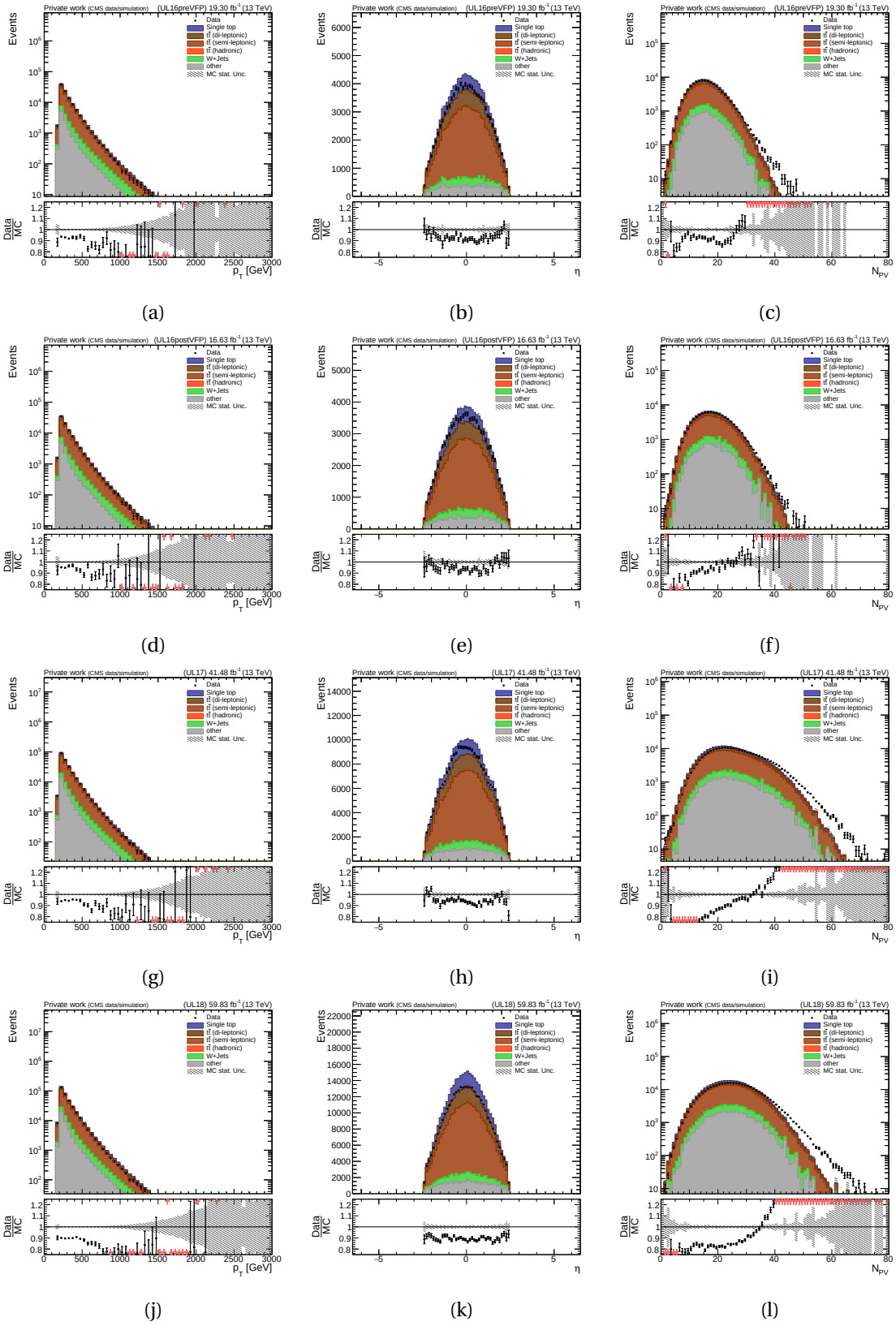


Figure B.1: Control plots of $t\bar{t}$ sample after the pre-selection.

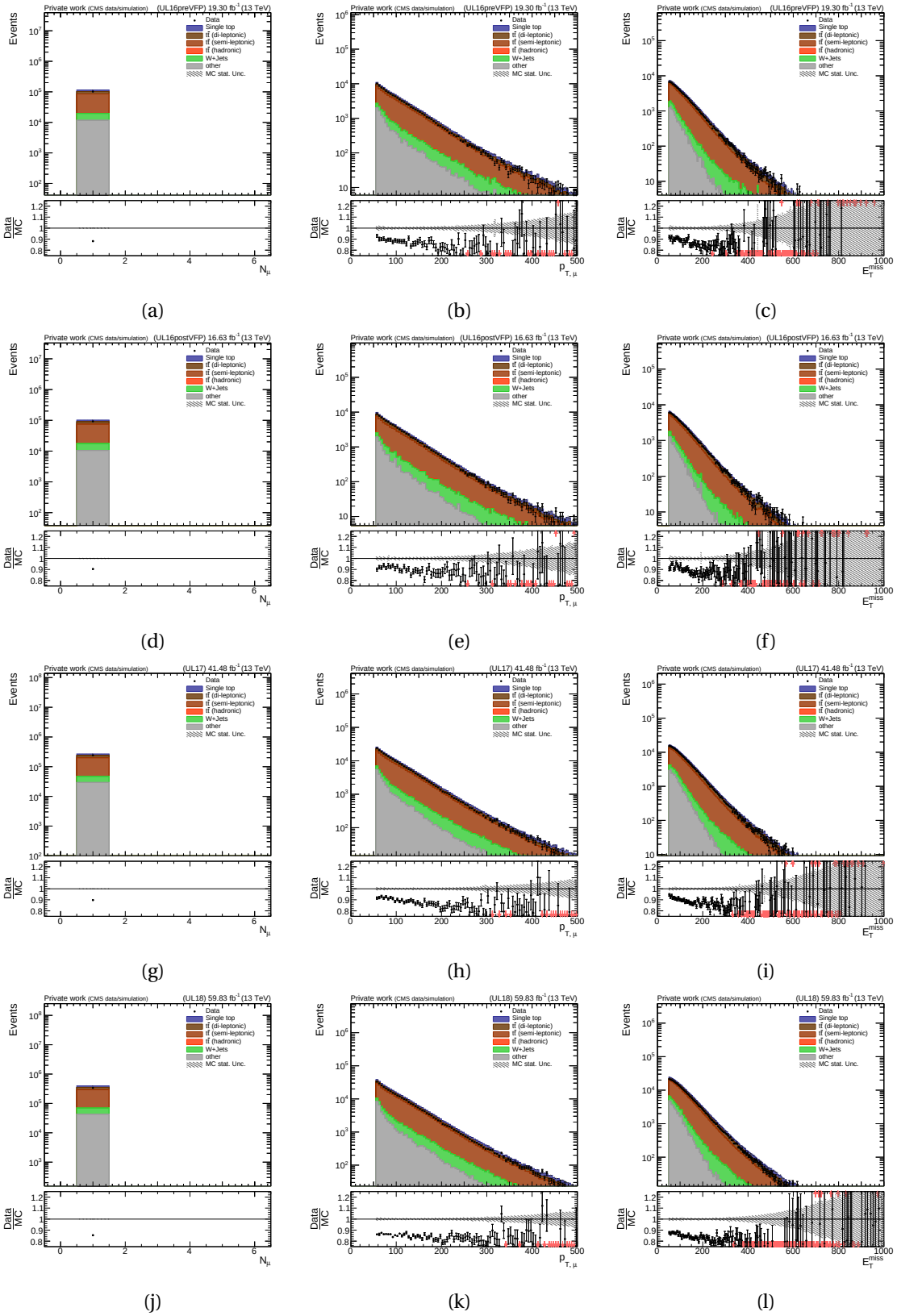


Figure B.2: Control plots of $t\bar{t}$ sample after the pre-selection.

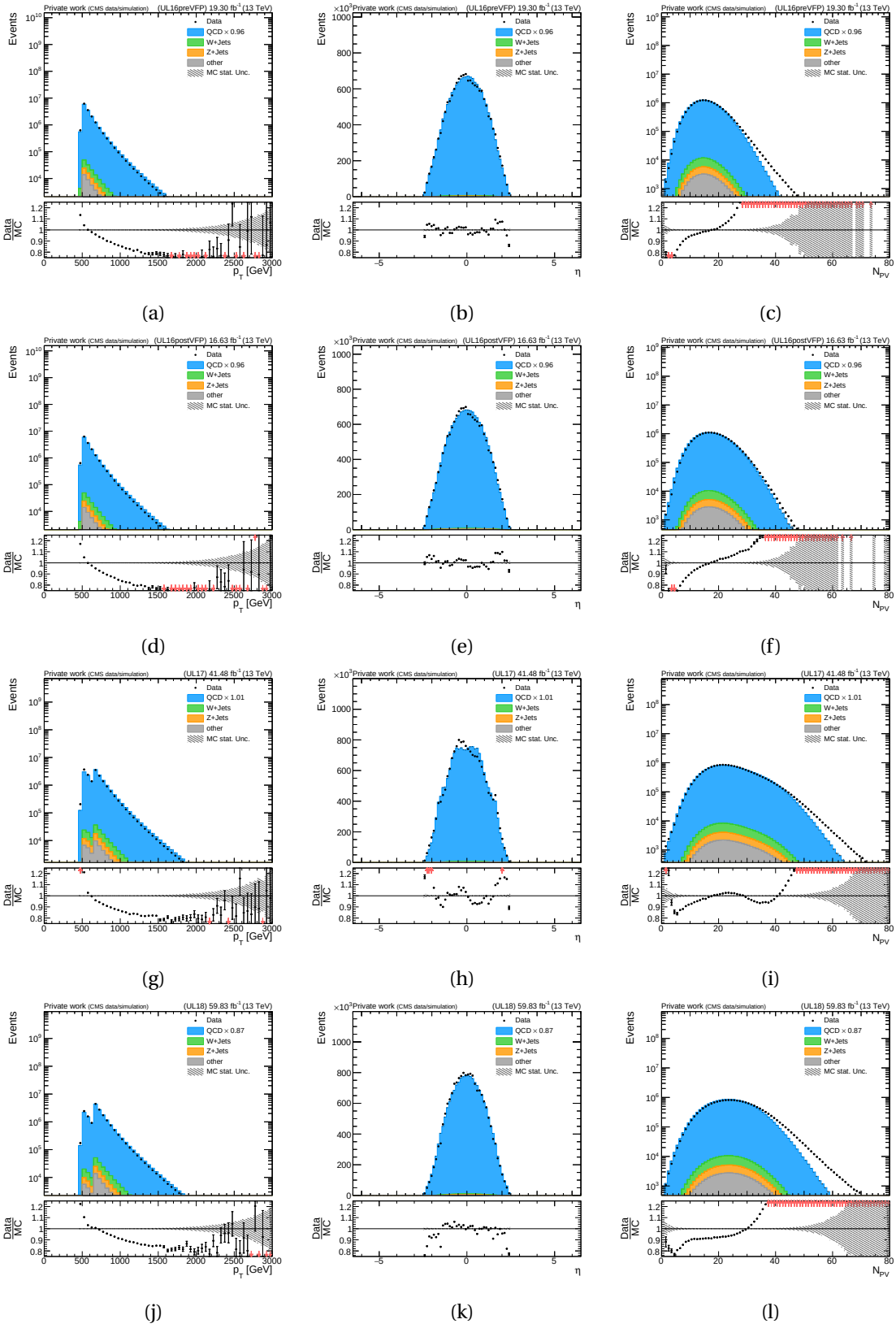


Figure B.3: Control plots of $W(q\bar{q})$+jets sample after the pre-selection.

B.2 Data vs. MC templates in fit regions

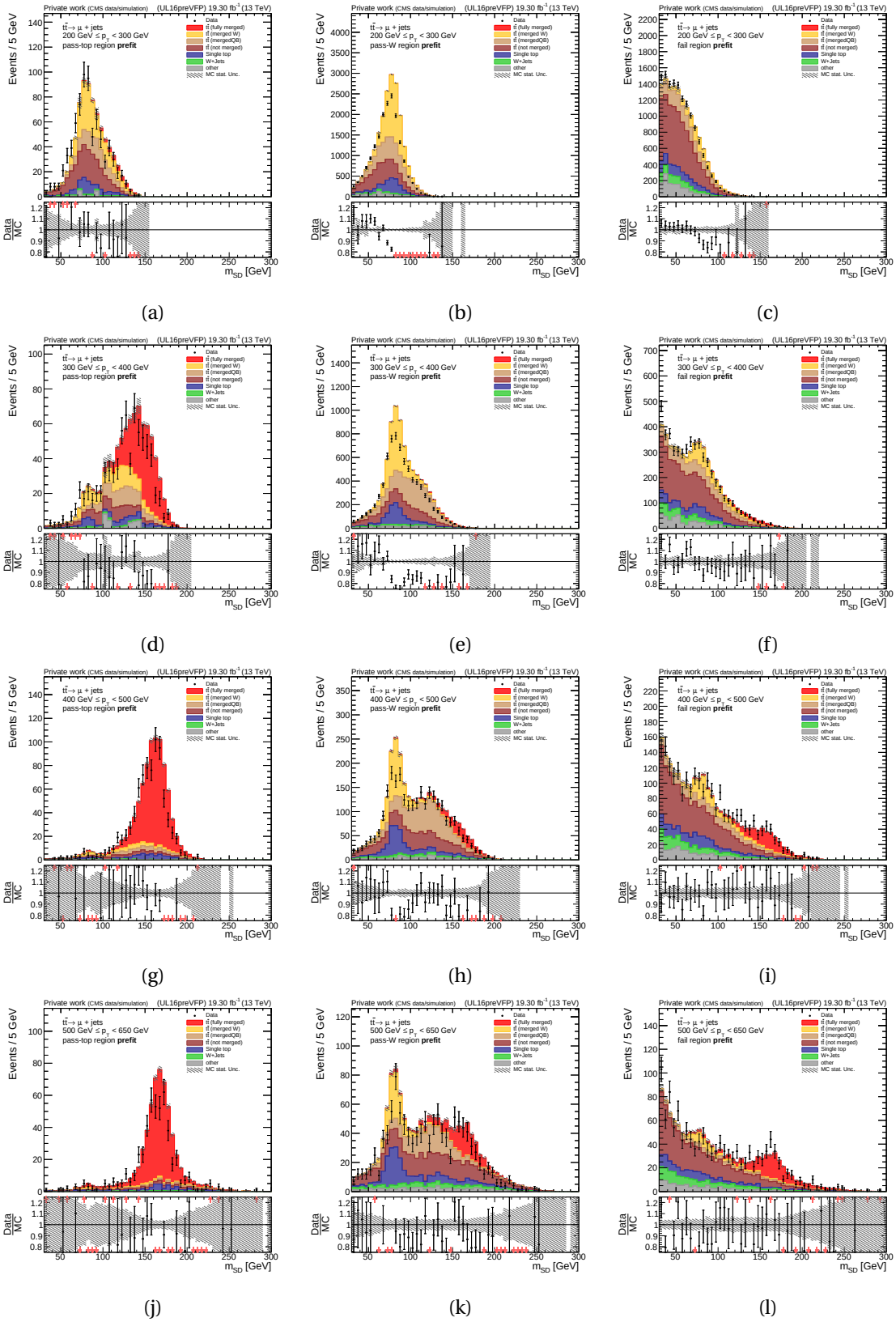


Figure B.4: Data and MC distributions for the tagger regions of the $t\bar{t}$ sample in the `UL16preVFP` dataset.

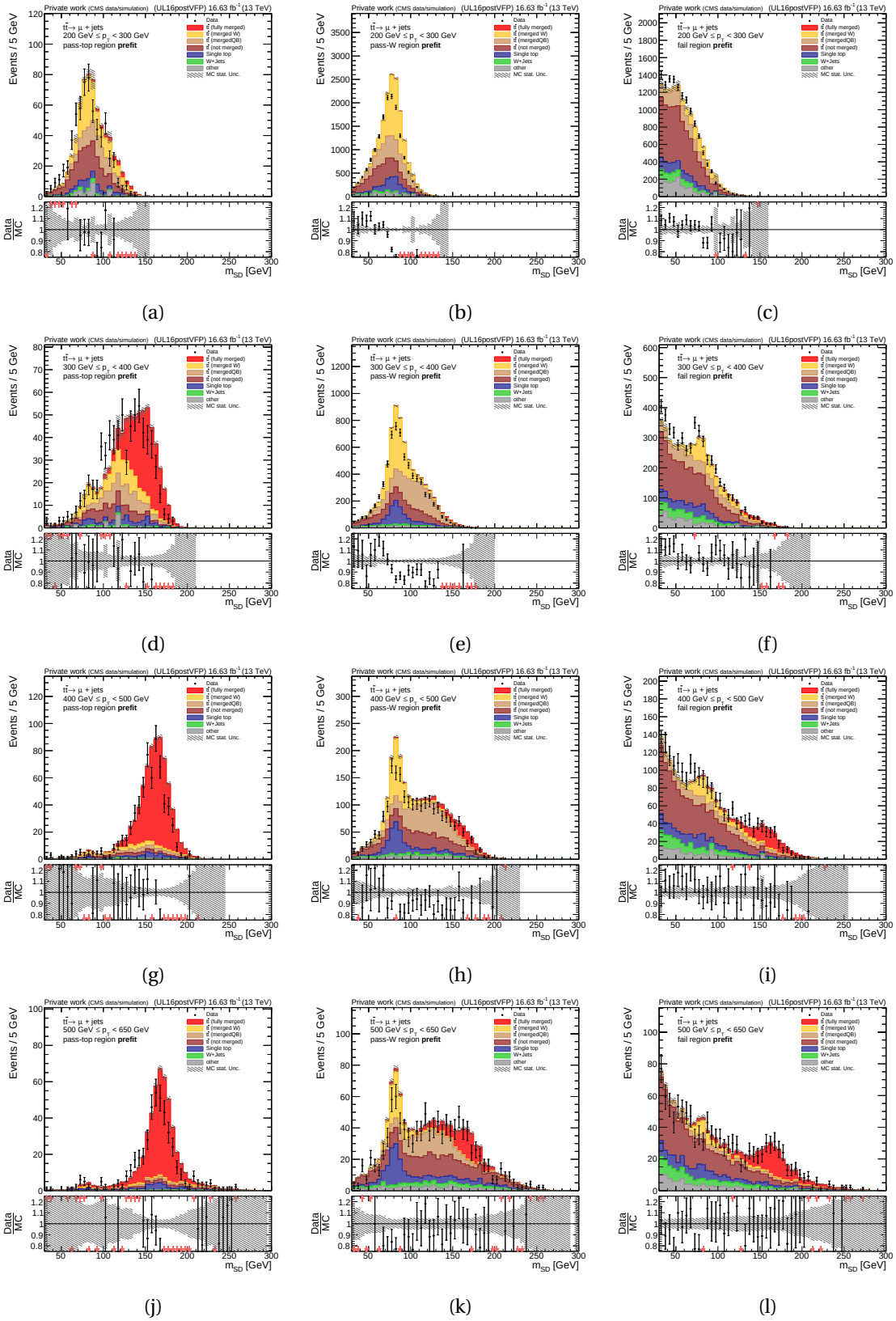


Figure B.5: Data and MC distributions for the tagger regions of the $t\bar{t}$ sample in the `UL16postVFP` dataset.

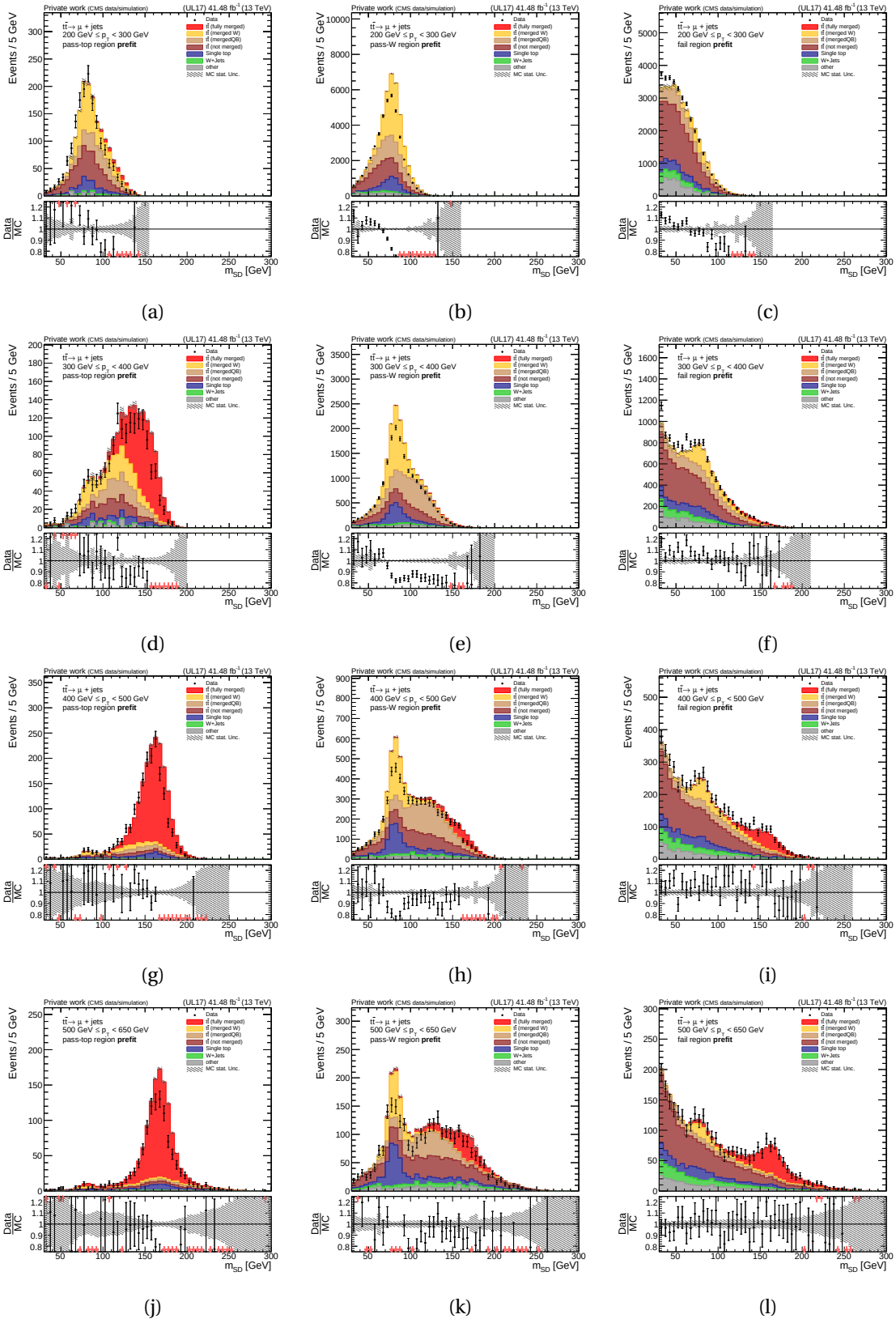


Figure B.6: Data and MC distributions for the tagger regions of the $t\bar{t}$ sample in the UL17 dataset.

