Measurement of the W → cq̄/W → qq̄ decay branching fraction ratio with the CMS detector at the LHC
and
Uncertainty estimation of machine learning models in particle physics

Universidad Autónoma de Madrid

Programa de Doctorado en Física Teórica

Tesis presentada por

Julia Vázquez Escobar

Directores

José María Hernández Calama y Miguel Cárdenas Montes

Madrid, España, 2024

A mi madre, mi padre y mi hermano

Agradecimientos

Me gustaría agradecer en primer lugar a mi directores, Chema, por la dedicación y orientación que me has dado, y Miguel, por el apoyo y las perspectivas brindadas. Hacer esta tesis con vosotros ha sido una oportunidad estupenda. Gracias también a Juan Pablo y Jorge, por sus inestimables aportaciones en la elaboración del análisis y el tiempo dedicado. Un agradecimiento especial a mis compañeros del CIEMAT, sin vuestro apoyo estos cuatro años habrían sido sensiblemente peores. Agradecer también al departamento de investigación básica, por ofrecer un entorno propicio para el desarrollo de nuestras actividades y fomentar un buen ambiente.

Gracias a mi familia y amigos, por su respaldo incondicional y por ser mi inspiración.

Abstract

Particle physics aims to understand the fundamental structure of matter and its interactions, relying on precise experimental measurements for its progress. The Standard Model (SM) provides the theoretical foundation for describing elementary particles and their interactions. The Large Hadron Collider and its experiments, particularly CMS, enable precise tests of the SM. This thesis focuses on testing weak universality within the SM, proposed for quarks by Cabibbo and extended by Kobayashi and Maskawa, to explain transitions between up-type and down-type quarks via weak interactions. The probability of these transitions, encoded in the CKM matrix, is determined experimentally. This thesis presents a measurement of the charm quark production rate in W decays relative to other quark flavors, quantified by the branching fraction ratio $R_c^W = \mathcal{B}(W \to cq) / \mathcal{B}(W \to q\bar{q})$. According to SM, R_c^W is expected to be 1/2, making its measurement a direct test of CKM unitarity and weak universality. The analysis utilizes a large sample of W bosons produced in $t\bar{t}$ events, where one of the W bosons decays leptonically and the other hadronically. Charm quark identification, or "charm tagging," is crucial for this measurement and involves identifying a muon within jets originating from charm hadron decays. This technique provides a clean sample with well-controlled systematics, enabling a more precise determination of R_c^W. The measurement presented reduces the uncertainty of the current world average by approximately half, marking a significant advancement in precision tests of the SM. Furthermore, this thesis explores uncertainty estimation in particle physics using machine learning techniques. By applying Bayesian neural networks, probabilistic random forests, and local ensembles to CMS open data, this work underscores the importance of reliable uncertainty estimation in enhancing the robustness of analyses using modern computational tools.

Key words: High energy physics, charm tagging, precision measurement, LHC, uncertainty estimation.

Resumen

La Física de Partículas trata de comprender la estructura fundamental de la materia y cómo interactúa entre sí, para ello, como ciencia experimental, debe apoyarse en medidas experimentales que hagan avanzar el campo. El modelo estándar (ME) proporciona un marco teórico para describir las partículas elementales e interacciones entre ellas. El LHC incluye una serie de experimentos, entre ellos CMS, que permite la realización de medidas de precisión del ME. Esta tesis se centra en poner a prueba la universalidad débil en el contexto del ME, una propiedad sugerida para quarks por Cabibbo y extendida por Kobayashi y Maskawa para explicar las transiciones entre quarks tipo "up" y "down" a través de interacciones débiles. Estas probabilidades de transición, que se encuentran en la matriz CKM, son determinadas experimentalmente. Esta tesis presenta una medida de la tasa de producción de quarks de tipo charm en decaimientos de bosones W, relativa al decaimiento a quarks de otros sabores, caracterizada por la razón de fracciones de desintegración $\mathcal{R}_c^W = \mathcal{B}(W \to cq) / \mathcal{B}(W \to q\bar{q})$. Según el ME, $\mathcal{R}_c^W$ debería resultar 1/2, por lo que medir esta predicción es una manera directa de poner a prueba la condición unitaria de la matriz CKM y la universalidad débil. El análisis desarrollado usa una muestra amplia de bosones W en el contexto de eventos $t\bar{t}$ semileptónica, en los que uno de los bosones W decae leptónicamente y el otro hadrónicamente. La clasificación de eventos de tipo charm es crucial para esta medida, se hace identificando muones dentro de jets, una característica de jets provinientes de quarks de sabor pesado. Esta técnica nos permite obtener una muestra charm limpia y controlar la sistemática asociada, concluyendo en una determinación de $\mathcal{R}_c^W$ precisa. Esta medida reduce la incertidumbre de la media mundial vigente a aproximadamente la mitad, haciendo de ésta un avance notable en precisión. Además, esta tesis explora la estimación de incertidumbre para el uso de machine learning en Física de Partículas. Se han aplicado técnicas de aproximación de redesneuronales bayesianas, random forest probabilístico y local ensembles a datos abiertos de CMS, subrayando la importancia de obtener resultados robustos estimando la posible incertidumbre que pueda ocasionar el uso de herramientas computacionales de inteligencia artificial.

Palabras clave: Física de altas energías, etiquetado charm, medida de precisión, LHC, estimación de incertidumbre.

Contents

Introduction

Particle physics, as an experimental field of knowledge, seeks to understand the fundamental nature of matter and its behavior, relying heavily on precise measurements for its advancement. The Standard Model (SM) serves as the theoretical framework that describes elementary particles and their interactions. Any measurable prediction derived from this model acts as a test of its accuracy and validity.

The universality of weak interactions is a key property of particle physics's SM. The generalization of the weak universality concept for quarks (by Cabibbo) and the extension of the Cabibbo scheme (by Kobayashi and Maskawa) opened a new chapter of physics. Up-type (u, c, t) and down-type quarks (d, s, b) transition between them through the weak interaction with the emission of a W boson. In the SM context, the probability of transition from an up-type quark to all down-type quarks together is the same for all three generations. This relation is called weak universality and was first indicated by Nicola Cabibbo in 1967. It has undergone continuous experimental testing.

The LHC, a circular particle accelerator, houses various experiments, including the CMS experiment. Its primary task is to create conditions that allow for precise testing of the SM, enabling predictions to be either corroborated or refuted. There are dedicated efforts to do both precision physics, analysis aiming to reduce the uncertainty of already measured quantities, and searches, analysis whose goal is to measure a statistically significant quantity of events contradicting SM. It may happen that a precision measurement has a deviation from the theory prediction revealing a beyond the SM occurrence, so notably improving the precision of the framework parameters is a valuable contribution.

The main result from this thesis consists of the measurement of the rate of charm quark production in W boson decays relative to the rate of W hadronic decays to different quark flavors, the branching fraction ratio $\mathcal{R}_c^W = \mathcal{B}(W \to c\bar{q}) / \mathcal{B}(W \to q\bar{q})$. $\mathcal{R}_c^W$ is related to the Cabibbo–Kobayashi–Maskawa (CKM) matrix elements, encoding weak universality for the transitions of the two first up-type quarks (u, c). Assuming the unitarity of the CKM matrix, $\mathcal{R}_c^W$ is expected to be equal to 1/2. Therefore, a measurement of $\mathcal{R}_c^W$ is a direct test of this assumption.

The large cross section of $t\bar{t}$ production at the LHC, each decaying into a W boson and a bottom quark, offers a sizeable high purity sample of W bosons. The final state used for the $\mathcal{R}_c^W$ measurement consists of events where one of the W bosons decays into a lepton (electron or muon) and a neutrino, while the second W boson decays hadronically into two jets. The high transverse momentum lepton provides an excellent signature for the online

selection of the events, while the identification of a charm jet enables the measurement of $^{W}_{c}$.

Charm tagging is therefore an instrumental part of the data analysis. The charm tagging technique used in this work involves identifying a muon within jets that originates from the decay of a charm hadron. This distinctive characteristic of heavy-flavor jets enables the isolation of a clean charm sample, with well-controlled associated systematics. In addition, the electric charge correlation of the tagged muon and the lepton from the W boson decay allows the characterization of most of the backgrounds using a data control region.

The $^{W}_{c}$ measurement presented in this thesis improves upon the current world average value by reducing its uncertainty by approximately half. This enhancement represents a notable step forward in precision.

In addition to this result, this thesis also offers insights into uncertainty estimation when addressing particle physics problems with machine learning tools. As these techniques become more widely adopted, assessing their reliability is crucial. This work demonstrates the application of three uncertainty estimation techniques, Bayesian neural network estimation, Probabilistic random forest and Local Ensembles, to a binary classification problem using CMS open data.

Introducción

La Física de Partículas, en su ánimo de entender la naturaleza de la materia y cómo esta interacciona entre sí, se sirve en gran medida de experimentos para su avance. El Modelo Estándar (ME) funciona como marco teórico describiendo las partículas elementales y las interacciones fundamentales. Cualquier predicción de este modelo susceptible de ser medida servirá como prueba de su precisión y validez.

La universalidad de las interacciones débiles es una propiedad clave del ME. La generalización del concepto de universalidad débil para los quarks (por Cabibbo) y la extensión del esquema de Cabibbo (por Kobayashi y Maskawa) abrieron en su momento nuevas vías en la Física. Los quarks de tipo "up" (u, c, t) y los quarks de tipo "down" (d, s, b) pueden transicionar entre sí a través de la interacción débil con la emisión de un bosón W. En el contexto del ME, las probabilidades de transición de un quark de tipo "up" a todos los quarks de tipo "down" sumadas es la misma para las tres generaciones. Esta relación se llama universalidad débil y fue indicada por primera vez por Nicola Cabibbo en 1967. Esta propiedad ha sido sometida continuamente a pruebas experimentales.

El LHC es un acelerador de partículas circular que alberga varios experimentos, incluido el experimento CMS. Es capaz de crear las condiciones necesarias para poder hacer medidas precisas del ME, pudiendo corroborar o refutar sus predicciones. Se dedican esfuerzos tanto a análisis de precisión, la realización de medidas que reducen la incertidumbre de medidas ya existentes, como a búsquedas, análisis cuyo objetivo es obtener una cantidad estadísticamente significativa de eventos que contradicen el ME. Puede ocurrir que una medida de precisión tenga una desviación de la predicción teórica, revelando un fenómeno más allá del ME, por lo que mejorar notablemente la precisión de los parámetros de este marco teórico es una valiosa contribución.

El resultado principal de esta tesis consiste en la medida de la tasa de producción de quarks charm en las desintegraciones de bosones W en relación con la tasa de desintegraciones hadrónicas de W a diferentes sabores de quark, $^{W}_{c} = \mathcal{B}$ W cq $\mathcal{B}$ W q$\overline{q}$. $^{W}_{c}$ está relacionada con los elementos de la matriz Cabibbo–Kobayashi–Maskawa (CKM), que involucra la universalidad débil para las transiciones de los dos primeros quarks de tipo "up" (u, c). Suponiendo que la matriz CKM es unitaria, el valor $^{W}_{c}$ tendría que ser igual a 1/2. Por lo tanto, una medición de $^{W}_{c}$ supone una comprobación directa de esta hipótesis.

El LHC ofrece una gran sección eficaz para la producción de pares top antitop, decayendo cada uno de éstos a su vez en un bosón W y un quark bottom. Este proceso ofrece una muestra pura de bosones W. El análisis para esta medida se ha hecho seleccionando un estado final con un leptón (electrón o muón), un neutrino, y al menos cuatro

jets, emulando un evento $t\bar{t}$ semileptónico en el que uno de los bosones W decae leptónicamente, proporcionando una característica apropiada para el trigger, y el otro bosón W decae hadrónicamente, haciendo posible la identificación de un jet charm para la medida $_{c}^{W}$.

Clasificar eventos charm es una parte clave de este análisis. La técnica utilizada para etiquetar jets como charm consiste en identificar un muón dentro de dicho jet, proveniente del decaimiento de un hadrón charm. Esta característica es particular de los jets que surgen de quarks de sabor pesado, lo que nos hace obtener una muestra charm limpia y poder controlar la sistemática asociada adecuadamente. Además, existe una correlación de signo de carga eléctrica entre este muón identificado dentro de un jet y el leptón originado en la desintegración del bosón W, a través de la cual podemos caracterizar la mayoría de los fondos usando una región de control de datos.

La medida $_{c}^{W}$ final mejora el valor promedio mundial actual al reducir su incertidumbre a aproximadamente la mitad. Esto supone un avance representativo en precisión.

Además de este resultado, esta tesis también aborda el tratamiento de datos en Física de Partículas con machine learning. Ofrece una perspectiva sobre la estimación de incertidumbre cuando se usan este técnicas de aprendizaje automático. Dado que su uso es cada vez más extendido, poder evaluar su fiabilidad es crucial. En este trabajo se aplican tres técnicas de estimación de incertidumbre: aproximación de redes neuronales bayesianas, random forest probabilístico y local ensembles, a un problema de clasificación binaria con datos en abierto de CMS.

Chapter 1

Theoretical background

The Standard Model of particle physics (SM) is the quantum field theory that describes the constituents of matter and their interactions. The measurement upon which this thesis is based explores one important aspect of the SM, the universality of the electroweak interactions. The determination of this prediction and other relevant aspects of SM are included in this chapter. The phenomenology of proton-proton collisions and their modeling are also discussed.

1.1 The Standard Model of particle physics

The SM is a theoretical framework that unifies electroweak and strong interactions. Quantum Electroweak Theory (QED) encompasses both electromagnetic and weak interactions, including neutral and charged currents, while Quantum Chromodynamics (QCD) describes the strong interaction. Together, these theories form the SM. It is a gauge invariant theory associated with symmetry group. The strong interaction is specified by and the electroweak interaction by . Particles in the SM are conceived as excited states of quantum fields, differentiated in fermions and bosons. Fermions are considered as matter constituents whereas bosons play the role of force mediators between fermions. The Brout-Englert-Higgs mechanism [1, 2] proposes a spontaneous rupture of the electroweak symmetry allowing for a unified theory of forces where bosons can have mass.

The SM has been put to test by multiple low and high energy experiments. CERN SPS experiment reported the observation of W and Z bosons, the mediators for the weak interaction, in 1983 [3, 4, 5, 6]. Particle accelerators have been crucial throughout history in testing existing theories and facilitating the development of new ones. LHC [7] experiments CMS and ATLAS have been able to conduct various measurements related to SM predictions. The discovery of the Higgs boson in 2012 [8, 9], the only postulated elementary particle that had not been observed yet, was a remarkable milestone. Although the SM has consistently aligned with experimental results, there are indications of phenomena beyond

it. Therefore, searching for anomalies and improving measurement accuracy remains of great interest.

The SM considers two types of particles: fermions, which have half-integer spins and constitute matter, and bosons, which have integer spins and mediate fundamental interactions. Fermions are categorized into quarks and leptons, with each type divided into three generations. Each generation of quarks includes two distinct quarks with electric charges – and – and two leptons with electric charges and , see Fig. 1.1. This results in six types of quarks and six types of leptons, each with a corresponding antiparticle. An antiparticle has the same mass and spin as its counterpart but opposite quantum numbers. The SM includes 8 gluons, 3 gauge bosons (, , and), and 1 photon ( ), totaling 12 gauge vector bosons. Additionally, there is the Higgs boson, associated with the Higgs scalar field, which enables fermions and bosons to acquire mass.

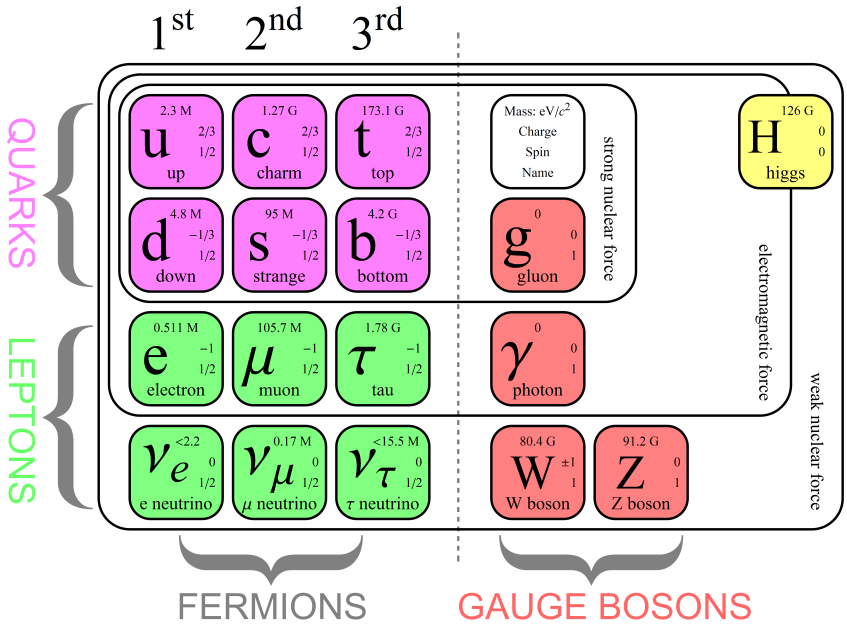


Figure 1.1: The depiction of SM particles, shown in the diagram from [10], includes both fermions (organized into leptons and quarks) and bosons.

The strong interaction, described by and known as Quantum Chromodynamics (QCD), only affects particles with color charge. Particles with color charge include quarks, which form color triplets, and gluons. Quarks are categorized into three generations: the first generation contains up (u) and down (d) quarks; the second generation includes charm (c) and strange (s) quarks; and the third generation comprises top (t) and bottom (b) quarks. They posses electric charge, mass, color charge, and flavor making them the only known elementary particles that engage in all four fundamental interactions. Gluons are massless vector bosons mediating the strong interaction. Color confinement is a property that prevents particles with non-neutral color charge from existing as free states. QCD governs how gluons bind quarks together to form color-neutral particles, such as protons and neutrons, known as hadrons.

There are three generations of leptons: electrons ( ), muons ( ) and taus ( ) paired

with their corresponding neutrinos (, and). Within the Standard Model, neutrinos are considered massless. However, the observation of neutrino oscillations [11] suggests that neutrinos do indeed have mass. Electroweak interactions are described by the gauge symmetry , affecting not only electrons, muons, and tau particles but also quarks. Neutrinos are only affected by weak interactions. The and bosons, along with the boson, mediate the weak interaction, while the photon ( ) mediates the electromagnetic interaction, affecting only particles that carry an electric charge.

1.2 The electroweak interaction

Electromagnetic and weak interactions are governed by the gauge symmetry group , making them gauge theories. Glashow, Weinberg and Salam formalised this theory, granting them the Nobel Prize in Physics in 1979 [12]. These interactions affect particles based on their weak isospin and hypercharge, with different effects on left-handed and right-handed fermions. Left-handed fermions form doublets, while right-handed fermions are singlets, each having distinct hypercharges. The chirality of a fermion is defined through the eigenvalues of the operator . The chirality of a particle is determined by the sign of its eigenvalue: right-handed particles correspond to +1, while left-handed particles correspond to -1. The fields associated with each chirality are:

– – (1.1)

An example of a fermionic representation is the first family of quarks (the up and down quarks). In this representation, the up and down quarks form a left-handed chirality doublet, with the right-handed chirality components represented as singlets:

(1.2)

Leptons and the other quarks follow a similar characterization. Neutrinos on the other hand do not have a right-handed component. The non-abelian group describes the weak interaction, with the weak isospin defined as the conserved charge associated with this symmetry. Left-handed fermions have for up-type quarks and for down-type quarks, whereas right-handed fermions, being singlets of the group , have . The Lagrangian term for the associated with weak interaction is defined as:

$\mathcal{L}$ – (1.3)

The term is the strength tensor and satisfies , being the structure constants of . The covariant derivative of said group has the expression with , and the coupling constant

of the weak force. It is relevant noting that Eq. 1.3 does not include a mass term of the type , a term like this would not be invariant under transformations.

The group has an associated Lagrangian expressed as Eq. 1.4. The term in the equation, that acts as as the covariant derivative of , is defined as , where the term is the coupling constant. The tensor field satisfies .

$$\mathcal{L} \quad - \tag{1.4}$$

This term has an associated conserved magnitude noted as the hypercharge . This quantity relates to the isospin and electric charge as . The framework presented so far is only compatible with massless fermions and bosons since, as stated before, mass terms break the gauge invariance under SM symmetry . However experiments show evidence of massive particles with these characteristics, except photons. We proceed to describe the Higgs mechanism, that involves the spontaneous breaking of the electroweak symmetry and allows for particles to acquire mass.

1.2.1 Electroweak symmetry breaking

The Higgs mechanism breaks spontaneously three of the four generators of . To accomplish this, a complex Higgs doublet of is defined:

$$— = \tag{1.5}$$

where $^{-}$ is the Higgs vacuum expectation value, the minimum value for the Mexican hat potential represented in Fig. 1.2, and . The Higgs Lagrangian is gauge invariant due to the presence of the covariant derivative, leading to interactions with the gauge fields:

$$\mathcal{L} \tag{1.6}$$

The term is added to the minimum , exciting the ground state function:

$$— = \tag{1.7}$$

where is the field associated to what we call the Higgs boson (H), a massive particle.

The Yukawa Lagrangian (Eq. 1.8) expresses the interaction of fermions with the Higgs field. This mechanism confers mass to fermions. The term in the equation represents the coupling strength between the Higgs boson and fermions. These interactions are

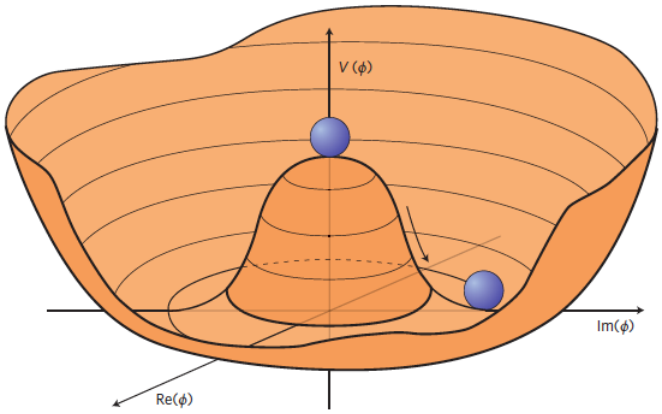


Figure 1.2: Higgs potential as a function of in the complex plane.

proportional to masses of the fermions, making the Higgs boson decay preferably to the heaviest particles that are kinematically accesible.

$$\mathcal{L}_{\text{Yukawa}} \tag{1.8}$$

For this model, neutrino masses are neglected, having only left-handed particle part for the leptonic case, that excludes neutrinos. This expression only considers one generation of fermions. In order to account for the three generations of quarks, the Lagrangian could be stated as follows:

$$\mathcal{L}_{\text{Yukawa}} \tag{1.9}$$

where Yukawa couplings, denoted as , are included. is a complex matrix. In principle, it is not diagonal [13]. Diagonalizing it would imply that W bosons only mediate weak interaction between quarks of different flavours. The Cabibbo-Kobayashi-Maskawa (CKM) [14] matrix values determine the strength coupling of these interactions. More details in the next section.

Electroweak Symmetry Breaking addresses the experimental requirement for gauge symmetry breaking, resulting in finite masses for fermions and some bosons. The introduction of the Higgs field into the electroweak Lagrangian enables these particles to acquire mass. The interactions between fermions and the Higgs boson are governed by Eq. 1.8, which is proportional to the fermion mass.

1.2.2 Charged current electroweak interactions

The interaction of quarks via the weak force can be compared to that of leptons, with the expectation of similar behavior. One might assume that the coupling strength between quarks and weak bosons is the same as for leptons, denoted by . Then, considering the two first generations of leptons and quarks

and (1.10)

they would have identical weak couplings, expecting . Experiments measuring and show that this is not the case. The reaction could occur since its quark constituents are . The reaction, however, has quark constituents and should not happen. Further measurements reveal that the coupling strength of quarks to weak bosons is weaker than that of leptons. This coupling becomes even smaller when it occurs between quarks from different families [15].

To model this quark flavour mixing, instead of adding new terms to account for the differences, the relation can be redefined as

u c
d s (1.11)

with

d d s (1.12)
s d s (1.13)

d d
s s (1.14)

where is the Cabibbo angle for quark flavour mixing [16]. This way, since , couplings within the same family of quarks are stronger than those between quarks from different families. These couplings can be expressed as $_{ud}$ W , $_{us}$ $_{cd}$ W . The decays ruled by these terms are represented in Fig. 1.3.

The matrix in Eq. 1.14 must be extended to consider all quarks. In 1973, Kobayashi and Maskawa expanded the Cabibbo matrix into the Cabibbo-Kobayashi-Maskawa matrix (CKM matrix) to incorporate CP-Violation into weak interaction theory [17]. This matrix is expressed as:

$$\begin{matrix} d \\ s \\ b \end{matrix} \quad \begin{matrix} d \\ s \\ b \end{matrix} \quad \begin{matrix} _{ud} & _{us} & _{ub} \\ _{cd} & _{cs} & _{cb} \\ _{td} & _{ts} & _{tb} \end{matrix} \quad \begin{matrix} d \\ s \\ b \end{matrix} \tag{1.15}$$

The CKM matrix represents the coupling strengths of quarks via the weak interaction. According to SM, this matrix should be unitary, that is to say, for every row and column, the sum of the squares of their elements should be equal to 1, .
The unitarity condition of the CKM matrix codifies the universality principle of the weak

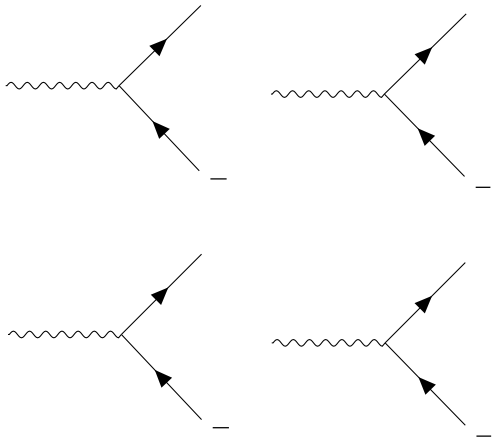


Figure 1.3: W boson hadronic decay modes for the first and second generations of quarks. Each transition is governed by the Cabibbo angle: the decays W ud and W cs are proportional to , while W us and W cd are proportional to . The most probable are then the decays corresponding to diagonal elements of matrix in Eq. 1.14.

interaction in the quark sector: the sum of the couplings of any up-type quark to all down-type quarks is the same. Theoretically it is a consequence of the fact that all doublets couple with the same strength to the vector bosons of weak interactions. A deviation from this unitarity would suggest the presence of physics beyond the SM.

As of 2023, the best determination of the individual magnitudes of the CKM matrix [18] elements is expressed as it follows:

ud	us	ub
cd	cs	cb
td	ts	tb

The rate at which W bosons decay hadronically to a pair of quarks is determined by Eq. 1.16, where is a constant dependent on , is the aforementioned weak coupling constant and is the mass of the W boson. The branching ratio of a W boson decaying to a charm quark relative to the rate of W hadronic decays to different quark flavors may be expressed as in Eq. 1.17 where contributions involving the top quark have been excluded since they are not kinematically accessible.

$$\mathcal{B}\ \mathrm{W} \quad \mathrm{q}\,\overline{\mathrm{q}} \quad \text{———} \quad \text{–} \tag{1.16}$$

$$\frac{\mathcal{B}\ \mathrm{W} \quad \mathrm{cq}}{\mathcal{B}\ \mathrm{W} \quad \mathrm{q}\overline{\mathrm{q}}} \quad \frac{\mathrm{cd} \quad \mathrm{cs} \quad \mathrm{cb}}{\mathrm{ud} \quad \mathrm{us} \quad \mathrm{ub} \quad \mathrm{cd} \quad \mathrm{cs} \quad \mathrm{cb}} \tag{1.17}$$

Assuming the unitarity of the CKM matrix, $^{\mathrm{W}}_{\mathrm{c}}$ is expected to be equal to 1/2. Therefore, a measurement of $^{\mathrm{W}}_{\mathrm{c}}$ is a direct test of this assumption.

1.3 The strong interaction

Strong interactions are described by Quantum Chromodynamics (QCD), the theory associated with the symmetry group. As discussed in Section 1.1, quarks and gluons cannot exist as free states due to color confinement. Instead, they bind together to form color-neutral composite particles called hadrons. Color charge is the conserved quantity associated with the symmetry of the strong interaction, and both quarks and gluons carry this property. Hadrons are formed by combining three quarks (baryons) or a quark and an antiquark (mesons). Protons and neutrons are examples of baryons, while pions and kaons are examples of mesons. The strong interaction also contributes to the stability of atomic nuclei by binding nucleons together. This interaction is sufficiently strong to overcome the repulsive forces between protons, which have the same electric charge.

The Lagrangian related to this interaction is expressed as:

$$\mathcal{L} \qquad - \tag{1.18}$$

where are the Dirac matrices, the quark field, the quark mass, is the coupling constant of the strong interaction and is the strength field tensor. This last term accounts for the strong interaction group to be non-abelian and models 3 and 4-gluons self-interactions.

The group representing this interaction, , has eight generators, , called the Gell-Mann matrices. They relate to the strong force mediators, the eight gluons, that are massless. The mentioned coupling constant is often expressed as . The renormalisation of the group, considering only one-loop Feynman diagrams, relates and , the interaction scale, as in Eq. 1.19. This means that as the energy scale decreases, the strength of the interaction increases. This behavior is crucial for color confinement.

$$\qquad \qquad \tag{1.19}$$

Although the strength of the strong force increases with distance between particles, at short distances, a phenomenon known as asymptotic freedom [19] occurs. In this regime, quarks experience a low coupling constant and can be treated as approximately free particles. This concept is crucial for modeling proton-proton collisions.

1.3.1 Proton-Proton collision phenomenology

The measurement performed in this thesis is based on data from proton-proton collisions. As mentioned, protons are not elementary particles but hadrons constituted by quarks and gluons. To model the hard scattering process of two hadrons the factorization theorem [20] is used. It assumes that quarks and gluons behave as free particles during the collision,

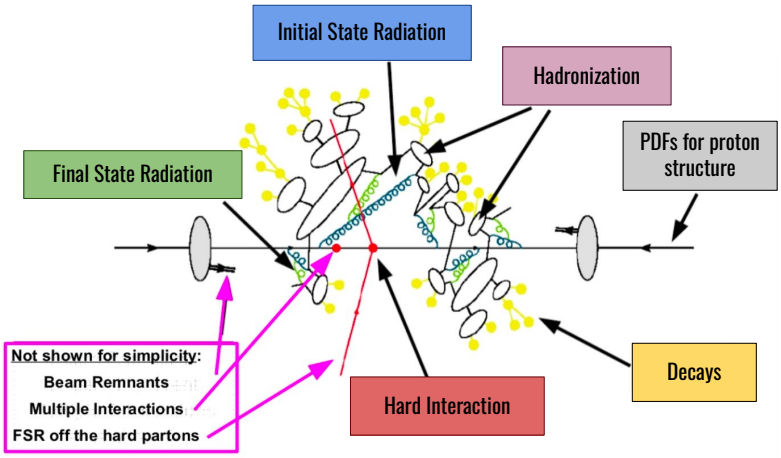


Figure 1.4: Depiction of the phenomena occurring in a proton-proton collision.

allowing the computation of the proton-proton cross section to be factorized into parton-level cross sections ( ), which represent the interactions between the elementary particles within the protons:

(1.20)

represents parton distribution functions (PDFs), which are probability density functions describing the momentum distribution of partons within the proton. Here, denotes the fraction of the proton's momentum carried by the parton, and is the energy scale of the interaction. PDFs are initially formulated and then refined based on experimental data.

The measurement of this thesis is done by comparing data collected at CMS and simulation generated to implement the theoretical calculations of the physics process and model the detector response. Proton-proton collisions are simulated by means of Monte Carlo (MC) generators, such as Pythiav8.2 [21], Powheg [22] and MadGraph5_amc nlo [23]. These generators use PDF sets, Nnpdf3.0 [24] for 2016 simulation samples and Nnpdf3.1 [25] for 2017 and 2018 simulation samples, to simulate the initial state of the interacting partons (quarks or gluons) in the proton. The hard interaction of the partons of the two protons is calculated up to certain order in perturbative QCD using Feynman diagrams of the possible outcomes of the interactions. Partons can radiate before (initial state radiation) or after (final state radiation) the interaction. In the so-called parton shower, quarks radiate gluons that can split into quark-antiquark or gluon-gluon pairs, which in turn can also further radiate creating a cascade effect.

As the energy of the quarks and gluons produced in the interaction decreases, hadrons are produced due to color confinement. The hadronization process is not yet fully described within QCD and a number of phenomenological models adjusted with experimental data are used.

Partons not involved in the hard interaction can also interact with each other (multiple parton interaction). The simulation of the underlying event, which includes the hadronization, the initial and final state radiations, and the underlying event, involves a tuning process using experimental data.

Lastly, the detector response to the produced particles must be modeled in the simulation. The tool used for this task is GEANT4 [26] that simulates the effect that the passing of particles has on matter.

Chapter 2

The CMS experiment at the LHC

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) Project was approved by CERN [27] Council in December 1994 [28]. The LHC is the world's largest particle accelerator, with 27 km diameter located in the France-Switzerland border, near Geneva, see Fig. 2.1. The tunnel accommodating the LHC previously housed LEP (Large Electron Positron collider) [29]. Located around 100 m underground, it is composed by two rings. Each of the rings contain a beam pipe where either protons or heavy ions circulate in opposite directions.

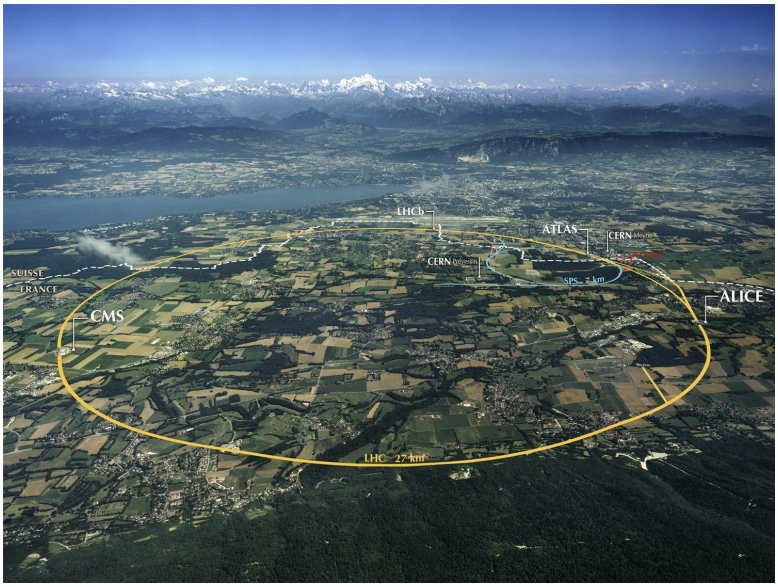


Figure 2.1: Air view of land where the LHC is buried in. The yellow line indicates its location.

They intersect at four points along the circunference, corresponding to four experiments: ATLAS (A tororidal LHC ApparatuS) and CMS (Compact Muon Solenoid), placed in opposite locations of the ring, are general purpose particle detectors with similar physics research goals. ALICE (A large Ion Collider Experiment) and LHCb (LHC beauty)

are smaller experiments with more specific physics purposes. ALICE studies heavy ions collisions for quark-gluon plasma research and LHCb focuses on CP-violation and matter-antimatter asymmetry by specilizing in b quark physics.

The LHC has 1232 superconducting dipoles magnets to assure the circular path of the beams creating a 8.3 T magnetic field and 392 additional quadrupole magnets to keep the beams focused and not spread because of the same sign electric charge of the accelerated particles. Around 96 tonnes of superfluid hellium at 1.9 K (-271.25ºC) are needed to cool down the magnets and maintaining them at operating temperature.

The work of this PhD thesis has been done studying collisions between protons. These are collided in a series of bunches of around protons at a 40 MHz rate. Before entering the LHC main ring, they are already accelerated by previous devices. Firstly, hydrogen atoms are ionized and injected into LINAC2 that accelerates them up to around 50 MeV, then a series of circullar accelerators and the system BOOSTER accelerates them to 1.4 GeV. Then protons are fed into the PS (Proton Synchroton), where they reach 26 GeV and SPS (Super Proton Synchroton), where they reach 450 GeV. Then they enter the LHC for acceleration to operating point. Fig. 2.2 showcases the accelerating set-up at CERN.

Even though this device was designed to have an energy at the center of mass, $\bar{}$, of 14 TeV, it has varied through out the years. During Run 1, a three year phase, 2010-2012, it had 7 TeV in 2010 and 8 TeV in 2011 and 2012. This work uses data collected during Run 2 (2015-2018), 2016, 2017 and 2018 specifically, at which $\bar{}$ TeV. Run 3 started in 2022 and reached 13.6 TeV.

2.1.1 Accelerator features

In order to keep record of collision rates a quantitative magnitude is needed. The cross-section of a process is interpreted as the probability of it occurring, it is typically measured in barns, a unit of area 1 b m . In particle physics, cross-sections are usually very small so units used are pb or fb. The collision rate will depend on the scattering cross-section of two protons colliding, Fig. 2.3, but also on the number of protons in the bunches subject to collide, the frequency of the collision and other parameters.

The instantaneous luminosity $\mathcal{L}$ [31] is defined in Eq. 2.1, where is the transverse emittance of the beam, defined as the relative distance of protons in the same bunch respresented in the position-momentum frame of reference, is the amplitude function, representing the spread of the beam, is the frequency with which the two beams collide and and are the number of protons of each beam.

$$\mathcal{L} \quad \text{———} \tag{2.1}$$

The collision rate, , is related to sigma and $\mathcal{L}$ through the expression — $\mathcal{L}$. $\mathcal{L}$ gives us, then, information about the accelerator performance as it does not depend on the physical process measuring the LHC production capacity.

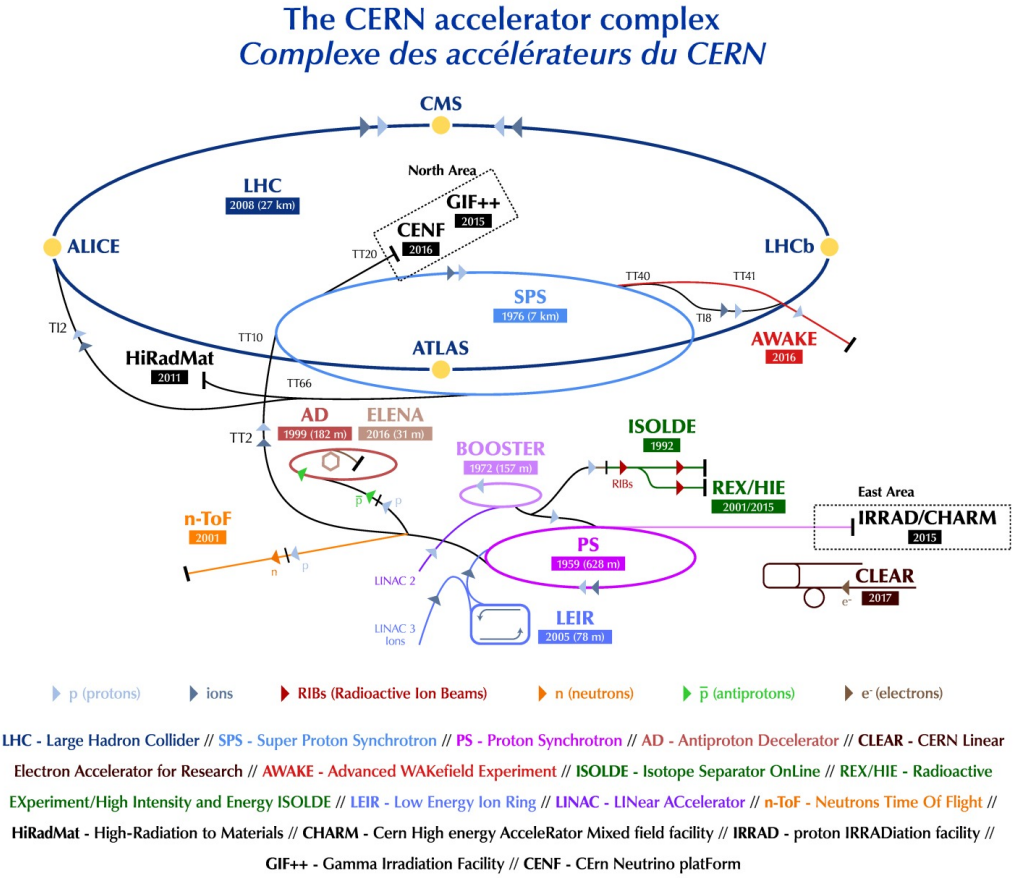


Figure 2.2: Scheme of the various accelerators present at CERN complex. The picture explains how the smaller devices feed the LHC. It was extracted from [30].