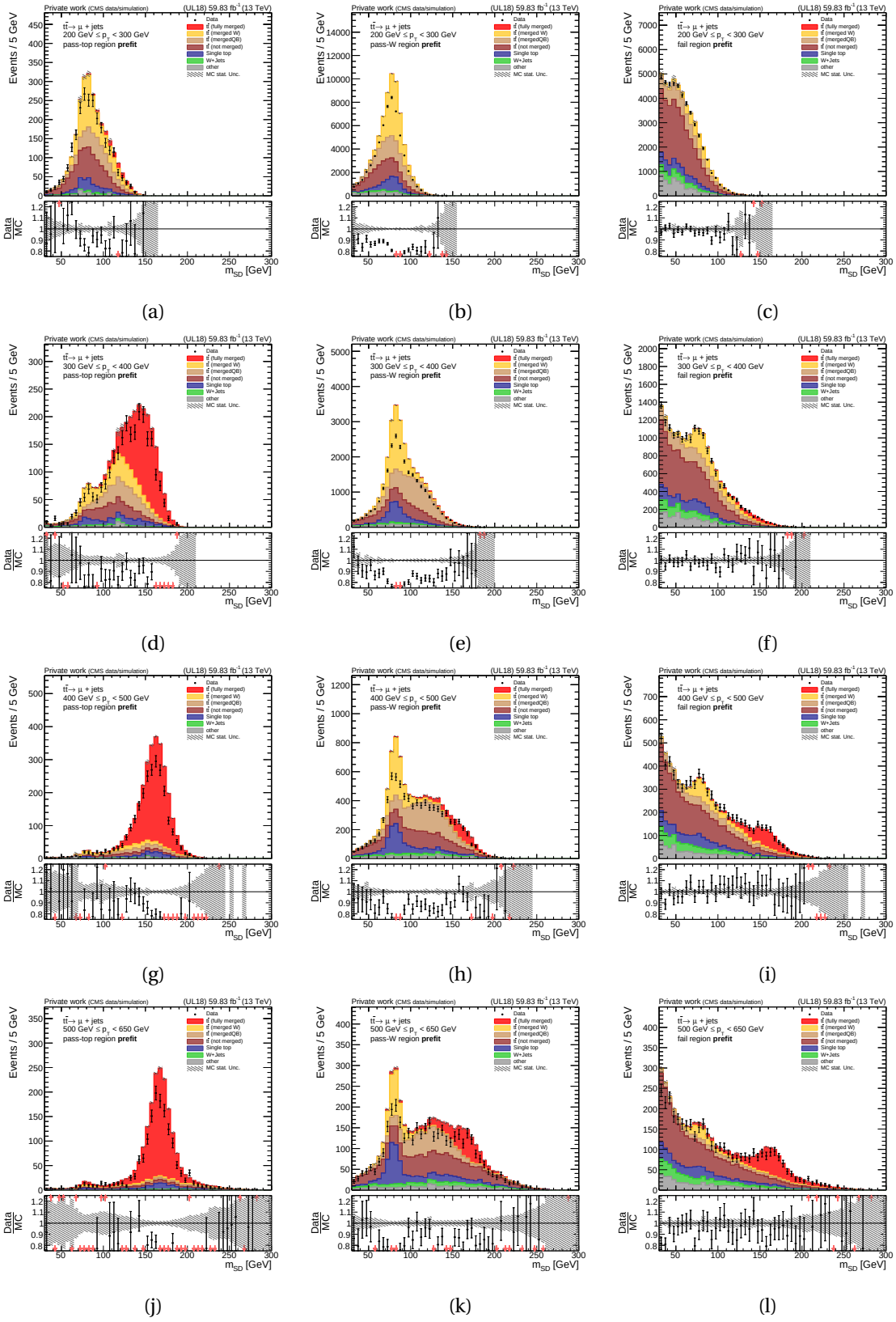


Figure B.7: Data and MC distributions for the tagger regions of the $t\bar{t}$ sample in the UL18 dataset.

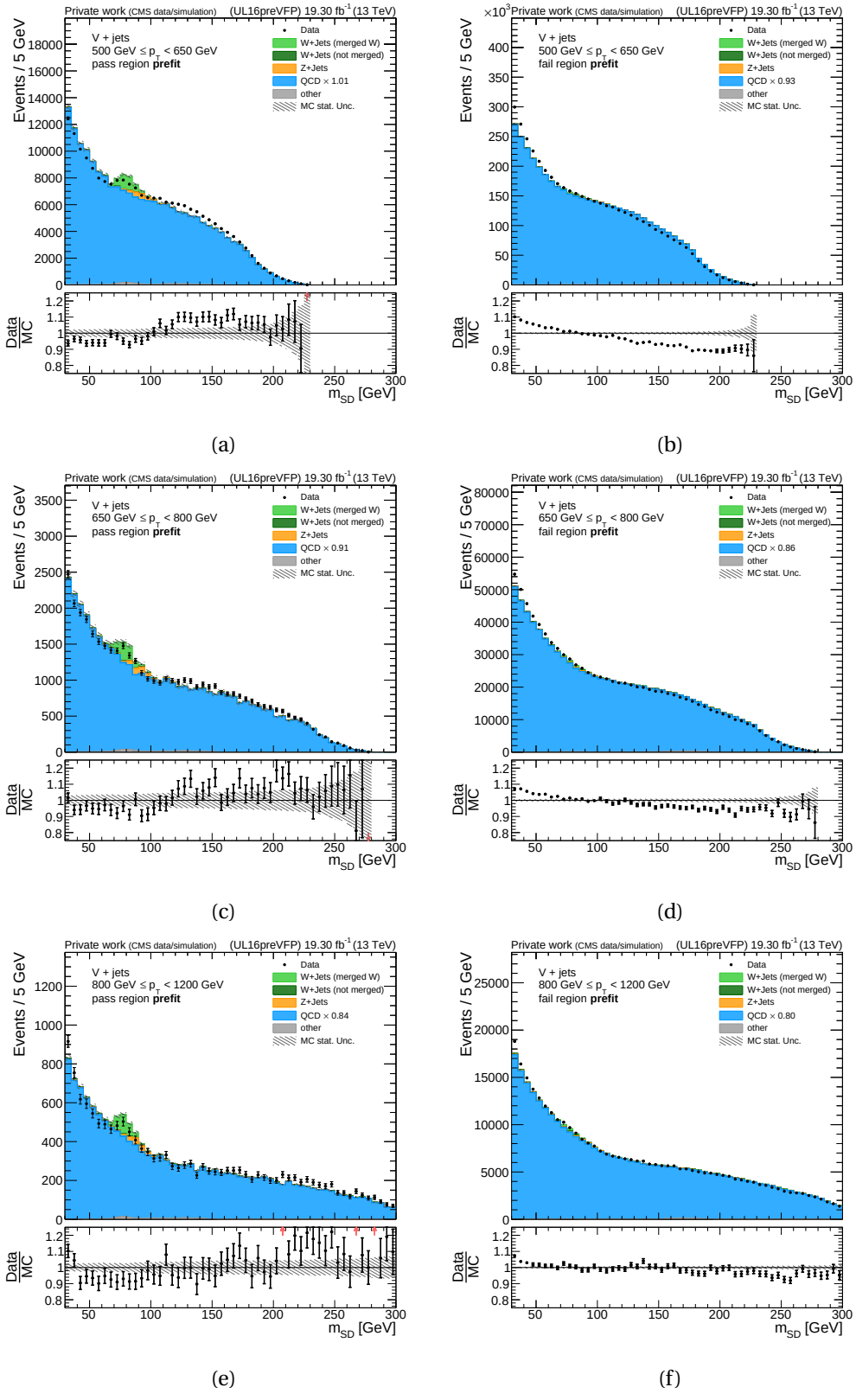
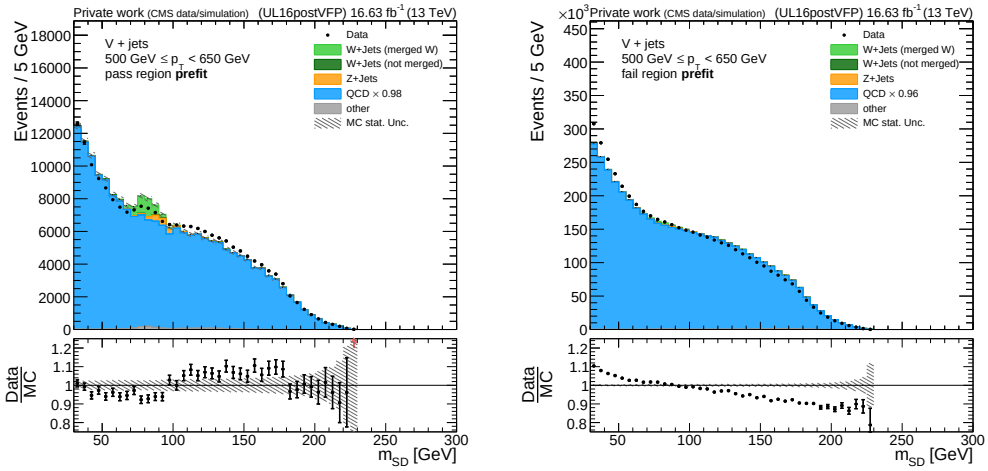
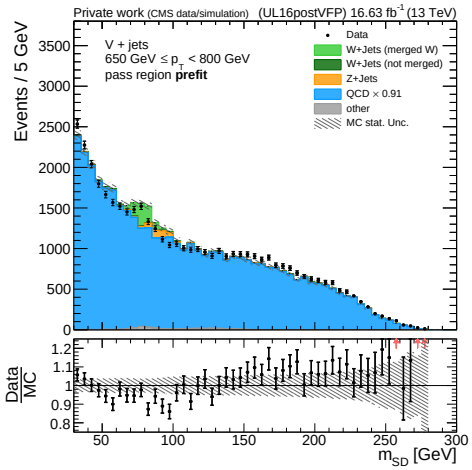


Figure B.8: Data and MC distributions for the W control region in the UL16preVFP dataset.

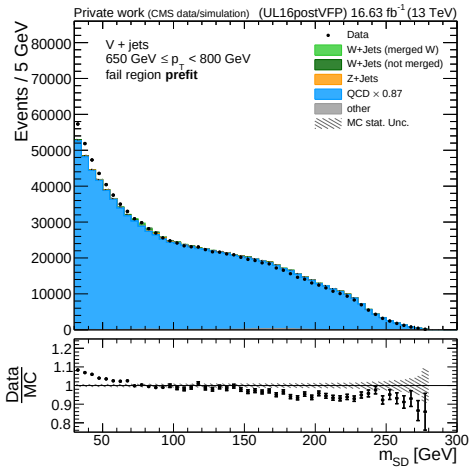


(a)

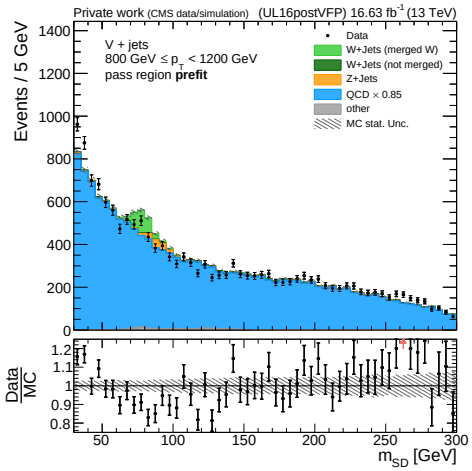
(b)



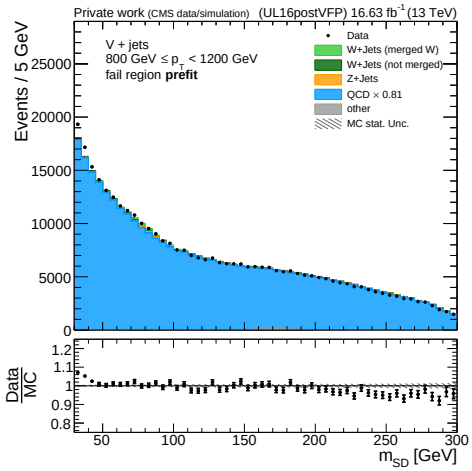
(c)



(d)



(e)



(f)

Figure B.9: Data and MC distributions for the W control region in the UL16postVFP dataset.

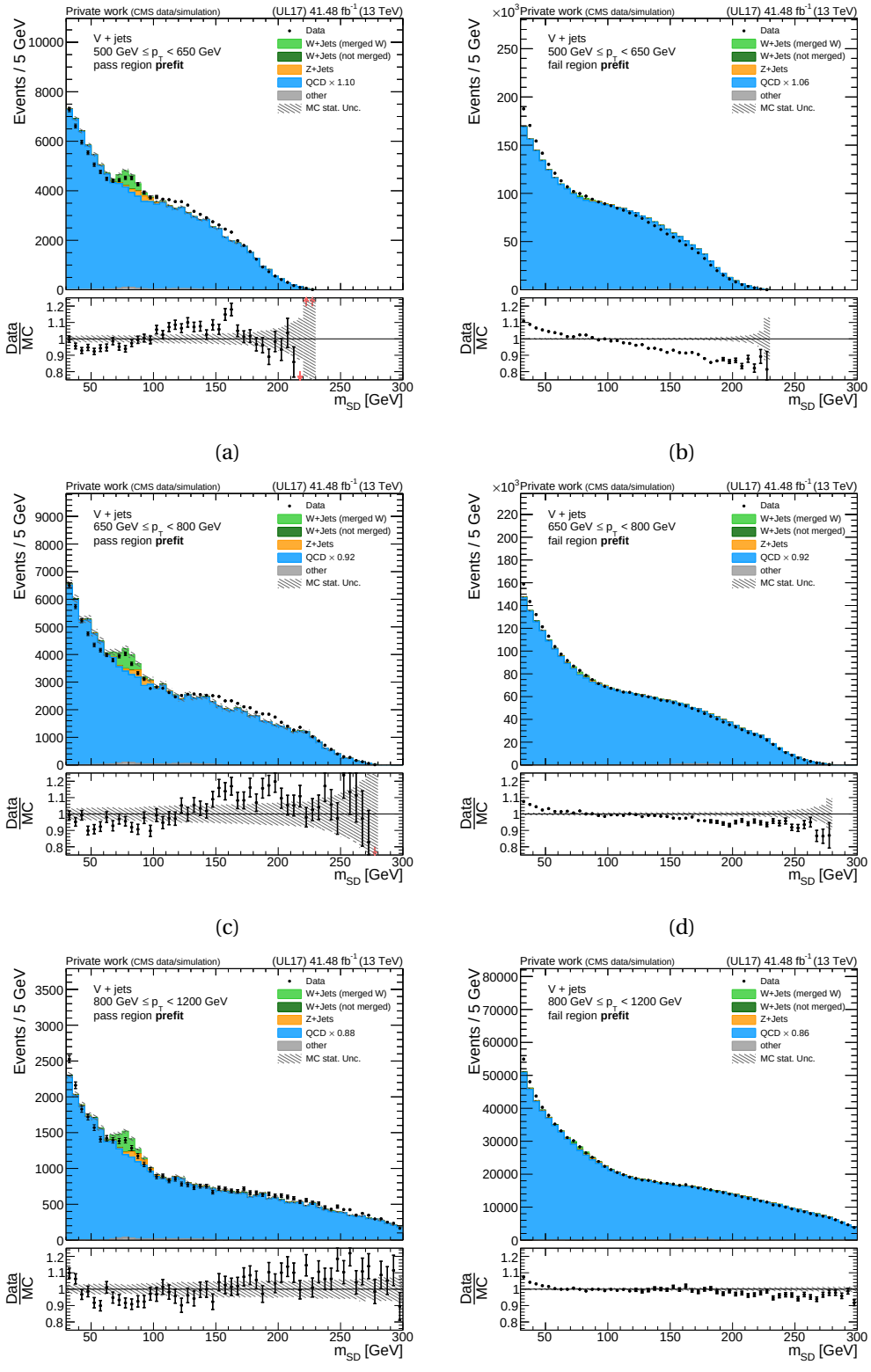
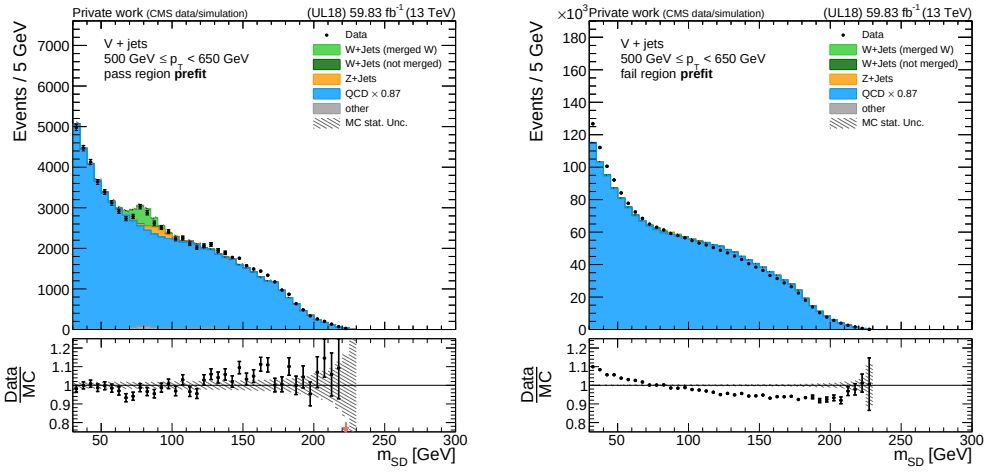
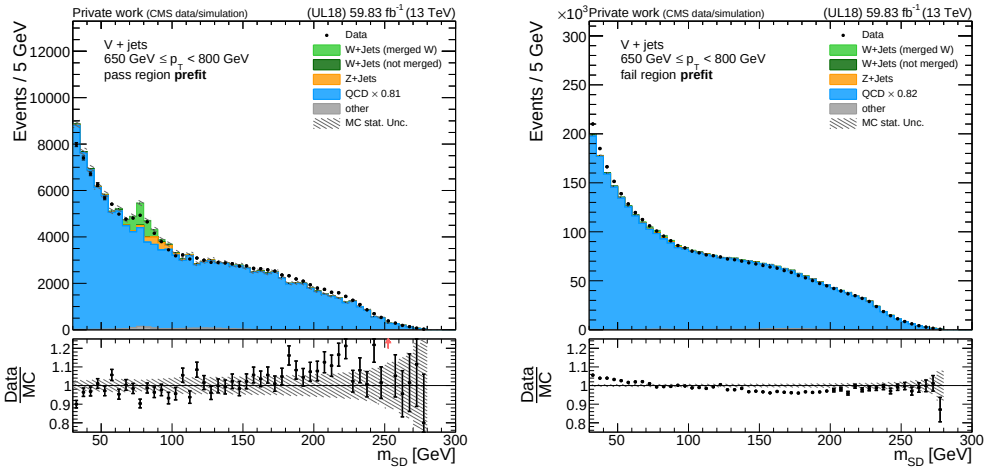


Figure B.10: Data and MC distributions for the W control region in the UL17 dataset.



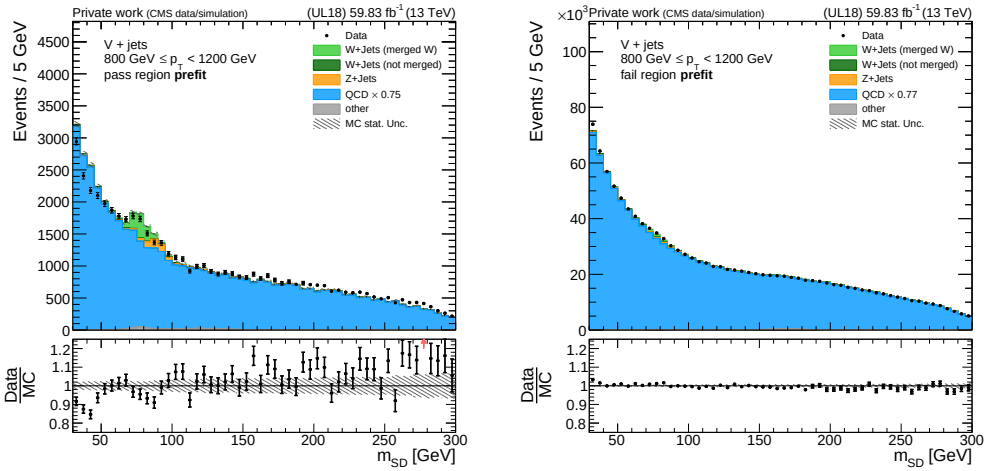
(a)

(b)



(c)

(d)



(e)

(f)

Figure B.11: Data and MC distributions for the W control region in the UL18 dataset.

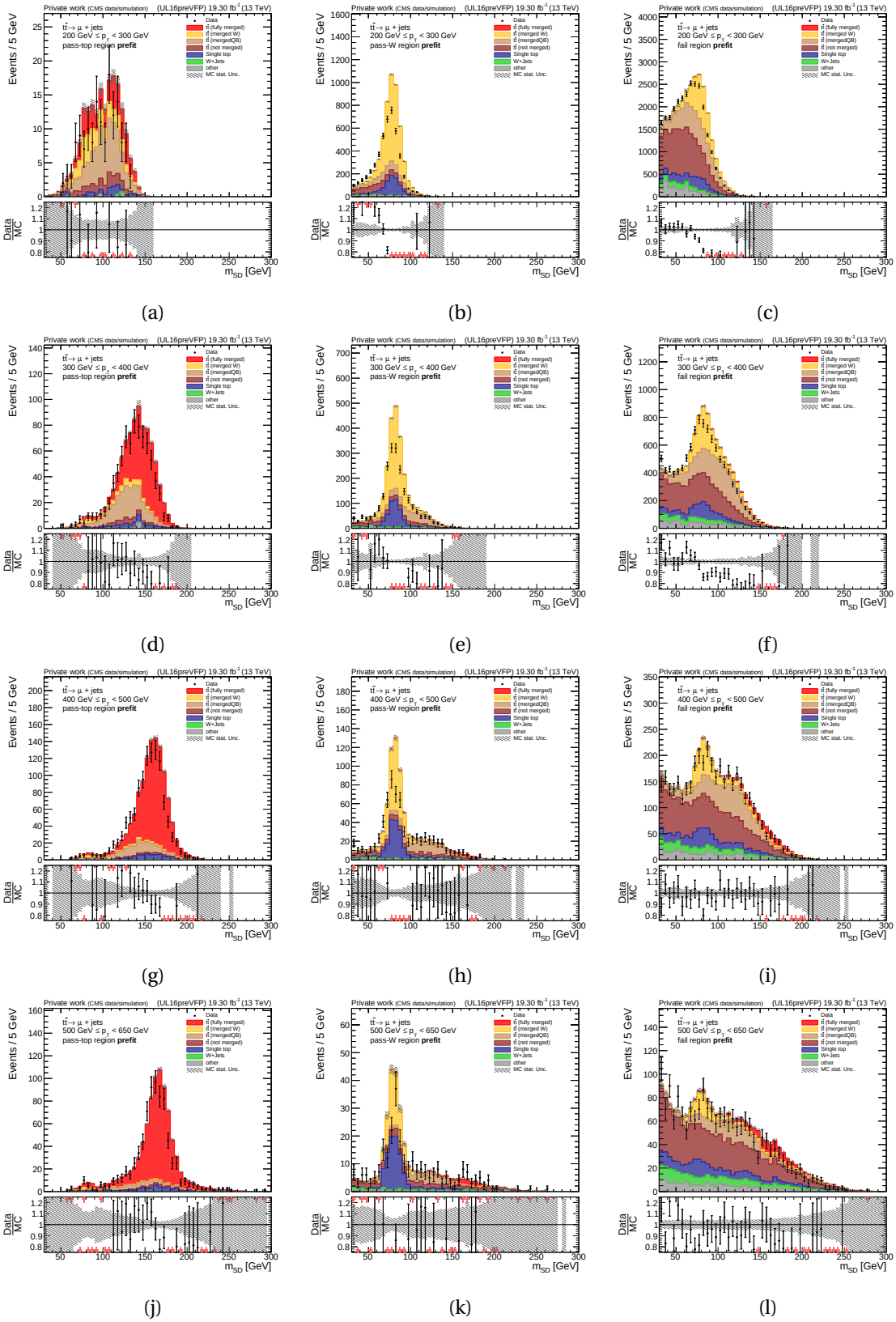


Figure B.12: Data and MC distributions for the tagger regions of the $t\bar{t}$ sample in the `UL16preVFP` dataset.

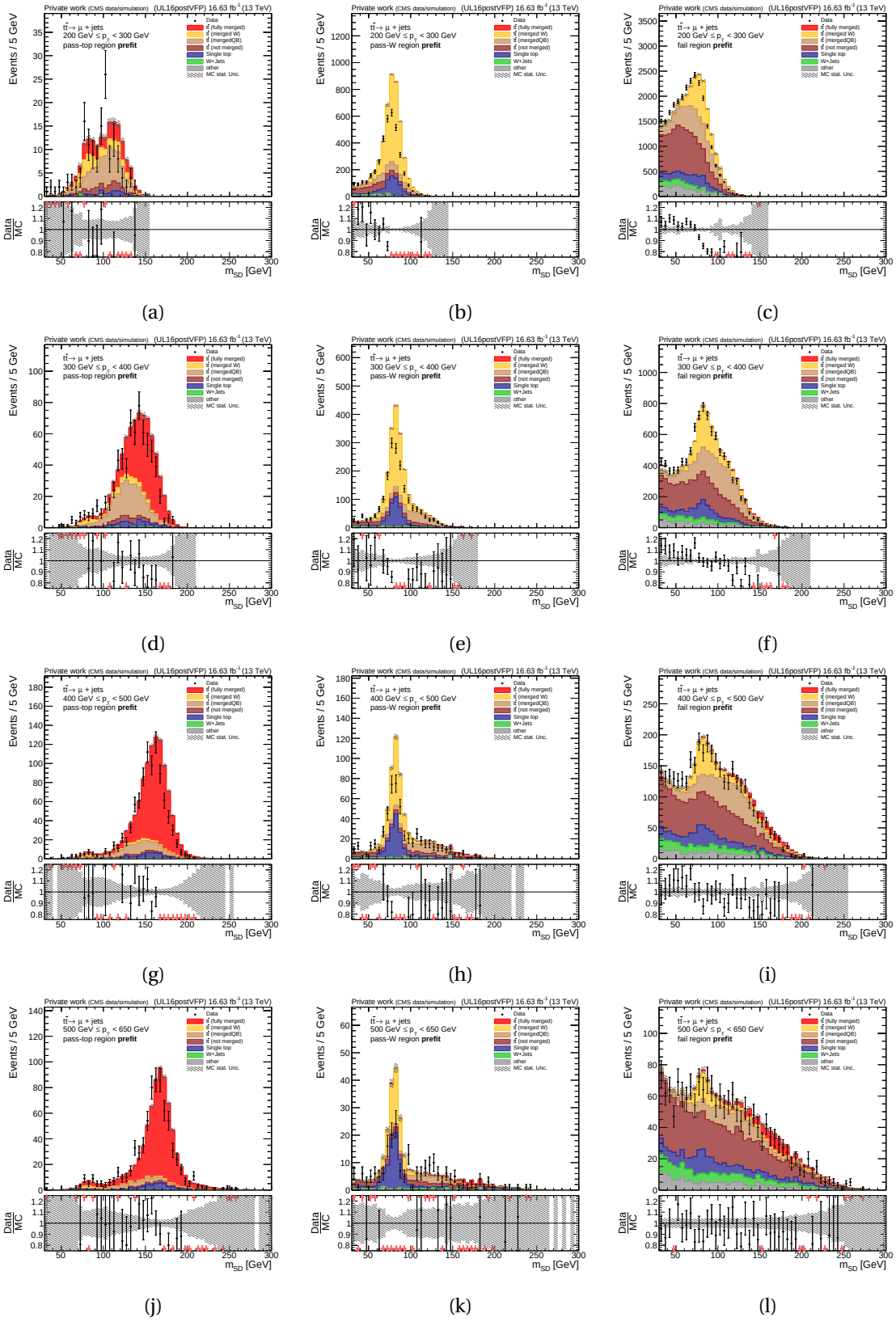


Figure B.13: Data and MC distributions for the tagger regions of the $t\bar{t}$ sample in the UL16postVFP dataset.

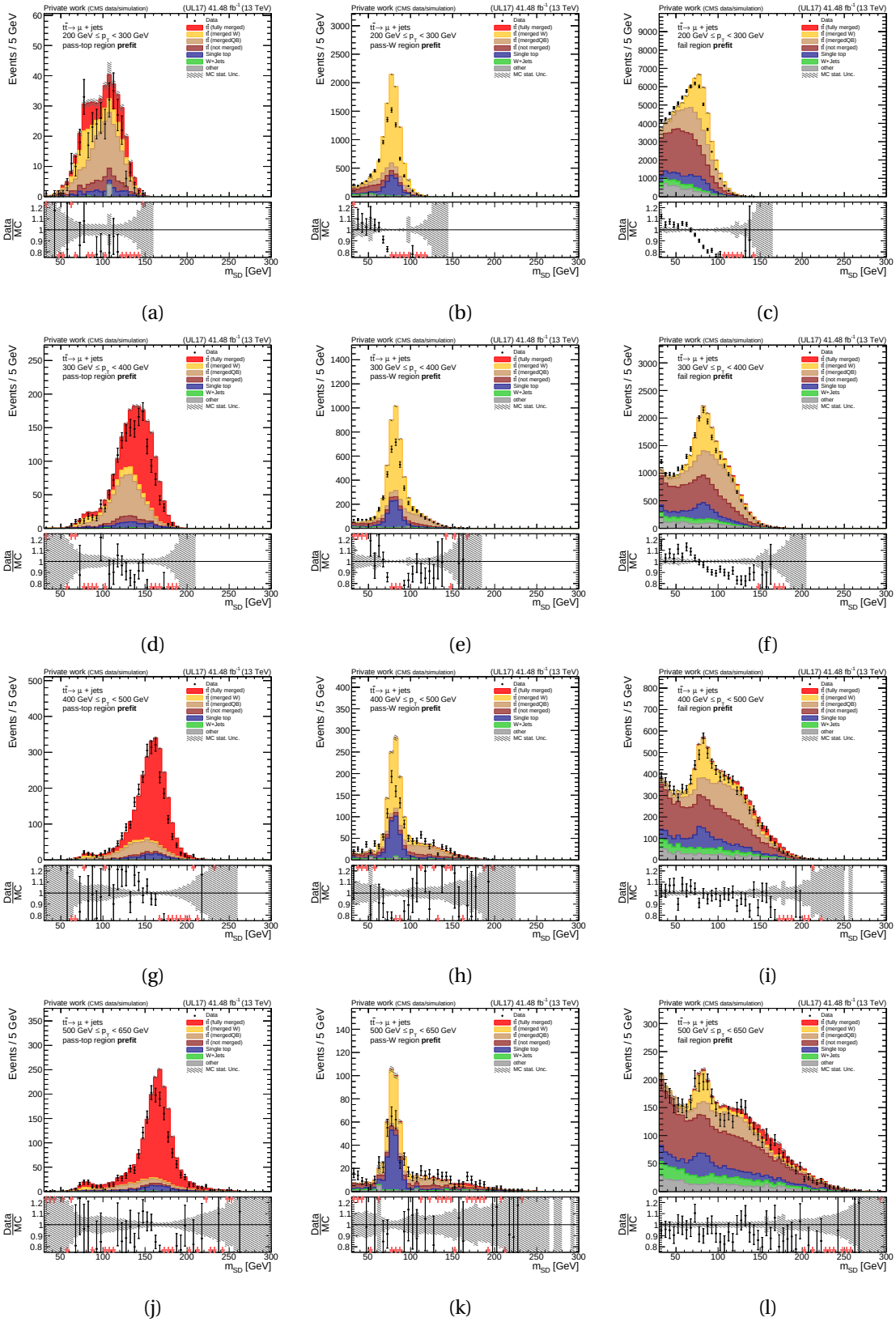


Figure B.14: Data and MC distributions for the tagger regions of the $t\bar{t}$ sample in the UL17 dataset.

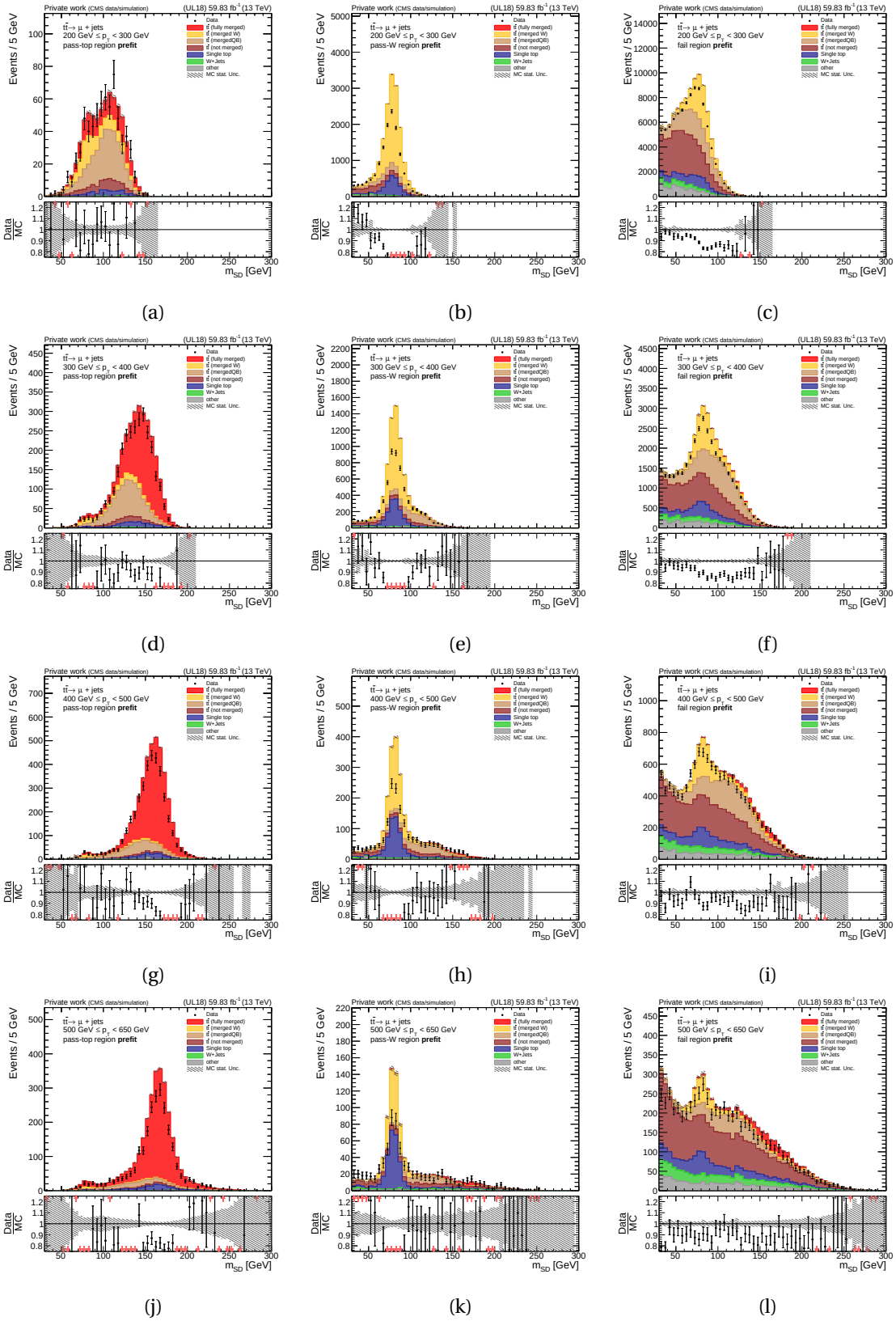
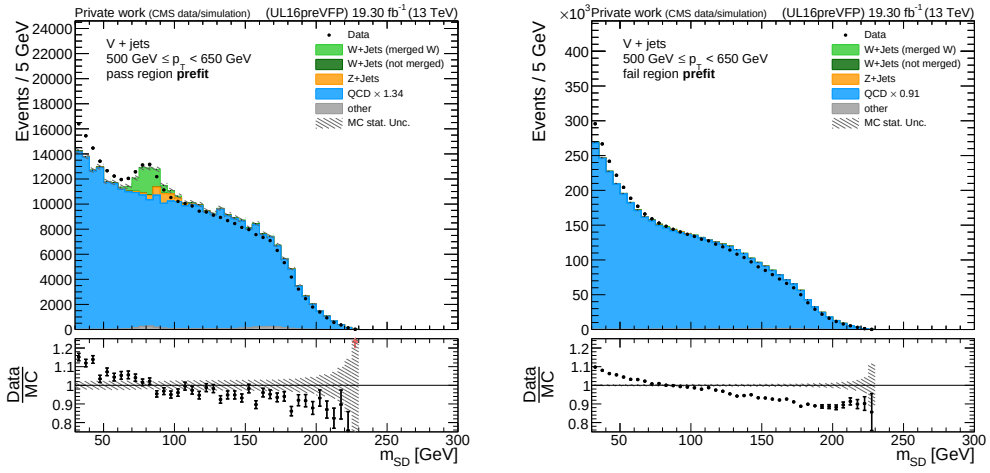
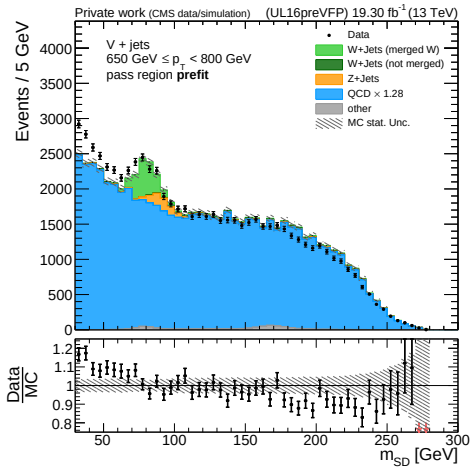


Figure B.15: Data and MC distributions for the tagger regions of the $t\bar{t}$ sample in the UL18 dataset.

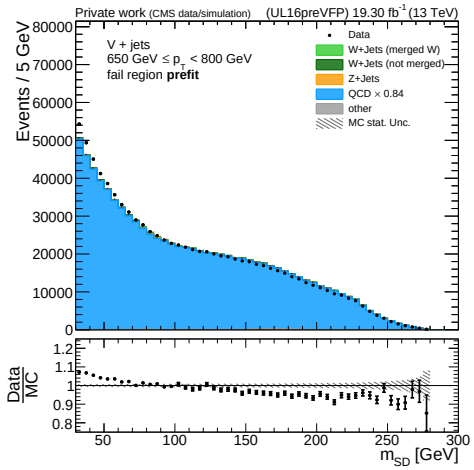


(a)

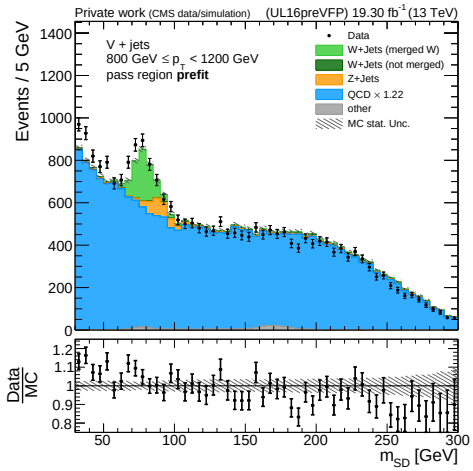
(b)



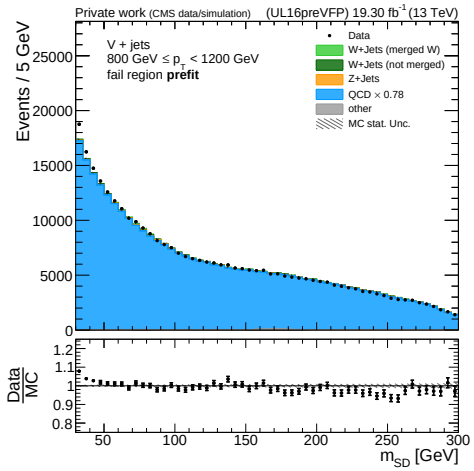
(c)



(d)

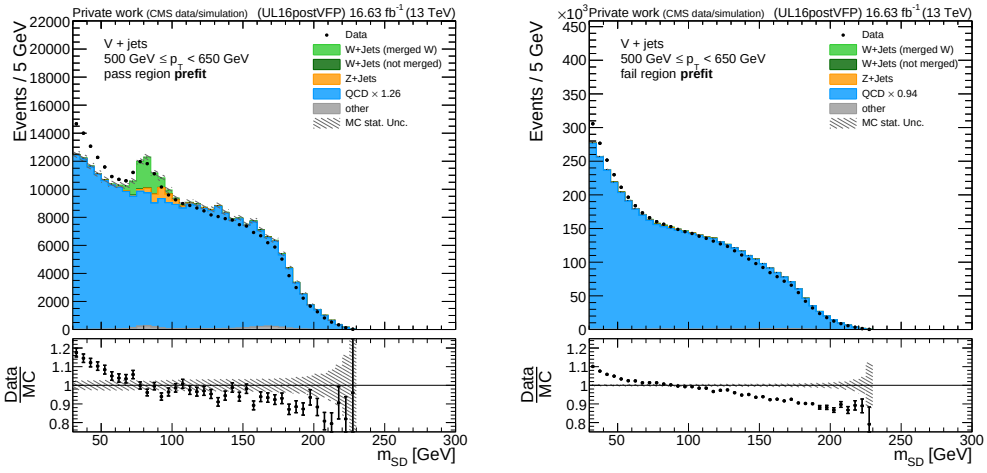


(e)



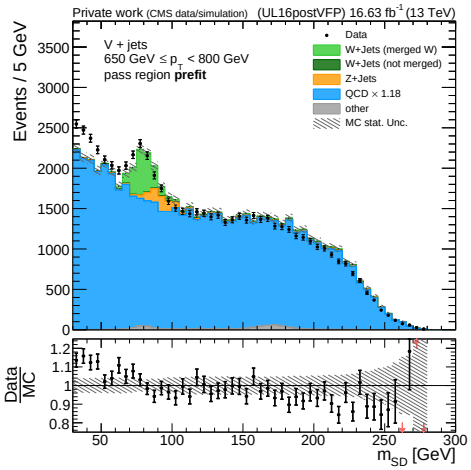
(f)

Figure B.16: Data and MC distributions for the W control region in the UL16preVFP dataset.