The integrated luminosity is the instantaneous luminosity integrated over time, resulting in the total number of events collected during that time period. It is expressed in units of inverse picobarns or femtobarns (pb or fb) while instantaneous luminosity is expressed in . Fig. 2.4 shows the integrated luminosity recorded in Run 2 per year, the total of data used in this work corresponds to 138 fb .

Multiple proton-proton collisions occur in every bunch crossing. This effect is called in-time pile-up (PU). The average value of the in-time pile-up can be expressed as in Eq. 2.2 where $^{\text{inel}}$ is the cross-section of inelastic proton-proton collisions, is the number of bunches (2556 is the maximum), and is the bunch crossing frequency (40 MHz). Figure 2.5 shows the distribution of the number of in-time interactions for the Run 2 data taking periods.

$$\mathcal{L}\frac{^{\text{inel}}}{} \tag{2.2}$$

Collisions occurring at nearby bunch crossings (separated by 25 ns) can also be integrated by the detector, since the time resolution of some of the subdetectors is not good

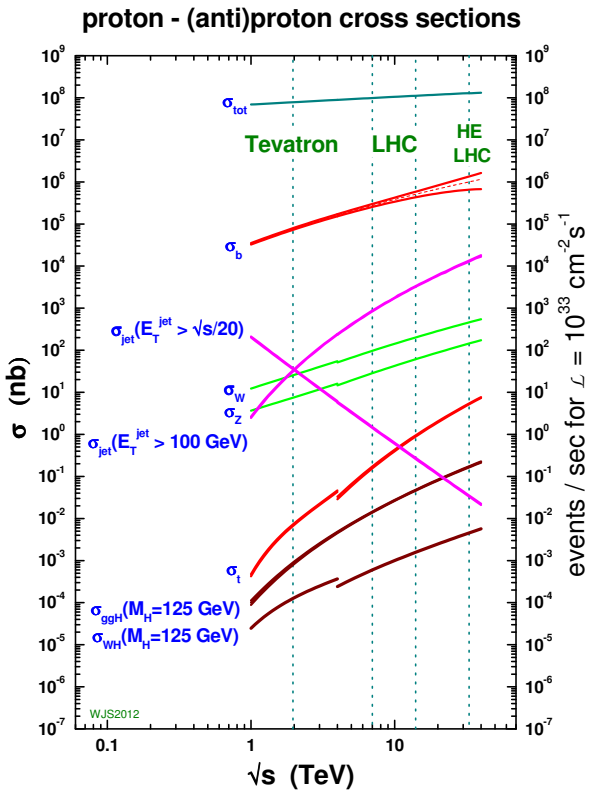


Figure 2.3: Cross section (and event rate) of pp collisions as a function of the center-of-mass energy. Production cross sections of various processes are also indicated. The image was obtained from [32]. The dashed lines indicate the working range of various accelerators (Tevatron, LHC, and a potential High-Energy LHC).

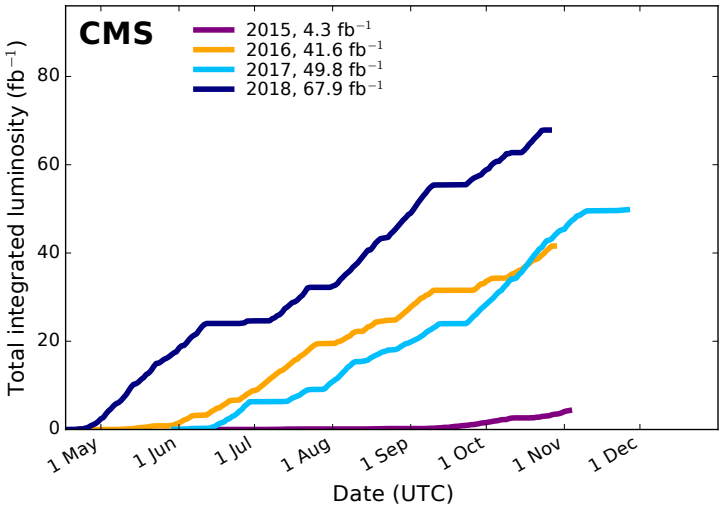


Figure 2.4: Integrated luminosity recorded at CMS for Run 2 operations, obtained from [33].

enough to distinguish nearby collisions. This effect is called out-of-time pileup. The pileup effect must be correctly modeled in the simulations, since large values result in a efficiency and resolution degradation of reconstruction of the information recorded.

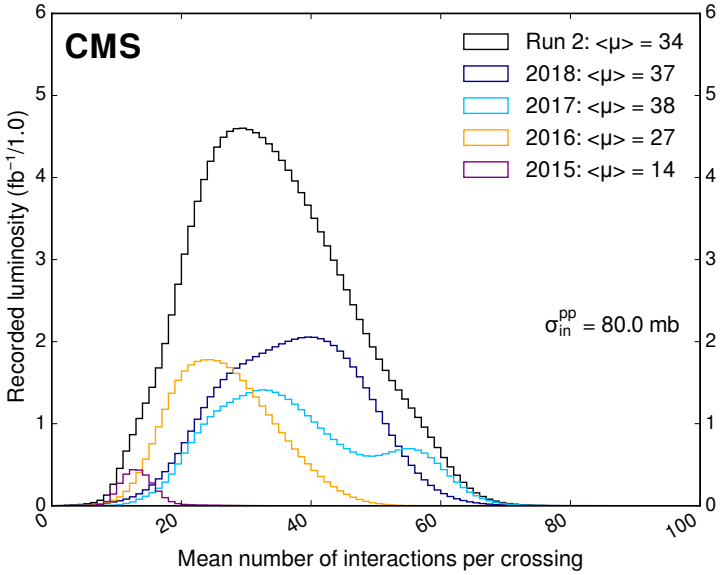


Figure 2.5: Distribution of the number of simultaneous proton-proton collisions for the Run 2 data-taking years, obtained from [33].

2.2 The CMS detector

The Compact Muon Solenoid (CMS) is a general-purpose detector at the LHC point 5 [34]. The 14,000-tonne detector gets its name from the fact that it really is quite compact for all the detector material it contains, it is designed to detect muons very accurately and it has the most powerful solenoid magnet ever made. It is a 21.6 m long cylindrical apparatus with 14.6 m diameter. The solenoid creates a 3.8 T magnetic field, it is a layer of 6 m internal diameter and 12.5 m of length. The muon system covers the solenoid and is composed by a cylindrical barrel and 2 planar endcaps. A photo of the real experiment is depicted in Fig. 2.6.

The coordinate system used at CMS is right-handed, referenced with respect to the nominal interaction point. As a coordinate system the orientation is as showcased in Fig. 2.7, with axis directed towards the center of the collider, axis normal to the plane defined by the collider circumference and axis tangent to the circumference, following the beam direction. Other used coordinates are , defined as the azymuthal angle in plane taking axis as reference and as the polar angle enclosed by the momentum vector with respect to axis.

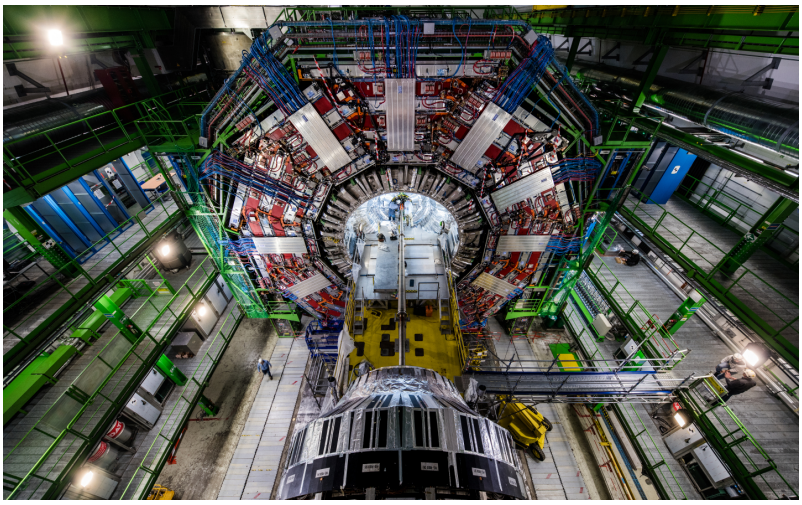


Figure 2.6: A real image of the CMS detector, shown in its open configuration, revealing the beam pipe. The muon barrel chambers are located on the sites of the red iron structures. It was obtained from [35].

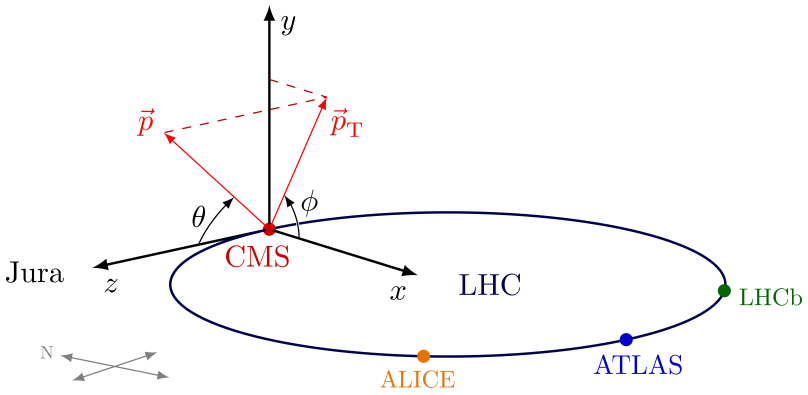


Figure 2.7: Coordinate system used at CMS. Using the beam line as reference, a cartesian set of variables are defined, being z tangent to the beam, and x and y perpendicular to it, pointing x to the center of the accelerator circle. The image source is [36].

A useful variable is the rapidity , expressed as:

– ——— (2.3)

where is the energy and the momentum along the axis for the detected particle. Since this quantity is invariant under Lorentz boosts along the z axis, it is of great interest in hadronic collisions studies. It is redefined for highly relativistic particles, an approximation, named pseudorapidity and denoted as is established. It is expressed as in Eq. 2.4. It satisfies not depending on the particle energy or momentum, and it ranges from 0 as to infinity when . We will speak of central region when $\lesssim$ and forward region otherwise. Left image in Fig. 2.8 shows the correspondence between and variables.

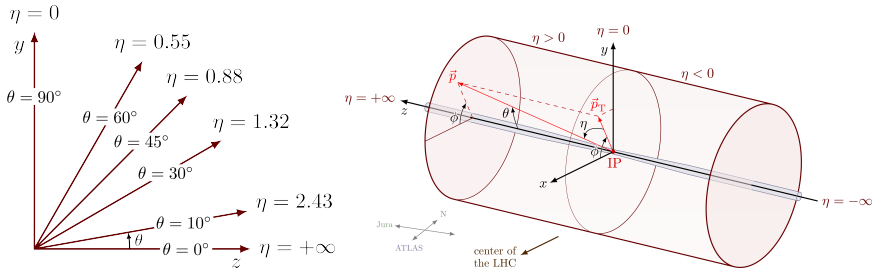


Figure 2.8: Depiction of the variable's correspondence to in the left. The right image illustrates this variable along the beam line. Images collected from [36].

ln tan− (2.4)

A particle's complete momentum information is defined by the variables , and the transverse momentum, denoted as , and defined as the momentum projected in the plane orthogonal to the beam axis. The relations between these variables and are formulated:

(2.5)

The right image in Fig. 2.8 illustrates transverse momentum in the reference system. Another magnitude that is widely used in the experiment context is . It is the angular distance between two objects making use of their coordinates and . Its expression is stated in the following expression:

(2.6)

2.2.1 Subdetectors

The CMS detector can be sectioned in different subdetectors, each specialized in detecting specific particles: electrons, photons, hadronic particles or muons. The detector can be divided in a central zone, referred to as barrel, divided in turn into 5 cylindrical slices, called wheels and two forward regions, denominated endcaps.

Figure 2.9 shows a scheme of the CMS detector structure, arranged around the interaction point. Ordered by distance to the interaction point, the silicon trackers are pixel trackers that reconstruct charged particles trajectories and their origin with respect to the reference system. The electromagnetic calorimeter (ECAL) is an array of scintillating crystals measuring the energy that electrons and photons deposit. The hadronic calorimeter (HCAL) is composed by passive material and plastic scintillators. It measures the energy of the showers of particles that hadrons produce in their interaction with the detector

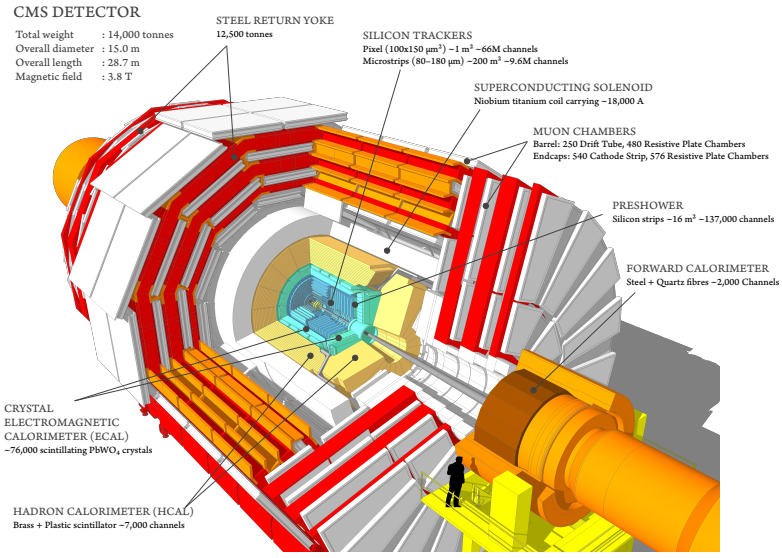


Figure 2.9: Scheme of the various subdetectors that CMS includes. Extracted from [37]

material. The superconducting solenoid creates a magnetic field large enough to bend the trajectory of high-energy charged particles, allowing the measurement of their transverse momentum. Lastly, muon chambers are formed by different detection technologies, drift tubes, cathode strip chambers and resistive plate chambers, making it possible to gather precise information of the muons produced in collisions. More detailed descriptions are included hereunder but further information can be found in [38].

The solenoid magnet

The superconducting solenoid magnet, made of niobiumtitanium cooled down to 4.7 K using liquid helium, provides a uniform magnetic filed of 3.8 T inside the solenoid, and up to 2 T outside of it. It weighs 220 tons, has a 6 m radius and is 12.5 m long. The muon chambers surrounding the magnet are placed in the iron magnet return yoke that confines the magnetic field.

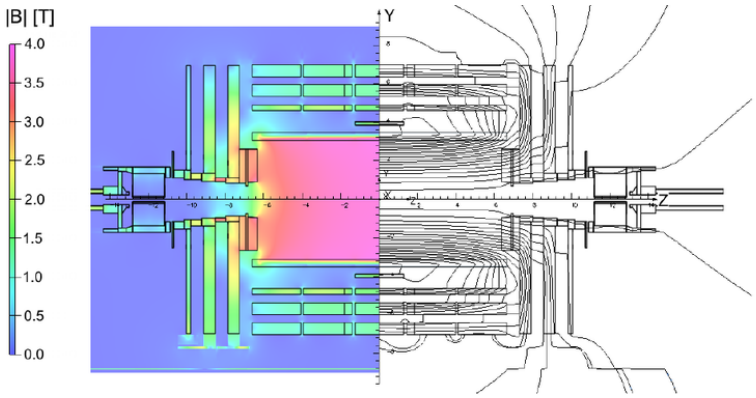


Figure 2.10: Map of CMS solenoid, with colors indicating the intensity of the magnetic field induced (left), and lines of the field (right). Image obtained from [39].

The trajectory of the charged particles produced in the collisions are bent by the magnet. The measurement of the tracks curvature allows the determination of their transverse momentum. The larger the magnetic field, the larger the bending and so the precision of the momentum measurement.

The silicon tracker

Charged particles, when going through the silicon tracker, ionize the silicon semiconductors creating electron-hole pairs. The induced currents, drifting to the electrodes, are measured and digitized for its posterior analysis.

The CMS tracker [40, 41] is the closest subdetector to the interaction point. Its dimensions are 5.8 m long, 2.5 m in diameter, a surface area of 200 , and an angular coverage of . It is made of silicon, making it resistant to the high radiation levels it is exposed to during the collisions while providing very good spatial resolution. Charged particles, when going through the silicon tracker, ionize the silicon semiconductors creating electron-hole pairs. The induced currents, drifting to the electrodes, are measured and digitized for its posterior analysis. The produced hits, containing position, time and energy information, are used to reconstruct the track of the particle and its origin using pattern recognition techniques.

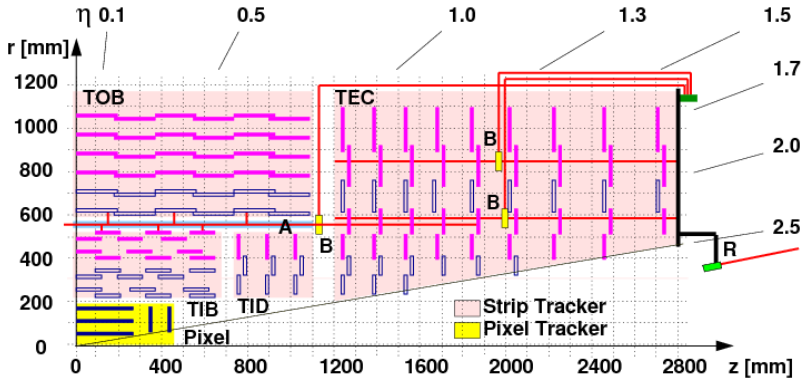


Figure 2.11: Section of the tracker indicating areas corresponding to the strip and pixel trackers. Extracted from [42].

The reconstructed track origin may match the main interaction point, also called primary vertex, or it may be displaced from it, creating a secondary vertex. Secondary vertices are produced from the decay of long-lived particles such as heavy-flavour quarks or leptons.

Figure 2.11 shows a schematic view of the tracker with its structure divided in two main parts. The silicon tracker consists of a pixel detector surrounding the beam pipe and a strip detector that is placed around the pixel detector. The pixel tracker is closer to the inner part of the detector, the particle flux originated by the collisions is higher in this region. Therefore, high granularity is required to obtain sufficient resolution to reconstruct individual tracks. In its initial configuration, the detector consisted of three detector layers in the barrel region of radius 4.4, 7.3 and 10.2 cm, and other two layers on the detector sides distanced cm and cm. This configuration is able to conduct 3-D measurements

from single hits following the path of the passing charged particles. There are a total of 124 million pixels, each of them measures m , providing a resolution for the hits of 10-40 m. Before 2017 data-taking period started, the detector was upgraded [43] in order to maintain a good tracking performance when instantaneous luminosities reached high values up to and the average pileup level exceeded 50 simultaneous collisions.

The strip detector is located around the pixel detector. It consists of million silicon micro-strip sensors placed with 15 different orientations depending on the distance to the interaction point. The number and size of the micro-strips are enough to reach a clear hit detection since the particle occupancy is lower in this region. This detector can be divided in four parts: the tracker inner barrel (TIB), the tracker inner disks (TID), the tracker outer barrel (TOB) and the tracker endcaps (TEC), see Fig. 2.11. The TIB covers the radial region cm and the TOB goes up to cm. They contain four and six layers respectively. The TID is located at the endcaps, it is made of three layers on each side, covering the TIB. The tracker endcaps (TEC), with nine layers, also covers the TIB. The spatial resolution in the transverse plane is m for the TIB and m for TOB. The resolution in the longitudinal direction is ten times larger. The overall efficiency for hit reconstruction in isolated muons is .

The Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) [44] is a hermetic and homogeneous apparatus formed by 70k scintillating crystals made of lead tungstate (PbWO). Its role is the identification of electrons and photons. They generate an electromagnetic shower when interacting with the ECAL material, leaving most of their energy in the ECAL detector. Knowing the position and magnitude of the energy deposit, the energy and direction of the particle causing the shower can be inferred. The characteristics of PbWO material, high density (g/cm), short radiation length (cm) and small Molière radius (cm), makes it appropriate for operating at the LHC conditions. This way it can provide measurements with fine granularity while remaining a compact calorimeter.

Figure 2.12 shows the ECAL layout, showing its division in three different regions: the barrel ECAL (EB) covering , two endcaps (EE) at and a preshower detector (ES) facing the endcaps ( ). The EB has 61200 trapezoidal crystals, each covering a surface in the front face, equivalent to in the plane, and at the rear face. The crystal is long, corresponding to radiation lengths. In order to avoid cracks between crystals, they are placed in specific ways so they align with the expected particle trajectories. Each EE is mounted with 7324 crystals, with a corresponding rear face cross section of , a front face cross section of and a length of (radiation lenghts).

The ES at the endcaps are composed by two layers of lead absorber equipped with orthogonal layers of strip sensors, helping with separation. The energy resolution of the ECAL [34] can be expressed as:

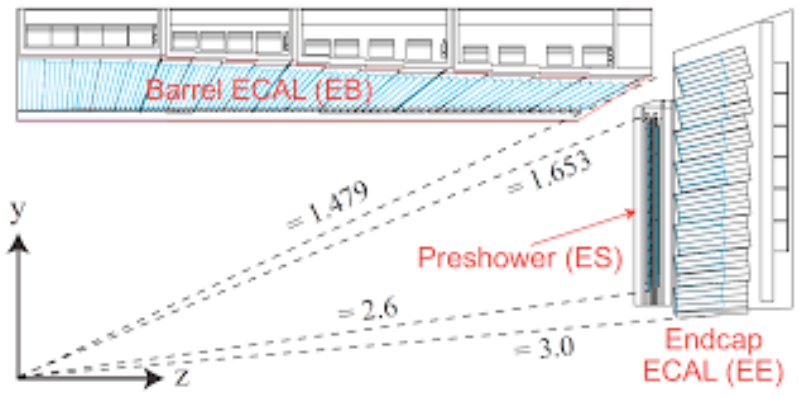


Figure 2.12: Schematic view of the ECAL detector, showing the barrel and the endcap zones. Obtained from [45].

— —= — (2.7)

where corresponds to a stochastic term modeling fluctuations in the shower, represents electronics, digitization or pileup noise and is a constant term for intercalibration errors or energy leakage from the back of the crystal.

Test beams confirm this relation by summing crystals, obtaining a resolution as:

— —= —— (2.8)

Hadronic calorimeter

The Hadronic CALorimeter (HCAL) [46] at CMS measures the energy of neutral hadrons and other non charged particles. It completely surrounds the ECAL, covering . Particles arriving deposit around of their energy after going through the ECAL. The HCAL is constitued by alternating layers of brass absorber) and plastic scintillator tiles (3.7 mm thick).

When hadrons interact with these materials, they create hadronic showers that induce light in the plastic scintillators. These hadronic-induced cascades are more difficult to measure than electromagnetic showers, this is because hadronic showers have larger fluctuations in the spatial development and energy loss and also nuclear interaction length is much larger than electromagnetic radiation length. Because of this, the HCAL is larger than the ECAL.

Figure 2.13 displays the HCAL structure, showing that it is an ensemble of different subdetectors: the barrel calorimeter (HB) covers the inner region and has a length of approximately 7 radiation lengths ; the endcap hadronic calorimeter (HE) covers the region 3.0 and has a depth of around ; the outer hadronic calorimeter (HO) complements the barrel and endcap detectors, expanding the detection

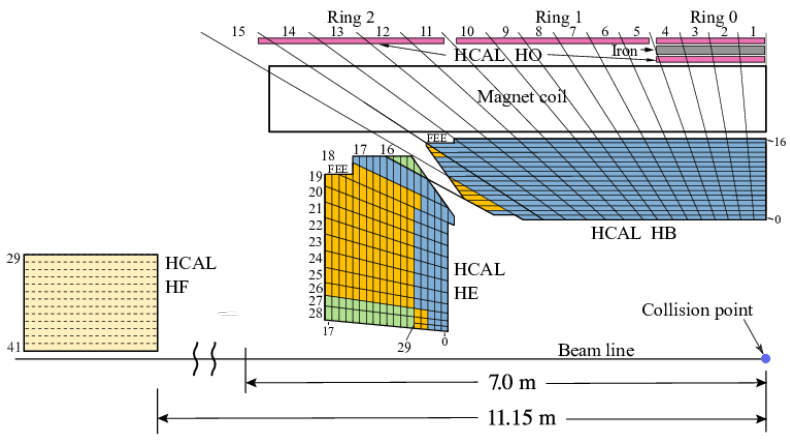


Figure 2.13: Schematic view of the HCAL detector showing its geometry, extracted from [46].

depth to approximately in the central region. It is made of scintillating material and is located outside the magnet. The HCAL Forward (HF) is located at the very forward region, which is situated 11.2 meters away from the interaction point, and expands the HCAL acceptance to the range. The HF are layers made of steel absorber and Cherenkov-emitting quartz fibers for energy measurement. The energy resolution expression for the HCAL is:

— —— (2.9)

Considering there are non-Poissonian events in hadronic cascades, neutrinos or electromagnetic components, the HCAL performance can be limited. However if HCAL and ECAL information is combined to get a better picture of the events, this inconvenience can be addressed. An expression for the HCAL energy resolution taking into account HCAL and ECAL data is shown:

— —— (2.10)

Muon chambers

Many physical processes of interest originated at LHC energies will produce muons. It is therefore vital to measure their energy precisely both for new physics analysis and SM studies. Muons escape the solenoid magnet because they hardly interact with calorimeters and the inactive material. CMS Muon detection [34, 47] is placed outside the solenoid, inside the iron return yoke of the magnet system. Muons trajectory is bent by the solenoid induced magnetic field so their charge and momentum can be deduced. This information is additional to the information recorded by the tracker detector.

The muon detection system is instrumented with three gas ionization devices: the Drift Tube chambers (DT), Resistive Plate Chambers (RPC) and Cathode Strp Chambers

(CSC). Figure 2.14 displays this configuration. The chambers contain gas so when muons go through them it is ionized, creating electric signals. Specialized electronic devices read out these signals, producing hit positions that can be used to reconstruct the muon trajectory. This information is combined for each of the three subsystems and then combined between subsystems to create muon candidates.

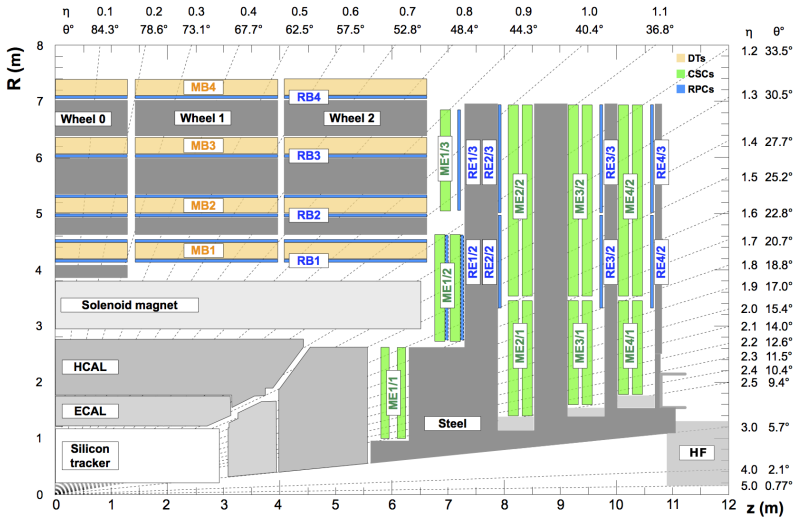


Figure 2.14: Muon chambers scheme, obtained from [47].

DTs are segmented into drift cells that are crossed by an anode wire. A muon passing through the cell creates an ionized cloud of electrons that travel towards the anode wire. The measured drift time is used to obtain the position of the passing muon. Muon chambers are organized in four sections and five wheels placed along the LHC beam pipe, covering the region . DT chambers are composed of rectangular cells with size cm and depth of 2.5 m. They are filled with gas, a CO (15%) and Ar (85%) mix. There are electrode strips on their top and bottom sides to create a constant electric field with a constant drift velocity of drift m/s. The drift cells are also equipped with cathode strips on the sides. Muons passing through the cell ionize the gas mixture and the electrons arising drift towards the wire. The maximum drift time, due to the size of the cells, is almost 400 ns. Knowing this drift time drift, the position of the muon hit can be deduced as drift drift drift. DT layers are placed so that their cells overlap at half their size. This allows to clear the left-right ambiguity with respect to the wire. The detection efficiency for a single cell is 99.8% and is spatial resolution around 180 m, so global position resolution is m.

Cathode Strip Chambers CSCs are arrays of positively-charged anode wires and negatively-charged copper cathode strips placed in the endcaps covering the region . The CSCs have faster response capacity and shorter drift paths to handle the stronger and non-uniform magnetic field present in the endcap region. These characteristics are also beneficial to cope with the higher background rates in this region. Their fine segmentation allows for a three-dimensional spatial resolution. They have trapezoidal shape and are filled with and gas mix. As for DTs, the detections occur

by the ionization of the gas when a muon crosses the chamber. The induced signal can provide a precise position measurement, with a spatial resolution of is .

Resistive Plate Chambers RPCs are constituted by two parallel plates, an anode and a cathode, whose in-between space is filled with a gas mixture. They are located in the barrell and the endcaps, covering the region . RPCs work in avalanche mode. When a muon crosses the chamber, an electromagnetic cascade is produced by the strong electric field inducing a signal that is read out by strips placed in the surface outside. The coarse segmentation in these devices results in a modest spatial resolution ranging from 0.8 to 1.2 cm. The time resolution, 3 ns, is however better than for the DTs and CSCs. They are useful then for measuring the muon track bunch crossing even when PU is high.

2.2.2 Trigger and data acquisition

Although LHC provides large amounts of pp collision data due to its high luminosity, events of interest, both for SM studies or new physics searches, can comprise only a small fraction of it. Storing all information is not possible due to physical limitations of data storage. The 40 MHz bunch crossing rate producing raw events of about 2 MB in size would generate a data rate 80 TB/s for CMS. Since not all data recorded can be saved, there needs to be some criteria for selecting events of interest. This is done by requiring kinematic constraints developed as a result of experience and accumulated knowledge in similar experiments of high energy physics. Most experiments at the LHC do this selective data taking method, with the exception of LHCb that is able to store all effective pp collision data since it runs at a lower luminosity than ATLAS and CMS and is smaller in size. CMS and ATLAS can not afford that due to heavier data, for example, a single event needs approximately 2 MB of disk space while an LHCb event only requires around 30 kB. Large quantities of data are produced at the LHC, there is a proton-proton interaction rate of 40 MHz. Data selection needs to be made online with the data taking, this way all the information collected can be stored permanently. CMS possesses trigger and data acquisition systems (TriDAS) that select events on the fly so the data flow reduces from 40 MHz to 1 kHz. This collection rate, 2 GB/s, 20 PB/year, is manageable to store and analyze offline.

Proton-proton collisions cross section pb is six orders of magnitude larger than cross-sections of most physics processes measured by CMS. Fig. 2.15 shows a summary of these cross-sections. It is then necessary an effective method for selecting events that will likely be a process of interest. To do so Trigger systems work in two steps, first the hardware-based Level 1 Trigger (L1 Trigger) takes action and then the software-based High-Level Trigger (HLT).

Level 1 trigger system

The L1 trigger reduces the 40 MHz readout bunch-crossing rate to a maximum of 100 kHz with a latency of less than 3.4 µs. This time interval is small for L1 to run the whole object reconstruction of an event, so it produces an L1 candidate, an event with low resolution physics objects with only calorimeters and muon systems information as input. The L1 trigger consists of hardware processors that have direct input data from

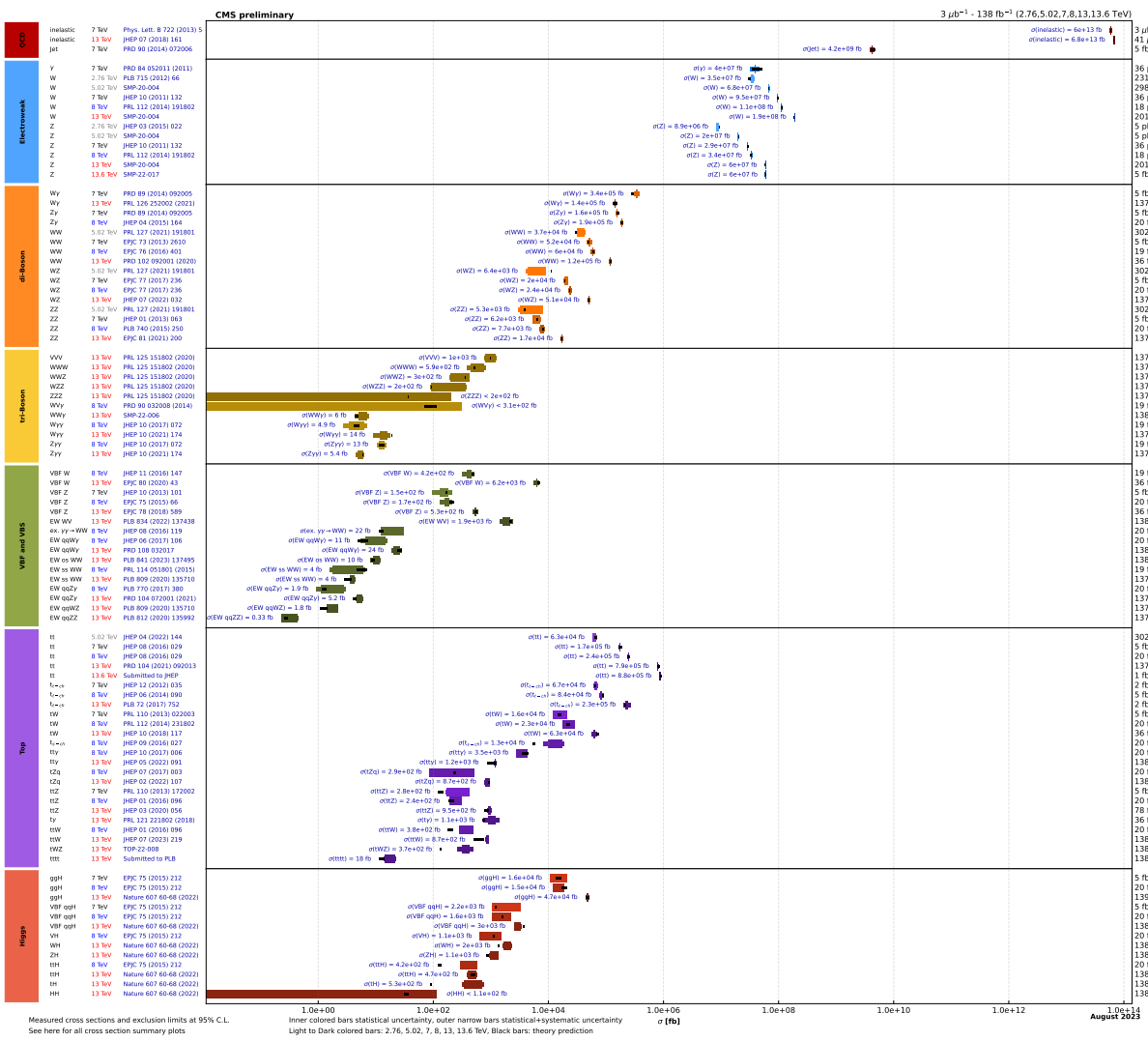


Figure 2.15: CMS cross section measurements summary. Obtained from [48].

several CMS subdetectors. A short processing time is possible by using FPGAs (Field Programmable Gate Arrays) and ASICs (Application Specific Integrated Circuits). During Run 2 phase of LHC, improvements were implemented by including new-generation FPGAs and optical connections.

A variety of objects are reconstructed as L1 candidates such as muons, electrons, photons, tau leptons, jets or energy computations. To do so information from the calorimeter and muon chambers is taken, see Fig. 2.16. This early data has lower quality, it consists of physics objects with irregular segmentation, because it has to be captured in a small period of time. For example, a L1 candidate corresponding to a muon is constructed using information from CSCs and DTs to then form segments and hits from the RPCs. There are tracks finders that reconstruct possible paths for the muon using pattern recognition techniques and compute the muon transverse momentum. The four best options

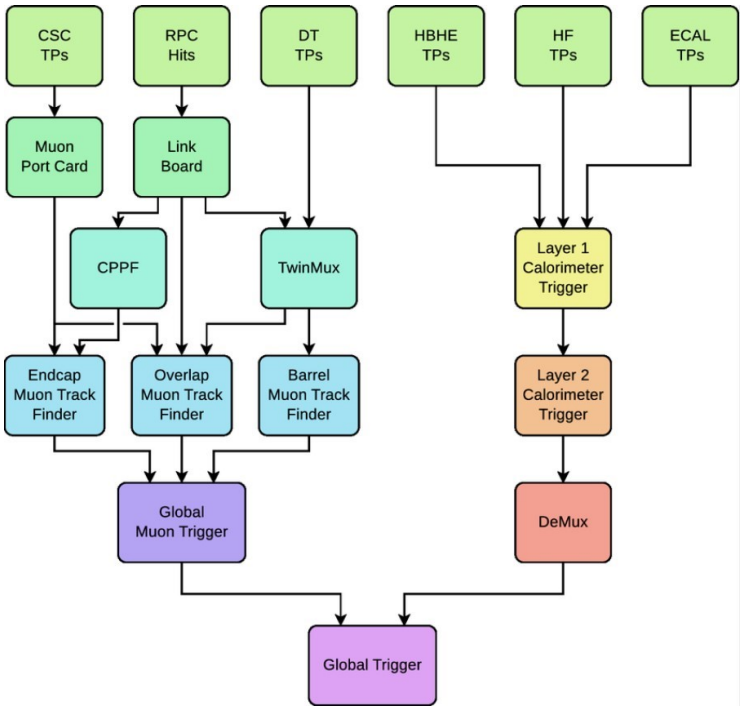


Figure 2.16: Scheme of the L1 trigger processing flow, extracted from [49].

for muon candidates are selected with the Global Muon Trigger and their information is transferred to L1 Global Trigger. This input together with the calorimeter trigger data is combined by the Global Trigger to make a decision on whether to accept or reject the event. The information of accepted events is transferred to the data acquisition (DAQ) system from the front-end electronics for further processing.