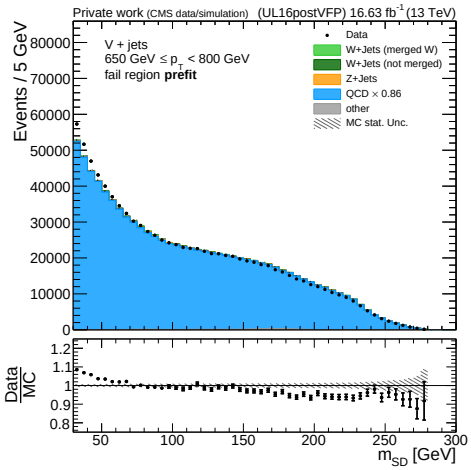


(a)

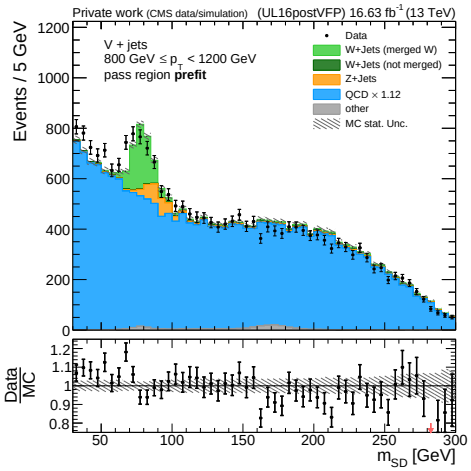
(b)



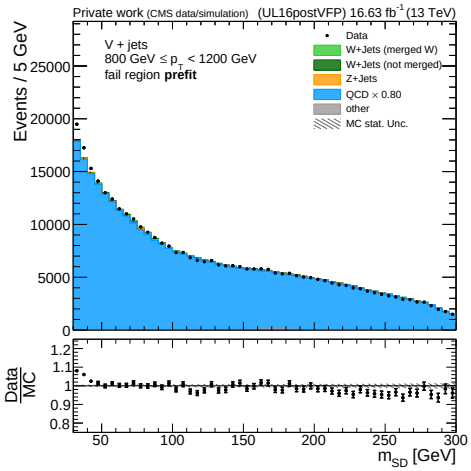
(c)



(d)



(e)



(f)

Figure B.17: Data and MC distributions for the W control region in the UL16postVFP dataset.

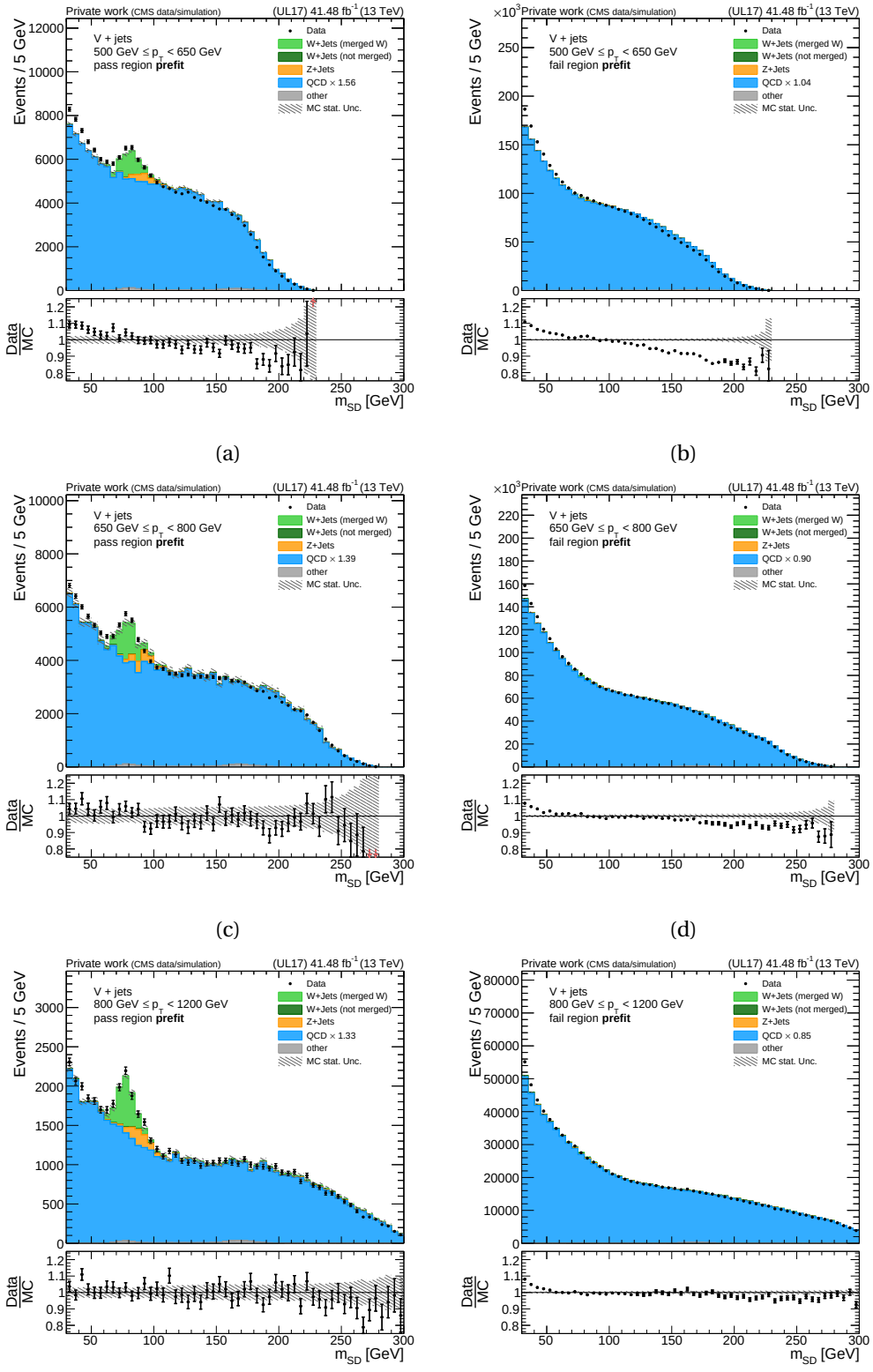


Figure B.18: Data and MC distributions for the W control region in the UL17 dataset.

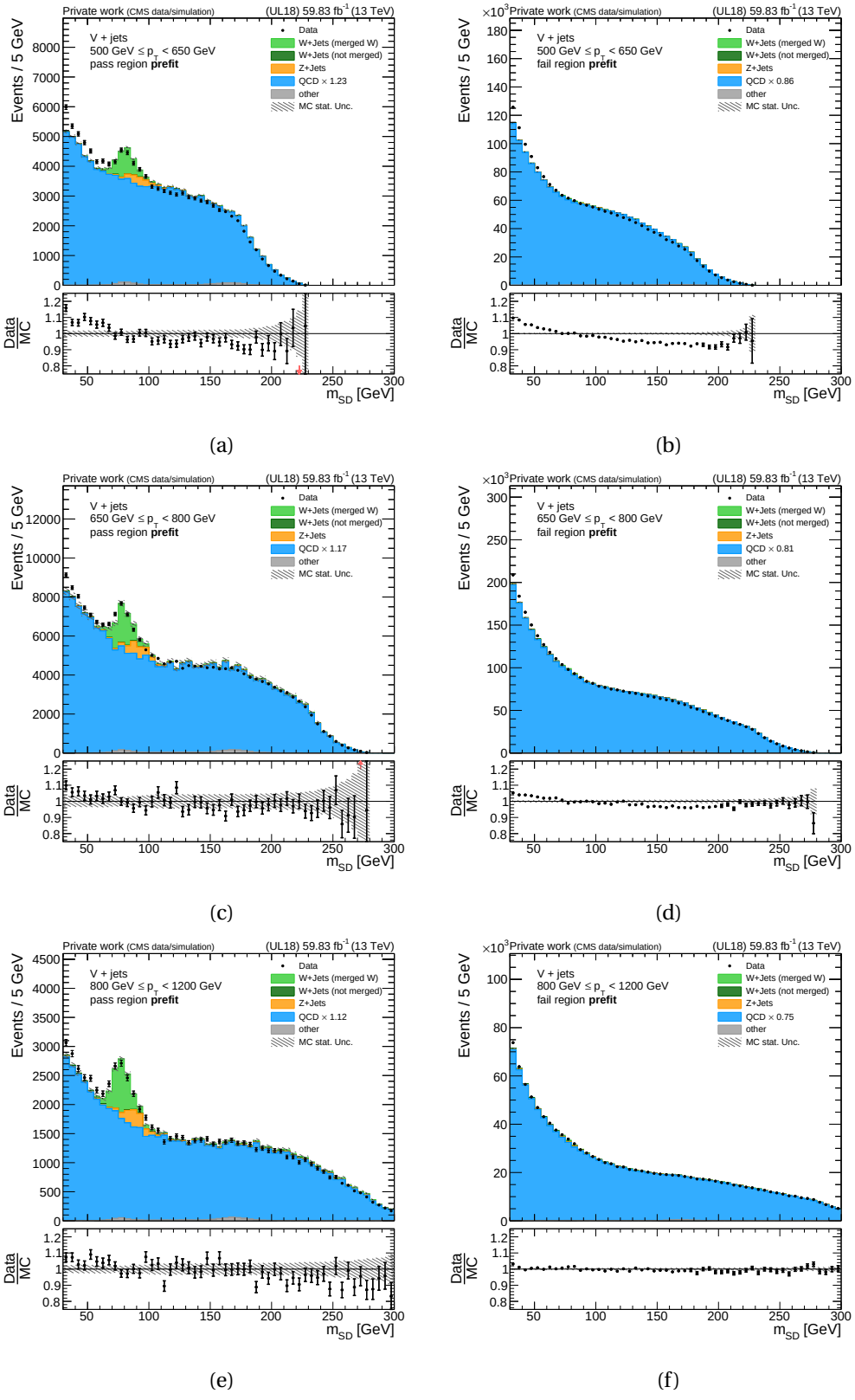


Figure B.19: Data and MC distributions for the W control region in the UL18 dataset.

C QCD background estimation

C.1 Mass-decorrelated tagger

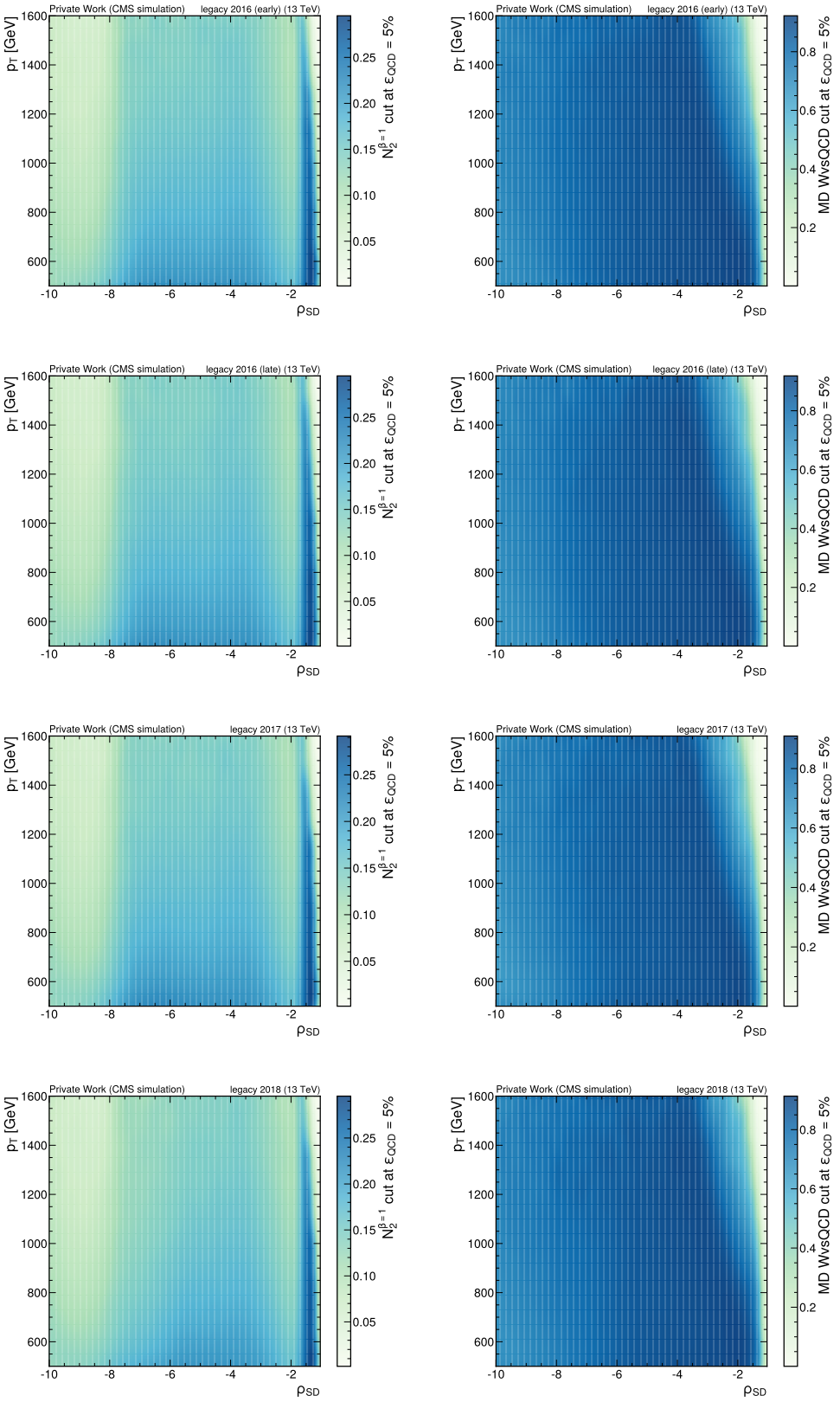


Figure C.1: Two dimensional distribution of 5% percentile in QCD efficiency for the N2 and ParticleNet taggers for all four periods of data-taking.

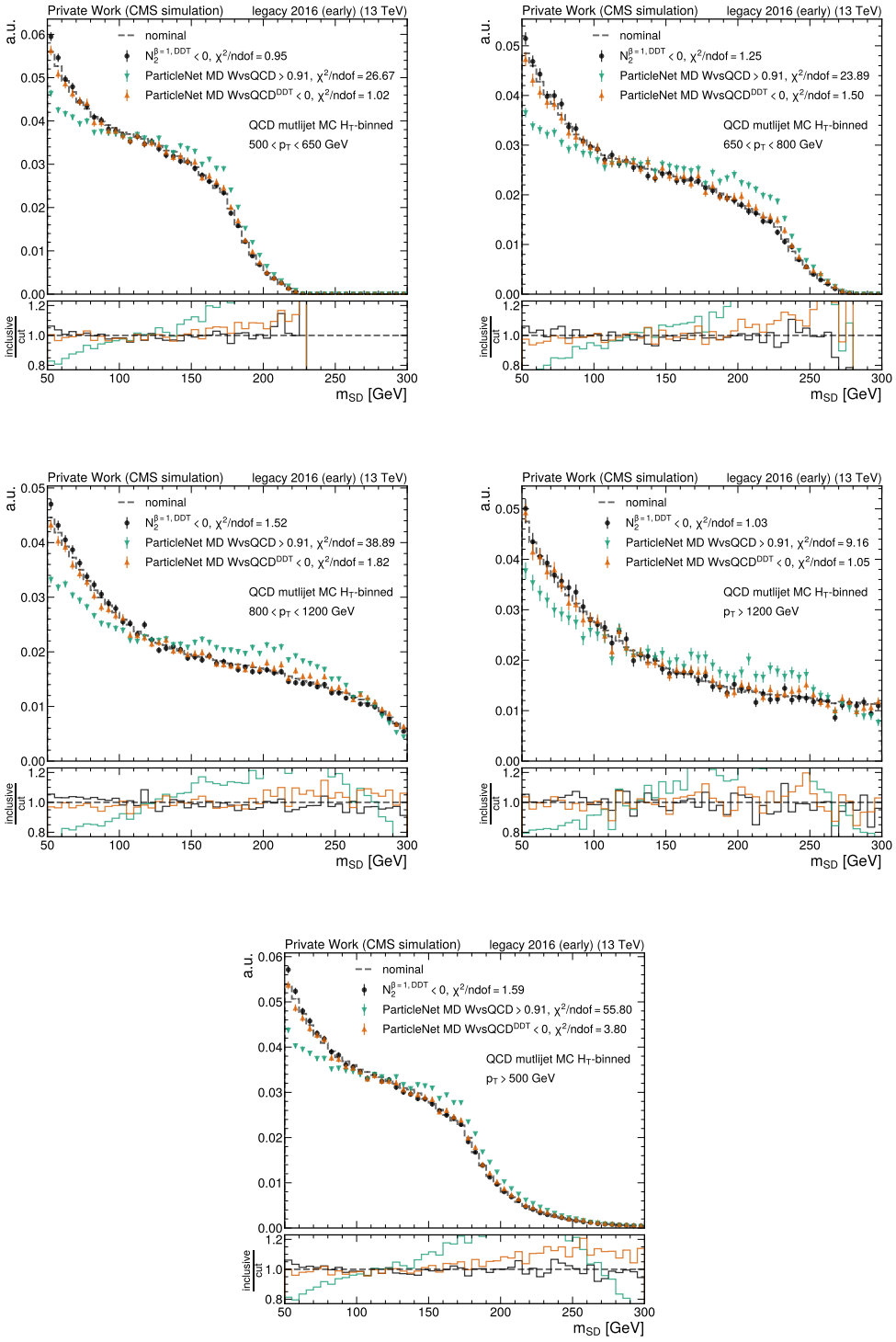


Figure C.2: Distribution of the soft drop mass of the leading AK8 jet in early 2016 QCD multijet events with $p_T > 500\,\mathrm{GeV}$ in different bins of p_T (first and second row) and p_T-inclusive bottom plot. The inclusive distribution is shown as black dashed line, while the marker show the distribution after a cut on the different W tagger. The black points correspond to $N_2^{\beta=1,\mathrm{DDT}} < 0$, the green triangles correspond to MDWvsQCD > 0.91 and the orange triangles correspond to MDWvsQCD$^{\mathrm{DDT}} < 0$. The χ^2/ndof quoted in the legend compares the weighted histograms.

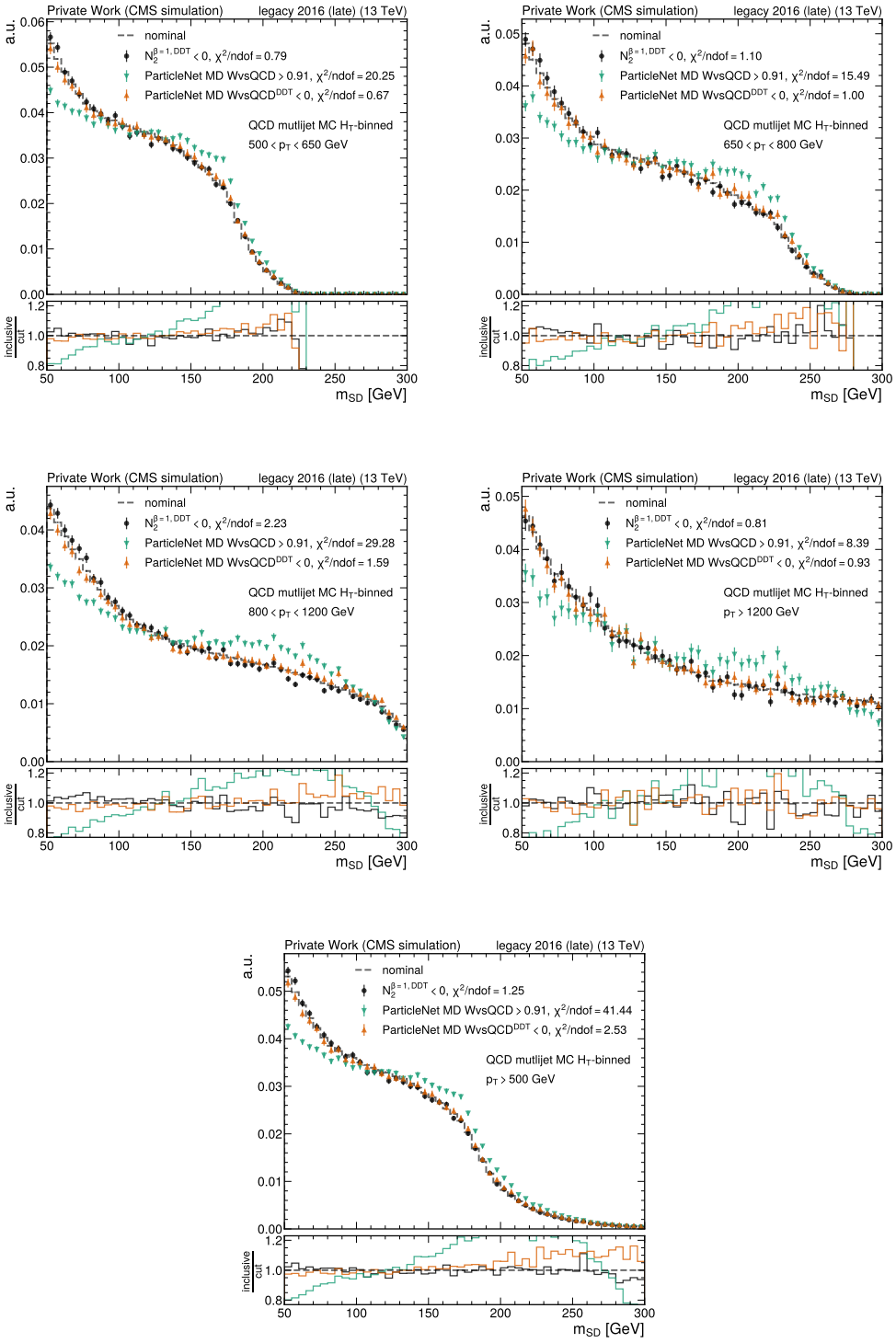


Figure C.3: Distribution of the soft drop mass of the leading AK8 jet in late 2016 QCD multijet events with $p_T > 500\,\text{GeV}$ in different bins of p_T (first and second row) and p_T-inclusive bottom plot. The inclusive distribution is shown as black dashed line, while the marker show the distribution after a cut on the different W tagger. The black points correspond to $N_2^{\beta=1,\text{DDT}} < 0$, the green triangles correspond to MDWvsQCD > 0.91 and the orange triangles correspond to MDWvsQCD$^{\text{DDT}} < 0$. The χ^2/ndof quoted in the legend compares the weighted histograms.

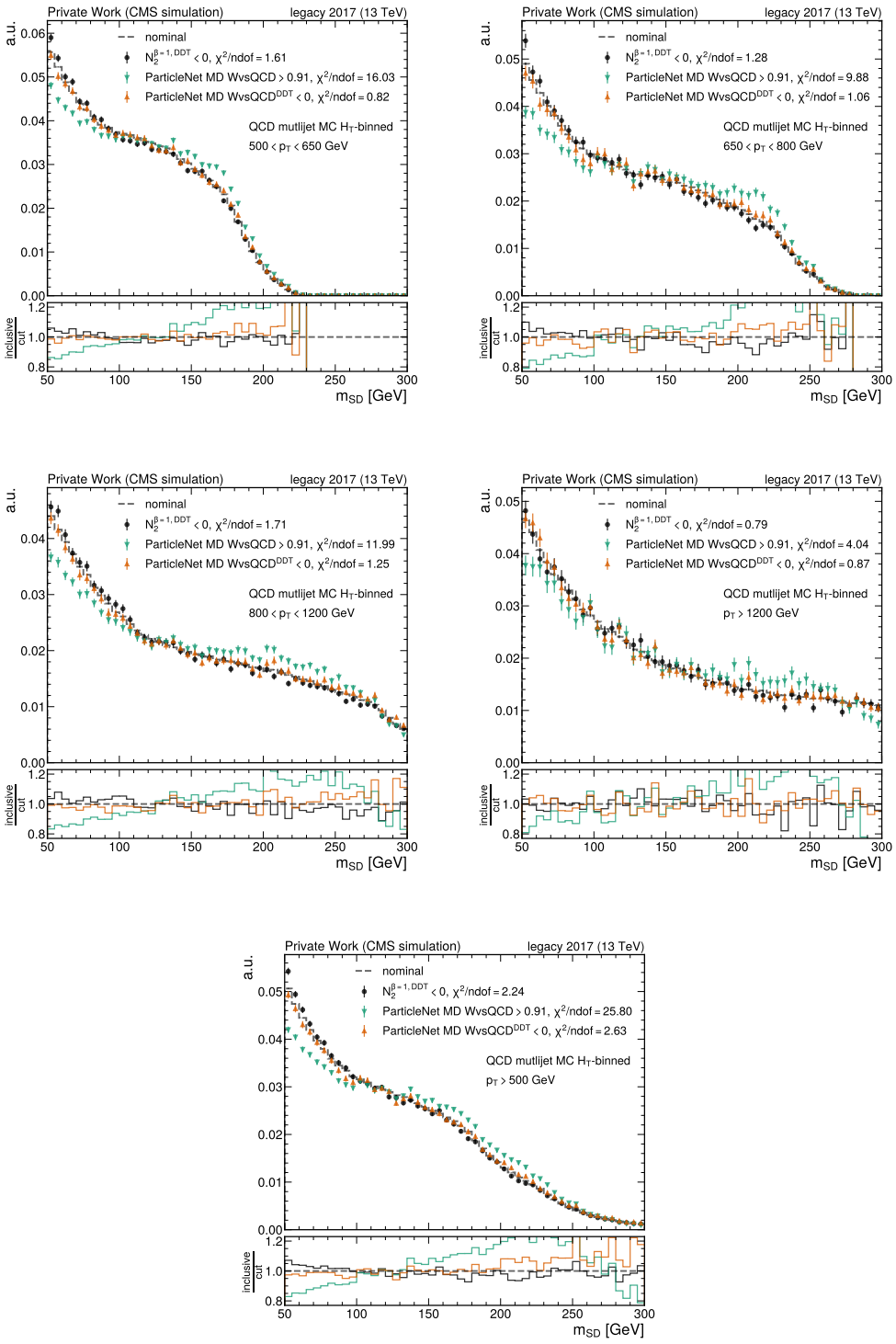


Figure C.4: Distribution of the soft drop mass of the leading AK8 jet in 2017 QCD multijet events with $p_T > 500$ GeV in different bins of p_T (first and second row) and p_T-inclusive bottom plot. The inclusive distribution is shown as black dashed line, while the marker show the distribution after a cut on the different W tagger. The black points correspond to $N_2^{\beta=1,DDT} < 0$, the green triangles correspond to MDWvsQCD > 0.91 and the orange triangles correspond to MDWvsQCD$^{DDT} < 0$. The χ^2/ndof quoted in the legend compares the weighted histograms.

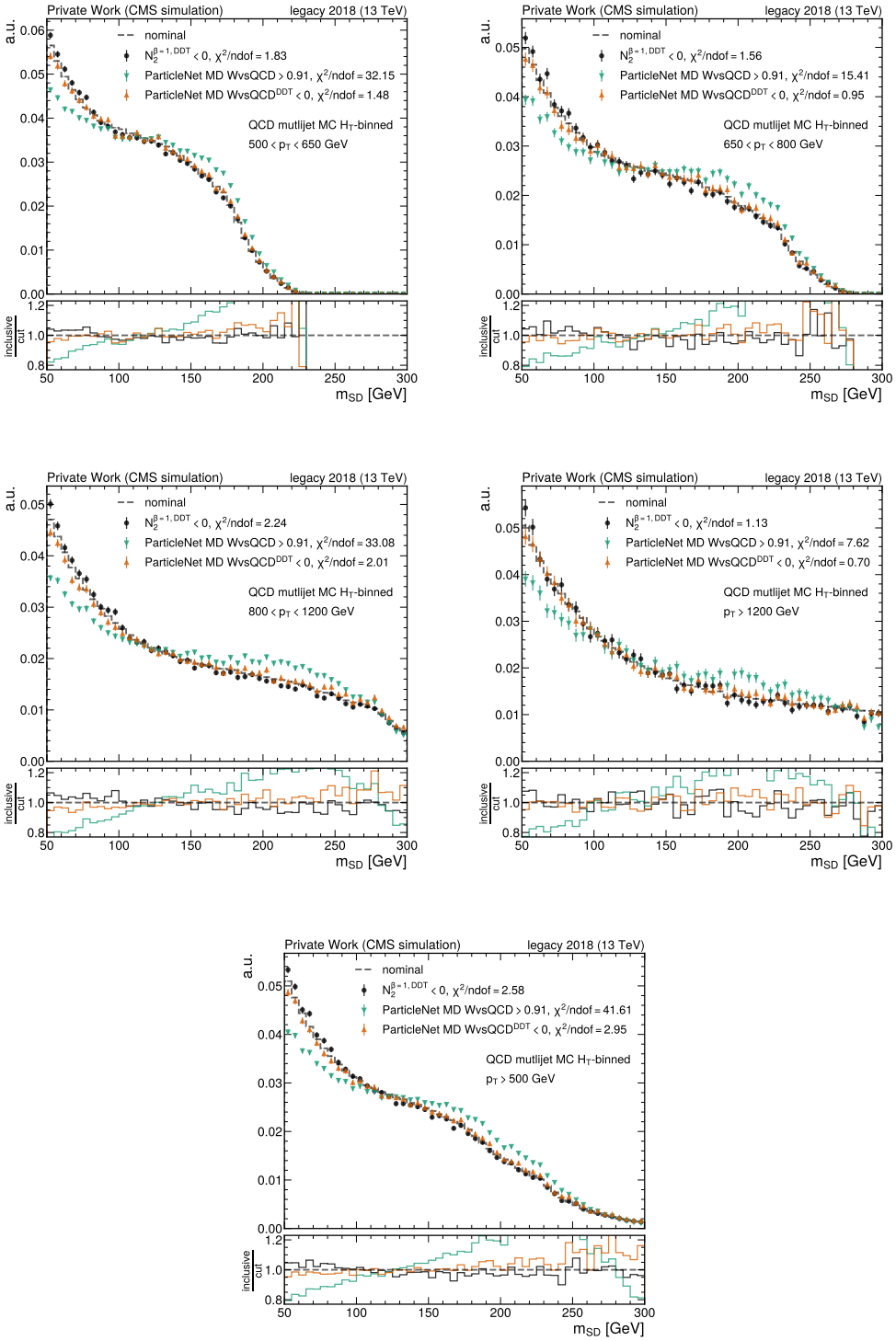


Figure C.5: Distribution of the soft drop mass of the leading AK8 jet in 2018 QCD multijet events with $p_T > 500$ GeV in different bins of p_T (first and second row) and p_T-inclusive bottom plot. The inclusive distribution is shown as black dashed line, while the marker show the distribution after a cut on the different W tagger. The black points correspond to $N_2^{\beta=1,\mathrm{DDT}} < 0$, the green triangles correspond to MDWvsQCD > 0.91 and the orange triangles correspond to MDWvsQCD$^{\mathrm{DDT}} < 0$. The χ^2/ndof quoted in the legend compares the weighted histograms.

C.2 QCD mistag efficiency in data

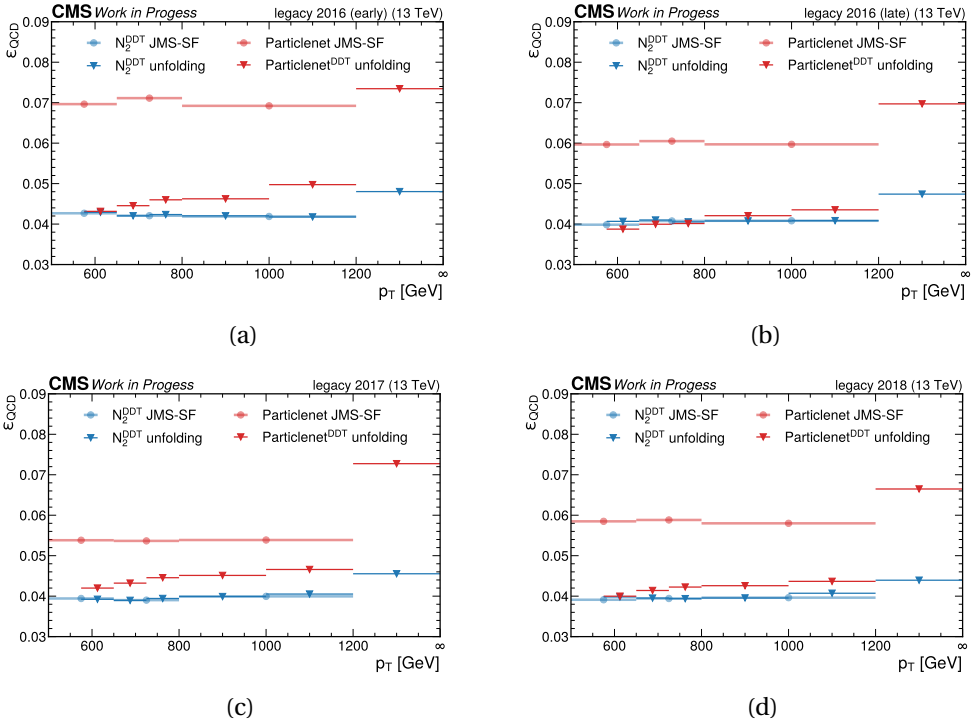


Figure C.6: Summary of the selection efficiency for QCD multijet background when using the different tagging approaches in the data-driven background estimation. The blue and red circles correspond to QCD mistag rates when using $N_2^{\beta=1,\mathrm{DDT}}$ and ParticleNet discriminators respectively as they are used for the jet mass calibration discussed in Section 7. Here the efficiency is measured in three p_T bins with the edges $[500, 650, 800, 1200]$. The blue and red triangles correspond to the QCD mistag rates when using the $N_2^{\beta=1,\mathrm{DDT}}$ and ParticleNet$^{\mathrm{DDT}}$ respectively. These are used in the measurement of the jet mass distribution discussed in Section 8. Here the efficiency is measured in six p_T bins with the edges $[575, 650, 725, 800, 1000, 1200, \infty]$.

C.3 F-Test results

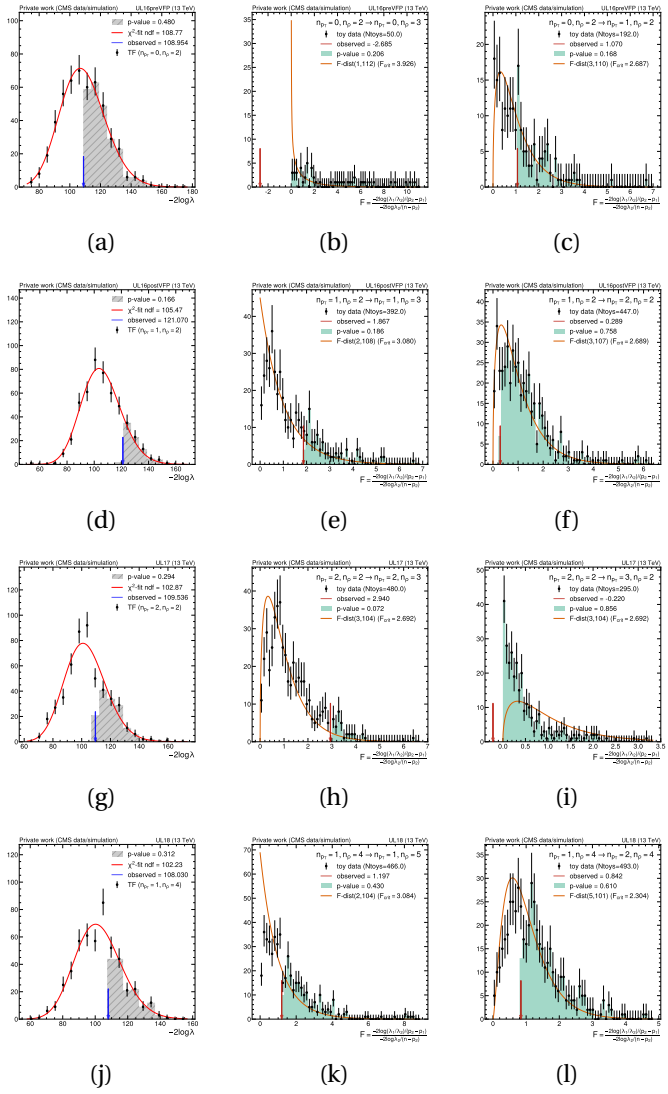


Figure C.7: Overview of metrics of the final iteration of F-Test for the tagger approach using $N_2^{\beta=1,\mathrm{DDT}}$ as W tagger in the QCD multijet background estimation. The rows correspond to the different periods of data taking, while the columns show from left to right: the Goodness of fit using the saturated model using the order combination under test (base model), the F-Test values and distribution testing the base model against more complex models by increasing $n_{\rho_{\mathrm{SD}}}$ by one (center) and increasing $n_{p_{\mathrm{T}}}$ by one (right).

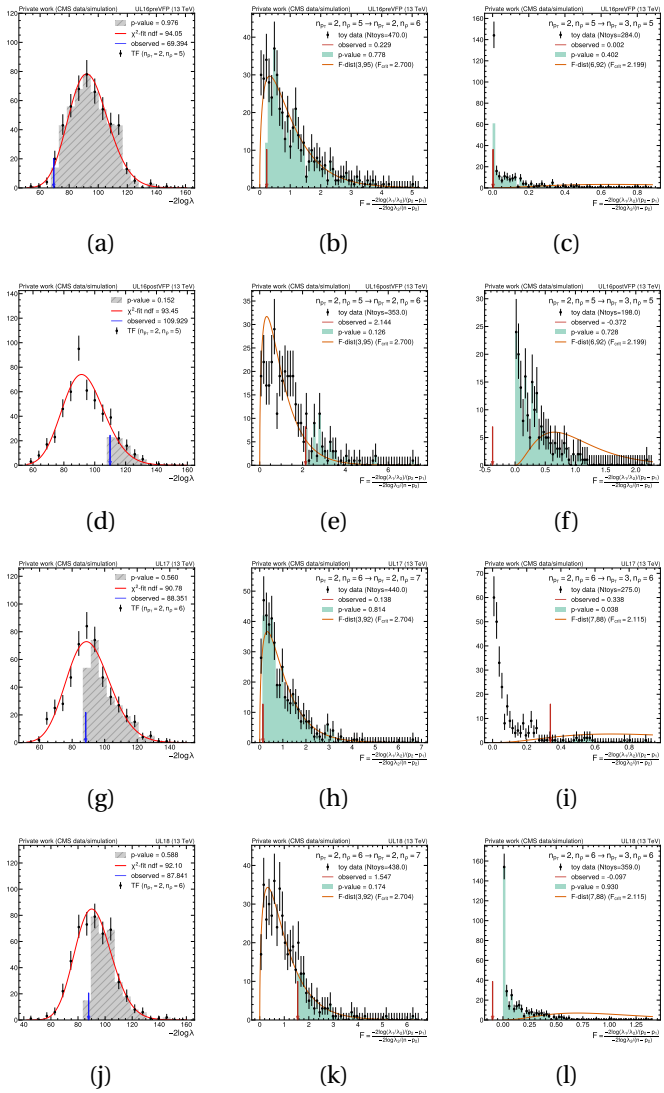
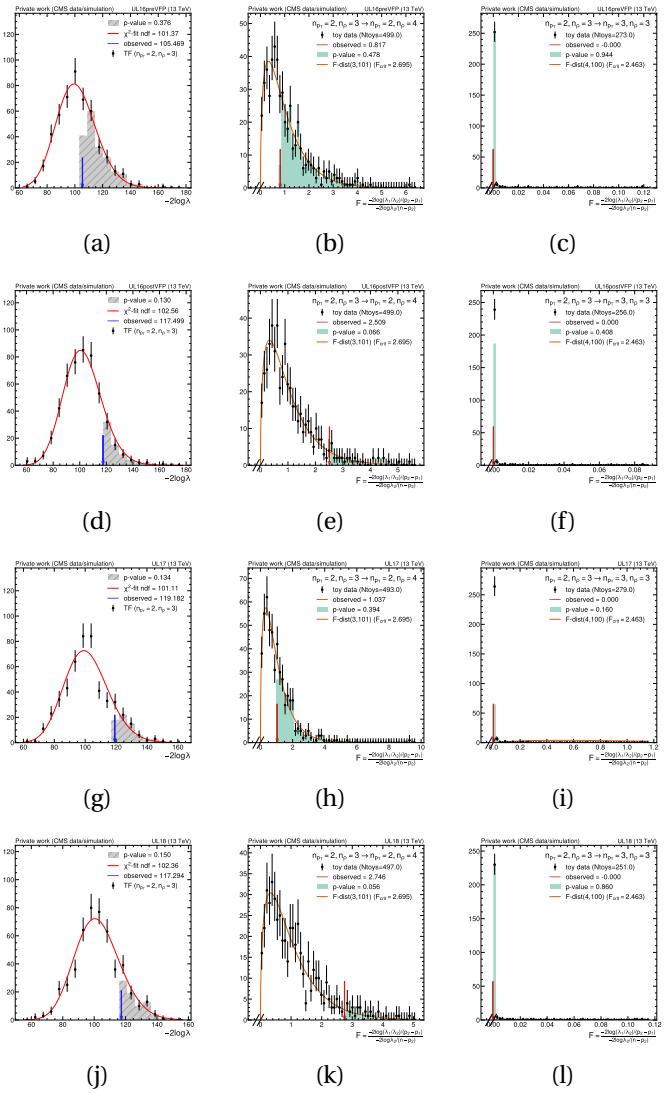


Figure C.8: Overview of metrics of the final iteration of F-Test for the tagger approach using the ParticleNet discriminator as W tagger in the QCD multijet background estimation. The rows correspond to the different periods of data taking, while the columns show from left to right: the Goodness of fit using the saturated model using the order combination under test (base model), the F-Test values and distribution testing the base model against more complex models by increasing $n_{\rho_{\mathrm{SD}}}$ by one (center) and increasing $n_{p_{\mathrm{T}}}$ by one (right).