Data Acquisition

After an event is selected by the L1 trigger, the CMS DAQ is responsible for gathering its corresponding data from all subdetectors. All the information must be properly assembled so the HLT can access to it. The DAQ system is designed to handle a maximum rate of input data of 100 kHz and an aggregated throughput of 100 GB/s [50]. Data processing happens in a computer farm called Event Filter Farm [51] where three types of applications are run: the Readout Unit (RU), connected to the detector front-end readout, performs a first data reduction, assembling pieces of information from a different subdetectors; the Builder Unit (BU) receives event fragments from the RUs and assembles full events; and the Filter Unit (FU) performs the final events selection. HLT algorithms operate within the FU and function over full granularity event data from all subdetectors, filtering about 1-10% of the events selected by the L1 trigger.

HLT Trigger

The HLT selection is more elaborate, working with events passing L1 trigger requirements. The rate of events is reduced from 100 kHz to 2-3 kHz, a frequency compatible with data storing capabilities. All detector information is used by HLT devices to perform online reconstruction of physics objects using a methodology very close to offline data treatment.

For events to be considered of interest and pass the HLT, different requirements, organised in steps of increasing complexity, need to be satisfied. These then go into a thorough reconstruction combining different objects information from the tracker to come up with a final decision for the event.

For Run 2 two new strategies were introduced in the HLT data treatment process, expanding the phase space for new physics studies. The first technique, denominated data scouting, is based on event-size reduction rather than event filtering. Only physics objects reconstructed by the HLT are saved, a few kBs per event, allowing the record of rates of several kHz without significantly increasing the total HLT bandwidth. The other technique is named parking, it directly records events without any previous reconstruction, so the information consists of raw data only. This method is limited due to DAQ bandwidth and disk storage capabilities.

Data Storage

The size of the output data from the HLT in the LHC Run 2 was about 1-2 GB/s leading to a total data volume of tens of petabytes per year. In order to store, process and analyze such a huge amount of detector data, and an equivalent volume of simulations, CMS uses the Worldwide LHC Computing Grid (WLCG) [52, 53]. The WLCG is a large scale distributed computing infrastructure optimized for data-intensive processing. It currently provides to the LHC experiments over a million CPU cores and two exabyte data archive capacity. The computing resources are distributed in more than 100 institutes around the globe, interconnected with high speed internet networks. The computing sites are organized in four tiers. There is a Tier-0 centre at CERN receiving data directly from the CMS DAQ system. It archives the raw data and distributes it among seven Tier-1 centres. One of them is the Spanish Tier-1 centre at PIC in Barcelona. These centres task is to perform reconstruction calibration and other data-intensive processing. The next level is formed by about 40 Tier-2 centres that produce simulated data samples and end-user data analysis activities. They conduct Monte Carlo simulations or data analysis activities for example. CIEMAT (Madrid, Spain) is a Tier-2 centre. Lastly Tier-3 centres accommodate interactive tools for their corresponding local research groups.

Chapter 3

Physics object reconstruction

The CMS detector integrates data from its various subdetectors to reconstruct a wide range of physics objects. Different kinds of particles have different signatures of detection. Figure 3.1 shows a scheme of how particles behave in each subdetector. Electrons and photons deposit all their energies in the ECAL. In addition, electron tracks can be reconstructed using the hits left in the silicon tracker. Hadrons pass through the ECAL, being only a small amount of their energies absorbed there, the relevant part of it is deposited at the HCAL. As explained in the previous chapter, the ECAL and HCAL energy information is combined to reconstruct the energy of hadrons with better accuracy and resolution. Muons do not leave much of their energy in the ECAL and HCAL, but since they are charged particles they are affected by the solenoid field and also detected by the tracker and the muon chambers. Neutrinos do not interact with the detector components exiting the device leaving no trace. To account for them, their momentum in the transverse plane can be inferred. Given that the total transverse momentum before the collision is zero and it must be conserved after the collision, the total transverse momentum of all produced particles is considered as the momentum of the neutrino (with opposite direction). This quantity is called missing transverse momentum and its module missing transverse energy.

This section describes how CMS reconstructs physics objects, explaining in some detail those that will be later used in the analysis. The physics objects of interest are the primary vertex, leptons (electrons and muons), jets, and missing transverse momentum. The reconstruction is carried out using the Particle Flow (PF) algorithm [54], which leverages the precise tracking system and detailed data from the highly-granulated calorimeters. It aims to reconstruct and identify each individual particle in an event, with an optimized combination of information from the various elements of the CMS detector. Photons are identified as ECAL energy clusters not linked to the extrapolation of any charged-particle trajectory. Electrons are identified as a primary charged-particle track and with many ECAL energy clusters corresponding to this track extrapolation to the ECAL and possible bremsstrahlung photons emitted along the path through the tracker material. Muons are identified as tracks in the central tracker consistent with either a track or several hits in the muon system, and associated with calorimeter deposits compatible with the muon hypothesis. Charged hadrons are identified as charged-particle tracks neither identified as electrons nor as muons. Finally, neutral hadrons are identified as HCAL energy clusters

not linked to any charged-hadron trajectory, or as a combined ECAL and HCAL energy excess to the expected charged-hadron energy deposit.

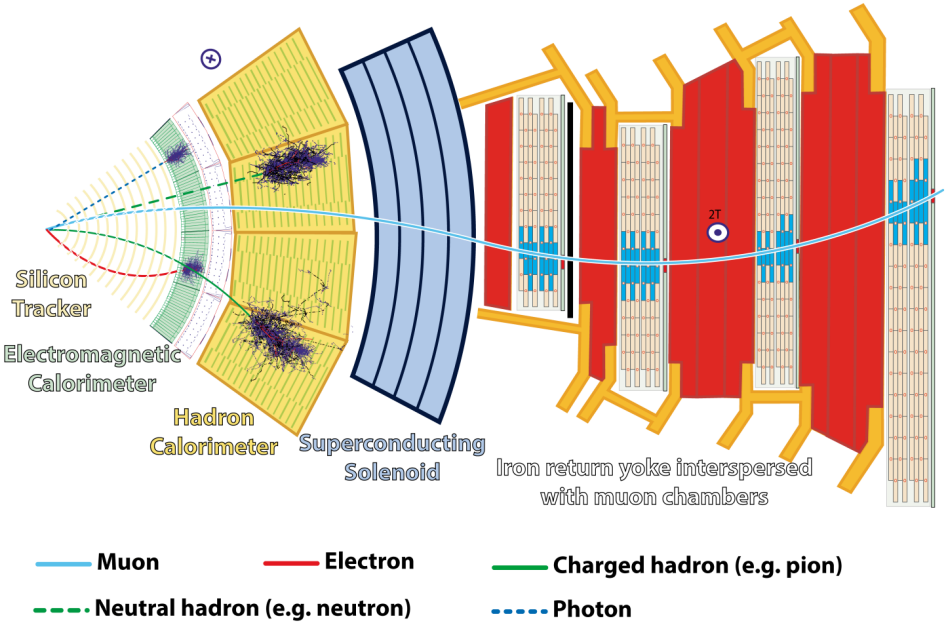


Figure 3.1: Transverse view of the detector sketching the signals left by the different particles. Extracted from [55].

3.1 Primary vertex

Information from multiple proton-proton interactions is recorded for each event. Each of these interactions are associated with a collision vertex. One of these is selected as the primary vertex of the event, associated with the hard interaction. This primary vertex is defined using an algorithm that takes the reconstructed tracks of all charged particles as inputs and asks for a series of conditions. Four or more tracks must be associated with the primary vertex candidate. The longitudinal position of the vertex must satisfy 24 cm along the beam line from the nominal centre of the detector and a radial distance of 2 cm in the transverse plane. The choice of primary vertex, once a series of vertexes are reconstructed, corresponds to the one that has larger quadratic sum of of the physics objects associated to it.

3.2 Electrons

Electrons leave a distinctive signal in the ECAL consisting of an isolated energy deposit also associated with a track detected in the silicon tracker [56, 57]. When electrons interact with the detector material, they may emit bremsstrahlung photons that in turn split into an electron-positron pair. In this case, the final detection signature is not a single path but a

shower of multiple electrons and photons. All these particles must be combined into a single object to compute the whole energy of the original electron. The curvature of the flying electrons in the tracker can be changed by the loss of momentum due to bremsstrahlung radiation. A dedicated tracking algorithm, based on the Gaussian Sum Filter (GSF), is used to estimate the track parameters for electrons [58]. Electron reconstruction in CMS is fully integrated into the PF framework and follows the same principles as other particles.

The first steps of the reconstruction algorithm consist of clustering ECAL crystals measurements with energies over a predefined threshold. The cluster where most of the energy has been deposited in any specific region is defined as the seed cluster. It has to record a transverse energy above 1 GeV. Superclusters (SC) are then arranged in a small window in and an extended window in surrounding the electron path to include photon conversions and bremsstrahlung losses near the seed cluster. Then the PF algorithm uses ECAL clusters, SCs, GSF tracks and generic tracks associated with electrons to form blocks associated with particles. If these objects meet loose selection requirements along with a GSF track, they are labeled as electrons, if not, as photons. Electron identification inside jets is more difficult to accomplish because the energy and position readings from supercluster devices are contaminated by other particle deposits.

Background sources associated with prompt electron identification can come from photon conversions, misidentified hadrons, and semileptonic decays of b or c quarks. Two techniques are used to detect these particles, one consisting of several sequential requirements of the electron characteristics, a cut-based method and the other using a multivariate discriminant where transformations are performed on the input information to construct a classifying model. Electron identification in this analysis is done using the multivariate method [56].

Prompt electrons, originating from the main collision taking place in the event, can be distinguished from those produced through the decay of the hadronized quarks (so-called jets) by looking at their isolation in space. The isolation of a particle is the scalar sum of the momenta of the surrounding particles within a cone of certain radius. The combined PF isolation uses information of the momentum of charged hadrons and energy from photons and neutral hadrons satisfying the condition with respect to the direction of the considered electron. It is defined as:

(3.1)

where is the isolation of charged hadrons, is the isolation of neutral hadrons, is the isolation of photons, and is the pileup (PU) contribution of neutral particles from pileup vertices. This last subtracted term is estimated with the jet area method described in Ref. [59]. Other variables that reveal themselves to be discriminant are the relative combined PF isolation or the hadronic over electromagnetic energy ratio (H/E), the ratio between HCAL deposit and electron energy, since the shower profile from two photons coming from the decay of neutral hadrons inside a jet is expected to be wider than that from a single incident electron. The quantity , defined as the second moment of the log-weighted distribution of crystal energies in , is also used, as well as , combining the SC energy and the track momentum at the point of closest approach

to the vertex and equal to , where is the position of the seed cluster in , and is the track extrapolated from the innermost track position.

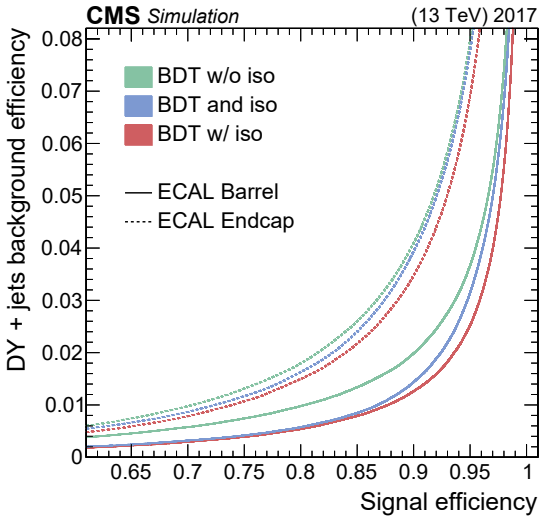


Figure 3.2: Performance of the MVA and cut-based methods for electron identification, differentiating between endcap and barrel detected electrons. Image obtained from [56].

The MVA approach used for electron identification was updated for Run 2 data treatment[56]. It consists of a model based in Boosted Decision Trees (BDT) fed with information such as the shower shape, the track quality, the track-cluster matching or the fraction of momentum lost due to bremsstrahlung. Two sets of models are trained, with or without adding isolation components to the input data. This analysis uses the model containing isolation information. Various trainings are conducted, depending on the electron and . The comparison between the MVA and cut-based methods is shown in Fig. 3.2. Three working points are defined, WP90 and WP80 , associated to a 90 and signal efficiency respectively, and a loose WP, with signal efficiency. The chosen WP for our analysis is WP80, the tightest working point, assuring higher purity for electron identification.

The electron momentum is estimated by combining the energy measurement in the ECAL with the momentum measurement in the tracker. The momentum resolution for electrons with 45 GeV from Z decays ranges from 1.6 to 5%. This variation depends on the electron , being generally better in the barrel region than in the endcaps, and also on the bremsstrahlung energy emitted by the electron as it traverses the material in front of the ECAL.

3.3 Muons

Muons are reconstructed using muon chambers, as well as the inner tracker to measure their momentum [60]. Different types of muons are defined, depending on the detector they are reconstructed in: standalone, global, and tracker muons. Standalone muons are built clustering hits from DTs or CSCs as seeds and gathering CSC, DT, and RPC hits along

the muon trajectory using a Kalman-filter technique. Tracker muons are built matching tracker tracks with GeV and a total momentum GeV to a muon segment of at least one layer of the DTs or CSCs. Finally, global muons are built by matching standalone-muon tracks with tracker tracks with the Kalman filter and checking if the parameters of the two tracks propagated onto a common surface are compatible. This work only uses global muons, taking advantage of the higher purity in reconstruction due to the activation of more than one muon detector plane and the inner track information.

A kink-finding algorithm tries to identify and remove muons from the decay in flight of long-lived particles, such as charged pions or kaons. To do so this algorithm splits the muon track in two separate tracks at several places along its path, each time making a comparison between the two separate parts, requiring that they are compatible with being a single track. Any kinks would be an indication of the decay of charged pions or kaons into a muon and a neutrino. The neutrino escapes the detector and the track of the new muon suffers a sudden change of direction.

Global muons can be classified depending on the selection efficiency or purity required in the analysis, using variables such as the track fit , the number of hits per track or the degree of matching between tracker tracks or standalone-muon tracks:

- *Loose* muon identification only requires the selected candidate to be either a tracker or a global muon. This selection aims to identify prompt muons and those from light and heavy flavour decays while maintaining a low rate of punch-through events which happen when a charged hadron does not interact with calorimeters, reaching muon chambers and being mistaken by a muon.

- *Medium* muon ID consists of a loose muon whose associated track has been reconstructed with hits from more than 80% of the inner tracker layers it traverses. This is optimized for prompt and heavy flavour decay muons.

- *Tight* muon ID requires a muon reconstructed as both tracker and global, with a tracker track that uses hits from at least six layers of the inner tracker including at least one pixel hit and a segment matching in at least two of the muon stations. The global muon fit must have goodness-of-fit per degree of freedom and include at least one hit from the muon system. It must be compatible with the primary vertex, considering impact transverse and longitudinal parameters cm and cm. This aims to suppress muons from decay in flight and from hadronic punch-through.

This analysis relies on a good background reduction from muons produced in the decay in flight of long-lived hadrons. Therefore, the tight muon ID has been used prioritizing signal purity over efficiency.

Similarly as it is done for electrons, the isolation quantity can be used to separate prompt muons from muons inside jets. This variable is defined as the sum of the transverse momenta of the charged hadrons and the transverse energy of the neutral hadrons and

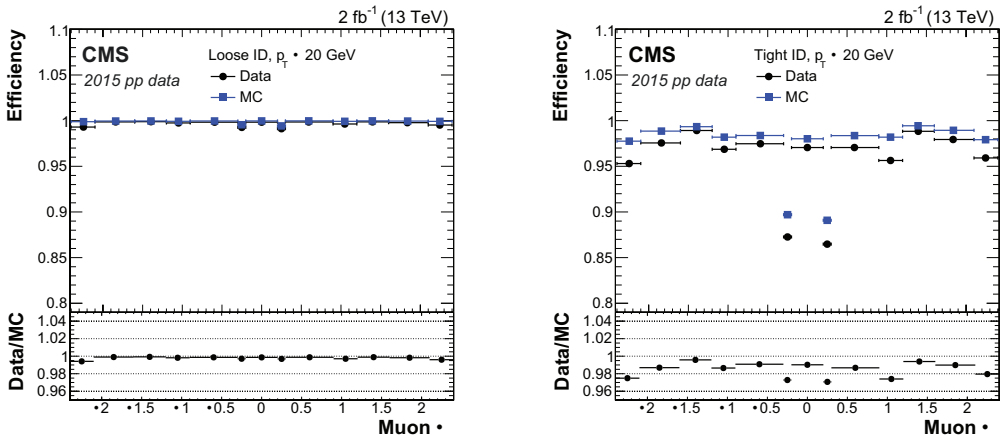


Figure 3.3: Muon identification efficiency for experimental data and simulated muons as a function of the muon pseudorapidity, on the left for the loose WP and on the right for the Tight WP. Extracted from [60].

photons detected around the muon within a cone of $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.4$, expressed as:

$$I_{comb} = \sum_{chhad} p_T + \sum_{nthad} E_T + \sum_{\gamma} E_T - \frac{1}{2}\sum_{PU} pT. \tag{3.2}$$

The last term subtracts the energy originated from PU particles. Muons are considered isolated if the isolation energy relative to the transverse momentum verifies $I_{comb}/p_T^{\mu} < 0.15$.

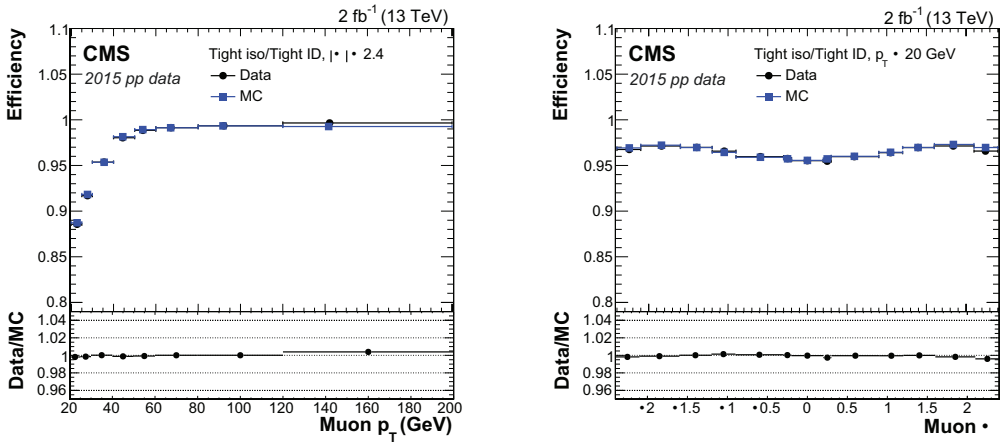


Figure 3.4: Muon tight isolation requirement efficiency for experimental data and simulated muons with Tight ID WP. It is displayed as a function of the muon transverse momentum (left), and muon pseudorapidity (right). Extracted from [60].

The muon transverse momentum resolution reaches 1% in the barrel and 3% in the endcaps for muons with $p_T < 100\,\text{GeV}$ and 7% in the barrel for muons with p_T up to 1 TeV.

3.4 Jets

Jets consists of showers of hadronic particles, originated by the emergence of quarks or gluons in the collisions. CMS reconstructs jets by clustering charged and neutral hadrons using an anti- algorithm [61]. This method groups PF candidates, forming a PF jet with a fixed size parameter determined by distance parameter . This analysis only uses jets constructed with , named AK4 jets.

Jets not arising from the primary vertex, produced in PU collisions, must be identified and removed since they do not come from the event of interest. CMS has developed several mitigation methods [62]. The Charged Hadron Subtraction (CHS) method is the most common used technique, removing all charged constituents associated to a vertex before the jet clustering procedure starts.

In order to reduce noise coming from bad jet reconstruction or instrumental noise of the calorimeters, jet are required to verify some quality criteria [63]. These conditions include thresholds in the fraction of charged and neutral hadrons, electrons, photons and muons or the multiplicity of charged and neutral hadrons. Three working points are defined, Loose and Tight, designed to remove jets originating from calorimeters noise, and Tight lepton veto, that suppresses fake jets from badly reconstructed electron and muon candidates background. Jets used in this analysis must pass the Tight WP.

The reconstructed jet energy is corrected using a factorized model to compensate for the non-linear and non-uniform response of the hadronic calorimeter [64, 65]. The basic corrections remove the pile-up energy (L1) and flatten the jet response versus p and (L2, L3). A residual correction is applied to data to improve the agreement between data and MC distributions. Jet energy corrections are derived from simulation to bring, on average, the measured response of jets to that of truth particle-level jets. In situ measurements of the momentum balance in dijet, multijet, photon+jet and leptonically decaying Z+jet events are used to account for any residual jet energy difference between data and simulation [66].

Since the jet energy resolution is different in the data and the MC simulation, the jet energy in MC is smeared to match the resolution observed in the data. Following the recommendations of the collaboration, the smeared jet 4-momentum is derived by scaling the jet 4-momentum with the resolution scale factor, . This factor is calculated using the hybrid method depending on whether the matched particle-level ( ) jet is found (scaling method) or not (stochastic smearing method). The jet energy resolution uncertainty is estimated by varying the data-to-simulation core resolution scale factor, , used in ——— calculation. If a particle-level jet is not found, stochastic smearing has to be used. In such case, the jet 4-momentum is rescaled with a factor where stands for a random number from a normal distribution with a zero mean and variance . In both cases (scaling method and stochastic smearing), the scaling factor is truncated at zero, i.e. if it is negative, it is set to zero. The jet energy resolution amounts typically to 15-20% at 30 GeV, 10% at 100 GeV, and 5% at 1 TeV [66].

3.4.1 Heavy flavour jet tagging

The CMS collaboration offers different tools for thematic physics studies, constructed by dedicated groups. The BTV group [67] is the bottom tagging and vertexing group, handling jet classification depending on the flavour of the originating quark, vertex fitting and secondary vertex finding within jets. Using the DeepJet algorithm [68], the BTV group provides both a bottom and charm tagging discriminant for jets, along with the necessary corrections to align simulation with data behavior. The DeepJet algorithm employs a deep neural network multi-classification algorithm that considers the full information of all jet constituents, charged and neutral particles, secondary vertices, and global event variables simultaneously. Three different operational working points (WP) are defined, based on the background suppression capability. The so-called Tight, Medium, and Loose WPs suppress the background of light-flavour jets to 0.1%, 1%, and 10% levels, respectively

3.5 Missing transverse momentum

The missing transverse momentum is a relevant piece of information. The momentum of each of the protons taking part in a collision is unknown but their component in the transverse plane, the projection on the plane perpendicular to the beams, can be considered as negligible. The total momentum in the transverse plane in the final state should be also negligible by conservation of momentum then.

The missing transverse momentum is defined as the sum of the transverse momenta of all particles that are reconstructed with the PF algorithm $^{\text{miss}}$. It is modified to account for corrections to the energy scale of the reconstructed jets in the event.

When speaking of $^{\text{miss}}$ we are referring to the missing transverse momentum modulus, which is a measure of the transverse momentum of particles leaving the detector undetected, such as neutrinos, whose presence can be inferred by this energy imbalance. For this quantity to be reliable most particles subject to be measured should be detected. CMS detector covers almost the entirety of the solid angle, nevertheless some tracking inefficiencies or calorimeter noise can affect this extrapolation. Events with large can be caused by the production of invisible particles as well as the detector noise, cosmic rays and detector miscalibration. A beam halo filter and filters against noise are used to avoid these bad occurrences [69].

Regarding missing transverse momentum, a variable of interest is the transverse mass, whose expression is:

(3.3)

where and are the azimuthal angles of the lepton and the $^{\text{miss}}$ vector. Since only the transverse component of the missing energy is considered, we compute the invariant mass with the transverse component of the lepton as well.

Chapter 4

Analysis of W cq production

This chapter details all the elements involved in the analysis of charm production in the decay of W bosons. It begins with a review of the motivation behind the measurement and its development. The following section describes the specific CMS data and simulation samples required to model both signal and background contributions. Moving into the analysis, the baseline selection criteria are defined, yielding a well-understood sample of events mostly composed by $t\bar{t}$ events with leptonic and hadronic W boson decays. The core of the measurement, the charm tagging method, is then introduced. This technique identifies signature muons within jets, indicating the presence of originating heavy-flavor quarks. The associated systematics are effectively controlled, resulting in a robust outcome. In addition, the electric charge correlation in semileptonic $t\bar{t}$ events of the tagged muon and the lepton from the W boson decaying leptonically allows the characterization of most of the backgrounds using a data control region. The chapter concludes with the discussion and evaluation of the systematic effects influencing the predicted yields in the simulations.

4.1 Analysis overview

As stated in the introduction, the motivation of this analysis is probing weak interaction universality in the quark sector. In the SM, the sum of all couplings of any up-type (u, c, t) quark to all down-type (d, s, b) quarks is the same for all three generations. We will explore the equivalence of the couplings of the up and charm quarks studying the hadronic decays of the W boson. A suitable sample for this analysis comes from $t\bar{t}$ production, as W bosons are produced in the decay of the top quark-antiquark pairs. The LHC offers a high cross section of $t\bar{t}$ production resulting in a large sample of W bosons. We will focus on what we denominate semileptonic $t\bar{t}$ events, where one of the W bosons decays leptonically, enabling the use of a lepton trigger, while the other decays hadronically, providing the sample for our study. Figure 4.1 depicts the leading order Feynman diagrams for the semileptonic $t\bar{t}$ process.

The signature features of this process include a lepton from the leptonic W decay, a neutrino, two bottom quarks, and an additional quark-antiquark pair from the hadronic

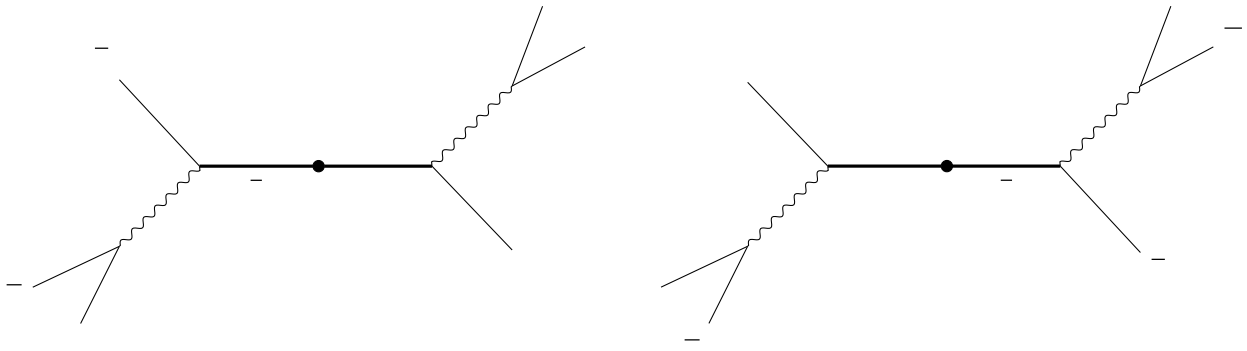


Figure 4.1: $t\bar{t}$ (semi-leptonic) decay diagram, illustrating on the left the case where the positive electric charged W boson decays leptonically and the negative one hadronically and on the right the opposite situation. This two scenarios show that the lepton arising from a W boson will have electric charge of opposite sign with respect to the quark carrying the electric charge sign of the other W boson.

W decay. For the lepton from the leptonic W decay, we will select a high- , isolated electron or muon in the final state, along with significant missing transverse energy to account for the neutrino. The four quarks will hadronize into jets. The two bottom jets will be identified using the standard CMS heavy-flavour identification algorithm. The two additional jets will be associated to the W boson. The proper identification of the four jets involved in the signal process is crucial. QCD radiation in the initial or final state can create additional jets, subject to be mistaken for signal jets. Kinematic constraints on the top and W boson masses will be applied to mitigate this issue. The selection criteria are detailed in Sec. 4.3. The resulting sample, denoted as jets, is mostly composed by semileptonic $t\bar{t}$ events.

The signal signature is also present in other processes, contributing to the background. Sources of background include the production of a Z or a W boson in association with jets (Z jets, W jets), top quark production (single top and dileptonic $t\bar{t}$), and diboson (WW, WZ, and ZZ,collectively denoted as VV) processes. For single top processes, the case where a W boson decays hadronically is treated as signal rather than background.

Once the jets sample is defined, including mostly events with a hadronically decaying W boson, we then apply charm tagging using a technique previously utilized at CMS in various publications [70, 71, 72, 73]. This method involves searching for muons within jets. These muons originate from the decays of charm hadrons within charm jets, with a branching fraction of approximately 9% for muon semileptonic decays of charm hadrons. Semileptonic decays into electrons are not considered because of the high background in identifying electrons inside jets. The muon-based charm tagging method provides a clean selection of charm jets with a low misidentification rate for light-flavour jets, which can be precisely determined from data, as described below.

The modeling of the muon-based charm tagging relies on the accurate simulation of the production of c hadrons from the hadronization of c quarks, and their leptonic and

semileptonic decays with a muon in the final state. As explained in detailed below, the charm fragmentation fractions (FF), defined as the probabilities for charm quarks to hadronize as particular charm hadrons, and the leptonic and semileptonic branching fractions (BFs) of the charm hadrons will be corrected in the simulation to match with measurements. Muons identified within jets are non-isolated, and existing calibration methods in CMS do not account for this type of muons. Therefore, we developed a dedicated calibration methodology, which corrects the muon identification efficiency in the simulation to align with the data.

The charm production process under study can be visualized by substituting $\bar{q}q$ by $\bar{c}q$ in the left diagram in Fig. 4.1, and by $c\bar{q}$ in the right diagram. In this process, the signs of the electric charges of the prompt lepton from the leptonic W decay and the charm quark from the hadronic W decay are opposite. The sign of the electric charge of the charm quark (or antiquark) is kept by the muon stemming from the semileptonic decay of the charm hadron containing the charm quark. We refer to these events as opposite-sign (OS) events, in contrast to the case where the charges of the prompt lepton and the muon in the jet are the same (same-sign events, SS). Signal events are OS, barring misreconstruction issues, while most of the background processes produce symmetric contributions of OS and SS events. Therefore, the subtraction of SS from OS events in the selected sample will heavily suppress the background contribution, while retaining most of the signal events.

4.2 Data and simulated samples

This analysis is performed using a data sample of pp collisions at $\bar{}$ TeV collected by the CMS experiment during the 2016 (36.3), 2017 (41.5), and 2018 (59.7) data-taking periods with a total integrated luminosity of 138 [74, 75, 76]. Given that the process of interest for the analysis (semileptonic $t\bar{t}$) includes a high- isolated electron or muon in the final state, the starting point will be the events corresponding to single electron and single muon triggers datasets. The trigger thresholds varies with the data taking year, being 24, 27, 24 (27, 32, 32) GeV for muon (electron) and 2016, 2017 and 2018, respectively. According to the simulation, a small fraction (about 4%) of the semileptonic $t\bar{t}$ events, where the W boson decays into a tau lepton which subsequently decays into an electron or a muon (plus the corresponding neutrinos) are selected by these triggers and are included in the analysis.

Samples of signal and background events are simulated using MC event generators based on a fixed-order perturbative QCD calculation, supplemented with parton showering and multiparton interactions. Samples of $t\bar{t}$ (with semileptonic, dileptonic and hadronic decays of W bosons) and single top (-, -, and channels) events are generated with POWHEG v2.0 [22] at next-to-leading-order (NLO) accuracy in quantum chromodynamics (QCD). The renormalization and factorization scales are set to the transverse mass of the top quark, where GeV is used. The diboson production is modeled with samples of events generated with PYTHIA v8.2 [21]. MC samples of W jets and Z jets events are produced with the MADGRAPH5_AMC NLO [23] matrix element

generator with up to four noncollinear high transverse momentum partons calculated at QCD leading order accuracy.

The parton distribution functions NNPDF3.1 and NNLO [25] are used. The output of the event generators is combined with the parton shower and hadronization simulation of PYTHIA v8.2 using the underlying event tune CP5 [77]. The jet matching and merging scheme for the MADGRAPH5_AMC NLO samples is MLM[78]. Pileup collisions are overlaid to each simulated event, and the generated distribution of the number of events per bunch crossing is matched to that observed in data. The detector response is simulated using GEANT4 [26]. All the simulated samples generated using Monte Carlo methods for the signal and main background processes are listed in Table 4.1.

Process	Generator (accuracy)	pb
$t\bar{t}$		
Semileptonic	POWHEG (NLO)	365.34
Dileptonic		88.29
Fully hadronic		377.96
Single (Single $\bar{}$)		
W-channel	POWHEG (NLO)	35.84
-channel		80.95 (136.02)
-channel		7.104 (3.549)
Dibosons		
WW	PYTHIAv8 (LO)	75.8
ZZ		12.14
WZ		27.6
W jets		
HT GeV	MADGRAPH5_AMC NLO (NLO)	1292
HT GeV		1627.45
HT GeV		435.237
HT GeV		59.181
HT GeV		14.5805
HT GeV		6.656
HT GeV		1.6081
HT GeV		0.0389
Z jets		
HT GeV	MADGRAPH5_AMC NLO (NLO)	208.977
HT GeV		181.302
HT GeV		50.4177
HT GeV		6.9839
HT GeV		1.6814
HT GeV		0.7754
HT GeV		0.1862
HT GeV		0.00438

Table 4.1: Monte carlo simulations used in the analysis. HT stands for the scalar sum of jets transverse momenta.

4.3 Baseline selection

In this section we describe the selection requirements of the baseline ℓ+jets sample. The signature features of these events are a high-p_T isolated lepton, muon or electron and four high-p_T jets, two of them identified as b-jets.

4.3.1 Basic selection

Events are required to have exactly one high transverse momentum isolated lepton, either a muon or an electron. Muons (Electrons) are required to have transverse momentum p_T^ℓ >30 (35) GeV, tighter than the trigger p_T condition, $|\eta|$ <2.4 (2.4) and tight ID (tight MVA). In addition, muons must satisfy the relative isolation condition I_{comb}/p_T^ℓ <0.15. In the case of electrons, there is no explicit isolation requirement since the isolation variable is included in the multivariate identification algorithm.

Figure 4.2 shows the p_T distribution at the generator level of the charged lepton from the decay of the W boson in semileptonic $t\bar{t}$ events. The efficiency of the lepton p_T requirement is about 2/3. The trigger p_T thresholds remove leptons with p_T lower than 24-32 GeV (depending on the lepton type and data taking year), and the offline selection requirement, p_T^ℓ >30 (35) GeV, higher than the trigger p_T condition, ensures that in that range the trigger efficiency is high and well understood.

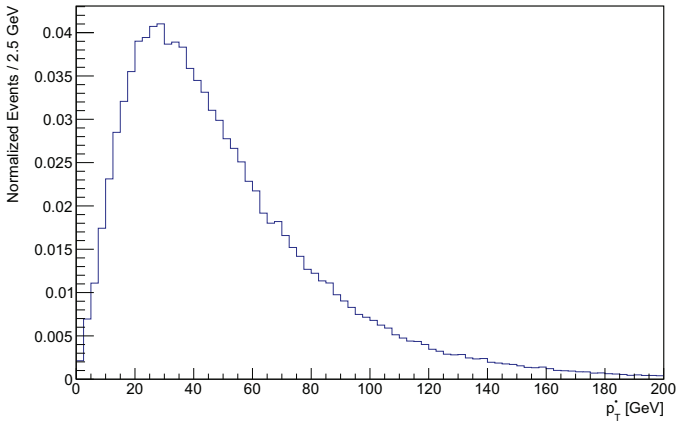


Figure 4.2: Distribution at the generator level of the charged lepton p_T produced in the decay of the W boson in semileptonic $t\bar{t}$ events.

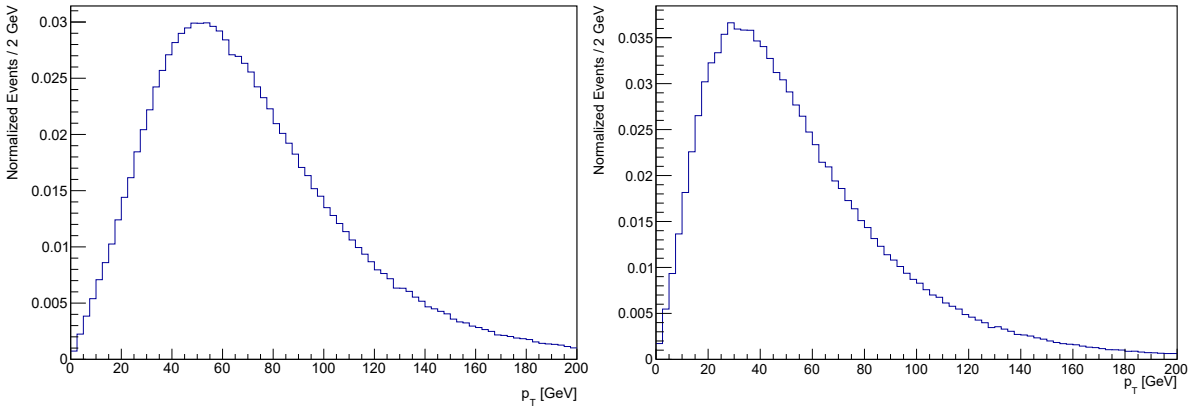
Events with a second isolated and Tight muon (electron) candidate with $p_T^\ell > 20$ GeV are removed. This reduces the background of $Z \to \mu^+\mu^-$ ($Z \to e^+e^-$) and dileptonic $t\bar{t}$ contamination. These processes include a high-p_T isolated lepton just as the signal, but also a second lepton of the same characteristics that the signal lacks. Therefore, the veto of a second lepton suppresses those backgrounds.

Since the process of interest includes a W boson decaying leptonically, producing a neutrino in its decay, events are required to have missing transverse momentum p_T^{miss} >20

GeV and transverse mass >20 GeV as well. The transverse mass is computed using Eq. 3.3, reconstructing the W boson as it did not have longitudinal motion.

The signal process exhibits four high- quarks in the final state. For this reason, the basic selection requires the presence of at least four jets with a minimum transverse momentum >25 GeV. The minimum jet threshold is dictated by jet reconstruction considerations. Figure 4.3a shows the distribution of b quarks at the generator level in $t\bar{t}$ events. As the figure illustrates, lowering the threshold would not benefit the selection, as the number of events decreases rapidly. While the jet requirement applied to b-jets removes only a small fraction of the signal events, it is less efficient for W jets, removing about 25% of the events, as shown in Fig. 4.3b.

In addition to the requirement, jets are required to be within the detector acceptance, <2.4 and well separated from the high- isolated lepton, verifying the angular condition (jet,)>0.4. These conditions account for the quarks characteristic of the signal, Fig. 4.1, being detected as jets.



(a) distribution for b quarks. **(b)** distribution for W quarks.

Figure 4.3: (a) Distribution at the generator level of the of b quarks produced in semileptonic $t\bar{t}$ events. (b) Distribution at the generator level of the of quarks produced in the decay of the W boson in semileptonic $t\bar{t}$ events.

These requirements conform the first steps of the jets baseline selection. Table 4.2 summarizes them, also including bottom tagging requirements for jets, that will be detailed in the next section. Depending on the flavour of the selected high- lepton, two main channels can be defined, muon W and electron W channels. The leptonic selection differs, but the missing transverse momentum and jet requirements are the same for both channels.

Channel	W			W		
Year	2016	2017	2018	2016	2017	2018
Lepton (GeV)	>30			>35		
Lepton | |	<2.4			<2.4		
Lepton isolation	<0.15			-		
Lepton ID	Tight ID			Tight MVA		
Number of leptons	1					
Extra leptons veto	Discard tight leptons with GeV					
(GeV)	>20					
W transverse mass (GeV)	>20					
Jet (GeV)	>25					
Jet | |	<2.4					
(jet,)	>0.4					
#Jets						
B-tagged jets	2 Medium WP					

Table 4.2: Summary for the selection requirements differentiating between muon and electron channel and year when appropriate.

4.3.2 Bottom quark tagging for jets

As mentioned before, two of the quarks produced in a semileptonic $t\bar{t}$ event are bottom quarks, see Fig. 4.1, creating jets susceptible to be detected as bottom flavoured jets. Tagging two of the reconstructed jets as b-jets will enable further reduction of background contamination while preserving signal events. To achieve this, we employ bottom tagging techniques provided by the CMS BTV group, as described in subsection 3.4.1. We use the DeepJet tool [68], which calculates a discriminating variable for each jet using a neural network model. This variable ranges from 0 to 1, where 0 indicates the lowest likelihood of the jet originating from a bottom quark, and 1 indicates the highest likelihood.

Once the jets are selected according to the criteria in Table 4.2, they are ranked by their bottom tagging score. The two jets with the highest scores, referred to as the b1-jet and b2-jet, are chosen as b-jets. A medium WP requirement is then applied to the b-tagging discriminant of these two b-jets to filter the events. These b-tagging requirements remove about 50% of the signal semileptonic $t\bar{t}$ events.

Along with the discriminant, the BTV group also provides corrections for the b-tagging and mistagging efficiencies in the simulation to make up for discrepancies between data and MC simulation for the b-tagging discriminant distribution of jets. These so-called scale factors (SF) depend on the jet flavour, the and of the jets, and the b-tagging WP being used. Corrections for the baseline selection are discussed in section 4.3.4 but those regarding b-tagging will be explained here since they are more intricate.

A general method for correcting the selection efficiency in the simulation after applying a b-tagging WP to jets is the computation of an event weight taking the following expression:

———— (4.1)

where the numerator an denominator are defined as:

not not

(4.2)

not not

.

The weight 4.1 is applied event by event in the simulation to match the b-tagging efficiency in the data. The formula in Eq. 4.2 is the generalized method for the case of using more than one b-tagging WP. It is a series of products over all jets considered depending on whether they meet the medium WP ( ), meet the loose WP but not medium (not), or do not meet the loose WP (not). The terms correspond to the scale factors for medium(loose) WP. The terms are the efficiencies with which jets in our selection tag the medium(loose) WP of the discriminator. These efficiencies are computed for simulated jets in our own selected sample as a function of the jets and distinguishing whether the jet contains bottom, charm or light flavour. As an example, Fig. 4.4 shows the b-tagging efficiency plots for simulated jets in year 2018.

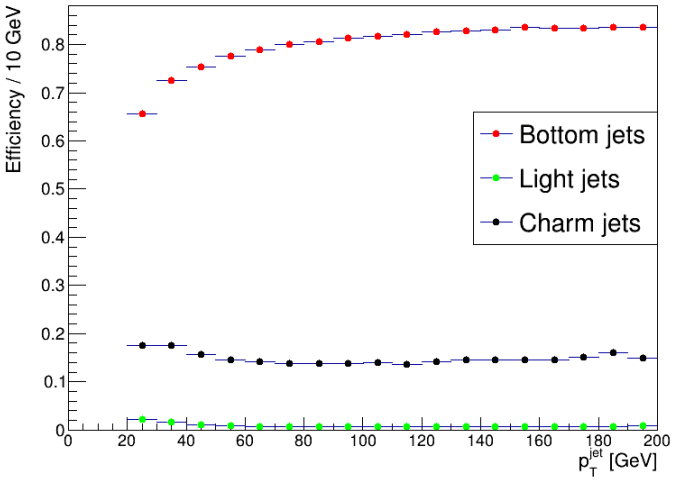


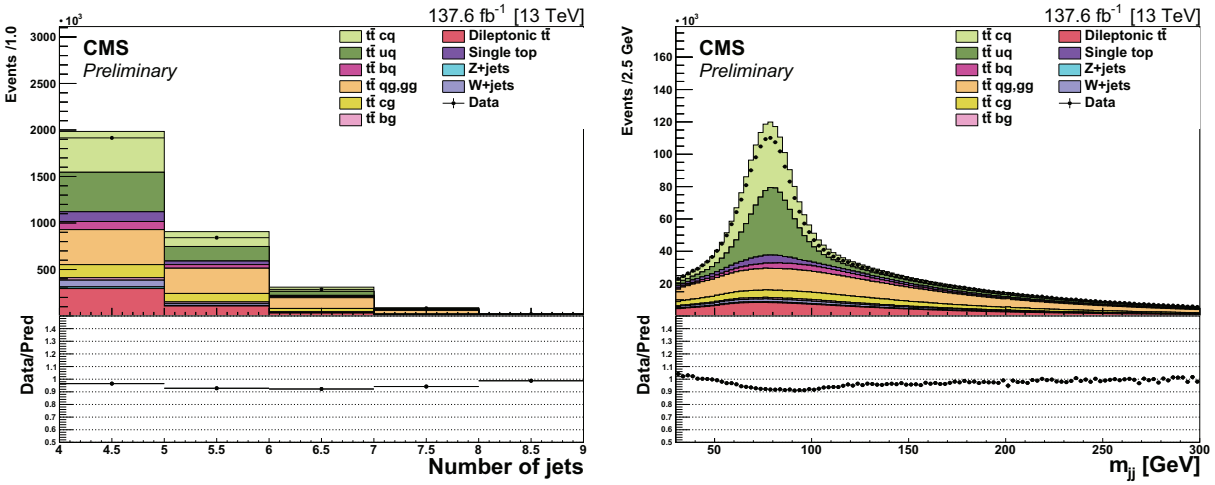
Figure 4.4: Efficiency of tagging a jet as originated by a bottom quark with a medium WP requirement, computed using the 2018 semileptonic $t\bar{t}$ simulation and the selection criteria outlined in the text. This efficiency is plotted as a function of the transverse momentum of the considered jet, with most jets in the analysis falling within the range of 25 to 80 GeV.

The implementation of the b-tagging events weights (Eq. 4.1) for MC events results in a total yield change of about 5%.

4.3.3 Kinematic requirements

In our analysis, accurately identifying the two jets originating from the hadronic decay of the W boson, among the multiple jets in each event, is crucial. The two jets with the highest p_T, excluding the two b-tagged jets, are associated with the hadronic W boson. Figure 4.5a displays the jet multiplicity in the event sample selected following the requirements discussed so far and outlined in Table 4.2. It can be seen that the fraction of events with more than 4 jets is large. In addition to the two b-jets and the two jets from the W boson decay, events can contain additional jets, mainly from the hadronization of gluon radiation.

The wrong assignment of the two W-jets leads to undesirable features in the dijet mass distribution, m_{jj}, displayed in Fig. 4.5b. The contributions in the plot are separated by process and true parton flavour of the two jets associated with the W boson. This distribution shows a large pedestal and a long tail of wrong two-jet combinations below the peaking distribution around the W mass corresponding to the correct W jet assignments. For semileptonic $t\bar{t}$ events, the contributions of correctly associated jets to the hadronic W boson are displayed in light green color (when the W boson decays to a charm quark) and dark green (when the W boson decays to a pair of light quarks). These events constitute the desired signal events, presenting a peaking distribution. The long tail mainly corresponds to combinatorial background of semileptonic $t\bar{t}$ events, and dileptonic $t\bar{t}$ production, where gluon or bottom jets are associated to the W boson.



(a) (b)

Figure 4.5: (a) Jet multiplicity distribution for data and simulation for the selected events. (b) Invariant mass distribution built from the two jets with the highest p_T, excluding the b-tagged jets. Events are classified according to the parton flavor of the two jets associated with the W boson (q=light quark, c=charm, b=bottom, g=gluon).

In order to improve the jet classification, some kinematic requirements are imposed, involving the following observables: the invariant mass of the high-p_T isolated lepton and one of the b-jets, named $m_{\ell b}$, the already mentioned invariant mass of the dijet reconstructing

the hadronic W boson, , and the invariant mass of this dijet and the other b-jet not chosen for the computation. The latter quantity, denoted as , reconstructs the mass of the top quark decaying to the hadronic W boson.

Figure 4.6 represents the bivariate distribution of and for the selected semileptonic $t\bar{t}$ simulated sample. It can be modeled with a normal bivariate distribution with parameters corresponding to the center and width of the observed distributions in the simulation. The corresponding expression is defined as:

$\mathcal{N}$ where (4.3)

The variances , expected values and correlation are the observed values for the simulation. Their resulting values are GeV, GeV, GeV and GeV and .

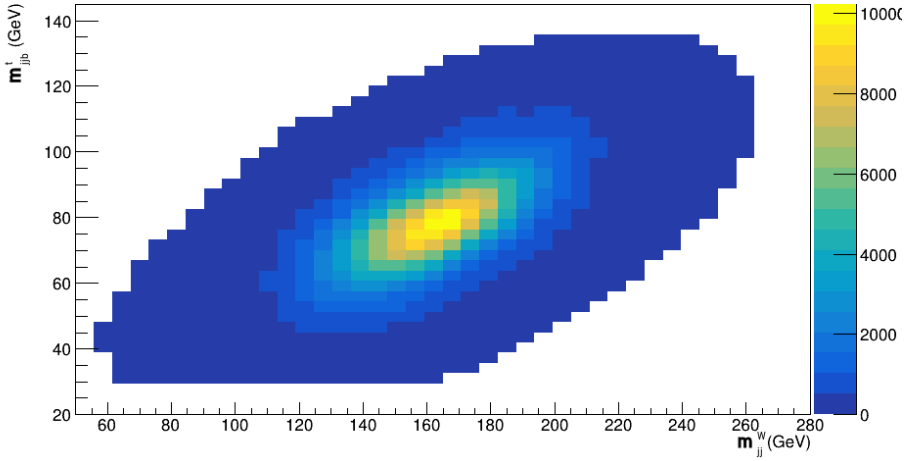


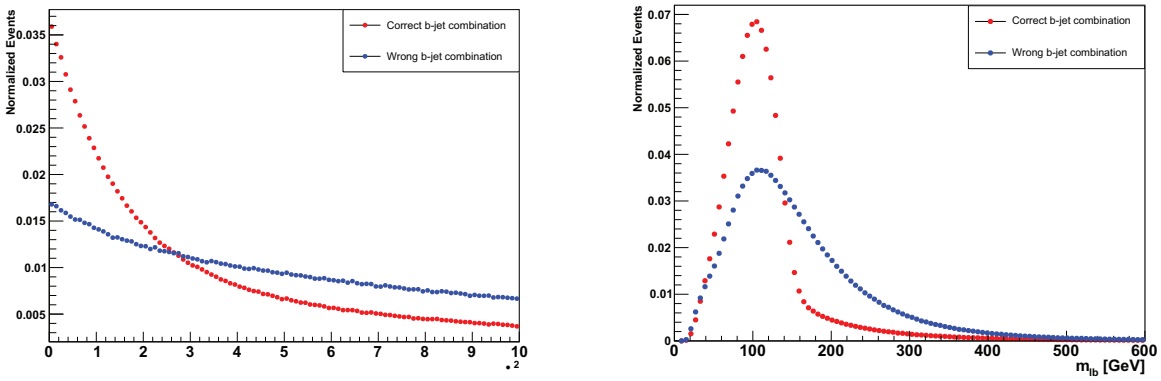
Figure 4.6: and joined distribution.

The compatibility of the observed values of and in data events with the hypothesis of Eq. 4.3 can be tested computing the associated value, expressed in Eq. 4.4, where and denote the reconstructed W and top quark masses in the data events.

(4.4)

Since there are two b-jets, b1-jet and b2-jet, one can compute two possibilities for and , using (,) or (,), ideally, one of the options corresponding to pairing the lepton with the b-jet arising from the same top quark and the dijet with the

other b-jet, arising from the other top quark. Figure 4.7 shows the distributions of χ^2 and $m_{\ell b}$ for correct and wrong b-jet combinations obtained with the semileptonic $t\bar{t}$ simulation. There are clear differences in shape. For $m_{\ell b}$, the correct combination distribution exhibits a sharp kinematic threshold at around 150 GeV. The χ^2 distribution is flatter for the wrong b-jet combination, while the b-jet correct combinations tends to score lower values.



(a) **(b)**

Figure 4.7: χ^2 (a) and $m_{\ell b}$ (b) normalized distributions for MC simulated events. The red and blue dots represent the correct and wrong b jet combinations. Distributions are normalized to area 1 so the comparison is to be made between shapes.

Considering these distributions, the following requirements are applied: it must be satisfied that $m_{lb} \leq 150$ GeV and $\chi^2 \leq 3.2$. If none of the b-jet combinations satisfy these conditions the event is discarded. If one of them satisfies them, the event is retained and that combination is chosen as the correct one. If both b-jet combinations satisfy the conditions, the event is also retained and the b-jet pairing with lower χ^2 is chosen as the correct one. Figure 4.8 represents this process.

These kinematic requirements complete the ℓ + jets selection. As shown in Table 4.3, over 0.9 million events are selected with the criteria described above. According to the simulation, the data sample is composed of 45.3% of semileptonic $t\bar{t}$ events with W $\rightarrow$ cq, 45.7% of semileptonic $t\bar{t}$ with W $\rightarrow$ uq, 5.6% of dileptonic $t\bar{t}$, 2.3% of single top, and 1% of V+jets. The VV contribution is negligible.

Table 4.4 shows the remaining fraction of selected events, according to the semileptonic $t\bar{t}$ simulation, after sequentially applying every selection requirement. The single electron/muon trigger only retains about one third of the total number of events, removing most of the $W \rightarrow \tau\nu$ events, and keeping about half of the $W \rightarrow \mu\nu$ and $W \rightarrow e\nu$ events. The increased requirement for the p_T of the lepton and the presence of at least 4 high-p_T jets further reduces the sample to a half compared to the previous step. Requirements related to the presence of a neutrino keep about 75% of the selected events. The b-tagging requirements further reduce the sample by half, while the kinematic requirements select one third of the remaining events, leaving in the end about 2.2% of the original semileptonic $t\bar{t}$ events.

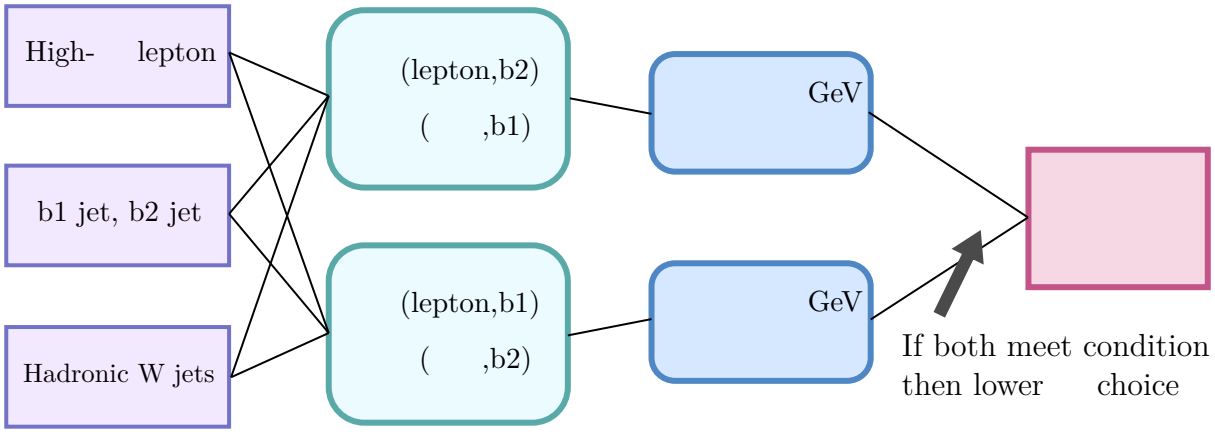


Figure 4.8: Process for choosing kinematic magnitudes for each event. Both combinations (,) and (,) are computed, then the cuts and GeV are applied to both combinations in order to reject or accept the event. If both combinations satisfy the conditions the one with lower is chosen as the correct one.

Process	Rate
$t\bar{t}$, W uq	45,7%
$t\bar{t}$, W cq	45,3%
Dileptonic $t\bar{t}$	5,6%
Single top, W cq	0,9%
Single top, W uq	0,8%
Single top, W channel	0,6%
	1,0%
Diboson	0,0%
Data	916680

Table 4.3: Number of selected events in data after the full selection and process composition according to the simulation.

	Cumulative efficiency	Relative efficiency
Single muon/electron trigger	35.6%	35.6%
1 high- muon/electron and 4 jets	17.5%	49.2%
and	12.8%	73.1%
2 b-tagged jets medium WP	6.0%	46.9%
Kinematic reconstruction	2.2%	36.7%

Table 4.4: Cumulative and relative efficiency of the selection requirements performed sequentially, according to the semileptonic $t\bar{t}$ simulation.

4.3.4 Simulation corrections

The MC simulation is not always able to emulate real data behaviour and the detector response perfectly, so corrections need to be applied to the simulated samples in order to better match the observed data. Corrections might change the global normalization and/or the shape of the distributions. Efficiency scale factors (SFs) are typically required to correct differences in reconstruction and identification efficiencies between data and simulation. Likewise, the calibration and the resolution of the measurement of certain observables sometimes need to be matched. Section 4.3.2 already covered the weight computation for b-tag scale factors, while jet corrections—commonly applied in any CMS analysis—were explained in Section 3.4 on object reconstruction. The remaining corrections will be detailed in this section.

Muon corrections

Regarding the trigger selection and the identification of the high- isolated muon, we need to apply scale factors accounting for differences in reconstruction efficiencies between data and MC. The total efficiency for muons is given by the expression in Eq. 4.5, which incorporates the trigger requirements, the identification efficiency of the used WP (Tight) and the efficiency of the isolation condition:

$$\text{Trigger} \quad \text{ID} \quad \text{ISO} \tag{4.5}$$

We use the SFs provided by the CMS Muon Physics Object Group (MUON POG), computed using clean samples of dimuon events in the Z mass peak with the tag-and-probe method [79]. The SFs depend on the muon and .

The impact on the global yield of our analysis is a reduction in the predicted rate by the simulation of 3% for 2016, and 2% for 2017 and 2018 due to the SFs applied for the trigger. The identification SF contributes approximately 1% in each case, while the effect of the isolation SF is minimal, being less than 1%.

Electron corrections

The online selection and identification of high- electrons also requires the use of SFs to reweight MC events to better match the observed data. The expression for electron efficiency is given in Eq. 4.6. The isolation requirement is applied by the identification algorithm, so no separate term is included.

$$\text{Trigger} \quad \text{ID} \tag{4.6}$$

The CMS EGM POG [80] centrally provides electron SFs calculated using dielectron samples in the Z boson mass peak with the tag-and-probe method. For both the Tight MVA ID and the trigger, the associated SFs depend on the electron and .

The application of the identification SFs results in a reduction in the global event yield of the simulated electron channel samples by 3% for 2016 and 4% for 2017 and 2018. The

reduction caused by the application of the trigger SFs is smaller, less than 1% for 2016, and 2% for 2017 and 2018.

L1 trigger pre-firing correction

The shape of ECAL pulses gradually shifted during operations in 2016 and 2017 [49]. This phenomenon could be accounted for by an increasing offset in the timing calibration of the pulses related the transparency loss of the ECAL crystals, mostly in the endcap. This offset was balanced out recalibrating the pulses, but this was not done for the ECAL trigger primitives. Ultimately, the endcap pulses were moved to a time region where the bunch crossing assignment was affected, causing the so-called Level-1 system prefiring, i.e. accepting collisions earlier in time than the collision of interest. Since Level-1 trigger does not allow for two consecutive bunch crossings to fire, the event of interest is actually lost. Simulations do not consider this effect, so corrections need to be applied additionally. The correction consists of event weights considering the probability of the event to prefire depending on the and of its forward jets and photons. The final weight is the product of all non prefiring probabilities of jets and photons present in the event:

pref (4.7)

photons, jets

In the 2018 simulation, events were adjusted to account for this effect by fixing endcap timing delays in the ECAL front-end, optimizing the pulse synchronization. The effect results in a 1% reduction of simulated event yield for the affected years.

Top quark reweighting

The modeling of top quarks in the simulated samples of $t\bar{t}$ events is not correct. The spectrum in data was significantly softer than the predicted by the various MC simulations based on either LO or NLO matrix elements interfaced with parton showers. There is a notion that it can be partially due to missing higher-order contributions. This acknowledged in the literature, for example, an approximate NNLO prediction by Kidonakis [81] is able to improve the level of agreement. Predictions by Mitov et al. at NNLO+NNLL [82] and later at NNLO+NLO EW [83] as well as at NNLO by Grazzini et al. [84] have provided a much-improved description, although some discrepancies still remain.

The TOP Physics Analysis Group in CMS has specific recommendations for how to tackle this problem. Based on most recent NNLO (+ NLO EW) calculations, corrections for top quarks momentum spectra in $t\bar{t}$ MC simulations are provided for different specific situations. The case of this analysis involves the computation of the following weight to be applied to each event, it is expressed as:

(4.8)

where and are the transverse momenta of the top and antitop particles generated in the simulation.

PU reweighting

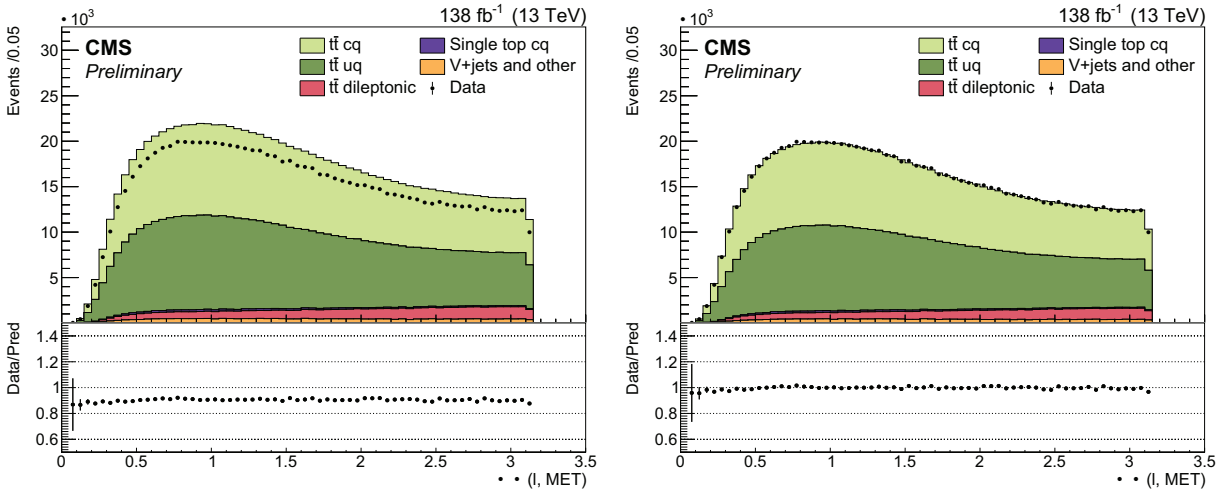
As explained in Sec. 2.1.1, events consists of a collision with a hard interaction together with multiple soft PU collisions. This effect needs to be correctly modeled, necessitating the inclusion of additional proton-proton interactions in the simulation. The probability distribution for the number of PU interactions may differ between data and simulation. Weights are applied to MC simulation, computed as the ratio between PU distributions of data and MC. This distribution is modeled before each data taking period and the expression for data is represented as:

$$\mathrm{PU} = \frac{L_{\mathrm{ins}} \cdot \sigma_{\mathrm{inel}}}{\mathrm{f}_{\mathrm{rev}}}, \tag{4.9}$$

where σ_{inel} is the total pp inelastic cross section (6.2 fb), L_{ins} is the instantaneous luminosity and $\mathrm{f}_{\mathrm{rev}}$ is the LHC orbit frequency (11246 Hz). This correction increases the global yield in the simulation relative to data in less than 1%.

Global normalisation factor

After applying corrections, a global normalization mismatch between data and MC is observed. To align the normalization of the simulation with data, the following global scale factors were applied: 0.92 for the muon channel and 0.90 for the electron channel. A similar mismatch in the data/MC normalization was found in the CMS publications of the measurements of the $\mathrm{t\bar{t}}$ production cross section and the top quark mass using the same ℓ + jets final state [85, 86]. Figure 4.9 illustrates this effect, which impacts the yield globally without altering the shape of the distributions. It is worth noting that the global normalization of the simulation is irrelevant for the measurement conducted in this thesis. This correction is made solely for presentation purposes.



(a) $\Delta\phi$ before adjusting global normalization. **(b)** $\Delta\phi$ after adjusting global normalization.

Figure 4.9: Illustration of the effect of the adjustment of the global normalization in the simulation: difference in azymuthal angle ($\Delta\phi$) between the missing energy transverse momentum and the high-p_T isolated lepton before and after applying the global normalisation correction.

4.4 Baseline sample kinematic distributions

The distributions for data and simulation of some observables of interest for the baseline jets sample are shown in Figs. 4.10, 4.11, 4.12, 4.13. The displayed distributions are related to the prompt lepton, the b-tagged jets, the jets associated to the W boson decaying hadronically, and some other interesting observables like the dijet invariant mass of the reconstructed hadronically-decaying W boson, the invariant mass of the three jets reconstructing the top quark, and the invariant mass of the lepton and the b-tagged jet corresponding to the W boson decaying leptonically. All corrections described in Sec. 4.3.4 are applied to simulated events. The data corresponds to the whole data taking period 2016-2028 (138 fb). The prompt muon and electron channels are added together. The overall trend shows excellent agreement between data and simulation, as evidenced by the distribution plots and the data-to-simulation ratio plots displayed in the lower panels of the figures.

Figure 4.14 displays the flavour of the four jets in the selection. CMS utilises an algorithm [87] to determine in the simulation the flavour of the quark that has originated a jet. It groups selected jet constituents that will most likely reveal its properties. If there is at least one matching b parton for the jet, it is classified as bottom-flavored (label = 5); if there are no matching b partons but at least one c parton matches, the jet is classified as charm-flavored (label = 4); if there are no matching b or c partons, but light partons do match the jet, it is classified as light-flavored and the label assigned corresponds to the hardest light-flavour parton clustered inside the jet (label=1, 2, 3, or 21, being 1 down, 2 up, 3 strange, or 21 gluon). From the upper plots in Fig. 4.14, it can be seen that b-jets are correctly tagged with a remarkably high efficiency of 98%. The small misidentification rate corresponds to mistagging c quarks and gluons as b jets. The lower plots in Fig. 4.14 classify the two jets associated to the W boson decaying hadronically according to their parton flavour. Events are separated according to the production process. The W cq production is dominated by W cs with a small contribution of W cd, as expected by the parameters of the CKM matrix. Likewise, the W uq contribution is dominated by W ud with a small contribution of W us. The two jets of the W boson are correctly assigned in close to 90% of the semileptonic $t\bar{t}$ events. The misidentification rate is dominated by gluon jets.

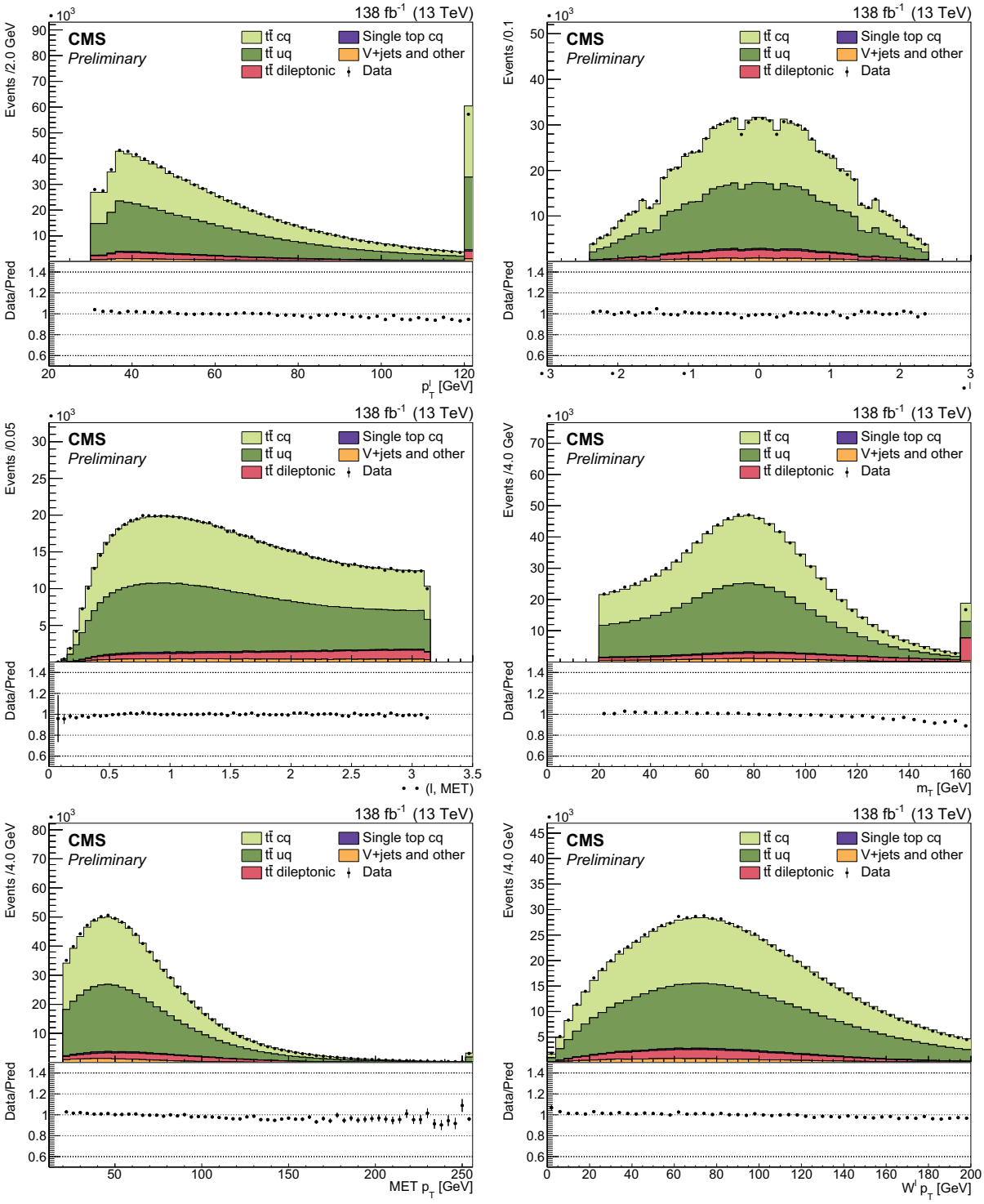


Figure 4.10: Distributions for the high-p_T isolated lepton. The top left image depicts the transverse momentum of the prompt lepton and the top right image its pseudorapidity. The center left image is the azymuthal angle difference between the prompt lepton and the missing transverse momentum. The center right image shows the leptonic W transverse mass. Bottom left image is the missing transverse momentum and bottom right image displays the leptonic W transverse momentum.

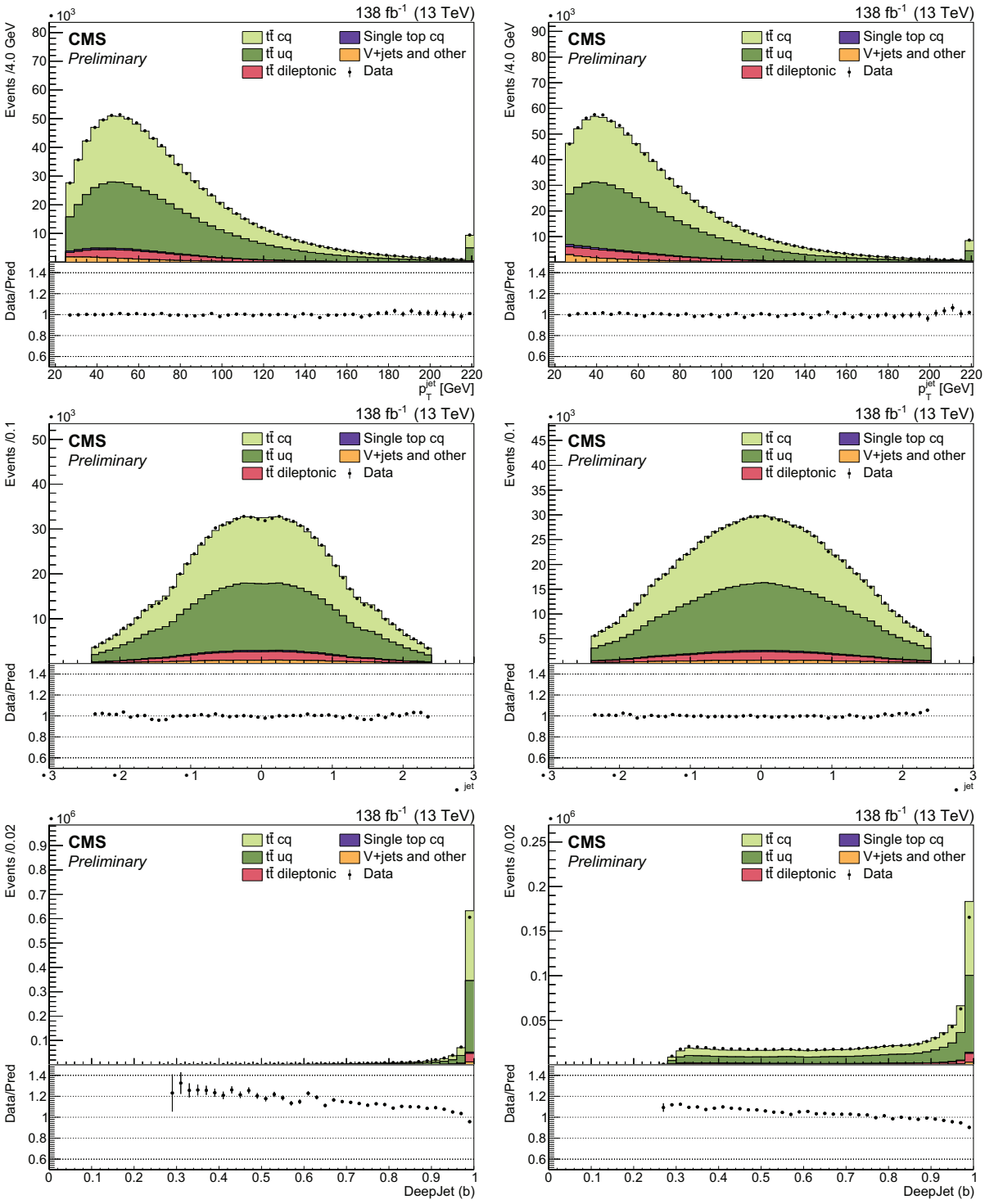


Figure 4.11: B-tagged jet distributions. The left column corresponds to b1-jet, that with the highest b-tagging discriminant, and the right column to the other, b2-jet. The distributions are, from top to bottom, p_T, η and b-tagging discriminant score.

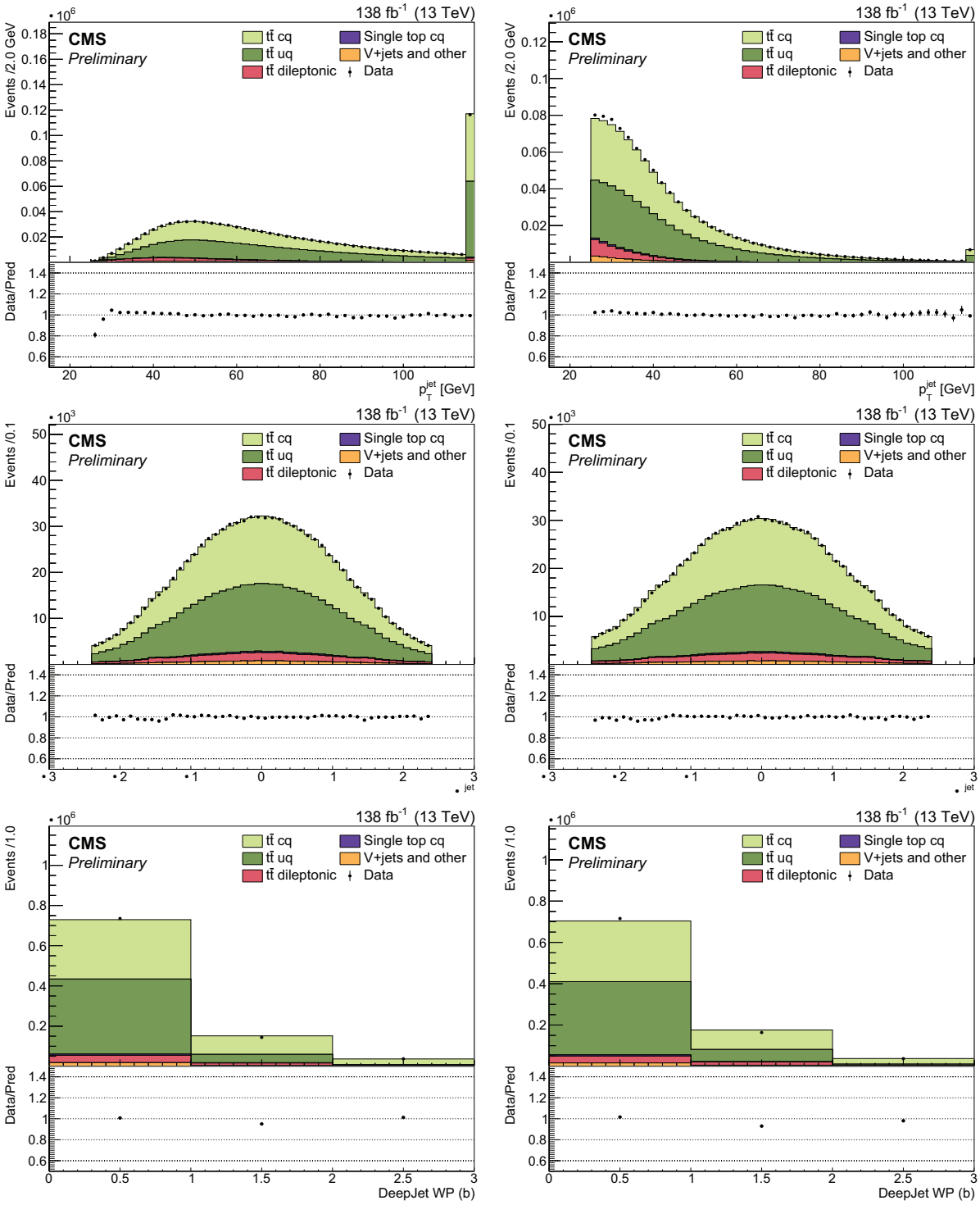


Figure 4.12: Distributions of the two jets associated to the W boson decaying hadronically. The distributions of the left column correspond to the leading-p_T jet, W jet 1, and the right column to the subleading, W jet 2. The distributions are, from top to bottom, the transverse momentum, the pseudorapidity and the b-tag discriminant binned in working points (the first bin corresponds to jets not passing the loose or medium WPs, the center bin for jets satisfying the loose WP but not medium, and the last bin for jets satisfying the medium WP).