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure C.9: Overview of metrics of the final iteration of F-Test for the tagger approach using ParticleNet$^{\text{DDT}}$ as W tagger in the QCD multijet background estimation. The rows correspond to the different periods of data taking, while the columns show from left to right: the Goodness of fit using the saturated model using the order combination under test (base model), the F-Test values and distribution testing the base model against more complex models by increasing $n_{\rho_{\text{SD}}}$ by one (center) and increasing $n_{p_{\text{T}}}$ by one (right).

Measurement of JetHT trigger efficiency in single muon dataset

D

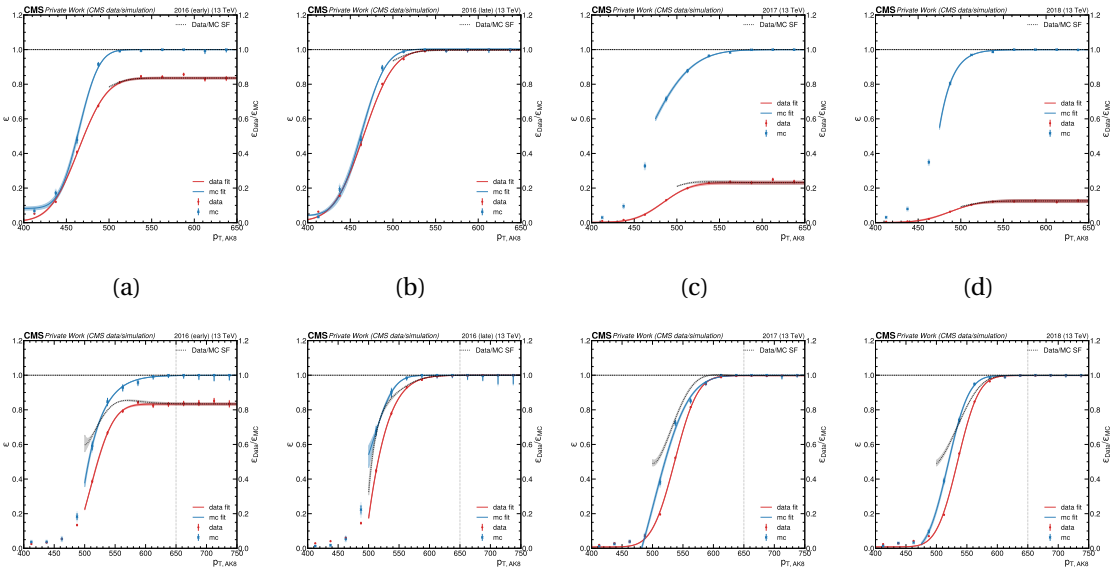


Figure D.1: Measurement of the efficiency of the two used jet triggers with p_T thresholds 450 GeV (top row) and 500 GeV (bottom row) in the four data-taking periods (columns). The efficiency is measured in data and simulation in bins of p_T of the reconstructed AK8 jet. The error bars represent the statistical uncertainty. The shaded band represents repeated measurements that take into account the statistical uncertainty in both the numerator and denominator histogram. The trigger scale factor is taken as the ratio of data and simulation efficiency and shown as grey dotted line.

Unfolding definitions

E

E.1 SVD regularisation with non-uniform binning

On particle-level the setup of this analysis consists of 4 bins in the p_T-axis with the bin boundaries $[500, 650, 800, 1200, \infty]$ GeV and 4 bins on the m_{SD}-axis with the bin boundaries $[10, 70, 80, 90, \infty]$ GeV.
To avoid the infinities here we substitute an arbitrary value for the last bin-boundaries. For the p_T-axis we choose 1400 GeV and for the m_{SD}-axis we choose 100 GeV.
From the bin boundaries, the bin centers follow as:

$$c = \begin{pmatrix} 325 & 725 & 1000 & 2100 \\ 35 & 75 & 85 & 195 \end{pmatrix} \tag{E.1}$$

Thus the vector of signal strength modifiers μ_i is given by:

$$\vec{\mu} = \begin{pmatrix} \mu_{p_T^0, m_{\mathrm{SD}}^0} \\ \mu_{p_T^0, m_{\mathrm{SD}}^1} \\ \mu_{p_T^0, m_{\mathrm{SD}}^2} \\ \mu_{p_T^0, m_{\mathrm{SD}}^3} \\ \mu_{p_T^1, m_{\mathrm{SD}}^0} \\ \vdots \\ \mu_{p_T^3, m_{\mathrm{SD}}^2} \\ \mu_{p_T^3, m_{\mathrm{SD}}^3} \end{pmatrix} \tag{E.2}$$

where the component $\mu_{p_T^i, m_{\mathrm{SD}}^j}$ scales signal contributions in the i-th bin in the p_T-axis and the j-th bin in the m_{SD}-axis.
The bin widths are:

$$w = \begin{pmatrix} 650 & 150 & 400 & 1800 \\ 70 & 10 & 10 & 210 \end{pmatrix} \tag{E.3}$$

where $w_{d,i}$ corresponds to the i-th bin-width along the dimension $d \in [1, 2]$ ($d = 1$ corresponds to p_T and $d = 2$ to m_{SD}).
In order to incorporate these into the curvature matrix we transform the derivatives to encapsulate the bin information. The first derivatives are transformed to:

$$(x_{j_2} - x_{j_1}) \rightarrow \frac{\Delta_d}{\delta^d_{j_1, j_2}} (x_{j_2} - x_{j_1}), \tag{E.4}$$

where j_1 and j_2 are two adjacent bins the 2D (p_t, m_{SD})-distribution, d is the dimension the

derivative is calculated for, Δ_d is the average bin-width of the dimension d and $\delta^d_{j_1,j_2}$ is the distance between the bin-centers of bin j_1 and j_2 in dimension d. The second-order derivatives are transformed similarly to:

$$(x_{j_3} - x_{j_2}) - (x_{j_1} - x_{j_2}) \rightarrow \frac{(\Delta_d)^2}{\delta^d_{j_2,j_1} + \delta^d_{j_2,j_3}} \left(\frac{x_{j_3} - x_{j_2}}{\delta^d_{j_2,j_3}} - \frac{x_{j_1} - x_{j_2}}{\delta^d_{j_2,j_1}} \right), \tag{E.5}$$

by rewriting equations E.4 and E.5 we get the factors needed to construct our curvature matrix:

$$\frac{\Delta_d}{\delta^d_{j_1,j_2}} (x_{j_2} - x_{j_1}) = +\vartheta'_{d,j_1,j_2} \cdot x_{j_2} - \vartheta'_{d,j_1,j_2} \cdot x_{j_1} \tag{E.6}$$

$$\begin{aligned}
\frac{(\Delta_d)^2}{\delta^d_{j_2,j_1} + \delta^d_{j_2,j_3}} \left(\frac{x_{j_3} - x_{j_2}}{\delta^d_{j_2,j_3}} - \frac{x_{j_1} - x_{j_2}}{\delta^d_{j_2,j_1}} \right) &= \left[\frac{(\Delta_d)^2}{\delta^d_{j_2,j_1} + \delta^d_{j_2,j_3}} \frac{1}{\delta^d_{j_2,j_1}} \right] \cdot x_{j_1} && \text{(E.7)} \\
&- \left[\frac{(\Delta_d)^2}{\delta^d_{j_2,j_1} + \delta^d_{j_2,j_3}} \left(\frac{1}{\delta^d_{j_2,j_3}} + \frac{1}{\delta^d_{j_2,j_1}} \right) \right] \cdot x_{j_2} && \text{(E.8)} \\
&+ \left[\frac{(\Delta_d)^2}{\delta^d_{j_2,j_1} + \delta^d_{j_2,j_3}} \frac{1}{\delta^d_{j_2,j_3}} \right] \cdot x_{j_3} && \text{(E.9)} \\
&= +\vartheta''_{d,j_2,-1} \cdot x_{j_1} - \vartheta''_{d,j_2,0} \cdot x_{j_2} + \vartheta''_{d,j_2,+1} \cdot x_{j_3}. && \text{(E.10)} \\
& && \text{(E.11)}
\end{aligned}$$

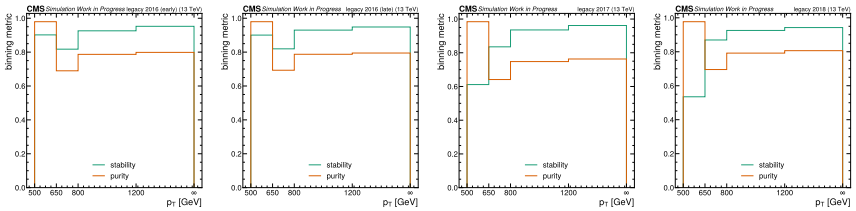
If we want to add penalty terms to regularise the unfolded results along the m_{SD}-axis we use the curvature matrix C as:

$$C = \begin{pmatrix} \sigma_{m_{\mathrm{SD}}} & & & \\ & \sigma_{m_{\mathrm{SD}}} & & \\ & & \sigma_{m_{\mathrm{SD}}} & \\ & & & \sigma_{m_{\mathrm{SD}}} \end{pmatrix} \tag{E.12}$$

where $\sigma_{m_{\mathrm{SD}}}$ is the penalty term for the one bin along the p_T-axis:

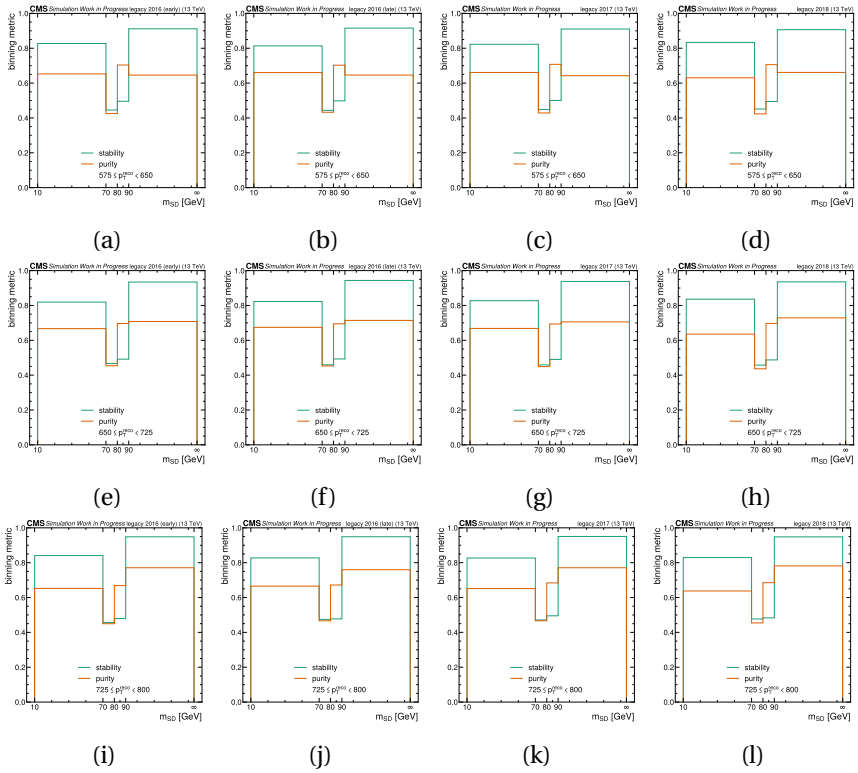
$$\sigma_{m_{\mathrm{SD}}} = \begin{pmatrix} -\vartheta'_{m_{\mathrm{SD}},0,1} & +\vartheta'_{m_{\mathrm{SD}},0,1} & 0 & 0 \\ \vartheta''_{m_{\mathrm{SD}},1,-1} & -\vartheta''_{m_{\mathrm{SD}},1,0} & \vartheta''_{m_{\mathrm{SD}},1,+1} & 0 \\ 0 & \vartheta''_{m_{\mathrm{SD}},2,-1} & -\vartheta''_{m_{\mathrm{SD}},2,0} & \vartheta''_{m_{\mathrm{SD}},2,+1} \\ 0 & 0 & +\vartheta'_{m_{\mathrm{SD}},0,1} & -\vartheta'_{m_{\mathrm{SD}},0,1} \end{pmatrix} \tag{E.13}$$

E.2 Stability and purity



(a) (b) (c) (d)

Figure E.1: Distribution of stability and purity of the final chosen particle-level binning for $p_{\mathrm{T}}^{\mathrm{ptcl}}$ for each $p_{\mathrm{T}}^{\mathrm{reco}}$ (rows) bin using simulation with early 2016 (left), late 2016 (second to left), 2017 (second to right) and 2018 (right) conditions.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure E.2: Distribution of stability and purity of the final chosen particle-level binning for $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ for the last three $p_{\mathrm{T}}^{\mathrm{reco}}$ (rows) bins using simulation with early 2016 (left column), late 2016 (second to left column), 2017 (second to right column) and 2018 (right column) conditions.

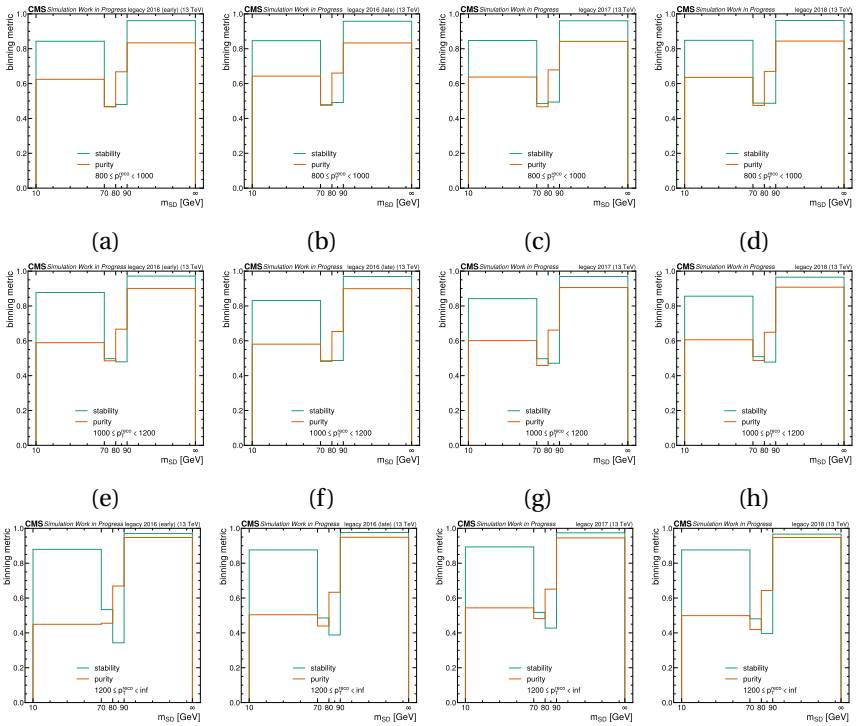


Figure E.3: Distribution of stability and purity of the final chosen particle-level binning for $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ for the first three $p_{\mathrm{T}}^{\mathrm{reco}}$ (rows) bins using simulation with early 2016 (left column), late 2016 (second to left column), 2017 (second to right column) and 2018 (right column) conditions.

E.3 Acceptance and Efficiency

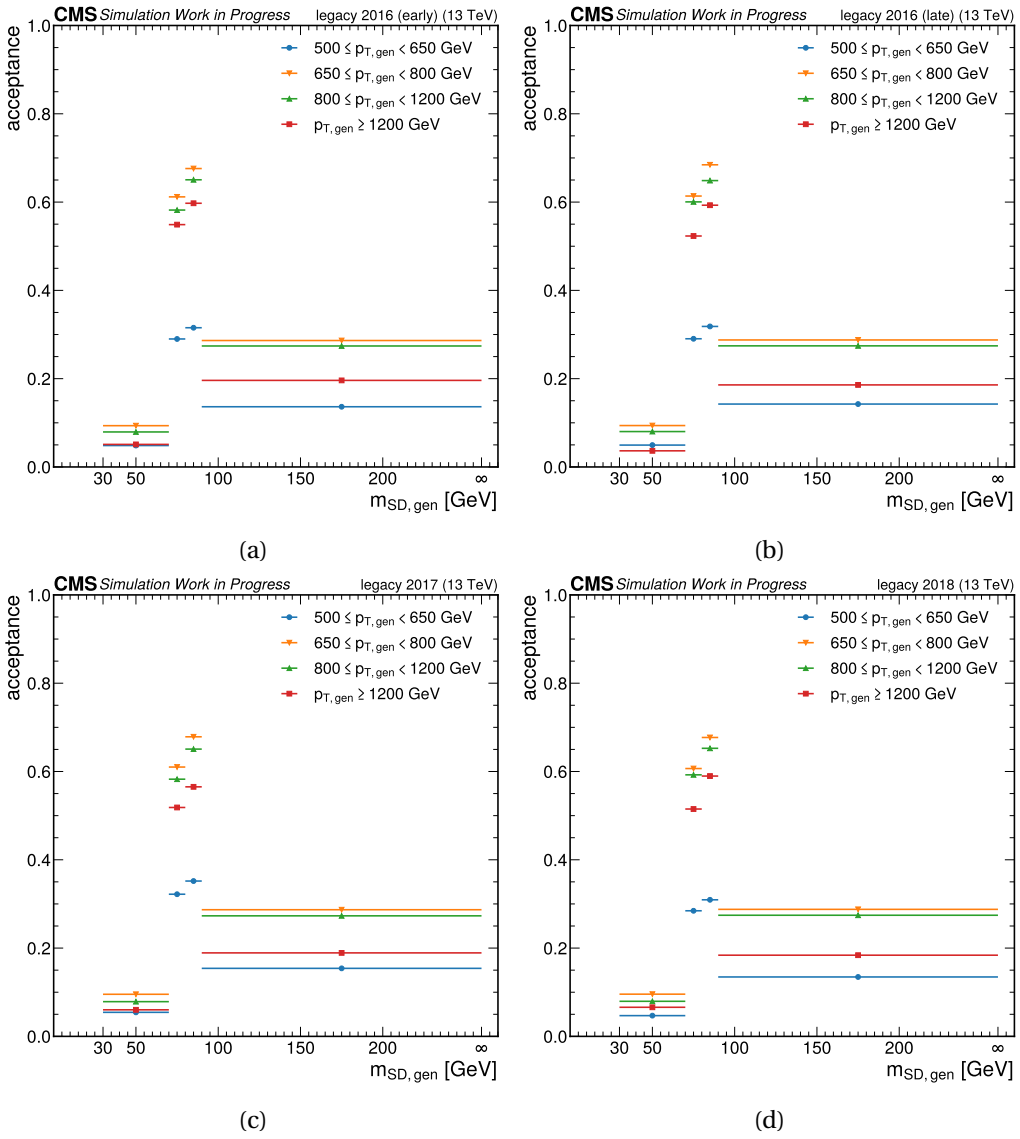


Figure E.4: Summary of the acceptance measured in the $W(q\bar{q})$+jets signal simulation sample with each data taking year conditions (from top left to bottom right: early 2016, late 2016, 2017 and 2018).

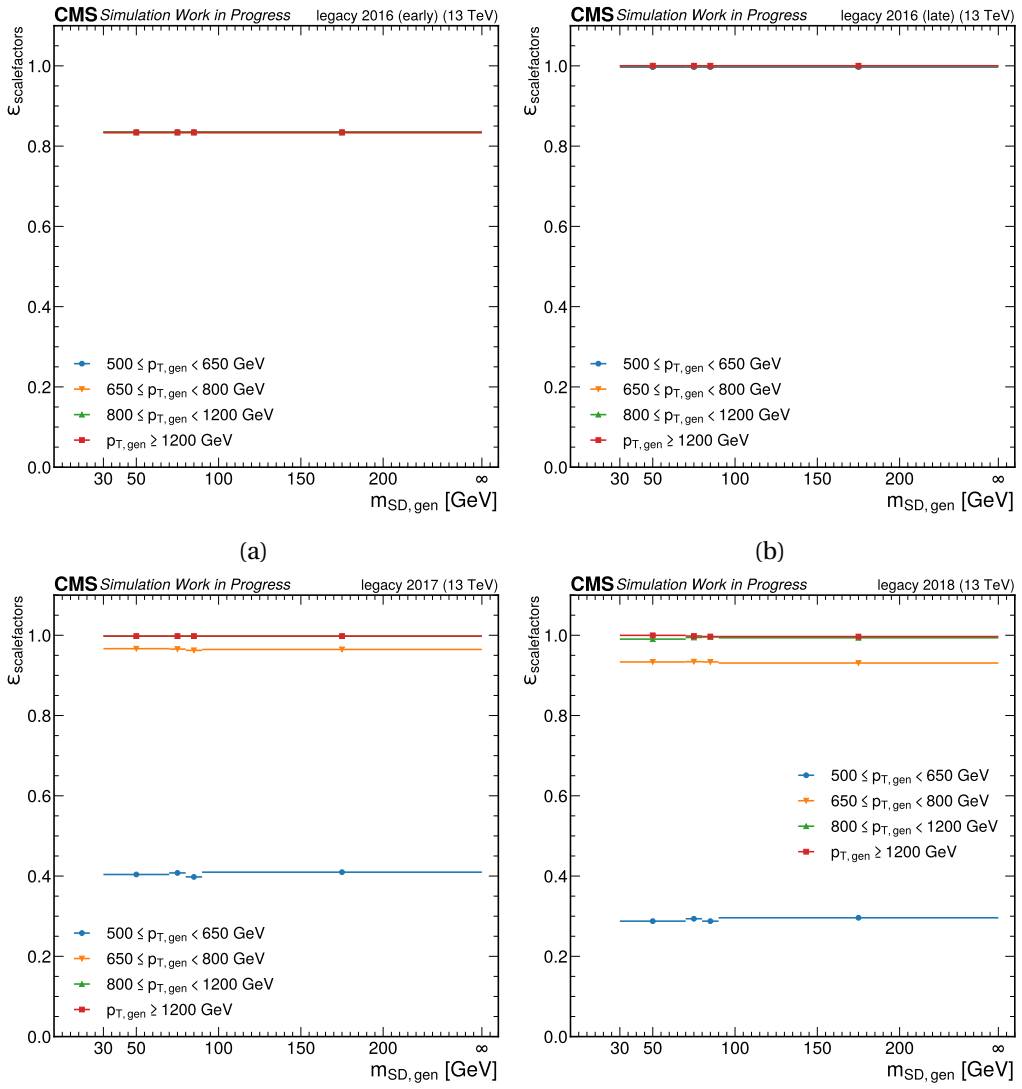


Figure E.5: Summary of the scale factor efficiency measured in the $W(q\bar{q})$+jets signal simulation sample with each data taking year conditions (from top left to bottom right: early 2016, late 2016, 2017 and 2018).

E.4 Closure and bias tests

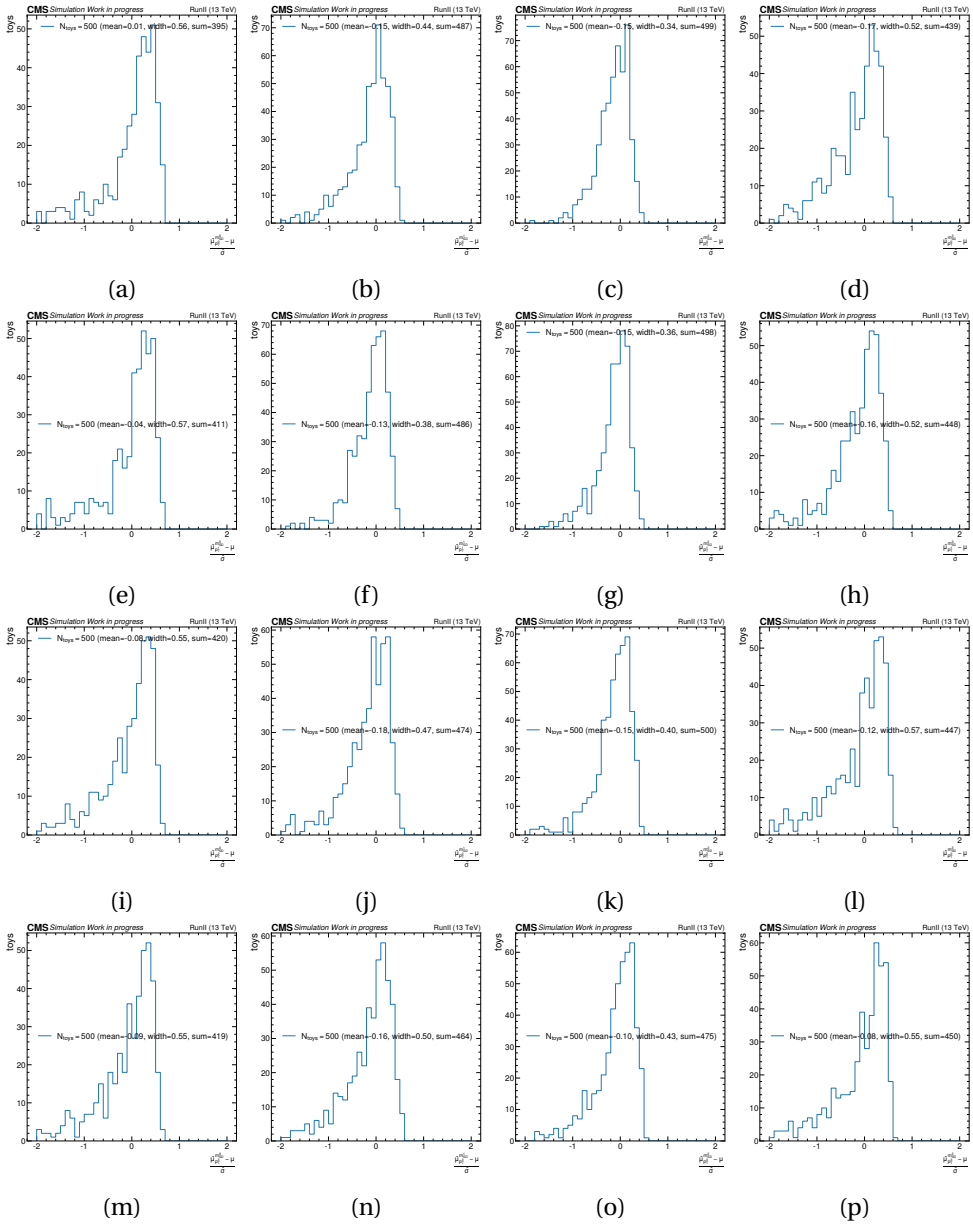


Figure E.6: Pull distribution of the signal strength modifiers of using toys derived from *asimov* data. The rows correspond to $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins and the columns correspond to $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins. $N_2^{\beta=1,\mathrm{DDT}}$ is used as tagger for the background estimation. The metrics in the legend are derived from either the total histogram (N_{toys}) or only from the visible range ([-2,2]) (mean, width and sum).

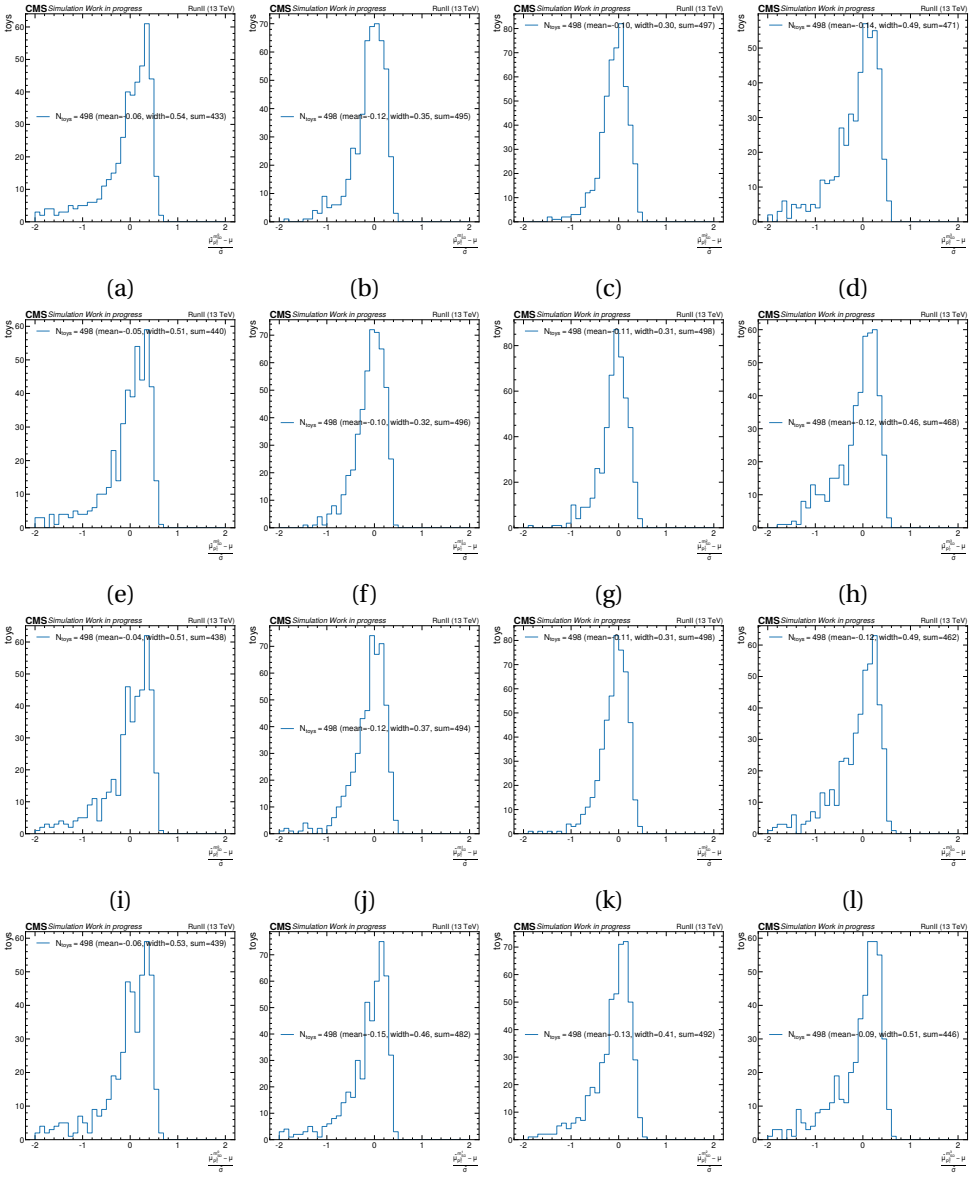


Figure E.7: Pull distribution of the signal strength modifiers of using toys derived from *asimov* data. The rows correspond to $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins and the columns correspond to $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins. ParticleNet$^{\mathrm{DDT}}$ is used as tagger for the background estimation. The metrics in the legend are derived from either the total histogram (N_{toys}) or only from the visible range ([-2,2]) (mean, width and sum).

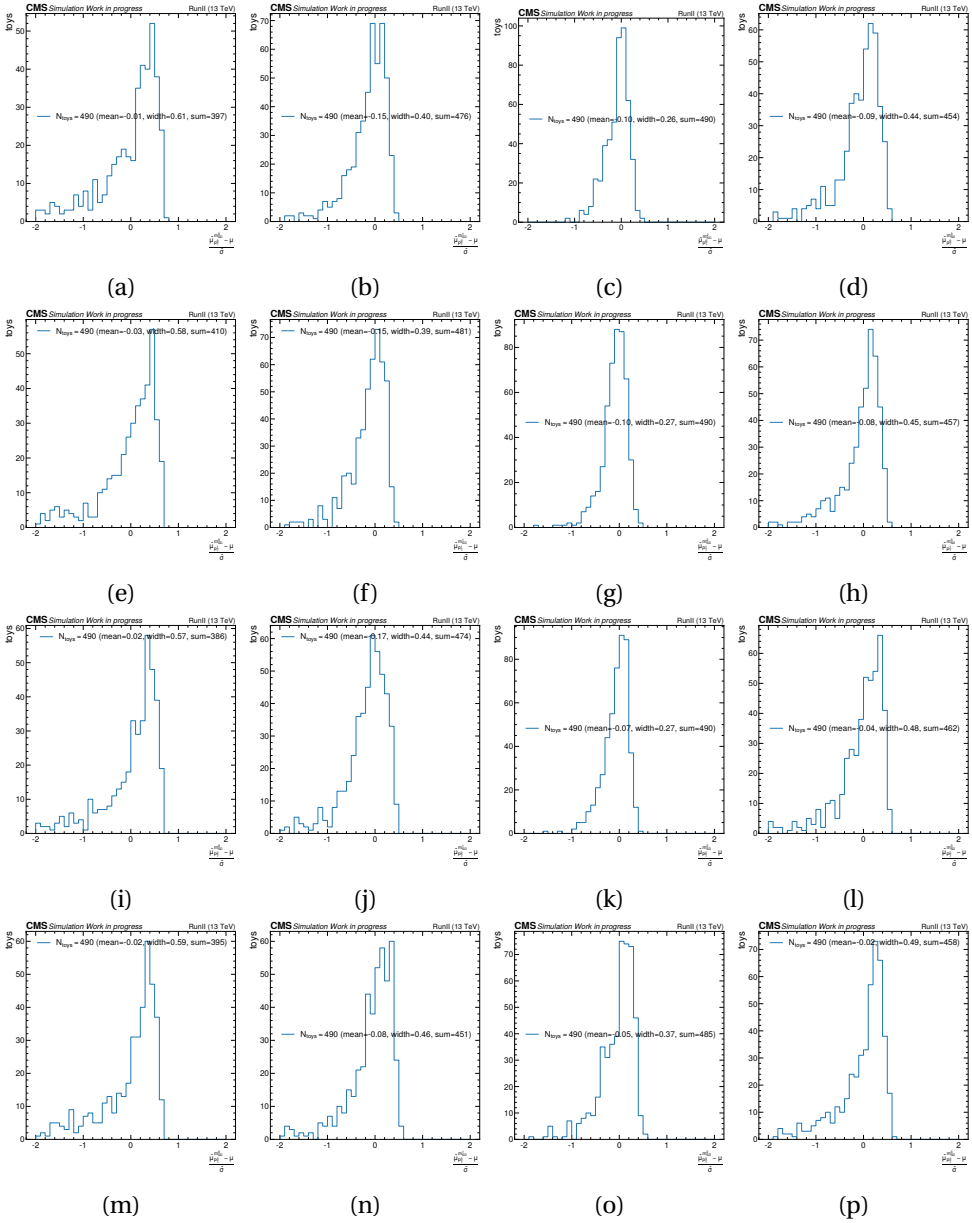


Figure E.8: Pull distribution of the signal strength modifiers of using toys derived from *statistically independent* pseudo-data. The rows correspond to $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins and the columns correspond to $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins. $N_2^{\beta=1,\mathrm{DDT}}$ is used as tagger for the background estimation. The metrics in the legend are derived from either the total histogram (N_{toys}) or only from the visible range ([-2,2]) (mean, width and sum).

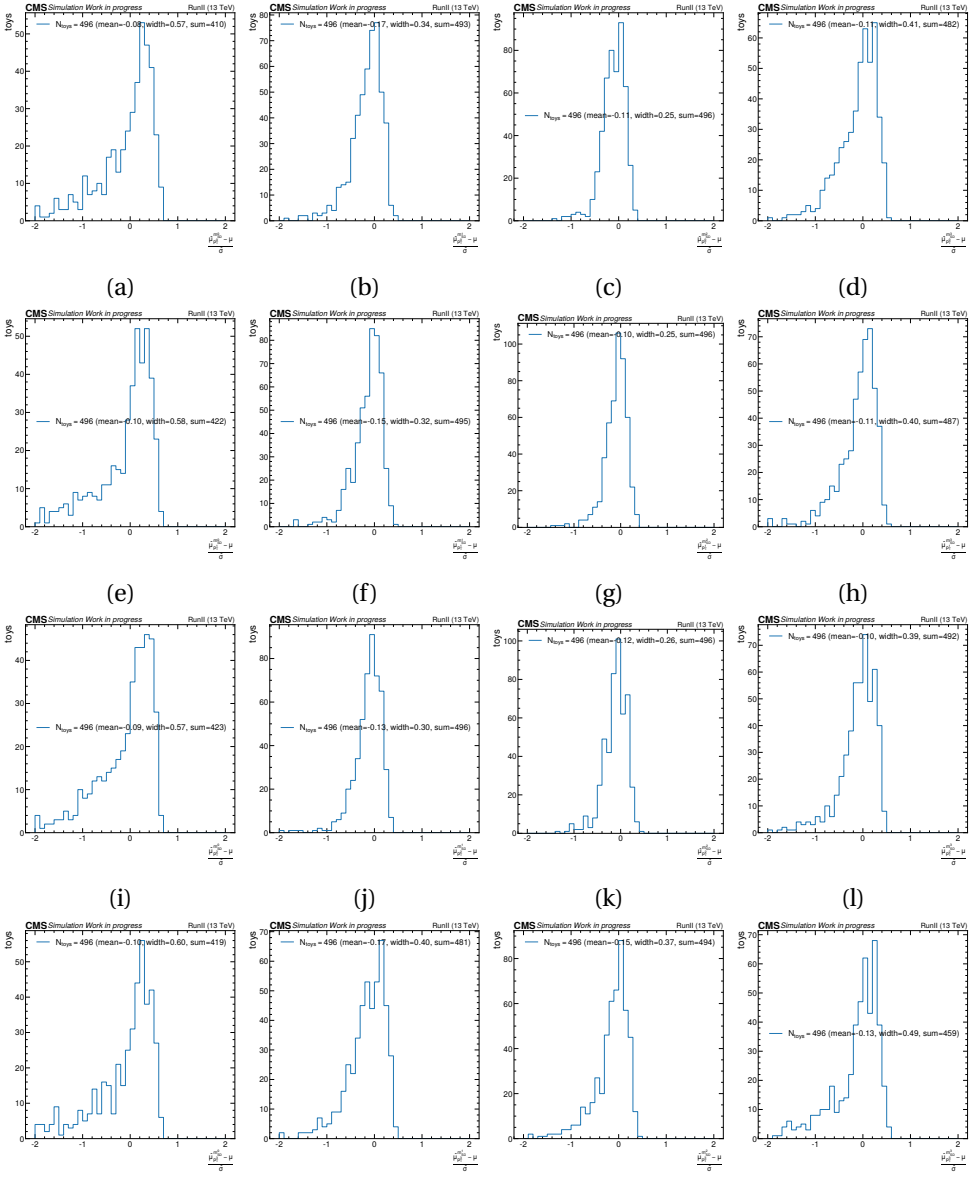
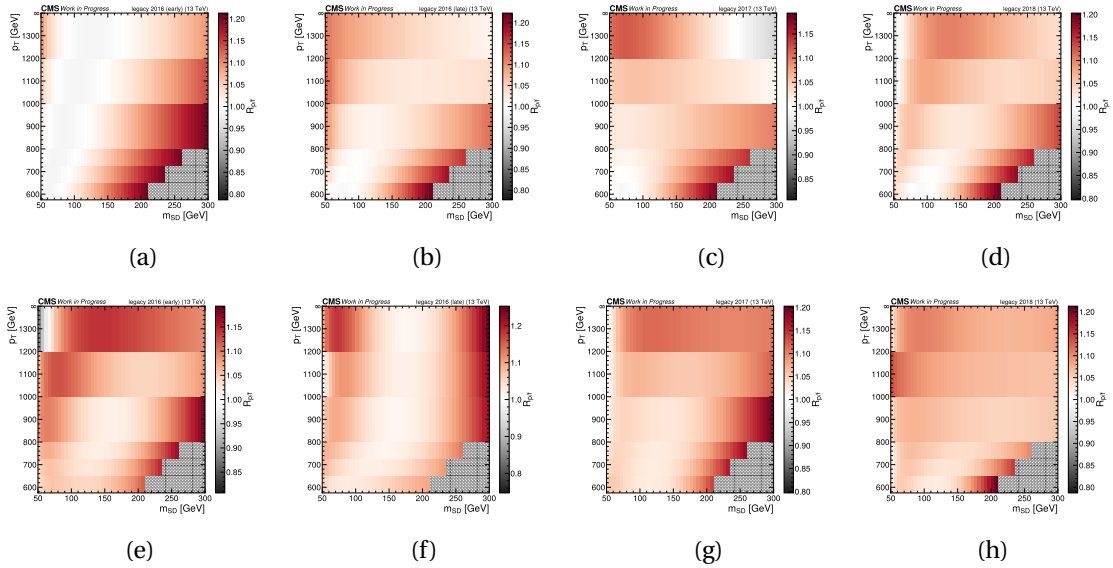


Figure E.9: Pull distribution of the signal strength modifiers of using toys derived from *statistically independent* pseudo-data. The rows correspond to $p_{\mathrm{T}}^{\mathrm{ptcl}}$ bins and the columns correspond to $m_{\mathrm{SD}}^{\mathrm{ptcl}}$ bins. ParticleNet$^{\mathrm{DDT}}$ is used as tagger for the background estimation. The metrics in the legend are derived from either the total histogram (N_{toys}) or only from the visible range ([-2,2]) (mean, width and sum).

Unfolding results

F



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure F.1: Post-fit transfer-factor of the data-driven QCD multijet background estimation, when using $N_2^{\beta=1,\mathrm{DDT}}$ and ParticleNet$^{\mathrm{DDT}}$ as tagger on the left and right respectively. Each row corresponds to the transfer factor of a different period of data-taking (early 2016, late 2016, 2017 and 2018 from top to bottom).