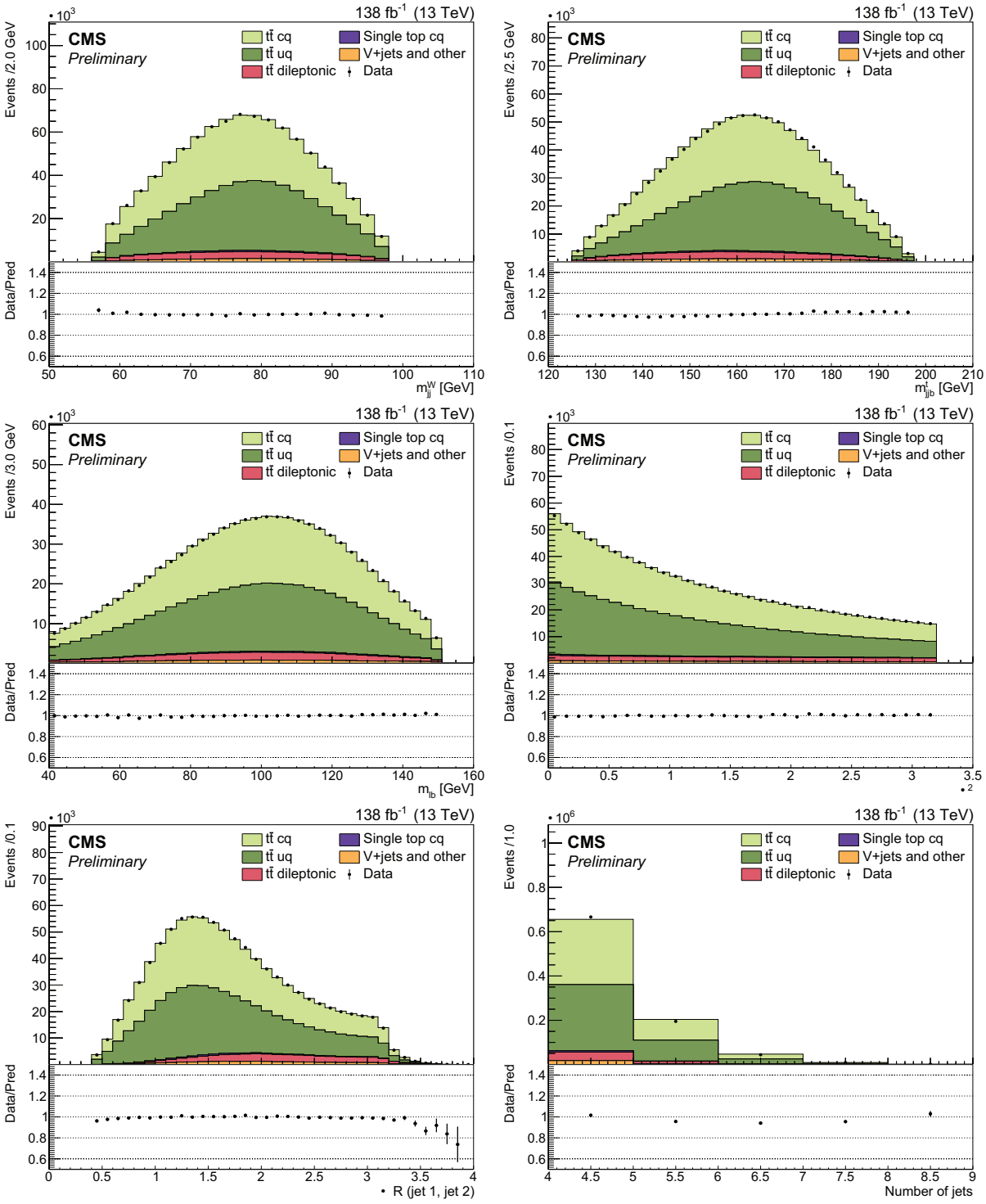
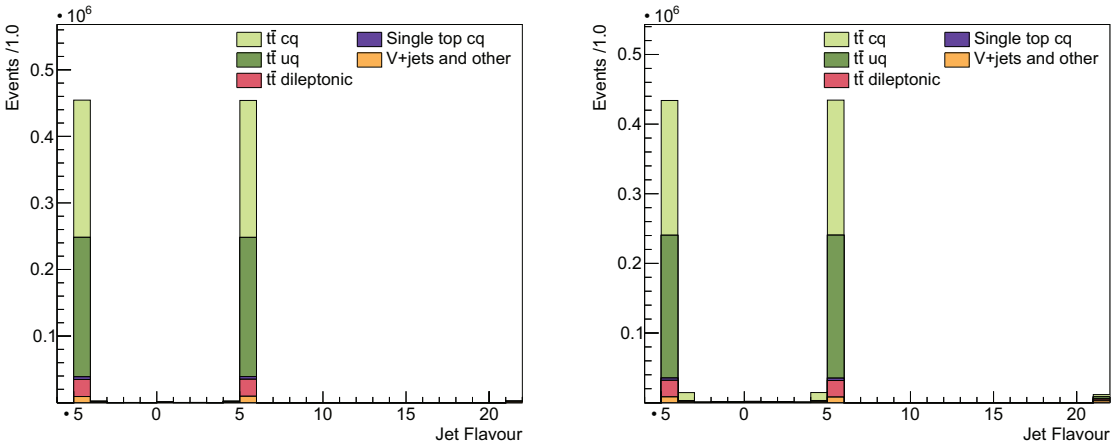
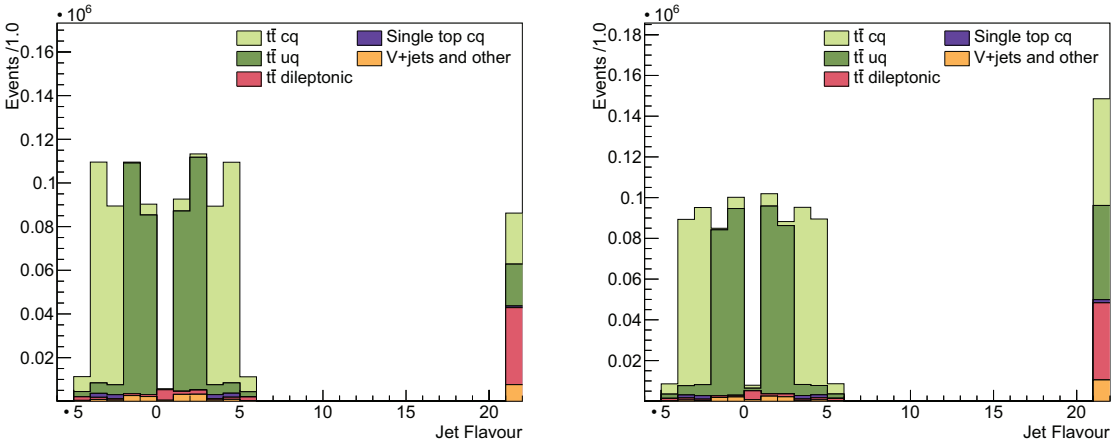


Figure 4.13: Some relevant kinematic distributions. The top left image corresponds to the invariant mass of the dijet reconstructing the hadronic W boson. Top right image displays the invariant mass of this dijet plus the corresponding b-jet reconstructing the top quark mass. The center left image is the invariant mass of the high-p_T isolated lepton and the other b-tagged jet associated to the W boson decaying leptonically. The center right image is the distribution of the χ^2 test value for the kinematic constraints. The bottom left image is the ΔR between the jets reconstructing the hadronic W boson and the bottom right image is the number of jets distribution.



(a) b1-jet flavour

(b) b2-jet flavour



(c) W jet 1 flavour

(d) W jet 2 flavour

Figure 4.14: Parton flavour for each of the four selected jets in the semileptonic $t\bar{t}$ simulation, top-left for b1-jet, top-right for b2-jet, bottom-left for W jet 1, and bottom-right for W jet 2. The various colors correspond to different production and W boson decay processes. The code for the flavour is the following [18]: 5 for bottom flavour, 4 for charm flavour, 1,2,3 for light down, up and strange flavours, 21 for gluonic and 0 is the case of no parton identified. The negative codes represent the corresponding antiquarks.

4.5 Charm tagging

After applying the baseline selection, we obtain a well-understood jets sample mostly composed of semileptonic $t\bar{t}$ events. To conduct the measurements presented in this work, it is essential to identify events where a W boson decays into a charm quark (denoted as W cq), distinguishing them from those where the W boson decays hadronically to light-flavour quarks (W uq). This section details the technique we used for charm tagging, how the charm tagging efficiency is calibrated using data, and how the mistagged background is determined from data as well. We also comment on the systematics affecting the chosen charm tagging method and how the associated systematic uncertainties are determined.

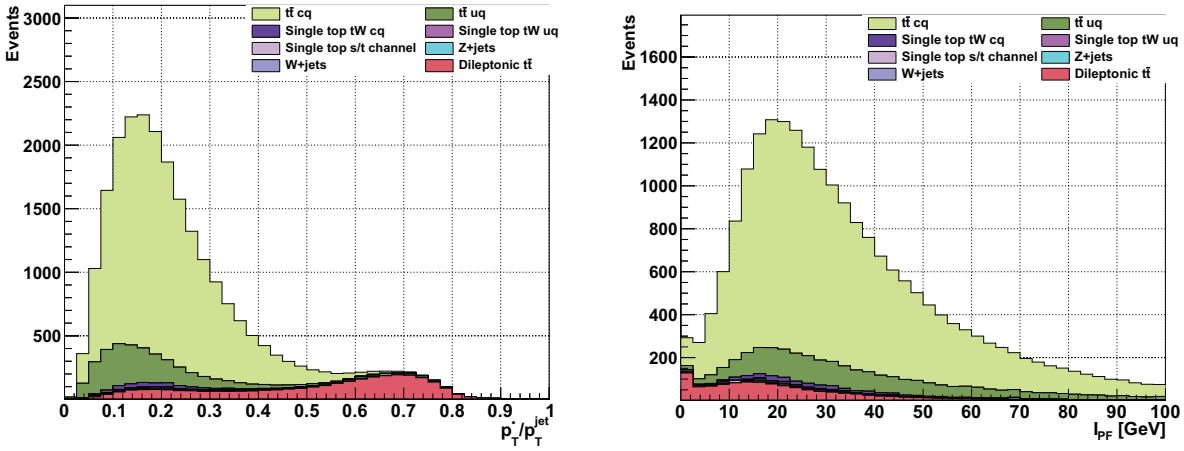
4.5.1 Muon identification in charm jets

The key aspect of this analysis is the charm tagging technique, together with its calibration and the estimation of associated systematic uncertainties. We employ a muon signature that is indicative of the presence of heavy hadron flavors within jets. When quarks form jets, the heavy-flavour hadrons arising from charm or bottom jets decay semileptonically, with a muon in the final state, in approximately 9% of the cases [18]. This feature makes muons a powerful signature for distinguishing these jets from light jets. Electrons are not considered for semileptonic c quark decays since the background for identification of electrons inside jets is really high. Since heavy-flavour hadrons have a short lifetime (0.4–1.5 ps), the muon is produced nearby the primary vertex. Muons inside jets stemming from the decay of pions and kaons (with lifetimes of 12–26 ns) are originated much farther in the detector and can be easily recognized by the characteristic kink in their track trajectory, the first part being the hadron track and the second part the muon track. The CMS muon reconstruction algorithm includes a kink finder to identify and remove such muons, leaving muons inside jets a signature for heavy flavour quarks. Muons from the decay of bottom hadrons behave similarly than those from charm hadrons, but this background, together with the remaining background of muons from light hadrons, can be determined from data, exploiting a charge symmetry that will be detailed below. This way it is possible to identify c jets from the decay of W bosons accurately. This muon-based charm tagging method enables a clean selection of charm jets and, as detailed below, allows for accurate determination of the background and mistag rates using real data. This capability facilitates a precision measurement of $^{W}_{c}$ with minimal and robust systematic uncertainties. This charm tagging technique has been employed in previous CMS publications [70, 71, 72, 73].

The CMS BTV group, already mentioned for b-tagging, also offers a DeepJet charm tagging discriminant for jets also based on neural networks. Using these taggers provides greater efficiency but results in lower purity. As demonstrated in appendix B, the DeepJet option significantly increases the uncertainty of the final result, so it has been discarded.

Charm tagging is performed on the two selected jets associated with the W boson candidate decaying hadronically. The selection of muons inside jets is done by angular

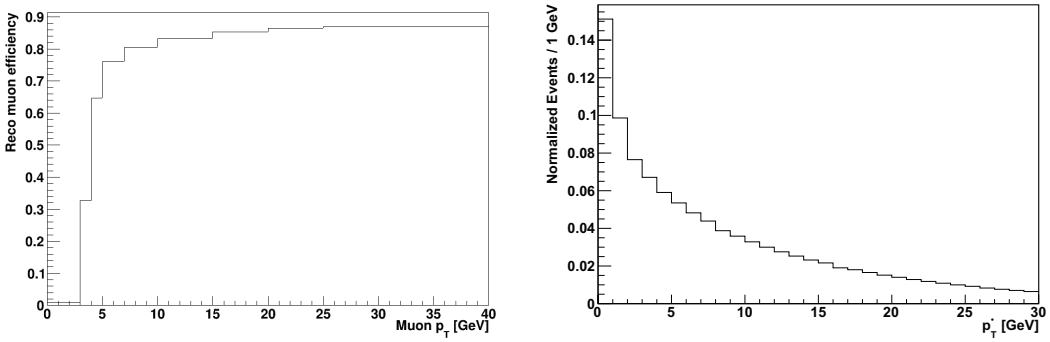
matching, $\Delta R(\mu, \text{jet}) = \sqrt{\Delta\phi^2 + \Delta\eta^2} \leq 0.4$. The muon inside the jet must satisfy the Tight reconstruction and identification quality criteria, with the goal of selecting a pure sample of muons, minimising background contamination. The Tight muon identification algorithm [60] features a kink finder in the tracker muon trajectory that largely suppresses the contamination of muons from the decay in flight of pions and kaons. A minimum non-isolation condition, $I_{PF} > 2.5$ GeV, is requested to the muon inside the jet. In addition, the muon must be reconstructed in the region $|\eta^\mu| < 2.4$, with transverse momentum $5 < p_T^\mu < 25$ GeV, and $z \equiv p_T^\mu / p_T^{\text{jet}} < 0.5$. If more than one such muon is identified, the one with the highest p_T is selected. The upper p_T requirement and the $p_T^\mu / p_T^{\text{jet}}$ condition significantly reduce the contamination from prompt muons overlapping with or misreconstructed as jets, mainly stemming from the dileptonic $t\bar{t}$ background (see Fig. 4.15a). The lower p_T threshold ensures a sufficiently high muon reconstruction efficiency ($\sim$70% at $p_T^\mu = 5$ GeV, see Fig 4.16a) since the muon must traverse the material in front of the muon detector and penetrate deep enough into the muon system to be reconstructed and satisfy the identification criteria. The $p_T^\mu > 5$ GeV requirement removes, however, about half of the events (see Fig. 4.16b).



(a) **(b)**

Figure 4.15: (a) Distribution of the $z = p_T^\mu / p_T^{jet}$ variable for the muon inside a jet, before the $z < 0.5$ requirement. (b) Distribution of the muon isolation variable I_{PF}, after the z requirement and before applying the condition $I_{PF} > 2.5$ GeV.

We will now introduce a crucial step of the analysis. Figure 4.1, where the Feynman diagram of a semileptonic $t\bar{t}$ signal charm event is displayed, shows that the sign of the electric charges of the lepton from the W boson decay and the charm quark are opposite. This translates into the semileptonic decay of the charm quark, with the resulting muon having electric charge of opposite sign with respect to the lepton from the W boson as well. Based on this property, we categorize events as opposite-sign (OS) or same-sign (SS), depending on whether the reconstructed muon within a jet has an electric charge that is opposite or identical to that of the high-pt isolated lepton. In addition to the selection requirements discussed so far, we will also ask for events to be OS.



(a) **(b)**

Figure 4.16: (a) Reconstruction efficiency for muons from the decay of a c hadron in the selected semileptonic $t\bar{t}$ MC events, as a function of the reconstructed muon p_T. (b) p_T distribution at the generator level of the muon arising from the charm jet.

Signal $t\bar{t}$ $W \to cq$ events will satisfy the OS condition whereas the background events in which the muon in the jet comes from the decay of a bottom or a light hadron will have symmetric OS and SS contributions. For the former case, $t\bar{t}$ events contain a bottom quark-antiquark pair that generates the charge symmetry. For the latter case, light jets formed in the decay of a W boson or from the hadronization of a gluon are charged symmetric with respect to the content of pions and kaons.

Using the OS = SS symmetry in the background events, we can determine the symmetric background contamination in data using the SS data sample, without relying on the simulation. In the OS selected data sample, there will be a fraction of events where the muon, although with opposite sign with respect to the prompt lepton, does not come from a charm hadron, but from the decay of a bottom or light (pion or kaon) hadron. This contribution is considered to be able to be modeled by the SS data sample. We can then subtract the SS events from the OS events in the data distributions to achieve a pure signal data sample. When comparing data with the predictions of the simulations, we will perform the same OS − SS subtraction in the MC samples. Equivalently, we can compare OS data with the sum of the OS − SS MC predictions and the SS data.

In Fig. 4.17 we illustrate the OS = SS symmetry using the simulation. The SS distribution has contributions from $W \to cq$ and $W \to uq$ decays, and from the rest of the background processes. All background processes have symmetric contributions in the OS and SS samples, except for a small fraction of the dileptonic $t\bar{t}$ events where the prompt muon from the decay of one of the two W bosons overlaps with a jet, producing an OS event. This can be seen by subtracting the SS events from the OS sample.

After charm-tagging, including the OS requirement, about 18000 events are selected. Figure 4.18 shows an event display of one of such events. It features a high-pt isolated muon, 4 jets, two of them tagged as b-jets, two additional jets compatible with coming from the decay of a W boson, with a muon inside one of them. According to the semileptonic $t\bar{t}$ simulation, the charm-tagged sample corresponds to about 2% of the total number of events in the ℓ + jets baseline sample, so around 0.04% of the produced semileptonic $t\bar{t}$

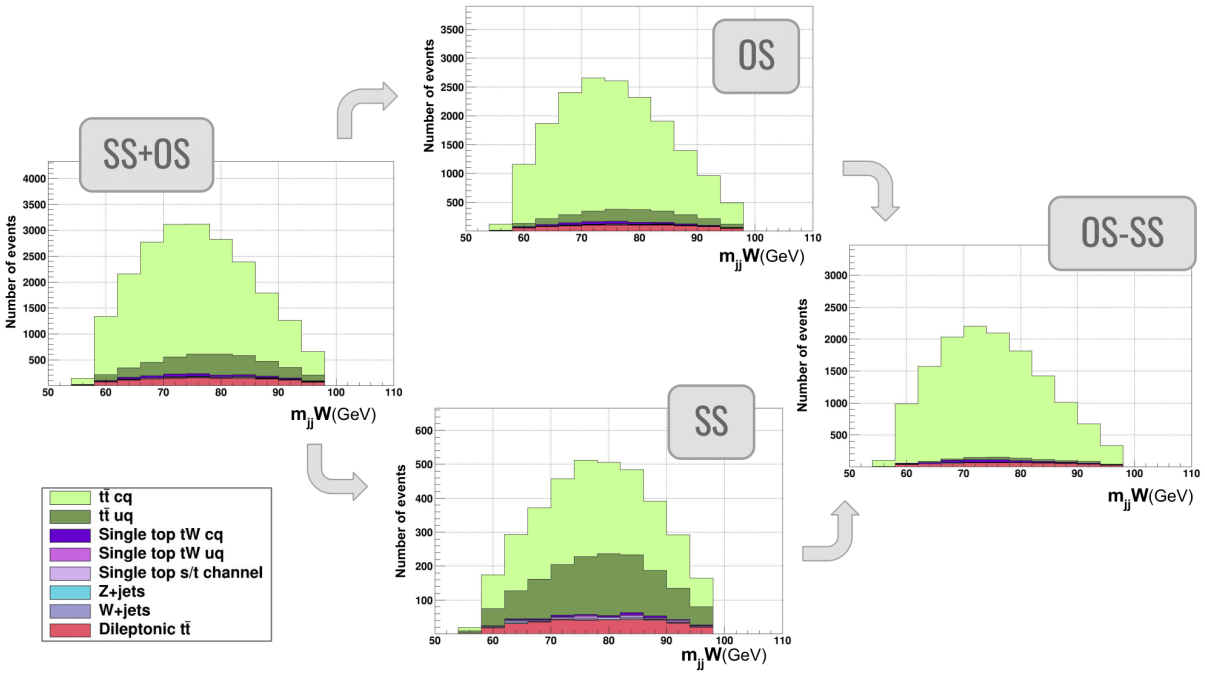


Figure 4.17: Illustration of the OS SS symmetry. The charm-tagged sample with a muon inside a jet (left plot) is divided between OS and SS events (middle plots). The distribution after subtracting OS and SS events is displayed in the right plot. All background contributions are canceled out except for a small fraction of the dileptonic $t\bar{t}$ events, that correspond to events where one of the prompt leptons from the decay of one of the W bosons overlaps with a jet producing an OS event since the charge of the lepton from the decay of the other W boson is opposite.

events. The selection efficiency is limited by the small charm semileptonic decay branching fraction (9%). The small charm tagging efficiency is not a concern for this analysis, given that the selected c-tagged sample is large enough to keep the statistical uncertainty below 1%. The advantage of the muon-based charm tagging is that it leads to a systematic uncertainty much smaller than that associated with the neural network-based CMS charm tagging method (see Appendix B).

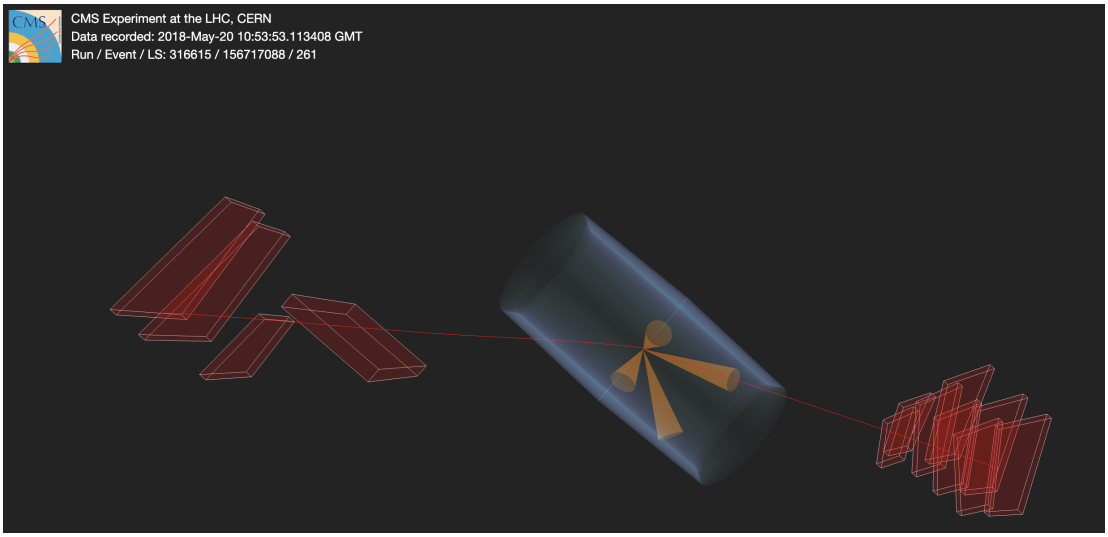


Figure 4.18: Event display of one signal candidate event: one high momentum isolated lepton and four jets. Two of the jets tagged as bottom jets and the other two jets being compatible with the W boson mass and with one muon inside one of them.

The composition of the sample, according to the simulation, is detailed in Table 4.5. The MC predictions in the table correspond to OS SS subtracted events, while the OS contamination is estimated using the SS data sample. The purity of c jets from the decay of a W boson in this sample is about 75%, most of the events coming from $t\bar{t}$ production with a small contribution from single top events. Around 3% of the selected sample consists of dileptonic $t\bar{t}$ events in which the muon from the decay of one of the W bosons overlaps with a light jet, resulting in a mistagged c jet. About 1% of the events correspond to semileptonic $t\bar{t}$ with W uq. The remaining contamination, about 20% of the sample, is represented by the SS data contribution.

4.5.2 Muon calibration

The muon-based charm tagging technique critically depends on the proper calibration of the reconstruction and identification efficiency of muons inside jets. CMS does not provide calibration corrections for non-isolated muons, so we have developed a method, described below, to calibrate the simulation with data.

The goal measurement, $^{W}_{c}$, also depends on how the simulation models the predicted rate of reconstructed muons from charm hadrons. For this reason the corresponding parameters are compared to most recent measured values and corrected accordingly.

Process	Muon channel	Electron channel	Total	Percentage
$t\bar{t}$, W cq	8172.2	4993.3	13165.5	72.0%
Dileptonic $t\bar{t}$	298.7	188.1	486.8	2.7%
Single top, W cq	133.3	93.3	226.6	1.2%
$t\bar{t}$, W uq	150.1	84.0	234.1	1.3%
Single top, W uq	1.7	1.9	3.6	0.0%
Single top, no W	14.5	8.9	23.4	0.1%
V+jets	43.0	8.7	51.7	0.3%
Diboson	1.3	1.0	2.3	0.0%
Data same sign (SS)	2551	1546	4097	22.4%
Total predictions	11365.8	6925.2	18291.0	
Data OS	11167	6806	17973	

Table 4.5: Composition of the sample of c tagged events selected from the baseline jets selection. Yields are given separately for the prompt muon and electron channels. The yields predicted by the simulations correspond to OS SS subtracted events. The SS contamination is estimated with data. The number of events in data correspond to OS events.

Rate of muon production from charm hadrons in the simulation

The production of c hadrons from the hadronization of c quarks, and their leptonic and semileptonic decays with a muon in the final state need to be modeled accurately. The QCD parton shower, hadronization, and particle decay in the simulated samples are done with PYTHIA V8.2. The charm fragmentation fractions (FF), defined as the probabilities for c quarks to hadronize as particular c hadrons, c , corresponding to D , D $\overline{\text{D}}$, D and hadrons, are reweighted in the MC samples in an event-by-event basis to match the measured values reported in Ref. [88]. In addition, the semileptonic branching fractions (BF) of the c hadrons, , are corrected to agree with more recent measurements [18]. Table 4.6 and Table 4.7 list the FF and BF measurements and their uncertainties, the corresponding values used in PYTHIA V8.2, and the event weights applied to correct the simulation.

Fragmentation fraction	Measurement	PYTHIA V8.2	Event weight
c D)		0.290	0.83
c D $\overline{\text{D}}$)		0.564	1.08
c D)		0.097	0.83
c)		0.036	1.74

Table 4.6: Charm quark fragmentation fractions measurements with uncertainties, values used in PYTHIA V8.2, and event weights applied to correct the simulation.

The overall correction reduces the rate of muons originating from c hadron decays in the simulation by . The uncertainty in the correction, calculated propagating the uncertainty in the fragmentation and decay fractions measurements, will be considered as a systematic effect in the $^{W}_{c}$ measurement, as detailed in Sec. 4.6.

Decay fraction	Measurement	PYTHIA V8.2	Event weight
D)		0.1664	1.06
D $\overline{\text{D}}$)		0.0665	1.01
D)		0.076	0.86
)		0.045	0.88

Table 4.7: Charm hadron semileptonic decay branching fraction measurements with uncertainties, values used in PYTHIA V8.2, and event weights applied to correct the simulation.

In addition to correcting the rate of muons from c hadrons in the simulation, the other important aspect of the modeling of muons inside c jets is the calibration of the reconstruction and identification efficiency in the simulation. The CMS Muon POG group does not provide data/MC SFs for non-isolated muons so an ad-hoc method is developed to calibrate the efficiency of such muons. The b jets, present in the jets selected sample, identified with high purity (98%), provide a suitable source of muons. Applying the selection requirements for muons inside jets to the b-tagged jets we can compare the rates of selected muons in data and MC and then extract the reconstruction and identification data/MC SFs.

Rate of muons from bottom hadrons in the simulation

Before embarking on the calculation of the SFs, the first step is to check that the rate of produced muons in b jets is accurately simulated in the MC samples so that any difference in the rates of muons can be directly attributed to the difference in reconstruction and identification efficiencies in data and MC. These muons either originate in the semileptonic decays of b hadrons or in the semileptonic decay of c hadrons stemming from the hadronic decay of b hadrons. The rate of muons, therefore, depends on the production rate of b hadrons from b quarks (the b quark fragmentation fractions b), the muon semileptonic decay fractions of b hadrons ( ), the decay fractions of b hadrons into c hadrons ( ), and the muon semileptonic decay fractions of c hadrons ( ). We have checked the modeling of all those parameters in the PYTHIA V8.2 simulation. Tables 4.8, 4.9, and 4.10 list the b , and values used in PYTHIA V8.2, together with the measured values and their uncertainties [18].

Fragmentation fraction	Measurement	PYTHIA V8.2
b B)		0.429
b B $\overline{\text{B}}$)		0.429
b B)		0.095
b)		0.047

Table 4.8: Bottom quark fragmentation fractions measurements with uncertainties, and corresponding values used in PYTHIA V8.2.

The total muon rate in b jets produced in the decay of heavy flavour hadrons, calculated

Decay fraction	Measurement	PYTHIA V8.2
B)		0.1109
B $\overline{\text{B}}$)		0.1024
B)		0.093
)		0.077

Table 4.9: Bottom hadron semileptonic decay branching fractions measurements with uncertainties, and corresponding values used in PYTHIA V8.2.

	Measurements				PYTHIA V8.2			
B D	D	D $\overline{\text{D}}$	D		D	D $\overline{\text{D}}$	D	
B					0.1076	0.7573	0.1171	0.0179
B $\overline{\text{B}}$					0.4441	0.4252	0.1102	0.0204
B					0.0814	0.1337	0.7672	0.0177
					0.0551	0.1098	0.0827	0.7524

Table 4.10: Measurements and uncertainties of the decay fractions of b hadrons into c hadrons, and corresponding values used in PYTHIA V8.2.

as,

b b b (4.10)

obtained from the measurements, , is only 0.5% smaller than the rate obtained with PYTHIA V8.2, 0.1944. No correction is then necessary for the simulation. Still, the uncertainty in the measured rate (2.3%), obtained propagating the various uncertainties of the measurements in Tables 4.7, 4.8, 4.9, 4.10, will be taken as a systematic uncertainty in the determination of the muon reconstruction and identification efficiency SFs.

Correction of the identification efficiency of muons inside heavy flavour jets

After having corrected the rate of muons in the simulation stemming from charm and beauty hadrons, in order to calculate the SF correcting the potential difference between data and simulation in the identification efficiency of muons inside heavy flavour jets, we start from the normalized jets sample, with the same number of data and MC events. We will then apply the muon selection requirements we use for charm tagging, and the resulting ratio data/MC will be the correction factors we are looking for.

Muon charm tagging requirements are applied to the b-tagged jets of the baseline selection but removing kinematic constraints to significantly increase the number of available events for calculating the correction. Those requirements were meant to optimise the identification of the hadronic W boson jets, and reduce by a factor of three the number of selected events. Reconstructed muons are associated with the b-tagged jets by angular matching with the jet axis (jet). Muons in b jets are required to verify the

same conditions as the muons identifying c jets, Tight ID, , GeV, , but we consider a wider range for their transverse momentum, GeV to enlarge the statistics. We obtain a sample of high-purity b jets three times larger than the baseline semileptonic $t\bar{t}$ one (3 million events).

We show in Fig. 4.19 kinematic distributions of the muons inside the b jets for data and MC, the transverse momentum , pseudorapidity , isolation , relative isolation , and ^{jet}. The contributions of muons from different sources are separated in the figures. About half of the muons come directly from the semileptonic decay of b hadrons, while the other half stem from the decay of c hadrons that come from b hadrons. There is a small contribution, less than 10%, of muons coming mainly from the decay in flight of pions and kaons inside b jets. The lower panel of the figures provides the ratio of the yields in data and MC, that is, the data/MC SFs for the muon reconstruction and identification efficiency as a function of the displayed variable. The or distributions reveal a global deficit of muons in data, compared to the simulation prediction, of about 8%, with no significant shape for the ratio data/MC. The absolute isolation ( ) distribution however shows a dependence in the ratio, so that as the muon becomes less isolated, the difference between data and MC identification efficiency becomes larger. The desired scale factors to correct the simulation are extracted from this data/MC distribution dependent on . These SFs are applied to the simulation to correct the reconstruction efficiency of the muons in our signal events (muons inside c-tagged jets). They will be applied to the identified muons in charm-tagged jets produced in the semileptonic $t\bar{t}$ events of our analysis. Given that the SFs depend on the activity around the muon, the calibration determined from a sample of events with the same topology is very suitable. Other distributions also exhibit a dependence in the ratio data/MC, the relative isolation and the ratio of transverse momenta of the muon and the associated jet ^{jet}. It was checked, however, that SFs calculated as a function of these variables introduce a distortion in the description of , due to their correlation. The best results are achieved when using SFs dependent on the absolute isolation .

To calculate the SFs from the distribution, a variable-width binning is used to account for the low event yield in the last bins, and a fit with a functional form is performed. Different fits are computed for five distinct regions of the reconstructed muon pseudorapidity ( ), [0, 0.45], [0.45, 0.9], [0.9, 1.2], [1.2, 1.8], and [1.8, 2.4]. The individual fits and all fits together are displayed in Fig. 4.20. The yellow band around the fitted curve represents the confidence interval at one sigma for the fit.

As a cross-check, Fig. 4.21 shows the muon distributions after applying the SFs to the sample used to derive them. They are to be compared to distributions in Figure 4.19. The distribution is fixed by construction, except for statistical variations since the binning of the distribution is not the same as the binning used to extract the corrections. The global mismatch for and is fixed, with no induced dependence. Overall, the ratio Data/MC becomes flat and centered at 1 for all the observables, which gives us confidence in the correction.

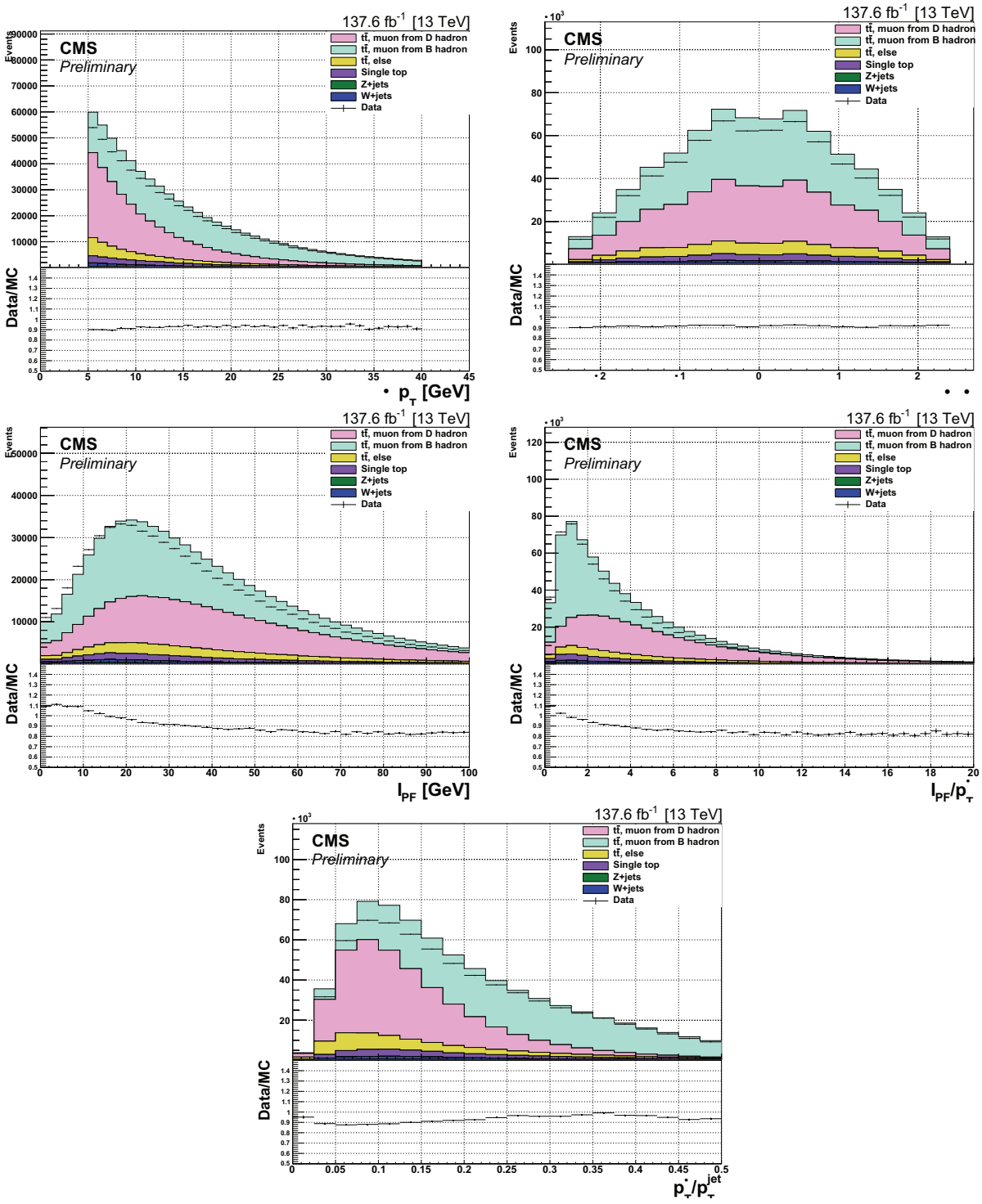


Figure 4.19: p_T^{μ} (top-left), η^{μ} (top-right), I_{PF} (middle-left), I_{PF}/p_T^{μ} (middle-right), and $p_T^{\mu}/p_T^{\text{jet}}$ (bottom) distributions of muons inside b jets before any correction.

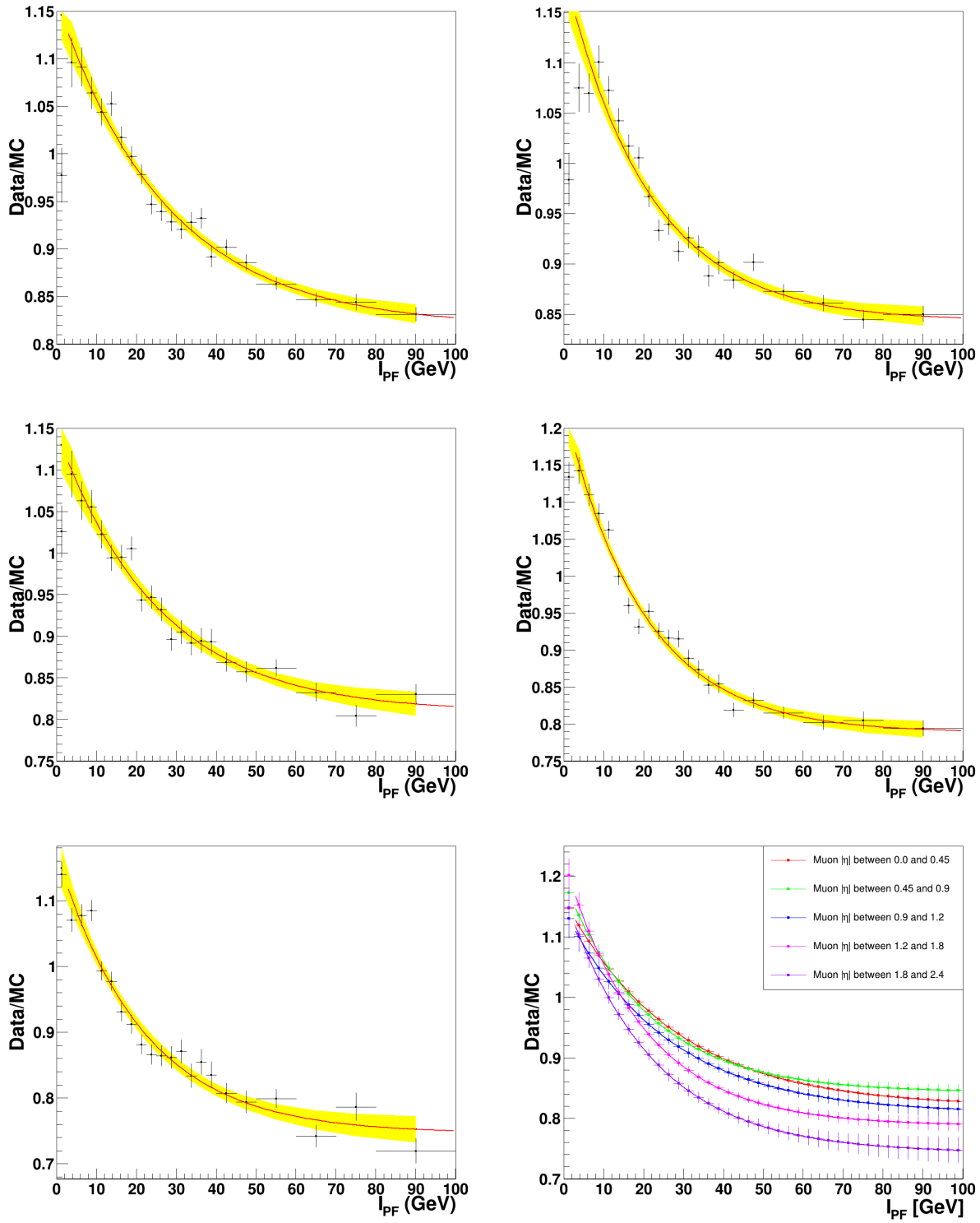


Figure 4.20: Fit of the muon yield Data/MC ratio as a function of muon isolation for the muon pseudorapidity intervals (top-left), (top-right), (middle-left), (middle-right), (bottom-left), and fits for all pseudorapidity intervals together (bottom-right).