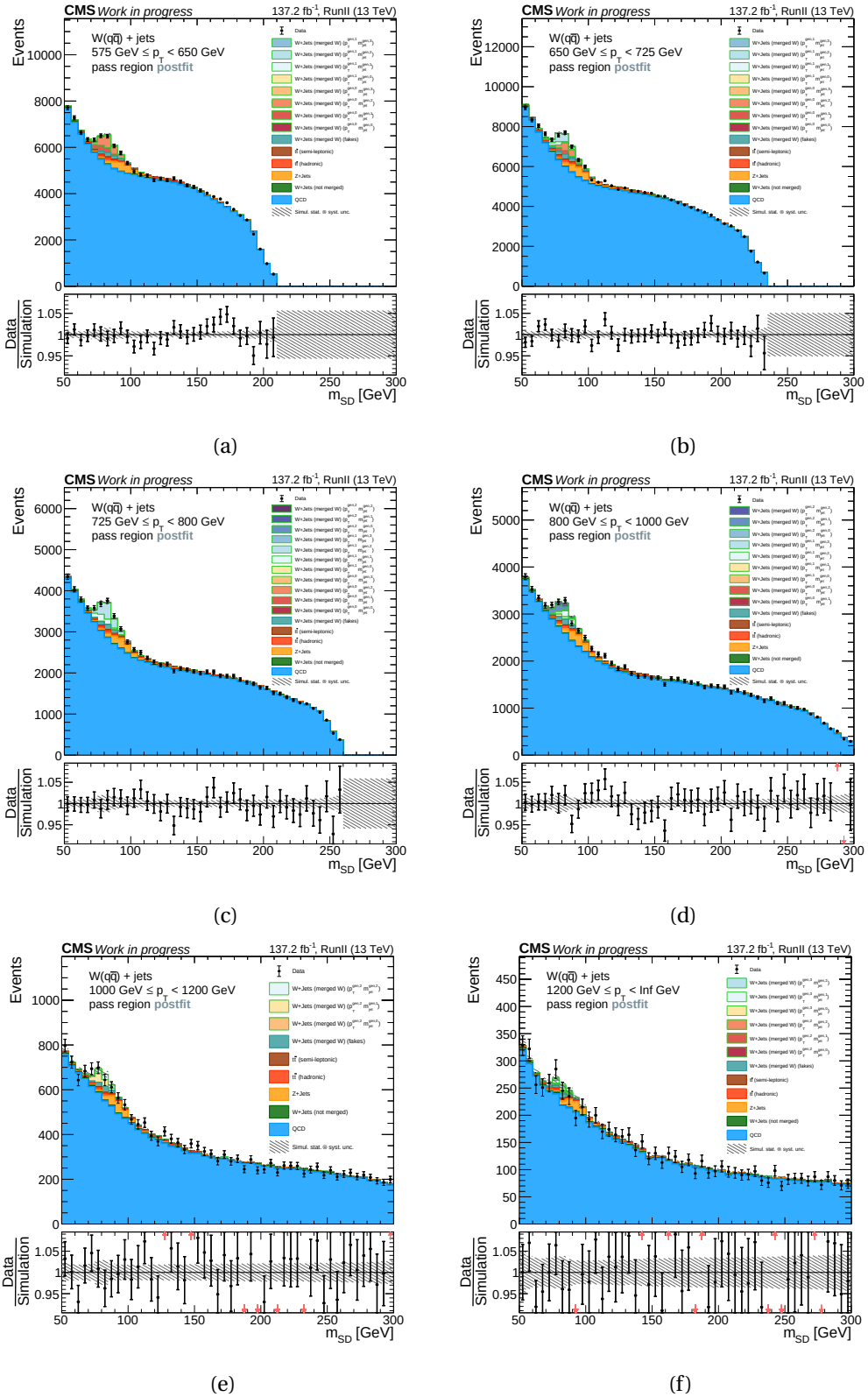


Figure F.2: Post-fit templates for all $p_{\mathrm{T}}^{\mathrm{reco}}$ bins for events passing the $N_2^{\beta=1,\mathrm{DDT}}$ W tagger. All data-taking periods (early 2016, late 2016, 2017 and 2018) are summed post-fit to reflect the full Run II distribution with the integrated luminosity $\mathscr{L}_{\mathrm{int}} \approx 137\,\mathrm{fb}^{-1}$.

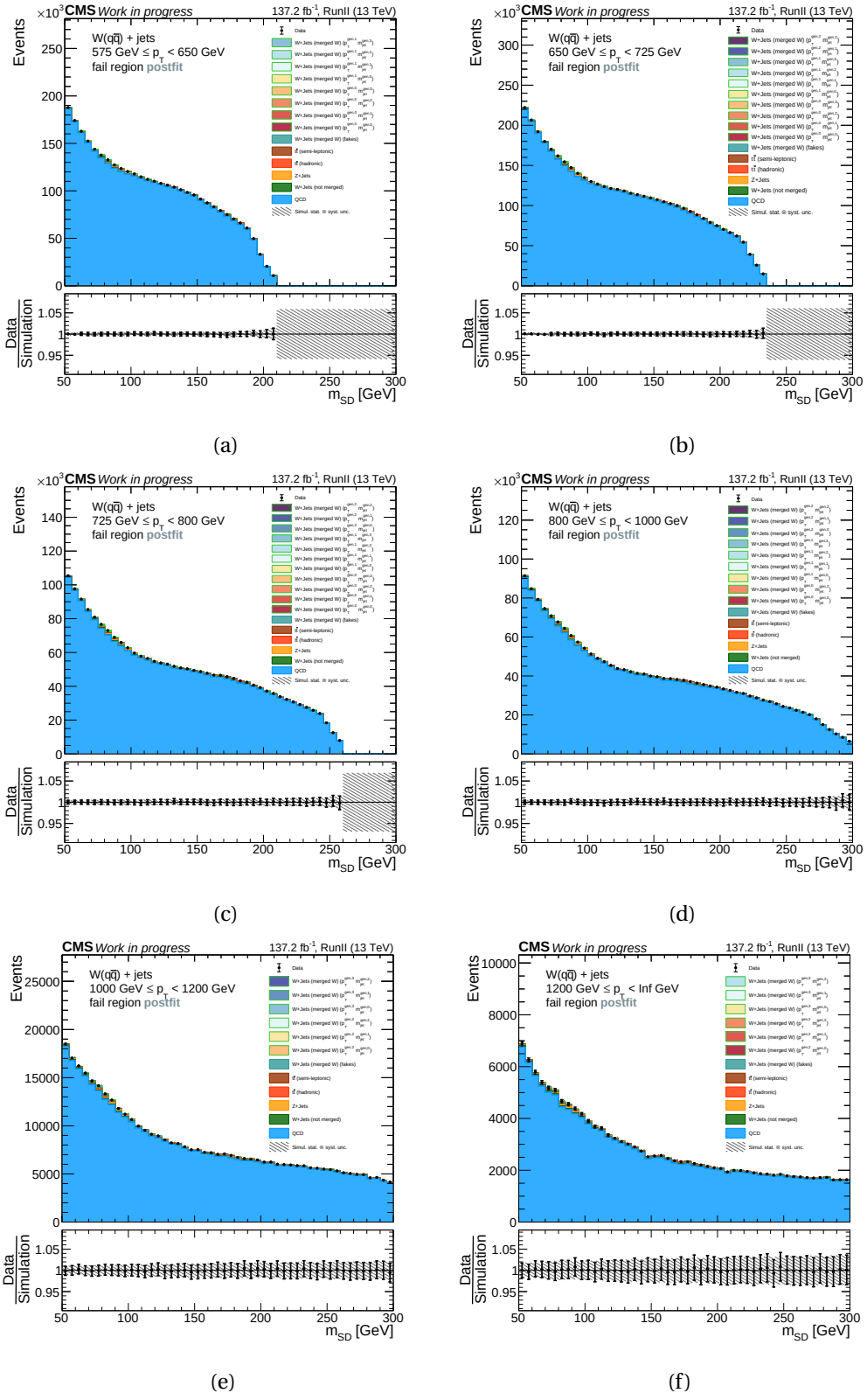


Figure F.3: Post-fit templates for all $p_{\mathrm{T}}^{\mathrm{reco}}$ bins for events failing the $N_2^{\beta=1,\mathrm{DDT}}$ W tagger. All data-taking periods (early 2016, late 2016, 2017 and 2018) are summed post-fit to reflect the full Run II distribution with the integrated luminosity $\mathscr{L}_{\mathrm{int}} \approx 137\,\mathrm{fb}^{-1}$.

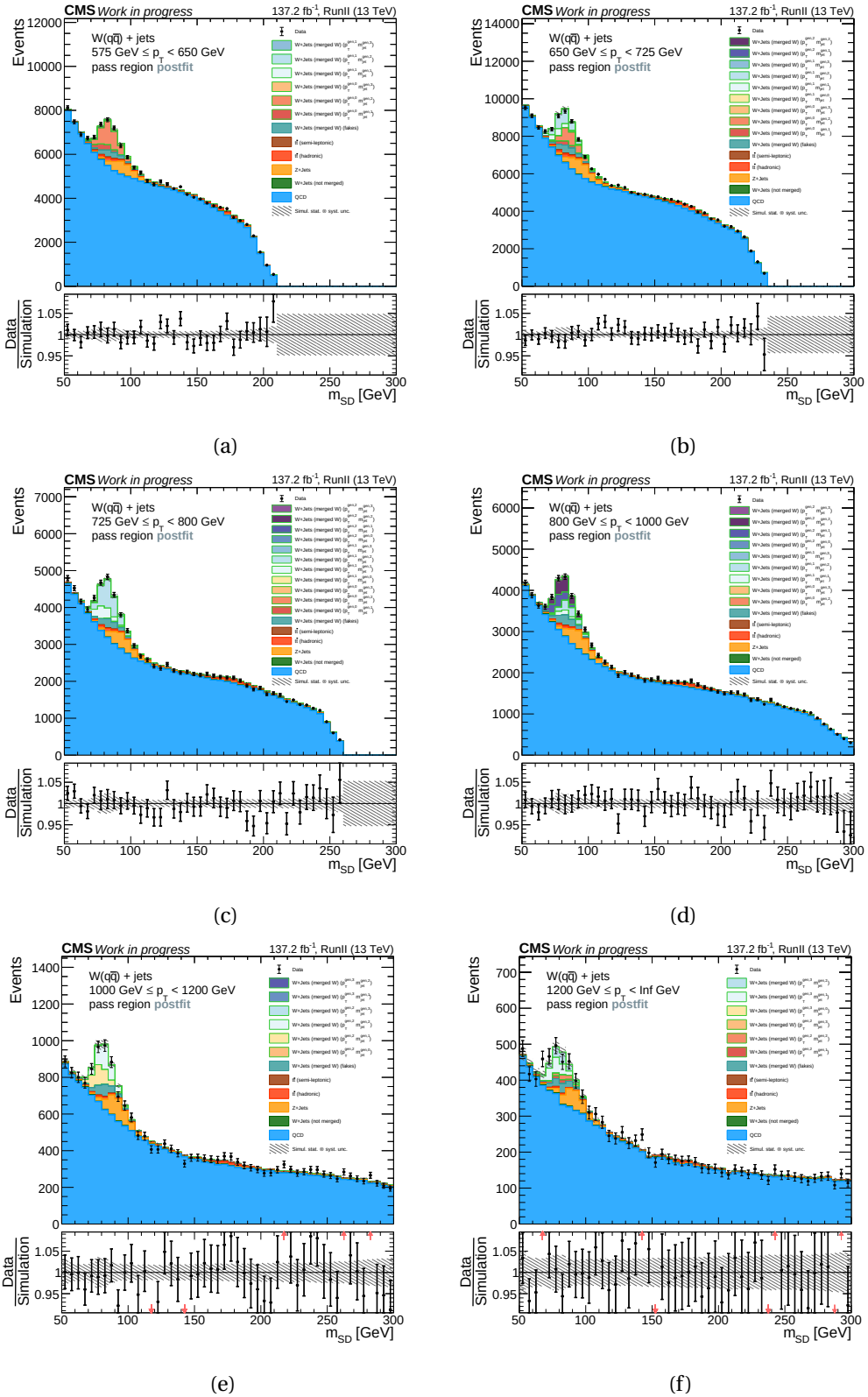


Figure F.4: Post-fit templates for all $p_{\mathrm{T}}^{\mathrm{reco}}$ bins for events passing the $N_2^{\beta=1,\mathrm{DDT}}$ W tagger. All data-taking periods (early 2016, late 2016, 2017 and 2018) are summed post-fit to reflect the full Run II distribution with the integrated luminosity $\mathscr{L}_{\mathrm{int}} \approx 137\,\mathrm{fb}^{-1}$.

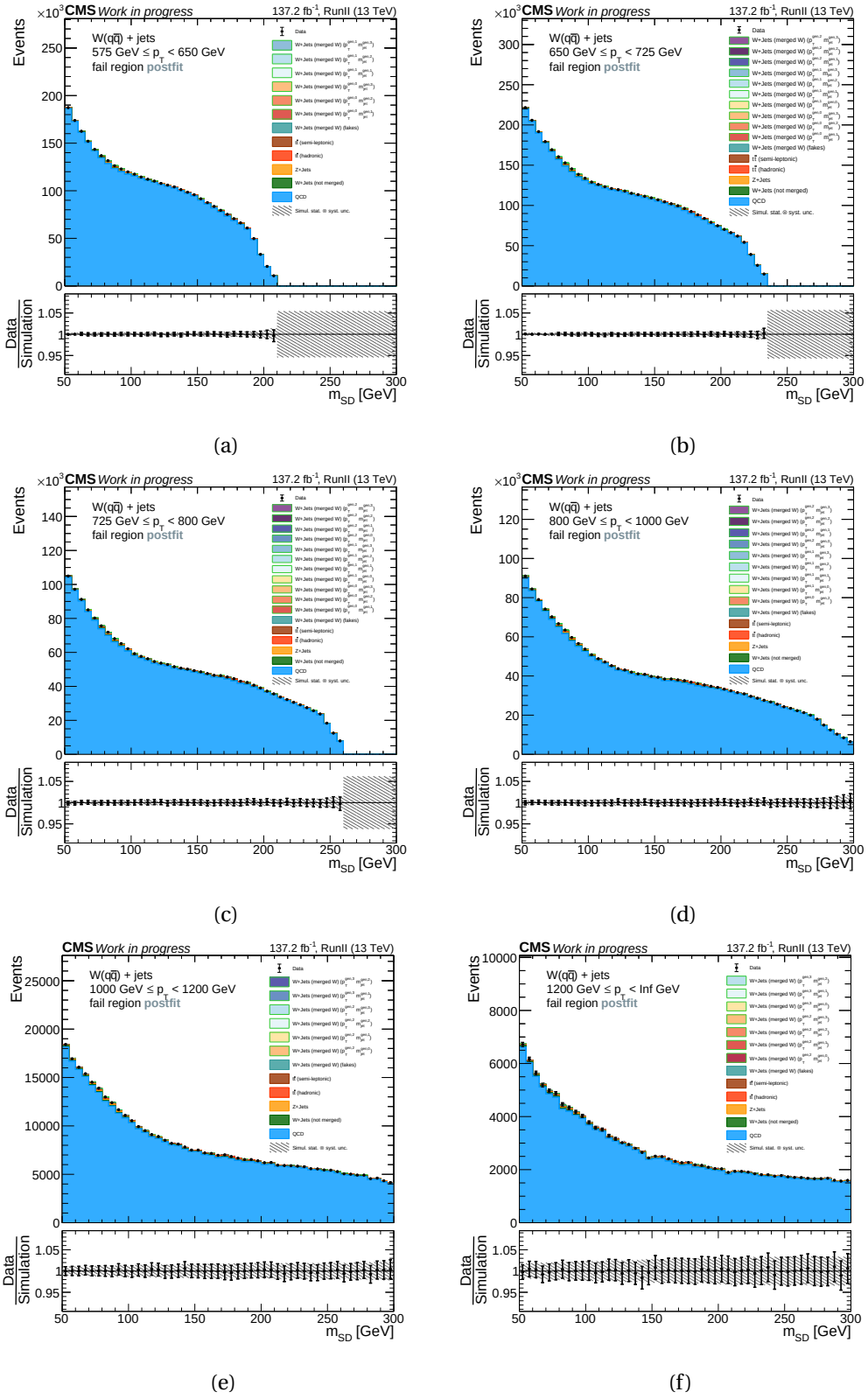


Figure F.5: Post-fit templates for all $p_{\mathrm{T}}^{\mathrm{reco}}$ bins for events failing the $N_2^{\beta=1,\mathrm{DDT}}$ W tagger. All data-taking periods (early 2016, late 2016, 2017 and 2018) are summed post-fit to reflect the full Run II distribution with the integrated luminosity $\mathscr{L}_{\mathrm{int}} \approx 137\,\mathrm{fb}^{-1}$.

Eidestattliche Erklärung

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

Ort, Datum

Unterschrift

Acknowledgements

At this point, I would like to thank all those who made this thesis and the related work and studies possible. First and foremost, I would like to thank Dr. Andreas Hinzmann for letting me join his working group in 2017 and providing excellent supervision, guidance, constant motivation, and inspiration throughout my Master's and subsequent doctoral studies!

I thank Prof. Dr. Johannes Haller for taking on the role of the second advisor and examiner and for integrating me into his working group since my Bachelor's thesis in 2015, as well as providing support over the years.

I express my gratitude to the other members of the examination committee, Prof. Dr. Gudrid Moortgat-Pick, Prof. Dr. Peter Schleper, and the chairman, Prof. Dr. Dieter Horns, for joining the committee and for the evaluation of my disputation.

I want to thank the people of the first floor of building 68. If you won't find your name, please assume I forgot and that I feel terrible about it.

I'd like to thank Torben Lange, Johannes Lange, Christoph Garbers and Hartmut Stadie for their excellent technical support (i.e. high tolerance against unnecessarily complicated questions) and for being good office neighbors during the first couple of months of my PhD.

A big thank you to all the members of the working group of Andreas Hinzmann and Andreas himself for the warm work environment! Anna Benecke, Irene Zoi, Robin Aggleton and Andreas I want to thank you for all your help during my master's studies and the first years of my PhD. A big chunk of the stuff I learned during my studies, I learned from you and your way of working!

The second era of AG Hinzmann namely Anna Albrecht and Ankita Mehta: I want to thank you for all the helpful discussions related to physics and analyses and especially also for the not-so-helpful discussions related to non-physics and gossip! I think we found the perfect balance between both! Also thanks for giving me a sense of purpose in my grid pack generation 'expertise'! A special thanks to Ankita also for proofreading!

I want to thank all of my office mates: Julian Zeyn, Robin Aggleton, Mathis Frahm, Ankita Mehta and Janek Moels who shared not only office space but also nerves regarding technical issues and laughs over the years.

For the entirety of my PhD, our working group was closely entangled with the group of Johannes Haller. I want to thank the entire group, all Bachelor, Master, PhD students and PostDocs who accompanied me over the years!

I want to especially thank Dennis Schwarz and Roman Kogler for their support, discussions and foundational work in the early months of my PhD.

I thank Alexander Paasch, Yannick Fischer, Andrea Malara and Daniel Savoiu for their endurance

regarding my JERC questions. The same goes for Finn Labe and Trigger questions. A special thanks to Andrea Malara for being a human color palette and helping me to actually like looking at my plots and to Matteo Bonanomi for making the impossible possible by shedding light into the abyss of the combine backbone.
I want to thank Christopher Matthies for the discussions on unfolding (i.e. for making me realize what I did wrong and that I still can't remember if it's purity or stability)! I thank Henrik Jabusch, Ksenia De Leo, Alexander Fröhlich, Daniel Hundhausen, Artur Lobanov, Matthias Schröder and all the aforementioned for the daily interactions on the Haller corridor. Lunch at CFEL was really a highlight on some grey days and coffee breaks and discussions will be dearly missed!
I want to thank Viktor Kutzner for answering all my questions about the organizational part of handing in and defending this thesis, as well as Andrea Bremer for helping with the printing of the thesis, handling all the room booking and most importantly for getting the glasses for the sparkling wine!
I'd like to thank the SMP and JetMET communities within CMS for all their feedback on my work. Here, I want to point out especially the conveners of JetMET Laurent Thomas, Henning Kirschenmann, Mikko Voutilainen and Anna Benecke and the SMP conveners Andrew Gilbert and Patrick Connor.
I owe a great thank you to the DAZSLE community for their foundational work on the background estimate and support over the years. I thank Phil Harris, David Yu, Cristina Mantilla Suarez, Jeffrey Krupa and Sang Eon Park for welcoming me at MIT for two weeks. Further, I'd like to express my gratitude to Nick Smith, Andrzej Novak, Ka Hei Martin Kwok, Javier Duarte, Nhan Tran and the aforementioned for the countless helpful discussions and pointers in my troubles with F-tests.
Finally, I want to thank the people in my life. Meiner Familie, insbesondere meinen Eltern möchte ich von ganzem Herzen sagen: DANKE! Ihr habt mir die letzten 10 (und ein bisschen) Jahre Studium ermöglicht, habt mich ermutigt und habt an mich geglaubt. Ohne euch wäre diese Arbeit nicht entstanden!

Und am Schluss bist und bleibst du die wichtigste Person, Sophie! Dir gebührt der größte Dank und all meine Liebe! Deine bedingungslose Unterstützung, Motivation, Verständnis und Liebe haben diese Arbeit möglich gemacht und mich dabei mental zusammen gehalten. Du glaubst gar nicht, wie dankbar ich Dir und dafür bin, Dich an meiner Seite zu haben!

Olee Olee Olee olay!