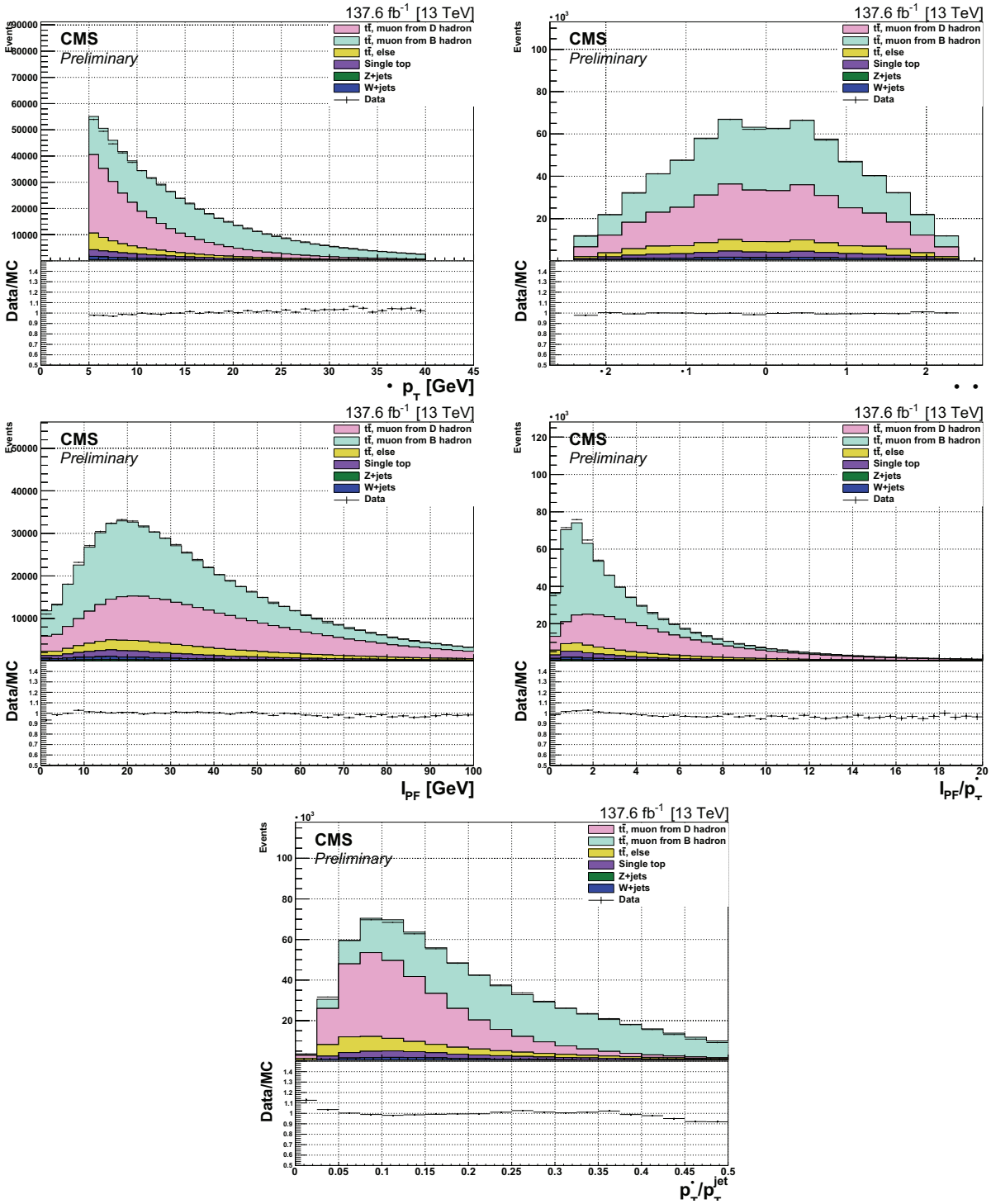


Figure 4.21: p_T^μ (top-left), η^μ (top-right), I_{PF} (middle-left), I_{PF}/p_T^μ (middle-right), and p_T^μ/p_T^{jet} (bottom) distributions of muons inside b jets after applying the I_{PF}-dependent SFs.

4.6 Systematic uncertainties

This section details the main sources of systematic uncertainties affecting the predicted event yields of the various contributions involved in the extraction of the $^{W}_{c}$ measurement, and how we have evaluated them. The impact of these systematic effects on the precision of the measurement is discussed in Section 5. Systematic effects are divided into two groups, those affecting the predictions for the W cq and W uq contributions in approximately the same way (global normalization effects), and those associated with the c tagging. The former have a minimal impact on the precision of the $^{W}_{c}$ measurement, while the latter propagate directly into the $^{W}_{c}$ uncertainty.

Systematic sources affecting global normalisation:

- All MC samples are normalized according to their respective SM cross section values, and uncertainties for the rates of each process are assigned using the precision of CMS cross section measurements or theoretical calculations. A 5% uncertainty is assigned to dileptonic $t\bar{t}$ [89], 1%, 2%, and 4% to single top quark -, -, and - channels, respectively [90, 91, 92], 2% to V [93], 6%, 5%, and 7% to diboson WW, ZZ, and WZ, respectively [94, 95, 96].

- The statistical uncertainty of the charm-tagged SS data sample, used as the prediction for the mistag contamination of the OS signal sample, is taken as the systematic uncertainty of this component (1.6%).

- The measured integrated luminosity values for the three data-taking years have uncertainties between 1.2 and 2.5% [74, 75, 76], while the overall uncertainty for the combined data set is 1.6%, affecting the predictions from the simulations.

- All MC samples are reweighted to match the pileup distribution in data, which is generated by using the instantaneous luminosity per bunch crossing for each luminosity section, with a total inelastic cross section of 69.2 mb; an uncertainty of 4.6% is applied to this value [97].

- The high- isolated prompt muons and electrons in the simulated samples have uncertainties associated with the high-level trigger, reconstruction, and identification. These uncertainties are uncorrelated across lepton flavors but correlated across years (except the HLT uncertainties) and are parameterized as a function of the and of the leptons. The overall effect is 1% (1.2%) for muons (electrons). The uncertainty associated with the possible misidentification of the sign of the lepton electric charge is negligible.

- To evaluate the JES and JER uncertainties, the procedure recommended by the JME group is followed [98], applying variations to the energy of the reconstructed jets in simulated events, considering various uncertainty sources split between detector regions and data-taking years [66]. For the JES, the recommended default procedure using a reduced set of uncertainty sources was followed. The JES variations are also

propagated to the miss. The resulting overall uncertainties are 4%(1%) for JES (JER).

- The b tagging identification and mistagging efficiencies in the simulation are calibrated to match the corresponding efficiencies in data. Separate uncertainties in the tagging and mistagging corrections are assigned. Uncertainties are evaluated following the BTV POG recommendations [67], resulting in an overall b tagging uncertainty of 2.6%.

- In addition to the experimental sources, we consider theoretical uncertainties affecting the MC simulations. The uncertainty from the choice of PDF is estimated from the Hessian NNPDF sets according to the procedure described in Ref. [25]. Renormalization and factorization scales at the matrix element level are varied by a factor of 2 or 0.5 to take into account the effect of higher-order corrections in the $t\bar{t}$ simulation. The effect is 1%.

- For the parton shower simulation, uncertainties are separately assessed for initial and final-state radiation (ISR and FSR), by varying the respective scales up and down by factors of two. The effect of the ISR uncertainty is negligible while the FSR uncertainty changes the predicted yields of the semileptonic $t\bar{t}$ selection by 4% while it changes the normalization of the predicted charm-tagged samples by 6%.

- The uncertainty in the correction to the top quark is evaluated as a one-sided variation computed from the difference between the top quark distribution with and without the correction [85]. The resulting systematic uncertainty in smaller than 1%.

Systematic sources affecting charm tagged sample:

- The correction of the muon rate in the decay of c hadrons in the simulation, described in section 4.5.2, has an uncertainty of 2.2%, reflecting the precision of the charm fragmentation fractions and semileptonic decay branching fraction measurements.

- We have considered several effects contributing to the uncertainty of the scale factor correcting the reconstruction and identification efficiency of muons from the decay of heavy flavour hadrons inside jets in the simulation. As described in Sec. 4.5.2, this correction is determined by comparing the rate of muons in b jets in data and simulation.

  - The rate of muons in b jets from the decay of b and c hadrons has an uncertainty of 2.3%. This value is calculated propagating the uncertainty in the measurements of the b quark FFs to B , B $\overline{\text{B}}$, B and hadrons, the semileptonic decay BFs of these b hadrons, the decay BFs of b hadrons into D , D $\overline{\text{D}}$, D and hadrons, and the semileptonic decay BFs of these c hadrons.
  - The uncertainty in the muon-in-jet scale factors, resulting from the uncertainty in the parameters of the fit procedure detailed in Sec. 4.5.2, is 1%.

– We assign an additional uncertainty of 1% to the muon reconstruction and identification scale factor. The sample of muons in b jets used to calculate the SF contains about 10% of muons coming from the decay in flight of pions and kaons. In the same sign (SS) control region in the analysis, dominated by this kind of muons, we observe a 10% mismatch in the normalization of the MC to the data. Taking the difference in normalization as a systematic effect, the resulting uncertainty in the muon SF is 1%.

Table. 4.11 summarizes the size of the systematic uncertainties.

Systematic source	**no charm sample [%]**	**charm sample [%]**
Dileptonic $t\bar{t}$ cross section	5	5
Single top t-channel cross section	2	2
Single top tW-channel cross section	4	4
Single top s-channel cross section	1	1
V+jets cross section	2	2
WW cross section	6	6
ZZ cross section	5	5
WZ cross section	7	7
SS data sample statistics	1.6	1.6
Measured integrated luminosity	1.6	1.6
Pile up inelastic cross section	4.6	4.6
Lepton trigger, reconstruction and identification	1 ( ) 1.2 (e)	1 ( ) 1.2 (e)
Jet energy scale	4	4
Jet energy resolution	1	1
Jet b-tagging	2.6	2.6
Mistag from b-tagging	10	10
Parton distributions functions	4	6
Initial- and final-state radiation	<1	<1
Top quark momentum reweighting	1	1
Muon rate from c hadron decays	-	2.2
Muon rate from b hadron decays	-	2.3
Muon in jet scale factors	-	1
Muon rate from /K decays	-	1

Table 4.11: Relative uncertainty on the global yield predicted by the simulation resulting from any of the systematic sources. Uncertainties are given for the selected samples with and without the application of charm tagging.

4.7 Charm-tagged sample kinematic distributions

In order to illustrate the high purity of the subtracted OS SS charm sample and the excellent description of the data by the simulation, some relevant distributions are shown in this section. The grey band in the MC distributions reflects the systematics uncertainties discussed in Sec. 4.6. After applying the muon-based charm tagging method described in Sec. 4.5.1 and muon calibrations detailed in 4.5.2, distributions in Figs. 4.22, 4.23, 4.24, 4.25, 4.26 show the result. The overall effect of the muon calibration is a reduction of 7.5% in the predicted yield in the simulation of the c-tagged selected sample.

Fig. 4.22 shows distributions related to the muon inside the tagged c jet: muon , , isolation, , transverse impact parameter , and longitudinal impact parameter . Figure 4.23 displays distributions related to the high- prompt lepton and the W boson decaying leptonically: the lepton transverse momentum , the lepton pseudorapidity , the invariant mass of the lepton and the b jet associated with the corresponding top quark , the difference in azymuthal angle of the lepton and , the missing momentum , and the transverse mass of the lepton and . Figure 4.24 shows distributions related to the two jets associated with the W boson, , , event jet multiplicity, and between the two jets. Figure 4.25 depicts distributions related to the two b-tagged jets, , , and b-tagging discriminant. Finally, Fig. 4.26 shows the invariant mass of the two jets associated with the W boson, and the invariant mass of the three jets associated with the top quark. Even though subtracted data is not used for the measurement computation, simulation distributions are displayed with data to show the great agreement achieved.

All distributions exhibit good agreement between data and MC simulation. The distributions separating the contributions of the electron and muon channels are featured in Appendix C. The distributions for the reconstructed muon inside the jet, as shown in Fig. 4.22, exhibit no significant discrepancies between data and MC, with the two matching within statistical uncertainties. No systematic behavior is observed in the ratio panel. The reconstructed muon, which serves as the signature for charm tagging, is well understood and accurately modeled. This confirms the robustness and reliability of the charm tagging technique used in this analysis.

Both leptonic, Fig. 4.23, and hadronic, Figs. 4.24, 4.25, 4.26, observables also show good agreement between data and MC, both in shape and normalization. Not only is the muon behavior well understood, but all the studied distributions are well described by the simulation. The selected charm sample is highly pure, with 94% of the contributing events corresponding to W cq signal events, as indicated by the simulation. The comparison of OS SS subtracted simulation with data further validates the suitability of these MC samples for accurately modeling the signal component, ensuring the successful completion of the measurement.

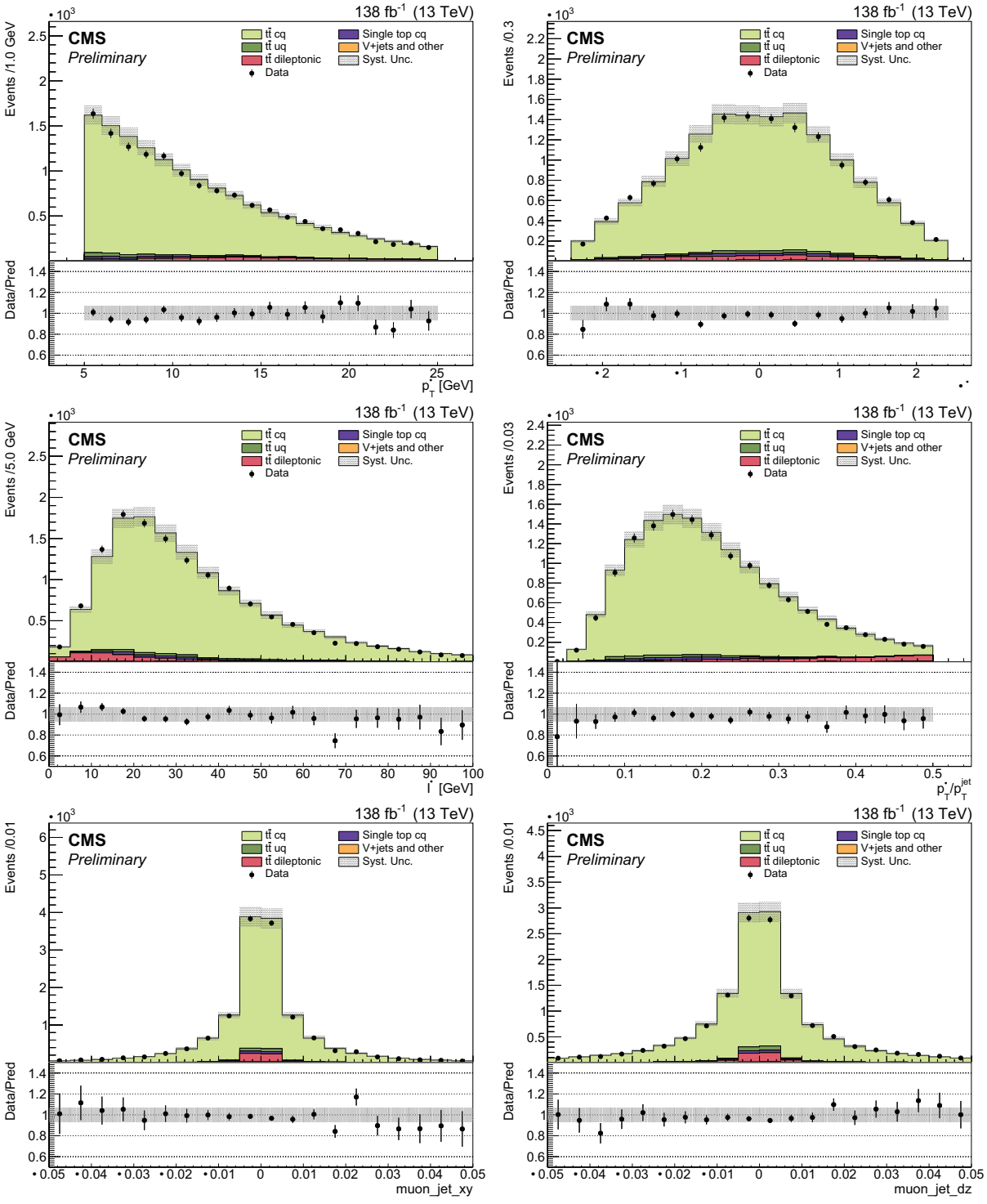


Figure 4.22: Charmed-tagged selected sample, adding prompt electron and muon channels, OS−SS distributions related to the muon inside the tagged c jet: Muon p_T (top-left), η (top-right), isolation (middle-left), $z = p_T^{\mu}/p_T^{jet}$ (middle-right), transverse impact parameter d_{xy} (bottom-left), and longitudinal impact parameter d_z (bottom-right).

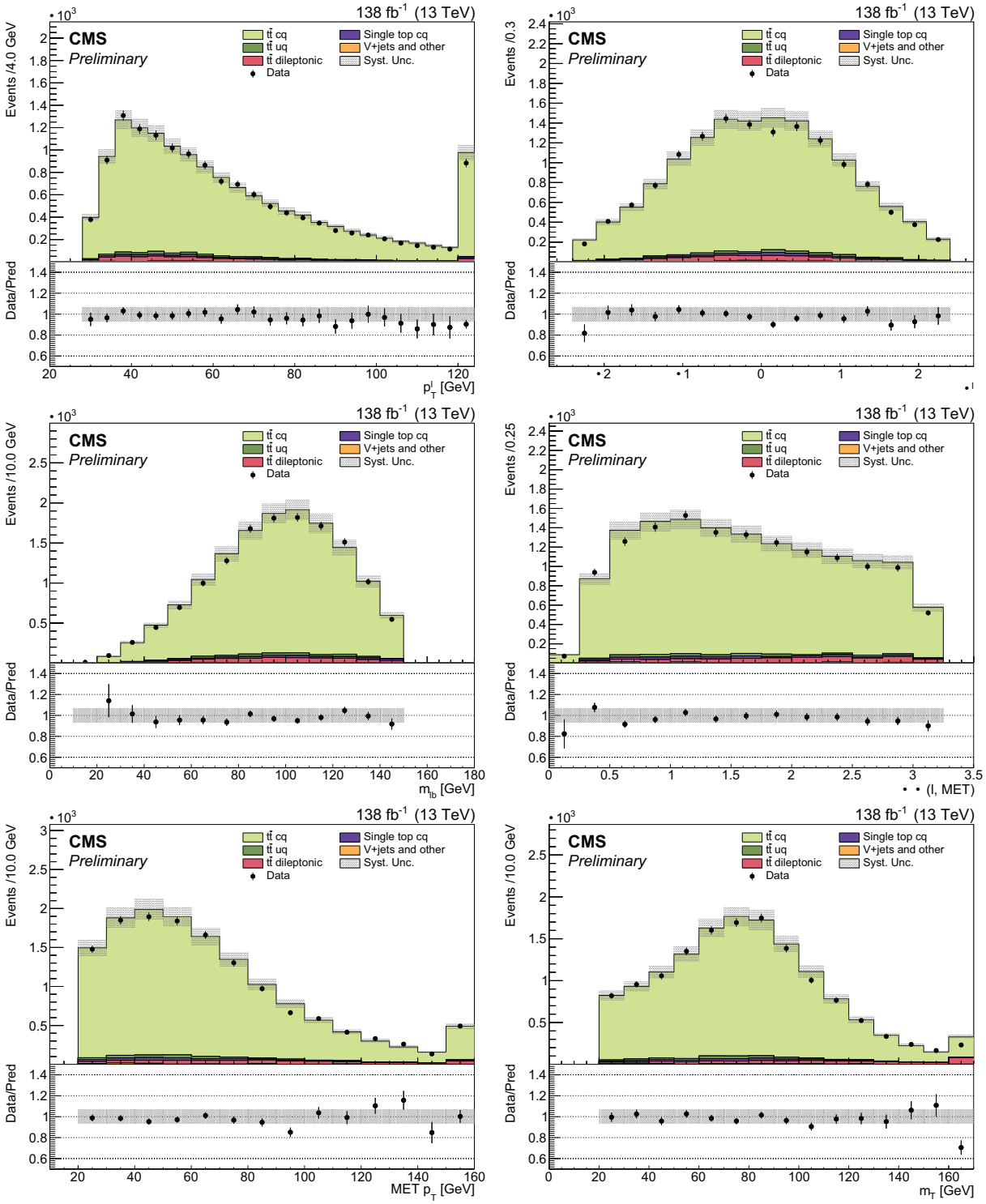


Figure 4.23: Charmed-tagged selected sample, adding prompt electron and muon channels, OS − SS distributions related to the prompt high-p_T lepton: lepton p_T (top-left), lepton η (top-right), invariant mass of the lepton and the b jet (middle-left), difference in azymuthal angle of the lepton and $\vec{p}_T^{miss}$ (middle-right), missing momentum p_T^{miss} (bottom-left), and transverse mass (bottom-right).

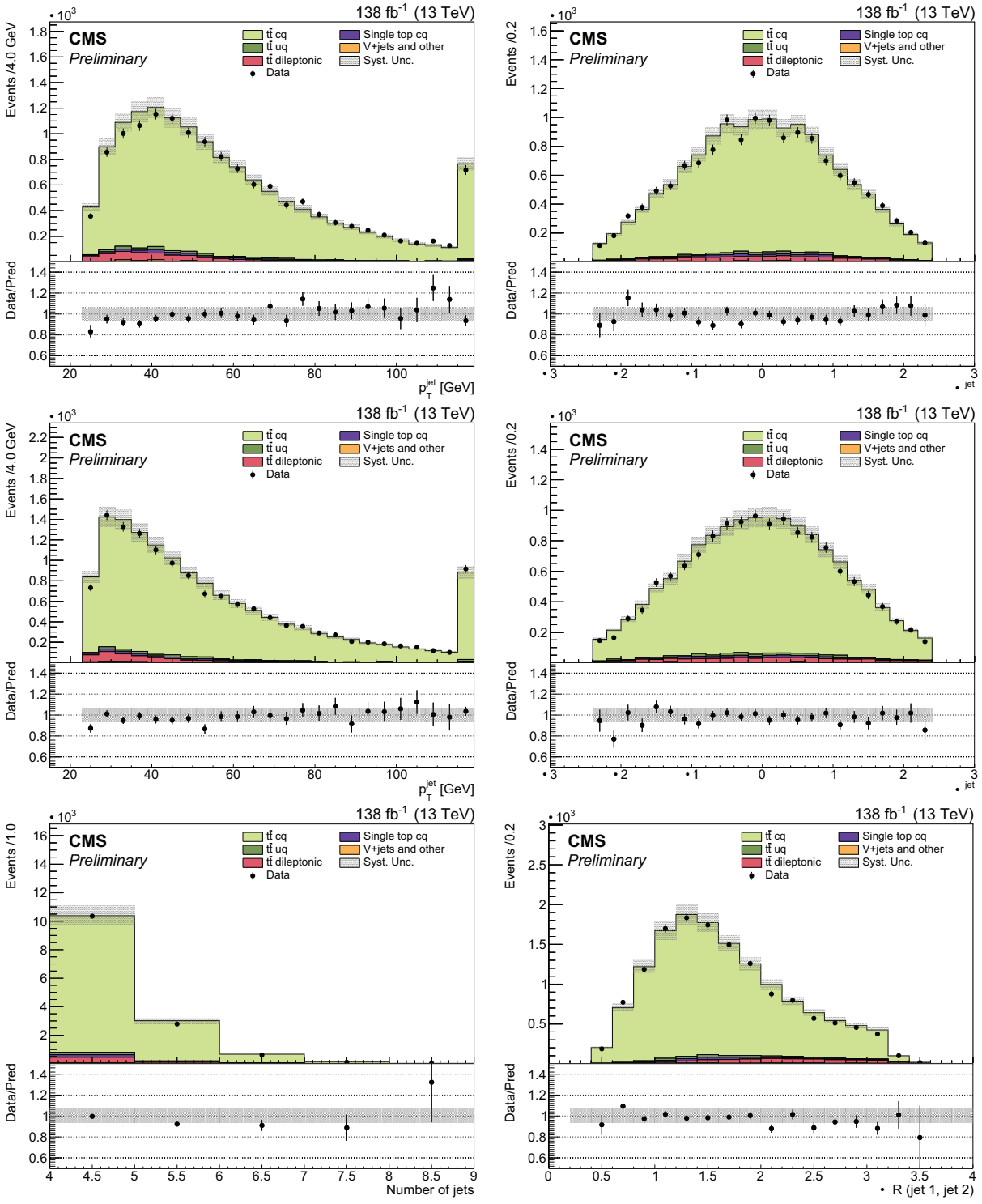


Figure 4.24: Charmed-tagged selected sample, adding prompt electron and muon channels, OS − SS distributions related to the jets associated with the W boson: jet p_T (top-left) and jet η (top-right) for the jet tagged as c-jet, jet p_T (middle-left) and jet η (middle-right) for the other jet, event jet multiplicity (bottom-left) and ΔR between the two jets (bottom-right).

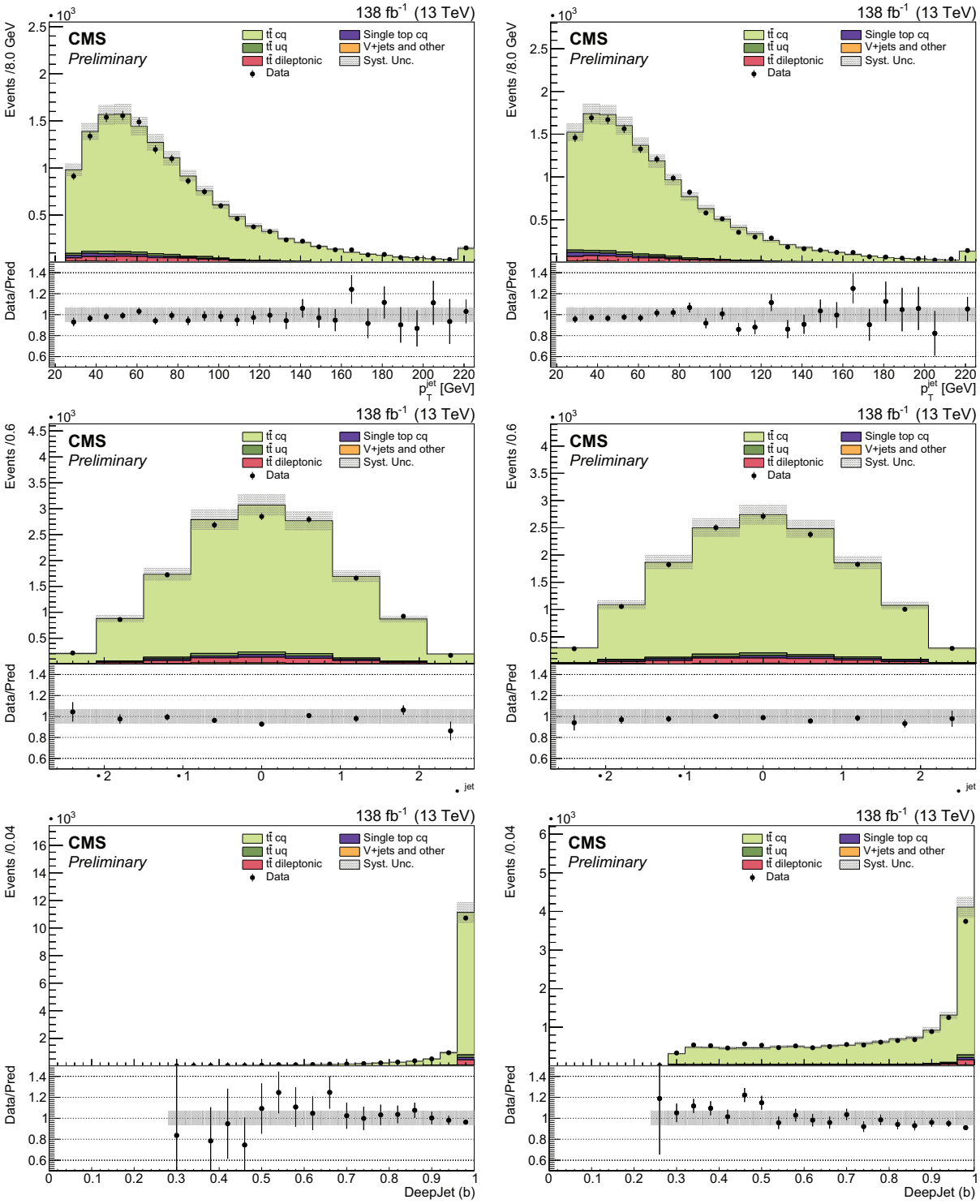


Figure 4.25: Charmed-tagged selected sample, adding prompt electron and muon channels, OS – SS distributions related to the b-tagged jets: jet p_T, η, and b-tag score for the jet with the highest b-tag score(left); jet p_T, η, and b-tag score for the other b-tagged jet (right).

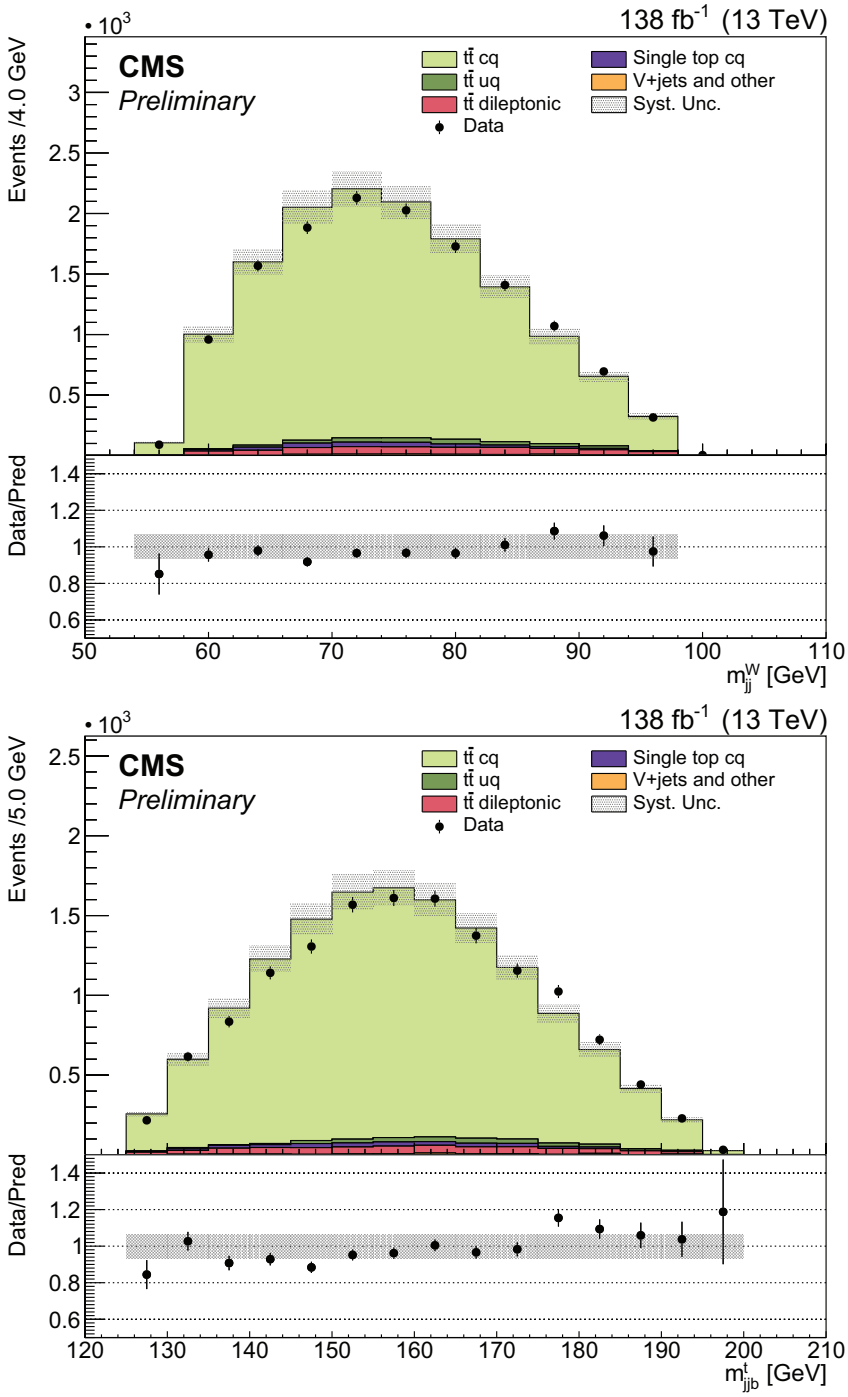


Figure 4.26: Charmed-tagged selected sample, adding prompt electron and muon channels, OS – SS distributions for the invariant mass of the two jets associated with the W boson (top) and invariant mass of the three jets associated with the top quark (bottom).

Chapter 5

Measurement of the R_c^W branching fraction ratio

Once the baseline event selection and the charm tagging method are defined, we can proceed to describe how the measurement is conducted. The measurement of interest $_{c}^{W}$ represents the ratio of the branching fractions of a W boson decaying to a charm quark and decaying hadronically $\mathcal{B}$ W cq $\mathcal{B}$ W q$\bar{q}$. The current world-average $_{c}^{W}$ value reported by the PDG [18] is (8% uncertainty). It is obtained from the combination of the ALEPH and OPAL measurements conducted in 1999 and 2000, yielding values of and respectively [99, 100].

Observed and predicted event yields in four different sub-samples (channels) will be used in a simultaneous fit. The four exclusive channels are defined depending on the flavour of the prompt high- lepton (muon or electron) and the charm tag (tag or no tag) of one of the jets associated with the W boson, as specified in Table 5.1. Starting with events with no charm tag, channel 1 (ch1) is formed by events in which the high- lepton is a muon, and channel 2 (ch2) by events with an electron. Likewise, events with a charm tag are separated in channels 3 and 4, with events with a high- muon (ch3) and a high- electron (ch4) (ch1 jet , ch2 jet , ch3 jet , ch4 jet).

	Channel 1	Channel 2	Channel 3	Channel 4
Prompt lepton	Muon	Electron	Muon	Electron
Charm tag (muon in jet)	No tag	No tag	Tag	Tag

Table 5.1: Categories used in the fit to extract the $_{c}^{W}$ branching fraction ratio.

The physical processes contributing to the predictions for each of the channels are the following:

- Semileptonic t$\bar{t}$ production where the hadronic W boson decays to a charm quark, denoted as ttWcq.

- Semileptonic $t\bar{t}$ production where the hadronic W boson does not decay to a charm quark (ttWuq).
- Single top production where there is a W boson decaying to a charm quark (stWcq).
- Single top where a W boson decays hadronically but not to a charm quark (stWuq).
- Single top where no W boson decays hadronically.
- jets processes.
- Dileptonic $t\bar{t}$ process.
- Diboson processes and fully hadronic $t\bar{t}$ events are combined into a single contribution, as their individual impact is minimal.

For the two categories with charm tag, as explained in Sec. 4.5.1, we can distinguish between opposite-sign (OS) and same-sign (SS) events, depending on the sign of the electric charges of the prompt high- lepton and the muon inside the jet. For W cq decays, true signal events are OS, while for the rest of the processes (W uq and backgrounds) the OS and SS yields are expected to be the same. In a fraction of the W cq OS events (around 10%) the tagged muon does not come from the decay of a c hadron, but mostly from the decay in flight of pions or kaons. For this background, the symmetry of the OS and SS yields also holds.

The CMS statistical analysis tool, Combine [101], is used to fit the predicted event yields to the observed event yields in the four channels simultaneously. Originally designed for Higgs searches, Combine has become a widely used statistical tool within the CMS collaboration to extract measurements based on a model of expected observations and a dataset.

The fit can be done for distributions, modelling shapes, but in this analysis it is performed using the total number of events of each process since none of the shapes of the distributions provides additional discrimination power. The procedure will be then modelling OS data with OS SS contributions of each process plus SS data sample, which models the OS background. For channels 3 and 4, the input for the fit consists of OS events for data, OS SS subtracted events for MC processes, and the SS data sample as an additional process. All baseline-selected data events are used. Since OS and SS events behave likewise for background processes we bypass the need to understand the MC description in this case and directly use SS data events. This procedure has the advantage that we do not rely on the simulation to model the background contributions, but it is directly extracted from data, this way limiting the systematic uncertainty to the statistical uncertainty of the SS data sample. Table 5.2 collects the predicted yields for the four categories separated by process. Observed yields are also given.

The approach of the fit consists of defining two free parameters, and , modifying the contributions of the W cq ttWcq stWcq and W uq ttWuq stWuq processes to fit the data minimising event yield differences. The measurement $^{W}_{c}$ will be

extracted by relating it to the fit parameters and . The parameter plays as the global rate modifier for the ttWcq, ttWuq, stWcq and stWuq processes, where a W boson decays hadronically,

(5.1)

The parameter is the rate modifier of the charm processes ttWcq and stWcq samples and it is also anticorrelated with the light processes, ttWuq and stWuq. We do so by setting the charm processes W cq rate to and the light processes W uq to (Eq. 5.2). This way, charm and light rate variations within the fit are constrained so that the total yield, light plus charm (W cq W uq), is not modified:

X = 2-r (5.2)

MC generated with

(5.3)

By including the parameter in the fit, the global normalisation of the related processes is left free to fluctuate, being the ruler for the relative proportions of W uq and W cq processes. Eq. 5.3 illustrates how W cq and W uq samples can be modified by the defined parameters. The branching fraction ratio $^{W}_{c}$ to be measured will therefore be $^{W}_{c}$ = , as deduced in Eq. 5.4:

$^{W}_{c}$ (5.4)

The results of the fit for the parameters of interest are the following:

(stat) (syst)
(stat) (syst)

The global uncertainty of the parameters is dominated by the systematic uncertainty. The statistical uncertainty for r is 1% and 0.1% for . This reflects the event counts of the channels used in the fit with charm-tagged events (about 20000) and without charm tag (close to 1 million events). The relative total uncertainties are 4% for and 7% for . Since represents the relative rate between W cq and W uq, systematic uncertainties affecting both equally cancel out. As a result, has a smaller relative total uncertainty compared to . On the other hand, the statistical error is larger for because channels

Process	Prompt no charm tag	Prompt e no charm tag	Prompt charm tag	Prompt e charm tag
$t\bar{t}$, W cq	245 816(7%)	151 570(7%)	8172(9%)	4993(9%)
$t\bar{t}$, W uq	257 789(7%)	159 146(7%)	150(9%)	84(9%)
Dileptonic $t\bar{t}$	31 343(7%)	19 219(7%)	299(8%)	188(8%)
Single top, W cq	5060(7%)	3085(7%)	133(10%)	93(10%)
Single top, W uq	4772(7%)	2948(7%)	2(50%)	2(50%)
Single top, no W $q\bar{q}$	3620(13%)	1884(13%)	15(20%)	9(50%)
V+jets	5005(12%)	3687(12%)	43(30%)	9(50%)
Diboson	299(12%)	142(12%)	1(50%)	1(50%)
Total predictions OS SS			8815(9%)	5379(9%)
Data SS			2551(2%)	1546(2%)
Total predictions	553 705(7%)	341 681(7%)	11 366(7%)	6925(7%)
Data OS	553 378	341 232	11 167	6806

Table 5.2: Observed and predicted event yields input to the fit for the four categories. Predictions are separated by process. For the two categories with charm tag, the yields predicted by the simulations correspond to OS SS subtracted events, the SS contamination is estimated with data, and the number of observed events in data corresponds to OS events. The relative uncertainties shown in parenthesis for the predictions are based on the statistical uncertainties of the MC samples and the systematic uncertainties and their correlations discussed in Sec. 4.6.

3 and 4, with significantly lower statistics, are only extra information for but key for .

Figures 5.1a and 5.1b display a scan of the likelihood values from the optimization process near the optimal point. The curves are presented for two scenarios: one considering only statistical uncertainty and the other incorporating all uncertainties, including systematic uncertainties.

The impact on the uncertainty of the fitted parameters due to the systematic effects taken into account in the fit, and detailed in Sec. 4.6, is displayed in Fig. D for and in Fig. 5.3 for . As expected, the systematic uncertainties with a higher impact on the total uncertainty of are those related to charm tagging, namely, the uncertainty in the reconstruction and identification efficiency of the muon in the jet (`charmtag_muonID`), and the uncertainty in the rate of muons from the decay of charm hadrons in the simulation (`charmtag_muonRate`).

Figure 5.3 illustrates that the largest impacts on correspond to systematic sources that show the most variation in yields, as detailed in Table 4.11 (e.g., `JetEnergyScale`, `FSR_syst`, `Bot_tag`, ...). Since serves as a global rate modifier for the ttWcq, ttWuq, stWcq and stWuq processes, it is inherently affected by these significant sources of uncertainty.

The obtained value is expected since the predictions of the MC simulations were normalized to the observed data prior to performing the fit. The obtained value reflects the difference in normalization between data and the predictions for the charm-tagged samples, as can be seen in the ratio panel of figures 4.22 to 4.26.

The resulting value for $\mathcal{R}_c^W$ is:

$\mathcal{R}_c^W$ (stat) (syst)

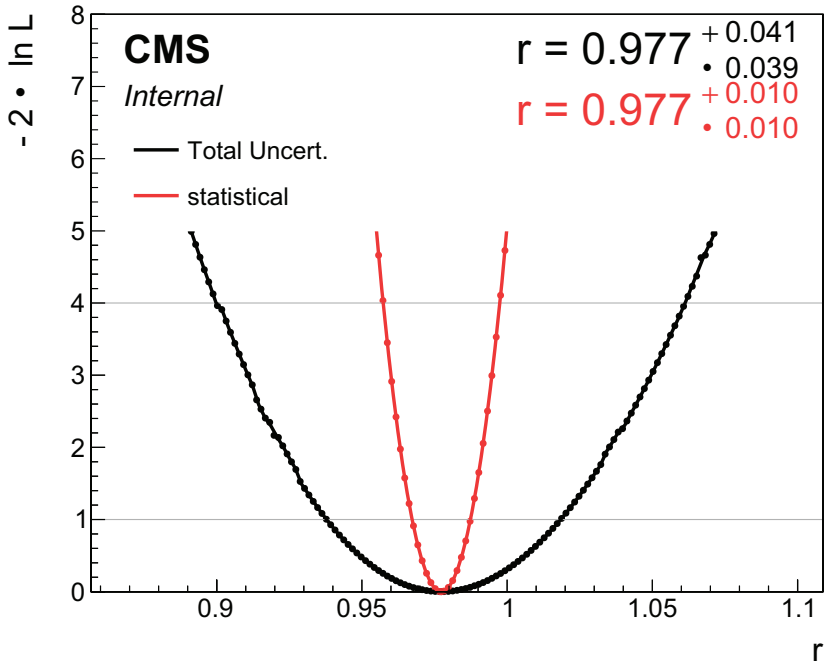
	Impact on $\mathcal{R}_c^W$ [%]
Charm tagging: muon identification	2.6
Charm tagging: muon rate in simulation	2.1
Parton shower final state radiation	1.9
Jet energy scale	0.6
SS data statistical uncertainty	0.5
Charm fragmentation modeling	0.3
Jet energy resolution	0.3
b tagging	0.2
MC background normalization	0.1
Integrated luminosity	0.1
Total	3.9

Table 5.3: Summary of the main impacts of the uncertainty sources, expressed in percentage of the measured $\mathcal{R}_c^W$ value.

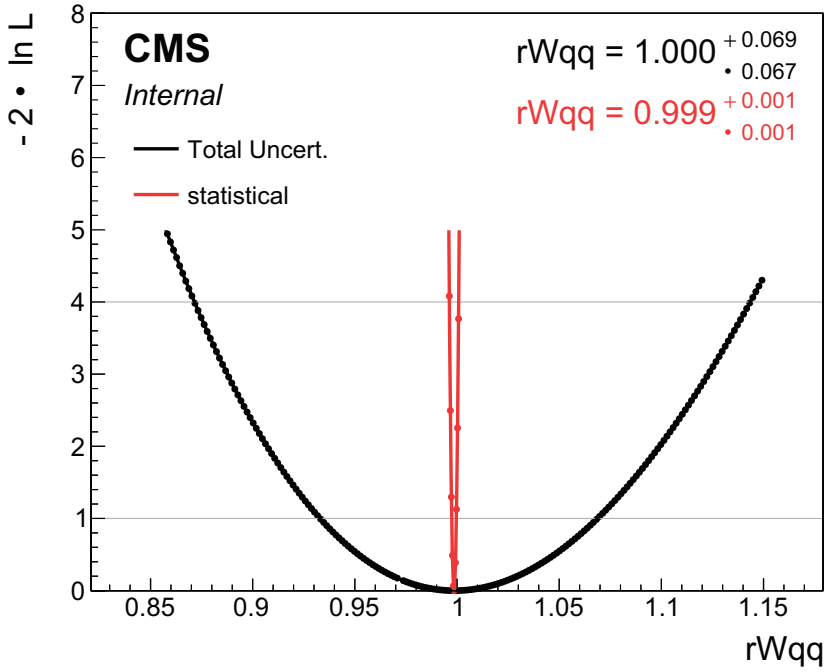
The total relative uncertainty in the determination of $\mathcal{R}_c^W$ is 4%. The systematic uncertainty dominates the precision. The statistical precision is 1%, resulting from the number of events in the charm tag categories used in the fit. The main systematic uncertainties in the measurement of $\mathcal{R}_c^W$, reflecting those of the r fitted parameter, are listed in table 5.3. This result is in good agreement with the prediction of the SM. The precision of the measurement, limited by the systematic uncertainty in the charm tagging efficiency, is improved by a factor of two compared to the world average value (see Fig. 5.4).

Using Eq. 1.17 and the measured value by CMS of the sum of squared elements in the first two rows of the CKM matrix (, determined from the measurement of the W boson leptonic decay branching fractions [102]), the sum of squared elements in the second row of the CKM matrix can be derived. The obtained value, , provides a consistency test of the CKM matrix unitarity. In addition, using the measured value of and the world average values of and [18], a value of is obtained.

For each of the four fit channels, we show in Figs. 5.5, 5.6, and 5.7, post-fit plots for relevant kinematic distributions. The data is well described by the simulation within systematic uncertainties.



(a)



(b)

Figure 5.1: (a) Scan of the likelihood function of the r parameter in the fit. Black dots represent the fit including all uncertainty sources and pink dots represent the case of freezing all systematic effects. (b) The same scan is computed for $r_{W_{qq}}$. In this case the statistical uncertainty is smaller.

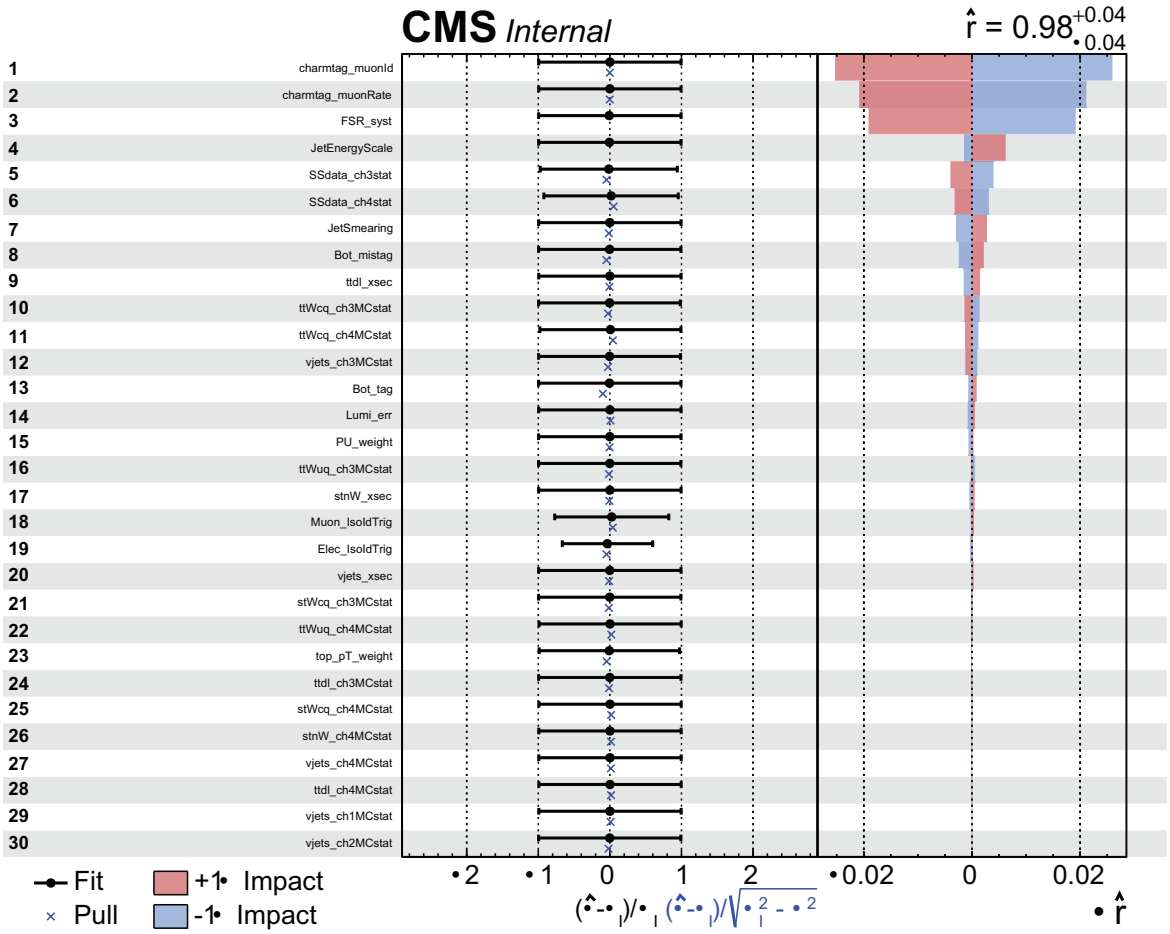


Figure 5.2: Impact of the various systematic uncertainties in the determination of r. The dominating systematics are the uncertainty in the reconstruction and identification efficiency of the muon in the jet (`charmtag_muonID`), and the uncertainty in the rate of muons from the decay of charm hadrons in the simulation (`charmtag_muonRate`).

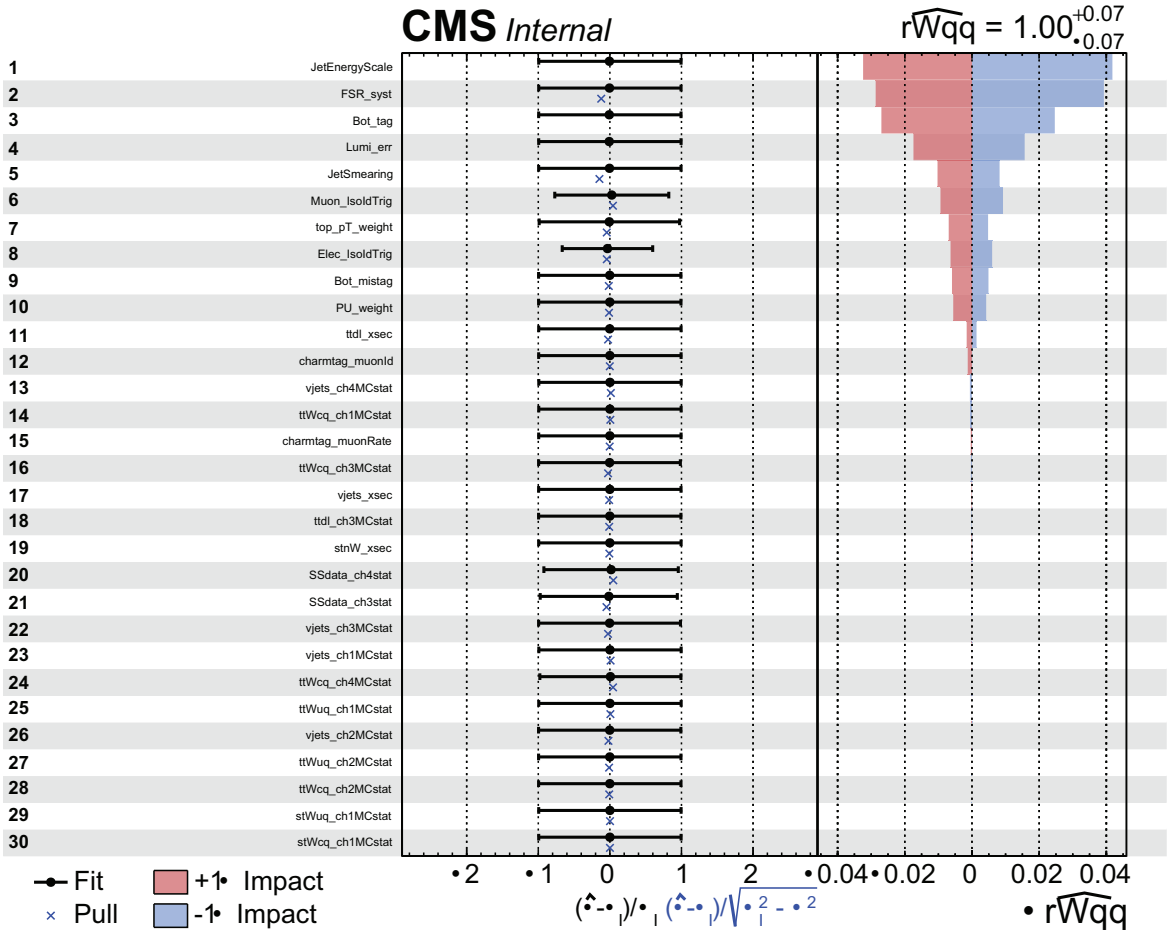


Figure 5.3: Impact of the various systematic uncertainties in the determination of $r_{W_{qq}}$.

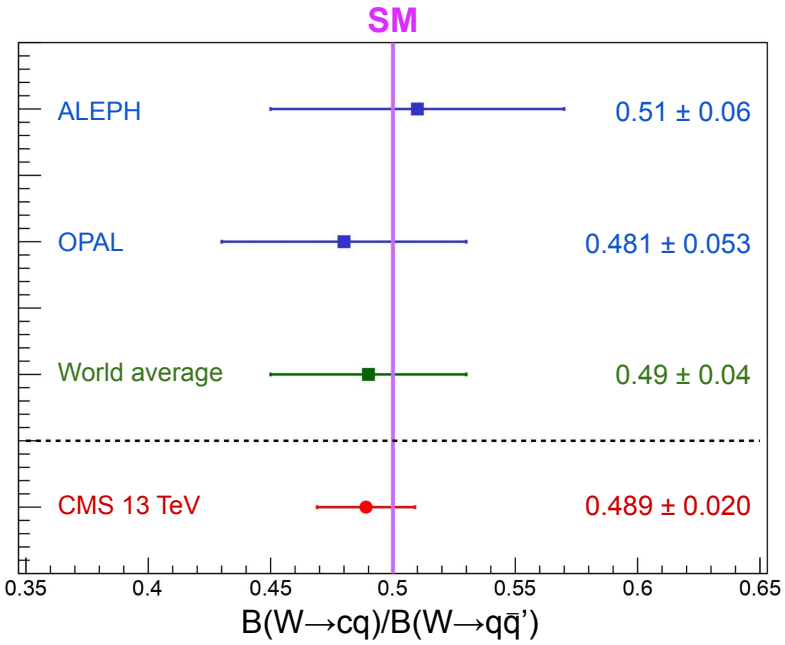


Figure 5.4: Comparison of the measured value of $\mathrm{W}_{\mathrm{c}}^{\mathrm{W}}$ with previous LEP2 measurements, and the world average value.

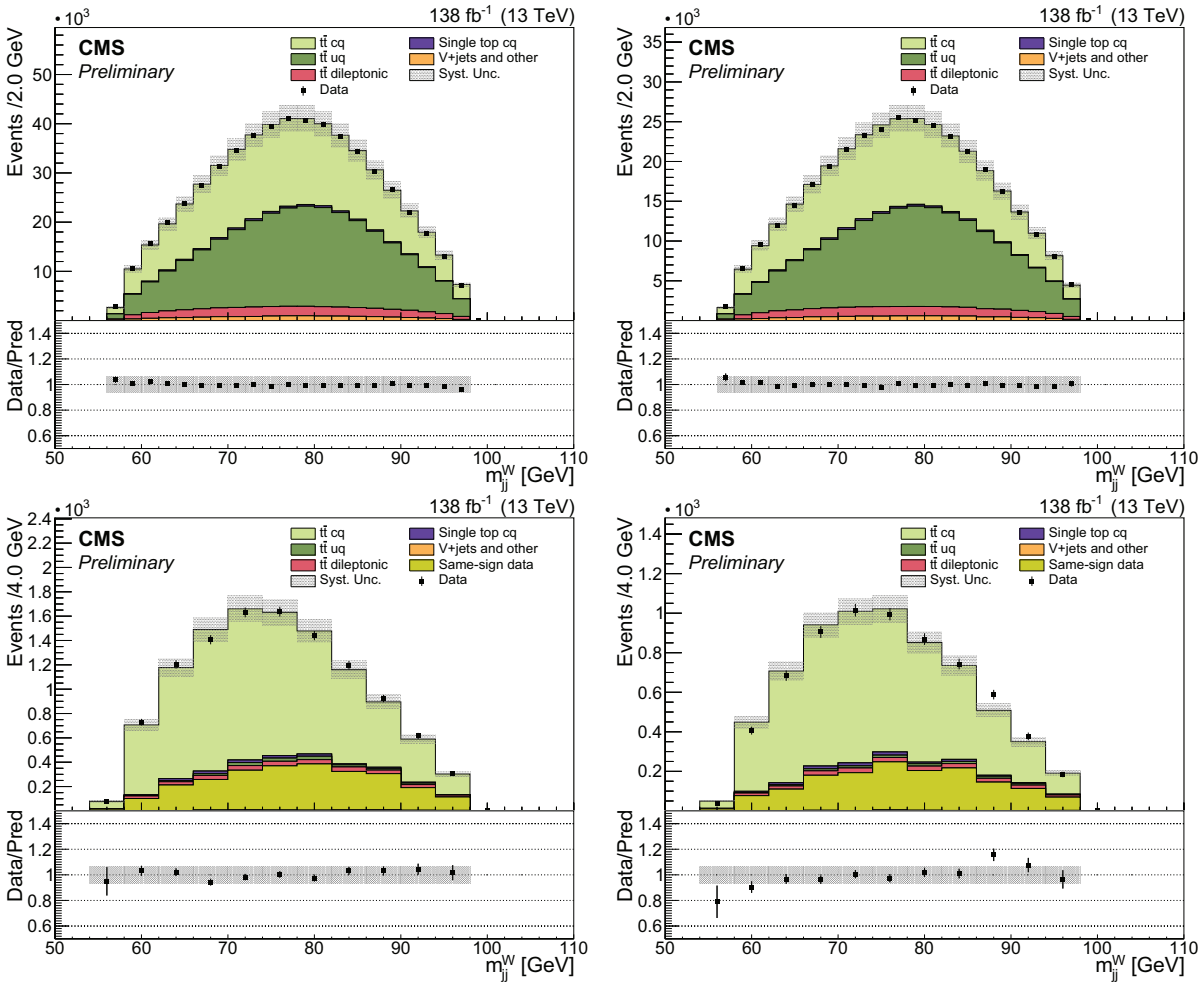


Figure 5.5: Invariant mass of the two jets reconstructing the W boson. The top left plot corresponds to channel 1, top right to channel 2, bottom left to channel 3, and bottom right to channel 4.

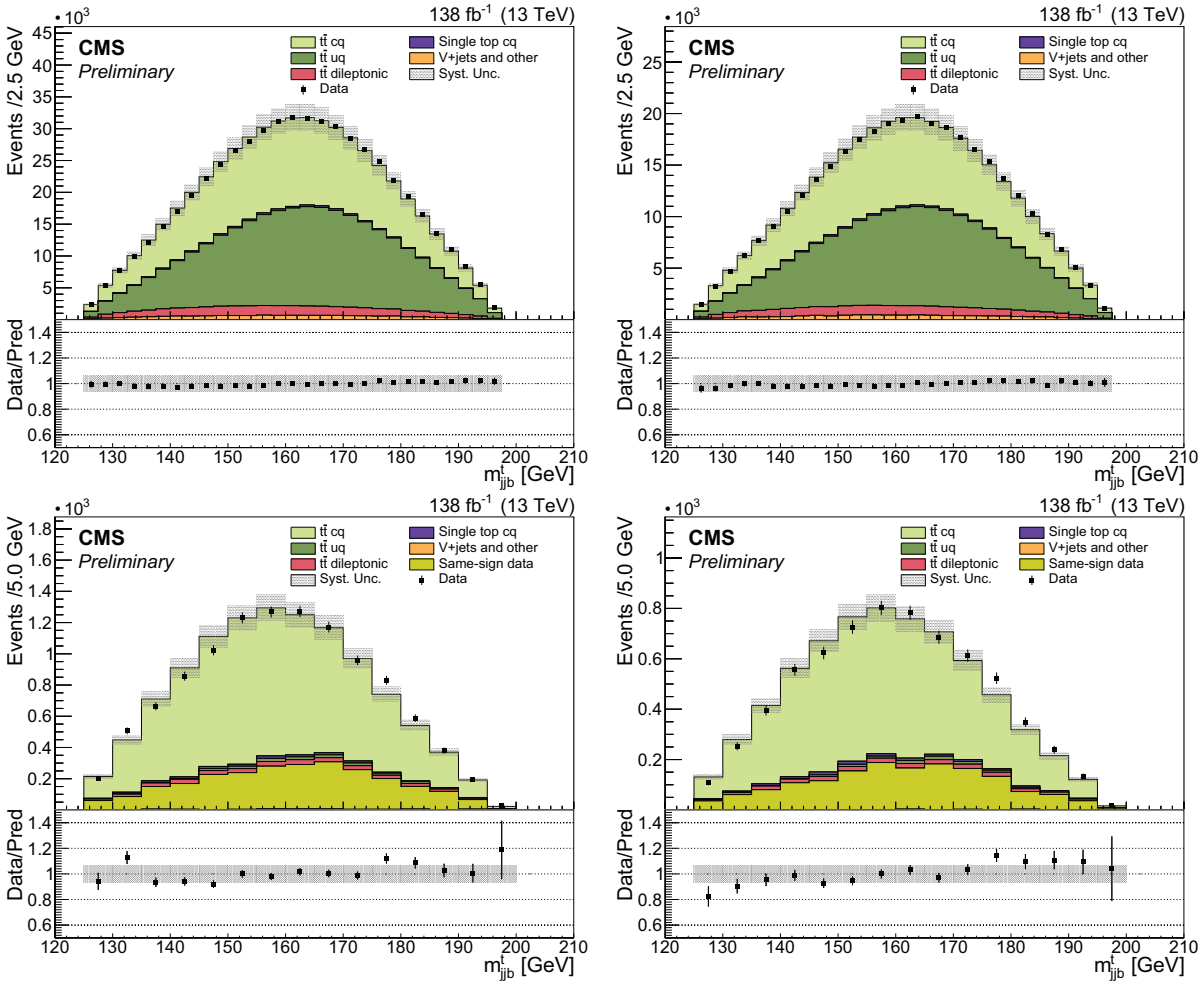


Figure 5.6: Invariant mass reconstructing top quark. The top left plot corresponds to channel 1, top right to channel 2, bottom left to channel 3, and bottom right to channel 4.

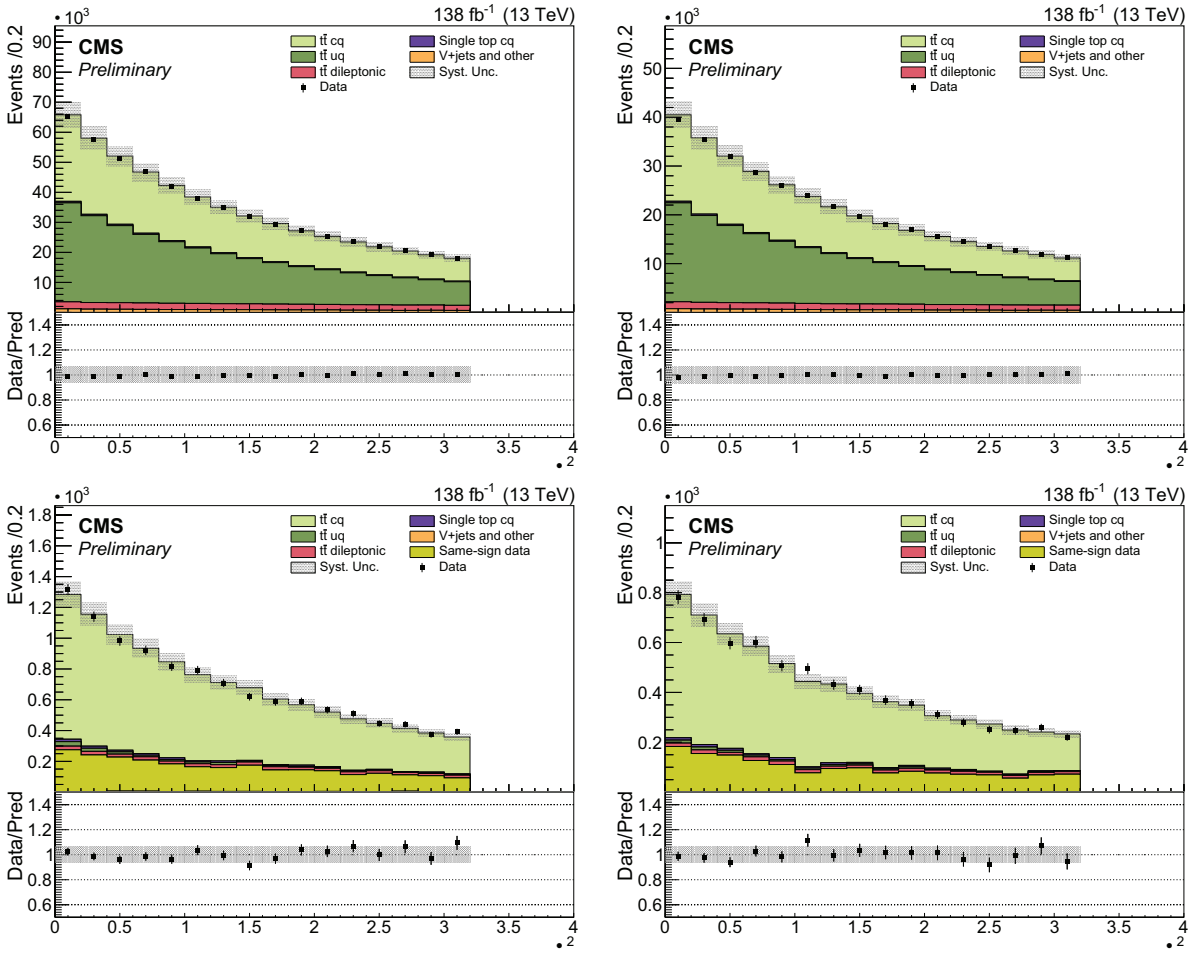


Figure 5.7: χ^2 distribution of the $t\bar{t}$ kinematic reconstruction. The top left plot corresponds to channel 1, top right to channel 2, bottom left to channel 3, and bottom right to channel 4.

Chapter 6

Uncertainty estimation of AI results for particle physics

An alternative option for data treatment in HEP are Machine Learning (ML) techniques. These were first applied to high-level physics analysis in the 1990s and have been popularised ever since [103, 104, 105, 106]. High energy physics experiments usually involve the treatment of huge quantities of information. The management of such large datasets utilises large amounts of computing resources. Studying specific processes in these big datasets makes scientist design intricate analysis techniques to strategically isolate events of interest. The proportion of these type of signals can be really low, for example, only one in proton collisions at the LHC produces a Higgs boson. Identifying these tiny quantities of signal involves careful examination of these large datasets, and the development of complex analysis tools. ML techniques have been used to tackle this task [107, 108, 109] successfully improving classification results.

Uncertainty values must be provided when conducting measurements in experimental sciences, this includes the use of AI. Estimating the uncertainty the use of ML entails is a work in progress since doing it analytically is computationally prohitive. This chapter presents the contribution of this thesis. It consists of applying AI uncertainty estimation techniques to a particle physics classification problem, that resulted in the publication [110].

However, it is rare to see results of these ML applications coupled with model related uncertainty measures. They are usually presented as point estimations with no evaluation of possible errors. In Physics, as in any other experimental science, it is of great importance calculating the uncertainty of a result. Three methods obtained in the literature are used to do this task, Bayesian neural network approximation [111], probabilistic random forest [112] and local ensembles method [113].

6.1 Bayesian neural network approximation

Bayesian neural networks (BNN) [111] are deep learning models whose parameters are considered random variables. This way it is possible to consider the computation of uncertainty measures for its predictions. The problem when using these type of networks is that architectures with more than one hidden layer generally become intractable.

However there are some results [114] that show a correspondence between optimising BNNs and optimising a multilayer perceptron (MLP) with dropout applied. A MLP is an artificial neural network (ANN), a deep learning (DL) model whose task is to find a function that best fits , given some data [115]. To avoid overfitting regularisation techniques are often applied to DL algorithms, they avoid for the model to be excessively complex. Dropout is one of these techniques[116], it randomly shuts down some connections between neurons in a MLP model, see Fig. 6.1.

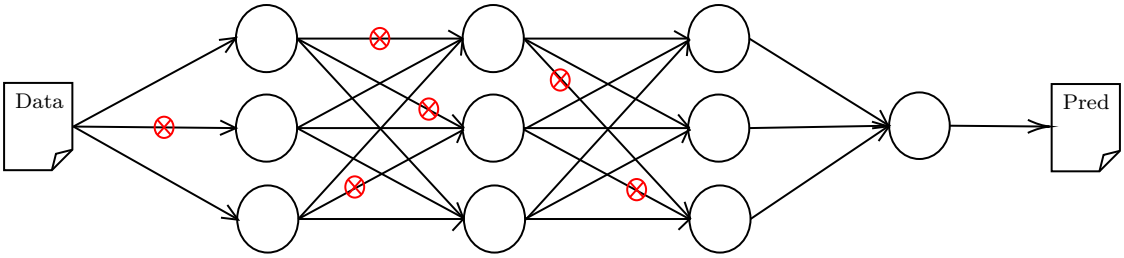


Figure 6.1: Illustration of a network with dropout applied. It consists of a MLP where the connections of the neurons are randomly shut down in the training process.

The loss function of a MLP with dropout and the approximation of the loss function of a BNN only differ in a constant, turning these optimisation processes equivalent. This way we can use an ordinary neural network to achieve the predictions with uncertainty that we would achieve with a BNN[1].

By applying dropout and training the MLP model various times, the expected predictions and variance associated to them can be obtained. These will be an approximation to those that we would have obtained by using a BNN [114][2].

Proposition 6.1.1 *Given* $\mathcal{N}$ [3] *for some can be estimated, if we train the MLP with dropout and make predictions for some test input times, with the unbiased estimator*

[1] This conclusion is reached in page 15 of the cited document [114], see equation 3.12.

[2] These propositions appear in page 19 of the cited document [114]. They correspond to propositions 2 and 3.

[3] where * * is the posterior probability distribution of the theoretical BNN prediction * conditioned to the output of the model * . We suppose that it follows a normal distribution with mean * and variances − , where is the identity matrix.

— (6.1)

where is the prediction of the dropout network for some new data , being the optimal weights in each training , and is the optimal predictive distribution for the BNN.

We can also estimate the second moment.

Proposition 6.1.2 *Following the notation from the previous proposition, given $\mathcal{N}$ for some , can be estimated, proceeding the same as before, with the unbiased estimator*

— (6.2)

Having the second moment we can obtain the variance of the prediction as:

— (6.3)

These two results imply that we can estimate the expected predictions and their variance of a BNN with a MLP with dropout applied. Using Bayesian neural networks to obtain uncertainty measures would led us to intractable operations. The alternative of using stochastic regularisation techniques, dropout in this case, offers an approachable solution.

The architecture used in this thesis is a network with 3 hidden layers, each of them consisting of 64, 32 and 16 neurons. The activation function for the hidden layers neurons is *relu* [117] whereas the activation function for the output layer is *sigmoid* [118]. The batch size was 128 and the number of epochs for training 50. L2 regularization is implemented as well, given that its cost function is differentiable and therefore computationally more efficient. The specific code can be found in the reference [119].

6.2 Probabilistic random forest

Another approach for obtaining uncertainty measures is probabilistic random forest (PRF). It is a variation on the concept of a random forest (RF) algorithm. RF is an ensemble learning method that trains several decision trees. Taking the predictions of each decision

tree the final RF prediction is computed with the mode for classification problems and averaging for regression. A decision tree takes this name because the model can be visualized as a succession of tree-like graph decisions, it is a non-parametric hierarchical model. The output of the model will be obtained by checking if a series of conditions are fulfilled or not by the input information. For example, if the input is an observation with real variables the conditions take the form , where is the input variable and is a fixed value for that variable to use as limit in the node of the tree. Fig. 6.2 shows a diagram of how it works, further details on RF models can be found in [120].

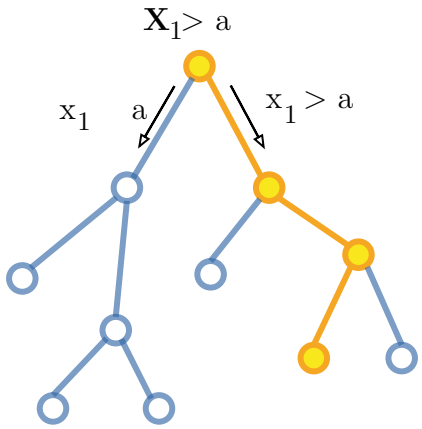


Figure 6.2: Decision tree illustration. It is displayed how the first node works if the associated condition applies for a real continuous variable.

In an ordinary RF process of modelling the output is deterministic, a unique result for each input observation. PRF reconsiders the procedure, drawing from a RF-model, in order to be able to offer better prediction abilities. The data used in RF are observations taking the form with . The goal is to learn a model sufficiently good so that predictions for new data can be carried out. For PRF case, instead of treating input variables as fixed single values, input data will be considered to be random variables with probability distributions associated. To do this observations needs to be accompanied by uncertainty values, in the form where and represent the error of the measurements. This way probability distributions can be constructed for each satisfying their expected value to be the provided single value and its variance the uncertainty squared. If the uncertainties of all data are sufficiently small, this PRF method would converge to an ordinary RF model.

To illustrate the mechanism of PRF we will use a classification problem as example. A regular RF algorithm takes an observation and tries to predict that is the class it belongs to. Depending on the fulfillment of conditions of the type for the variable at the node the event will propagate through specific branches discarding the others, as it is displayed in Fig. 6.3. PRF on the contrary computes the said probability distributions, so for a condition of the type its probability is computed integrating the area of the function, and also . This is illustrated in Fig. 6.3.

Since no branch is discarded, instead of choosing one path or the other two quantities

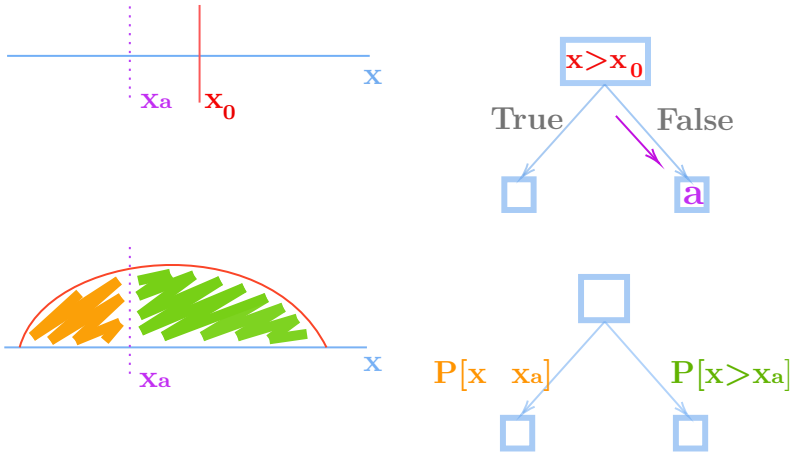


Figure 6.3: Scheme of the different variables treatments, adapted from [112]. Above the classic point variable value with the condition imposed in a regular decision tree. Below the input value treated as a random variable and the process in the probabilistic decision tree, where instead of choosing one branch or another, the probability of propagation through both branches is calculated.

are computed, all nodes are taken into account. These values go through the whole tree and in each node again two probabilities are computed and multiplied accordingly. The final result consist of a probability in each end of the tree, all adding 1. RF determines a unique path ending in a prediction of a class for an event, PRF results in a series of probabilities, one must sum all for each class and we will have the probability of belonging to each class. Fig. 6.4 displays this distinctions in methodology. More information about the specifics can be found in [112]. This particular technique can become inefficient so some approximations are added, calculating the probability of each final node of a tree can be computationally expensive so some branches are removed. Branches with a probability considered negligible[4] are discarded in order to moderate computation time.

By having the probability of an event belonging to a class or another we can compute the expected class and its variance[5]. The input uncertainty information is successfully transferred to the model output, creating a more complete consideration of data. This method also allows for obtaining the desired predictions with uncertainties when resorting to ML.

[4]Imagine that for a node the probability of fulfilling the condition is 0.9999999 and not then $^{-}$. If the results we want do not need to have $^{-}$ precision, then it is reasonable to not propagate further that section of the tree since any probability arising from it will be $^{-}$.

[5]For example, in a binary signal (class 1) vs background (class 0) classification problem we have probability of belonging to class 1 and for class 0, so the expected class will be with an uncertainty of .

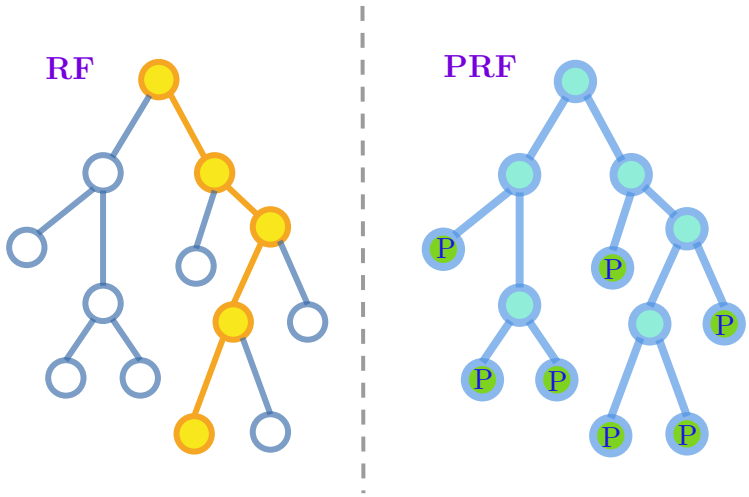


Figure 6.4: Illustration of the different procedures in a regular Random Forest and in the probabilistic model. Adapted from [112].

6.3 Local ensembles

The last method used for this study is Local Ensembles [113]. It also based in the use of artificial neural networks. The input information for the selected neural network is named , where is a feature vector and the goal label to predict. We then define a loss function $\mathcal{L}$, dependant on this input and some parameters characterizing the prediction model. Additionally will be the parameters values such that arg $\mathcal{L}$ given and is the prediction of model given . Test feature vectors that have not been used for training are noted as .

When predicting the target for test observations a formula, the *Extrapolation Score*, is able to quantify the extrapolation ability of the model. To do so the algorithm obtains a set of models varying an already trained model, consistent with the training data, and then the variability of outcomes is computed. The *Extrapolation Score* is applied to a test observation and has the expression in Eq. 6.4 where is the gradient of the prediction of with respect to the model's parameters evaluated in the optimal point .

Being the Hessian matrix of the loss and its spectral decomposition, the matrix of orthonormal eigenvectors of , then is defined as a matrix of Hessian eigenvectors corresponding to a subset of eigenvalues small enough so the training loss is not changed significantly. These eigenvectors form an *ensemble subspace* that represent directions of low curvature of the model.

$$\mathcal{E} \tag{6.4}$$

This extrapolation score turns out to be proportional [113], in first order, to the standard deviation of the prediction of the corresponding observation 6.3.1.

Proposition 6.3.1 *Let be the projection of a random perturbation with mean zero and covariance proportional to the identity into the ensemble subspace spanned by . Let be the linearized change in prediction induced by the perturbation*

(6.5)

then $\mathcal{E}$.

To implement this method a practical computational method developed by the main authors is used. Given a training dataset this method returns the extrapolation score for input test events. This *Extrapolation score* is interpreted as a way of estimating the variance of the predictions if we had trained the model ordinarily.

The calculation of Hessian matrices and their spectral decomposition is computationally expensive when treating large matrices. For that reason, while the author's implementation was made in an efficient way, we chose a simpler model than the one used for the Bayesian approximation method. We expect then to yield a worse nominal classification performance. Nonetheless the goal of this exercise is to evaluate the uncertainty of the result, not the result itself.

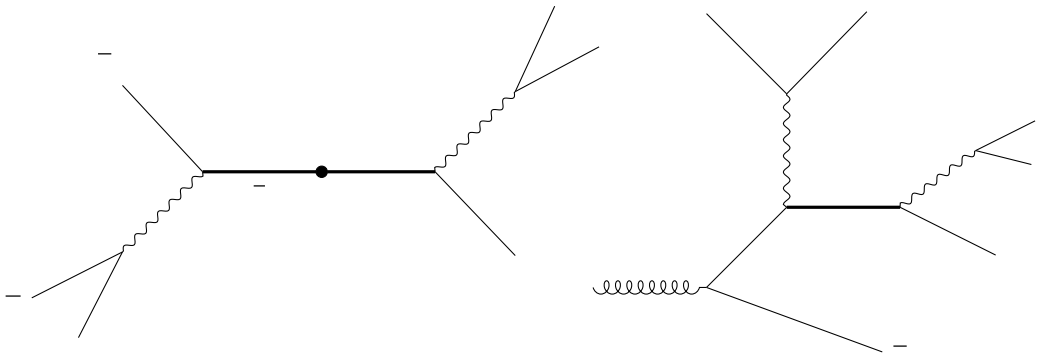
6.4 Results

These three methods are tested using a public CMS dataset [121]. It consists of simulation events of a series of processes that generate similar features at detection level. The process of interest will be $t\bar{t}$ events, the goal is to label observations as signal or background, a binary classification problem. The signature features of a top-antitop signal event, see Fig. 6.5a, are the production of two jets originated from bottom quarks, and lepton or extra quarks depending on the decay of two W bosons also produced in the process.

The other processes treated as background include single top events, see Fig. 6.5b, vector bosons (W or Z) plus jets and dibosons (WW, WZ and ZZ).

A set of variables is selected for the model input. The chosen features are kinematical and topological magnitudes constituting distinctive characteristics of the processes. The number of leptons and jets detected is included as well as their transverse momentum and their pseudorapidity . Additionally, a discriminating variable identifying bottom jets is included since the signal process involves the production of two b-jets. If one of the W bosons of a top-antitop event decays leptonically the produced lepton should be detected with little activity surrounding it, resulting in an isolated particle. For this reason an isolation variable for reconstructed leptons also takes part in the dataset. Neutrinos are also taken into consideration by adding the missing transverse energy of the event. The final dataset comprises 35 training variables.

As mentioned the goal of this exercise is to classify events whether they are $t\bar{t}$ or not. The target variable is then the label of the event, 1 if it is signal or 0 if it is background. An additional variable is considered, the *weight* of each event, when combining all events



(a) Signal event. **(b)** Background event.

Figure 6.5: Feynman diagrams displaying processes described by the dataset used.

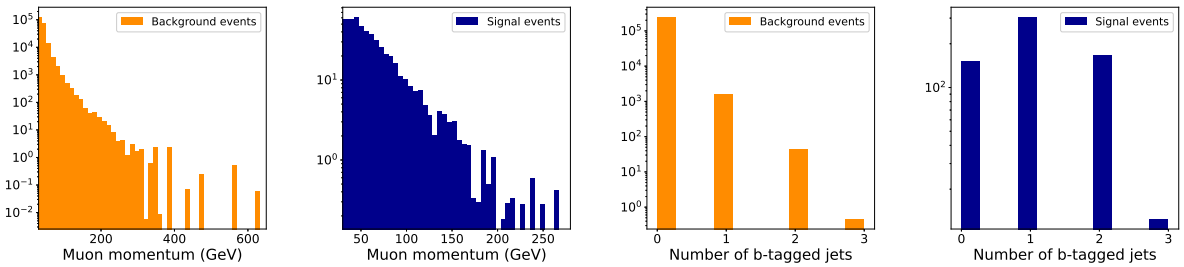


Figure 6.6: Pairs of histograms differentiating between signal and background. From left to right, the muon momentum distributions for background events, then the same for signal events, number of b-tagged jets for background and lastly for signal events. The shape of these variables is different for signal and background, these images display signatures subject to be exploited by the classification algorithms.

it corrects proportions between them to match those dictated by nature. Signal $t\bar{t}$ events occur about one thousand times less often than background events. This weight also correct background processes rates, they have different production probabilities. This weight is not used for training the models, it will be only used when displaying results of testing such as plots or final efficiencies.

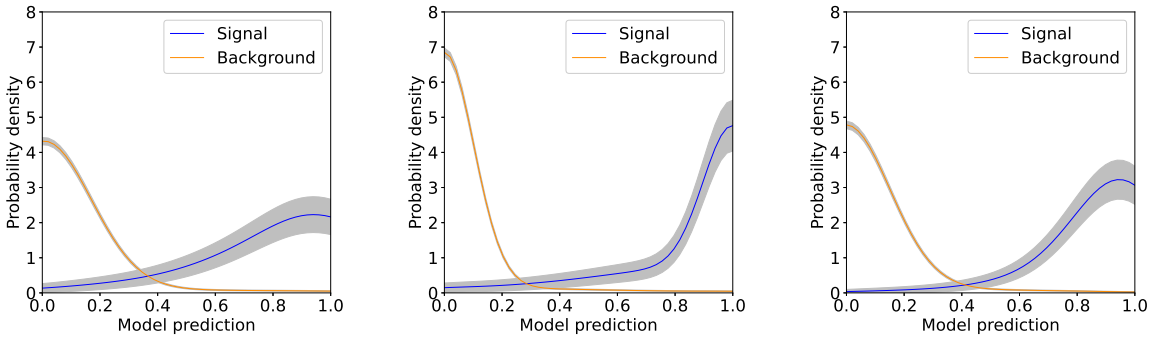
The dataset contains a total of 200 events, divided into a training set ( ) and a test set ( ). With the initial proportions this results in the training set having only 3.5 of signal events. This can entail a poor training because the neural network works in batches of observations, so in many of them there might not be any signal events. If this happens frequently enough, the network's weights are updated so that the identification of signal events is disfavoured. To correct this the background is undersampled in the training set. Events labeled as background are removed randomly to obtain a set consisting in 40 signal and 60 background events. By training the models with this undersampled dataset allows for obtaining more reliable predictions, solving the explained problem.

The three methods mentioned before are applied to this binary classification problem

Figure 6.7: Scheme illustrating the undersampling method, where the orange part symbolizes background and blue corresponds to signal. The complete dataset is split in test and train sets, maintaining the signal-background proportion and then background events are dumped from the training set to balance the signal-background proportion.

using this dataset and the results are presented. In the case of the Bayesian approximation method we developed an implementation, it can be found in [119]. For PRF and LE methods we used the already available implementations [122, 123].

The test set is used to asses the classification performance of the three algorithms and the estimation of its associated uncertainty. Plots shown from now on are computed used the mentioned weight of each event to match relative probability of occurrence in nature of the physics processes. This ensures that the results, obtained with simulated events, are representative of the real data recorded by the CMS detector and conclusions can be trusted when testing real events with the classifiers. Details on the calculations appearing for performance metrics can be found in the appendix D.



(a) Local Ensembles **(b)** PRF **(c)** Bayesian approximation

Figure 6.8: Distribution of label values for the different methods normalised to unit area. Distributions are plotted separately for signal ($t\bar{t}$) and background events with the model predictions using the test set. The shaded areas correspond to the predictive uncertainty.

Fig. 6.8 shows the probability density distributions of the classification parameter returned by the model. The distributions are divided in signal, blue, and background, orange. The prediction of the models is considered a continuous random variable , these plots represent then their associated density functions computed for the whole test set.

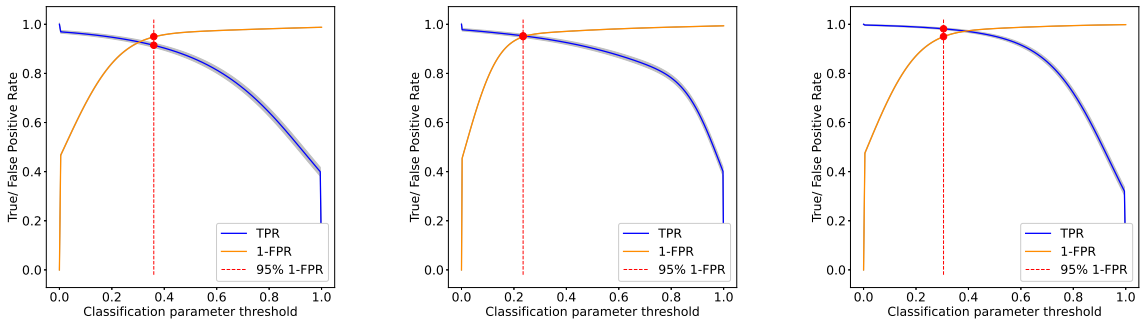
The figures show a clear separation between signal and background in the three cases. This indicates a great discrimination power of the models, the outcome of this work demonstrates an excellent performance of the algorithms. True signal events predictions accumulate towards value 1 whereas true background events tend to score close to 0. The groundings for this uncertainty study is a problem with a successful solution, what variance arising from the use of ML do we expect from a classification problem that ML tools can solve with ease? These good results happen for each case, having the LE method the largest overlap between signal and background distributions. Local ensembles model, compared to the Bayesian approximation neural network, had to be constructed in a simpler form for computation time reasons. This circumstance produces a worse differentiation capability for LE.

The shaded bands in the plots correspond to the variance of the values obtained from the uncertainty estimation of each method. It is worth noting that the variance in signal distributions is larger than that of the background. The calculation of the variance of the bins depends on the test set. In this case the proportion of signal events is really low hence the disparity of variance size. In order to check the source of this effect, we repeated the inference process but using a balanced test set instead (the original test set has a signal proportion of after weights are applied). For this case the variances in signal and background curves become similar when using a similar number of signal and background events in the test set. We conclude that the extent of the uncertainties in this specific plot depends on the test set configuration.

Even with uncertainties the curves have enough separation to differentiate between signal and background. These uncertainties associated to the classification performance metrics derived from the plotted distributions turn out to be arguably small, being also quite similar for each of the three methods tested.

True positive rate (TPR) and false positive rate (FPR) values [124] can be computed if we fix a threshold in the prediction distributions. One defines a event selection, keeping only those having a prediction above this threshold, then TPR is the quotient between true signal events classified as such by the model and the total number of true signal events and FPR is the quotient of true background events classified as signal (i.e., background events that score higher than the threshold value) over the total number of background events. Fig. 6.9 showcases TPR and 1-FPR values plotted against the value of the classification parameter threshold chosen. The estimated uncertainties are represented as the grey shaded bands. These are really small, the possible shifts for the curves are almost unnoticeable. Details on the way TPR and FPR and the associated uncertainties are evaluated can be found in the appendix D.

A useful reference is the TPR value corresponding to a threshold satisfying suppression of the background (FPR). Table 6.1 displays the achieved values for every method with associated uncertainties. This variability is computed using the TPR uncertainty grey band in Fig. 6.9. Results show that the three methods perform exceedingly well, retaining above 90% of the signal events when asking for a maximum of 5% of background contamination levels. The best model appears to be the Bayesian approximation (BA) with a retention of of signal events, PRF model maintains and



(a) Local Ensembles **(b)** PRF **(c)** Bayesian approximation

Figure 6.9: TPR and 1-FPR values as a function of the cut classification parameter, for the three different models. Vertical red dotted lines corresponding to 95% suppression of the background (FPR) are also plotted. The intersection of those lines with the TPR curve gives the corresponding signal preservation efficiency.

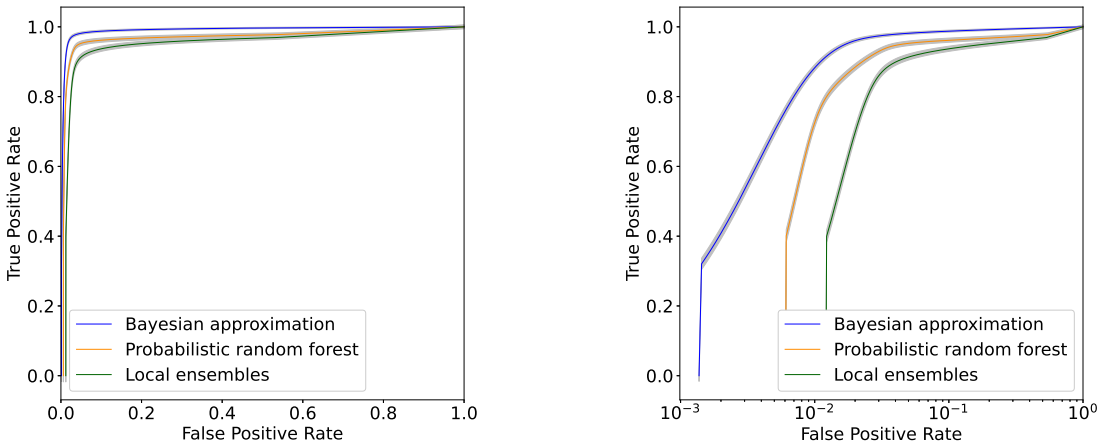
LE . All the variances accord to be very small (below 1%) for every method even when they do not share the same baseline model, one is random forest based and the other two are neural network implementations. This situation indicates high reliability for the results, for a problem expected to behave good such as this, the uncertainty related to using a ML tool is small.

Model	TPR for FPR=
Probabilistic Random Forest	
Local Ensembles	
Bayesian Approximation	

Table 6.1: TPR values (signal preservation efficiency) corresponding to a background suppression of 95% (FPR , or False Negative Rate equal to 5%). The uncertainty in the values is extracted from the shaded bands in Fig. 6.9.

The diagnostic ability of the chosen clasifying models can be evaluated resorting to ROC (receiver operating characteristic) curve [124]. This function consists of plotting TPR values against FPR ones, pairing them for the same threshold. Results are shown in Fig. 6.10a, displaying the ROC curve of each of the three methods. Since they are quite similar the same curves with X-axis in logarithmic scale are included, this way small differences for low FPR values are better assessed. The uncertainty of these curves correspond to the grey bands in the plot. These, just as other metrics mentioned, turn out very small, they are also considered robust against variations of the models. The area under the ROC curve (AUC) measures the quality of the models as classification tools. We can interpret AUC as the probability with which a model predicts a higher value for a randomly chosen signal event than a randomly chosen background event, when the case is that the class label of a signal event is higher than that of a background event. This testing metric provides an objective measure of the quality of our binary classifier. A theoretical perfect classifying model would have the curve with maximum area underneath, area 1. The worst

case would be a straight line between the points and , area 0.5, corresponding to a completely random classifier. Table 6.2 reflects AUC values for the three methodologies. They all have a shape close to the ideal case, AUC values close 1, which makes them great classifiers according to this metric. Just as before best results are obtained for BA case along all values of the discrimination threshold, showing its excellent discrimination power.



(a) ROC curve. **(b)** ROC curve with x-axis in logarithmic scale.

Figure 6.10: ROC curves for the three models, in linear scale (a) and using logarithmic scale for the x-asis (b).

Model	AUC	
Probabilistic Random Forest	0.969	0.005
Local Ensembles	0.951	0.006
Bayesian Approximation	0.990	0.001

Table 6.2: Area under the ROC curve (AUC metric) for the three different models. The uncertainty in the values is extracted from the shaded bands in Fig. 6.10a.

Uncertainties obtained from the grey band of the plot are also captured in the table, presenting AUC values with their error. Again, this uncertainty is small, interpreting that when ML models are able to solve a classification problem with ease the uncertainty associated with the use of the model, the epistemic error, is small. Other sources of uncertainty that could affect the final model results would be those associated with the data used to train the models, but this work only considers the uncertainty arising from the models themselves.

Conclusions

The Standard Model provides a comprehensive framework for understanding the fundamentals of matter and its interactions, attempting to offer a view of the universe at the most basic level. The LHC and its experiments are crucial in testing and refining this theoretical framework. The LHC's high-energy environment allows for precise exploration of the Standard Model's predictions and the potential discovery of new phenomena.

The CMS experiment serves an important role in testing and refining the Standard Model, leveraging its advanced instrumentation and sophisticated analysis capabilities. With high-resolution tracking detectors, precise energy measurement through its electromagnetic and hadronic calorimeters, and a robust muon detection system, CMS excels in detecting and analyzing particle interactions. Its efficient trigger system selects significant events from the vast data generated by LHC collisions, while its cutting-edge data analysis tools enable detailed event reconstruction and particle identification. These capabilities are crucial for achieving precise measurements and exploring new physics beyond the Standard Model, as this thesis demonstrates by making use of the resulting data.

The study presented in this thesis probes the universality of the weak interaction in the quark sector, a fundamental aspect of the Standard Model. This principle asserts that the sum of all couplings of any up-type quark to all down-type quarks is the same for all three generations. Weak universality is verified by measuring the rate of charm quark production in W boson decays relative to the rate of W hadronic decays to different quark flavors, $_{c}^{W} = \mathcal{B}$ W cq $\mathcal{B}$ W q$\overline{q}$. $_{c}^{W}$ encodes weak universality for the first two up-type quarks (up and charm) and its expected value is 1/2. Therefore, a measurement of $_{c}^{W}$ is a direct test of weak universality. The measured value, $_{c}^{W}$, consistent with the Standard Model prediction, represents a significant advancement in precision. The achieved uncertainty for the measurement improves by a factor of two the precision of the world-average value, determined combining the measurements of LEP2 experiments.

This analysis is performed using a data sample of pp collisions at $^{-}$ TeV collected by the CMS experiment during the 2016–2018 data-taking periods with a total integrated luminosity of 138 fb . The large cross section of t$\bar{t}$ production at the LHC, each top quark decaying into a W boson and a bottom quark, offers a sizeable high purity sample of W bosons. The final state used for the $_{c}^{W}$ measurement is carefully designed to maximize the content of events where one of the W bosons decays leptonically into a lepton (electron or muon) and a neutrino, while the other W boson decays hadronically into two jets. The high transverse momentum lepton provides an excellent signature for the online selection of the events, while the identification of a charm jet enables the measurement of $_{c}^{W}$.

The simulations of the various processes contributing to the selected data sample have been meticulously studied and calibrated, resulting in an exceptional level of precision in matching the simulated predictions to the observed data.

Charm tagging plays a crucial role in the data analysis. We employed a charm tagging method pioneered by the CIEMAT analysis group, which is both simple and highly effective. This technique relies on identifying a muon within a jet, allowing for the selection of a pure sample of charm jets with minimal background. The background is well-characterized using a control sample derived from the charge correlation between the prompt lepton from the W boson decay and the muon inside the jet. We developed a calibration method to accurately determine the reconstruction and identification efficiency of muons within jets using data. The resulting systematic uncertainties associated with charm tagging, which dominate the precision of the $^{W}_{c}$ measurement, are well understood and smaller than those from the standard CMS charm tagging algorithm. The charm tagging technique employed in the analysis is crucial for improving the precision of the measured $^{W}_{c}$, advancing beyond the current world-average value.

Additionally, this thesis highlights the increasing importance of machine learning techniques in particle physics. By applying various methods—Bayesian neural networks, probabilistic random forests, and local ensembles—this work assesses their effectiveness in distinguishing between top-antitop signal events and various background processes. The results show strong performance across all methods, with the BA model achieving the highest efficiency, then PRF and lastly LE methods. This demonstrates the models' robust discriminative power and their ability to provide reliable predictions in challenging scenarios. Importantly, the uncertainty estimation methods used throughout the analysis reveal that the epistemic uncertainty arising from model construction is minimal. This small uncertainty enhances the reliability of the results and indicates that the models are well-suited for this classification task. This results reinforces the utility of machine learning in particle physics research, paving the way for future explorations in this dynamic field and illustrating how these tools can improve the accuracy of analyses while providing valuable insights.

Overall, the results and methodologies presented in this thesis represent an advancement in both precision measurements and the application of machine learning techniques in particle physics. The ongoing efforts of the LHC and the CMS experiment continue to refine our theoretical models and push the boundaries of our understanding, driving the field into new territories.

Conclusiones

El Modelo Estándar nos proporciona un marco teórico para poder entender los fundamentos de la materia y cómo ésta interactúa entre sí, en un intento de visualizar el universo en sus niveles más básicos. Los experimentos del LHC son cruciales a la hora de poner a prueba y refinar esta teoría. El LHC ofrece un entorno de altas energías que permite la exploración precisa de las predicciones del Modelo Estándar y el potencial descubrimiento de nuevos fenómenos.

El experimento CMS juega un papel importante en esta tarea, aprovechando su avanzada tecnología y sofisticadas sus capacidades de análisis. Al contar con detectores de alta resolución, medidas de energía precisas gracias a sus calorímetros electromagnéticos y hadrónicos y un sistema de detección de muones robustos, CMS sobresale en la detección y análisis de la interacción de partículas. Su eficiente sistema de trigger selecciona eventos de interés entre las grandes cantidades de colisiones que se producen en el LHC, a la vez que sus herramientas avanzadas de análisis de datos permiten la reconstrucción detallada de eventos y la identificación de partículas. Estas características son de vital importancia para poder obtener medidas precisas y explorar la nueva física más allá del Modelo Estándar, tal como esta tesis demuestra al hacer uso de estos datos.

El trabajo presentado en esta tesis pone a prueba la universalidad de la interacción débil en lo que se refiere a quarks, un aspecto fundamental del Modelo Estándar. Este principio sostiene que la suma de todos los acoplamientos de cualquier quark tipo *up* es la misma para todas las generaciones de quarks. La universalidad débil puede ser verificada si se mide la tasa de producción de quarks charm en decaimientos de bosones W relativa a la tasa de decaimientos hadrónicos de bosones W a cualquier quark, $\mathcal{R}_c^W = \mathcal{B}(W \to c\bar{q}) / \mathcal{B}(W \to q\bar{q})$. La medida $\mathcal{R}_c^W$ codifica la universalidad débil para los primeros dos quarks de tipo *up* (up y charm), siendo su valor esperado de 1/2. Por lo tanto, una medida de $\mathcal{R}_c^W$ es una prueba directa de la universalidad débil. El valor medido, $\mathcal{R}_c^W$, consistente con la predicción del Modelo Estándar, constituye un avance significativo en precisión. La incertidumbre final de la medida mejora en un factor dos la precisión del valor promedio mundial, determinado con una combinación de medidas de experimentos de LEP2.

Este análisis se ha llevado a cabo con datos de colisiones entre protones a $\sqrt{s}$ TeV tomados en CMS durante el periodo 2016-2018 con una luminosidad integrada total de 138 fb^{-1}. La sección eficaz de producción de $t\bar{t}$ en el LHC, donde cada quark top decae en un bosón W y un quark bottom, al ser grande, nos permite tener una muestra de bosones W de alta pureza. El estado final usado para la medida $\mathcal{R}_c^W$ está cuidadosamente diseñado para maximizar la cantidad de eventos en los cuales uno de los bosones W decae

leptónicamente a un leptón (electrón o muon) y un neutrino, mientras que el otro bosón W decae hadrónicamente, formando dos jets. El leptón de alto momento proporciona un excelente indicador para la selección de eventos, mientras que la identificación de un jet charm permite llevar a cabo la medida.

La simulación de varios procesos que contribuyen a la muestra de datos escogida ha sido estudiada con detalle y calibrada, resultando en un alto nivel de precisión para el acuerdo de ésta con los datos observados.

El etiquetado de eventos de tipo charm es crucial en este análisis. Se ha utilizado una técnica desarrollada por el grupo CIEMAT, que es simple y efectiva. Esta técnica se basa en identificar muones dentro de jets, permitiendo una selección de jets charm pura con fondo mínimo. El fondo está bien caracterizado usando una muestra de control derivada de la correlación de carga existente entre el leptón proveniente del bosón W y el muon dentro del jet. Se ha desarrollado un método de calibración para poder reconstruir e identificar la eficiencia de muones dentro de jets usando datos. La incertidumbre resultante asociada a esta técnica de etiquetado charm, que es dominante a la hora de calcular la precisión de $^{W}_{c}$, está bien comprendida y resulta menor que aquella correspondiente a usar las herramientas existentes de CMS para etiquetado charm. La técnica de etiquetado de charm empleada en el análisis es crucial para mejorar la precisión de la medición de $^{W}_{c}$, superando el valor promedio mundial actual.

Además, esta tesis destaca la creciente importancia de las técnicas de aprendizaje automático en física de partículas. Aplicando varios métodos—redes neuronales bayesianas, random forest probabilístico y el método local ensembles—este trabajo evalúa su eficacia a la hora de clasificar eventos top-antitop y varios procesos de fondo. Los resultados muestran un desempeño sólido en todos los métodos, con el modelo BA logrando la mayor eficiencia, seguido de PRF y, finalmente, LE. Esto demuestra el poder discriminativo robusto de los modelos y su capacidad para proporcionar predicciones fiables. Es importante destacar que los métodos de estimación de incertidumbre utilizados en el análisis revelan que la incertidumbre epistémica derivada de la construcción del modelo es mínima. Al ser pequeña la incertidumbre de los resultados, indica que los modelos son adecuados para esta tarea de clasificación. Este resultado refuerza la utilidad del aprendizaje automático en la investigación de física de partículas, allanando el camino para futuras exploraciones en este campo en constante cambio e ilustrando cómo estas herramientas pueden mejorar la precisión de los análisis al tiempo que proporcionan valiosas perspectivas.

En general, los resultados y metodologías presentados en esta tesis representan un avance tanto en medidas de precisión como en la aplicación de técnicas de aprendizaje automático en física de partículas. Los esfuerzos continuos del LHC y el experimento CMS siguen mejorando nuestros modelos teóricos y ampliando los límites de nuestro conocimiento.

References

[1] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (9 Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[2] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (16 Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[3] G. Arnison et al. "Experimental observation of isolated large transverse energy electrons with associated missing energy at s=540 GeV". In: *Physics Letters B* 122.1 (1983), pp. 103–116. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)91177-2.

[4] G. Arnison et al. "Experimental observation of lepton pairs of invariant mass around 95 GeV/ at the CERN SPS collider". In: *Physics Letters B* 126.5 (1983), pp. 398–410. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)90188-0.

[5] M. Banner et al. "Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN pp collider". In: *Physics Letters B* 122.5 (1983), pp. 476–485. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)91605-2.

[6] P. Bagnaia et al. "Evidence for Z at the CERN pp collider". In: *Physics Letters B* 129.1 (1983), pp. 130–140. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)90744-X.

[7] Lyndon Evans and Philip Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[8] ATLAS collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020.

[9] CMS collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pp. 30–61. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.021.

[10] *SM scheme*. https://www.physik.uzh.ch/groups/serra/StandardModel.html.

[11] Y. Ashie et al. "Measurement of atmospheric neutrino oscillation parameters by Super-Kamiokande I". In: *Phys. Rev. D* 71 (11 June 2005), p. 112005. DOI: 10.1103/PhysRevD.71.112005.

[12] T. W. B. Kibble. *History of electroweak symmetry breaking*. 2015.

[13] Ivo de Medeiros Varzielas and Stephen F. King. "Origin of Yukawa couplings for Higgs bosons and leptoquarks". In: *Phys. Rev. D* 99 (9 May 2019), p. 095029. DOI: 10.1103/PhysRevD.99.095029.

[14] M. Bargiotti et al. "Present knowledge of the Cabibbo-Kobayashi-Maskawa matrix". In: *La Rivista del Nuovo Cimento* 23.3 (Mar. 2000), pp. 1–71. ISSN: 1826-9850. DOI: 10.1007/bf03548883.

[15] K. Kleinknecht. "Quark and lepton mixing in weak interactions". In: *Particles and Detectors: Festschrift for Jack Steinberger*. Ed. by Konrad Kleinknecht and Tsung Dao Lee. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, pp. 149–164. ISBN: 978-3-540-39768-7. DOI: 10.1007/BFb0041271.

[16] Nicola Cabibbo. "Unitary Symmetry and Leptonic Decays". In: *Phys. Rev. Lett.* 10 (12 June 1963), pp. 531–533. DOI: 10.1103/PhysRevLett.10.531.

[17] Makoto Kobayashi and Toshihide Maskawa. "CP-Violation in the Renormalizable Theory of Weak Interaction". In: *Progress of Theoretical Physics* 49.2 (Feb. 1973), pp. 652–657. ISSN: 0033-068X. DOI: 10.1143/PTP.49.652.

[18] R. L. Workman et al. "Review of Particle Physics". In: *PTEP* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097.

[19] Frank Wilczek. "Nobel Lecture: Asymptotic freedom: From paradox to paradigm". In: *Rev. Mod. Phys.* 77 (3 Sept. 2005), pp. 857–870. DOI: 10.1103/RevModPhys.77.857.

[20] Sidney D. Drell and Tung-Mow Yan. "Partons and Their Applications at High Energies". In: *Annals of Physics* 281.1 (2000), pp. 450–493. ISSN: 0003-4916. DOI: https://doi.org/10.1006/aphy.2000.6014.

[21] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (June 2015), pp. 159–177. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2015.01.024.

[22] Stefano Frixione, Paolo Nason and Carlo Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". In: *Journal of High Energy Physics* 2007.11 (Nov. 2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070.

[23] 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". In: *Journal of High Energy Physics* 2014.7 (July 2014). ISSN: 1029-8479. DOI: 10.1007/jhep07(2014)079.

[24] Richard D. Ball et al. "Parton distributions for the LHC run II". In: *Journal of High Energy Physics* 2015.4 (Apr. 2015). ISSN: 1029-8479. DOI: 10.1007/jhep04(2015)040.

[25] Richard D. Ball et al. "Parton distributions from high-precision collider data: NNPDF Collaboration". In: *The European Physical Journal C* 77.10 (Oct. 2017). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-017-5199-5.

[26] S. Agostinelli et al. "GEANT —a simulation toolkit". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8.

[27] https://home.cern/about.

[28] Oliver Sim Brüning et al. *LHC Design Report*. CERN Yellow Reports: Monographs. Geneva: CERN, 2004. DOI: 10.5170/CERN-2004-003-V-1.

[29] Thomas Taylor and Daniel Treille. "The Large Electron Positron Collider (LEP): Probing the Standard Model". In: *Technology Meets Research*. Chap. Chapter 7, pp. 217–261. DOI: 10.1142/9789814749145_0007.

[30] Esma Mobs. "The CERN accelerator complex - August 2018. Complexe des accélérateurs du CERN - Août 2018". In: (2018). General Photo.

[31] Deepak Kar. *Experimental Particle Physics*. 2053-2563. IOP Publishing, 2019. ISBN: 978-0-7503-2112-9. DOI: 10.1088/2053-2563/ab1be6.

[32] J. Stirling. https://mstwpdf.hepforge.org/plots/plots.html.

[33] *Public CMS Luminosity Information*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[34] The CMS Collaboration et al. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[35] *CMS images gallery*. https://home.cern/resources/image/experiments/cms-images-gallery. 2017.

[36] I. Neutelings. *CMS coordinate system*. https://tikz.net/axis3d_cms/.

[37] Tai Sakuma. *Cutaway diagrams of CMS detector*. https://cds.cern.ch/record/2665537?ln=es.

[38] G. L. Bayatian et al. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. https://cds.cern.ch/record/000922757?ln=es. 2006.

[39] CMS Collaboration. "Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays". In: *Journal of Instrumentation* 5.03 (Mar. 2010), T03021. DOI: 10.1088/1748-0221/5/03/T03021.

[40] V Karimäki et al. *The CMS tracker system project: Technical Design Report*. https://cds.cern.ch/record/368412. Geneva, 1997.

[41] *The CMS tracker: addendum to the Technical Design Report*. https://cds.cern.ch/record/490194. Geneva, 2000.

[42] S Chatrchyan et al. "Alignment of the CMS silicon tracker during commissioning with cosmic rays". In: *JINST* 5 (2010), T03009. DOI: 10.1088/1748-0221/5/03/T03009.

[43] A Dominguez et al. *CMS Technical Design Report for the Pixel Detector Upgrade*. Ed. by A Dominguez. https://cds.cern.ch/record/1481838. 2012.

[44] *The CMS electromagnetic calorimeter project: Technical Design Report*. https://cds.cern.ch/record/349375. Geneva, 1997.

[45] Andrea Benaglia. "The CMS ECAL performance with examples". In: *Journal of Instrumentation* 9 (Jan. 2014). DOI: 10.1088/1748-0221/9/02/C02008.

[46] *The CMS hadron calorimeter project: Technical Design Report*. https://cds.cern.ch/record/357153. Geneva, 1997.

[47] J. G. Layter. *The CMS muon project: Technical Design Report*. https://cds.cern.ch/record/343814. Geneva, 1997.

[48] CMS Collaboration. *Summaries of CMS cross Section Measurements*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsCombined.

[49] A.M. Sirunyan et al. "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of Instrumentation* 15.10 (Oct. 2020), P10017–P10017. ISSN: 1748-0221. DOI: 10.1088/1748-0221/15/10/p10017.

[50] G Bauer et al. "The data-acquisition system of the CMS experiment at the LHC". In: *Journal of Physics: Conference Series* 331 (Dec. 2011), p. 022021. DOI: 10.1088/1742-6596/331/2/022021.

[51] J-M Andre et al. "File-based data flow in the CMS Filter Farm". In: *Journal of Physics: Conference Series* 664.8 (Dec. 2015), p. 082033. DOI: 10.1088/1742-6596/664/8/082033.

[52] K. Bos et al. *LHC computing Grid: Technical Design Report. Version 1.06 (20 Jun 2005)*. https://cds.cern.ch/record/840543. Geneva, 2005.

[53] I Bird et al. *Update of the Computing Models of the WLCG and the LHC Experiments*. https://cds.cern.ch/record/1695401. 2014.

[54] A.M. Sirunyan et al. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003.

[55] David Barney. *CMS Detector Slice*. https://cds.cern.ch/record/2120661.

[56] Albert M Sirunyan et al. "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC". In: *JINST* 16 (2021), P05014. DOI: 10.1088/1748-0221/16/05/P05014.

[57] CMS Collaboration. *ECAL 2016 refined calibration and Run2 summary plots*. https://cds.cern.ch/record/2717925. CMS Detector Performance Summary. 2020.

[58] W Adam et al. "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC". In: *Journal of Physics G: Nuclear and Particle Physics* 31.9 (2005), N9. DOI: 10.1088/0954-3899/31/9/N01.

[59] Matteo Cacciari and Gavin P. Salam. "Pileup subtraction using jet areas". In: *Phys. Lett. B* 659 (2008), p. 119. DOI: 10.1016/j.physletb.2007.09.077.

[60] A.M. Sirunyan et al. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at √s=13 TeV". In: *Journal of Instrumentation* 13.06 (June 2018), P06015–P06015. ISSN: 1748-0221. DOI: 10.1088/1748-0221/13/06/p06015.

[61] Matteo Cacciari, Gavin P. Salam and Gregory Soyez. "The anti- jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063.

[62] A.M. Sirunyan et al. "Pileup mitigation at CMS in 13 TeV data". In: *Journal of Instrumentation* 15 (Sept. 2020), P09018–P09018. DOI: 10.1088/1748-0221/15/09/P09018.

[63] *Jet algorithms performance in 13 TeV data*. https://cds.cern.ch/record/2256875. Geneva, 2017.

[64] A. Tumasyan A.M. Sirunyan et al. "Performance of missing transverse momentum reconstruction in proton-proton collisions at ¯ = 13 TeV using the CMS detector". In: *JINST* 14 (2019), P07004. DOI: 10.1088/1748-0221/14/07/P07004.

[65] Vardan Khachatryan et al. "Pileup mitigation at CMS in 13 TeV data". In: *JINST* 15 (2020), P09018–P09018. DOI: 10.1088/1748-0221/15/09/p09018.

[66] A.M. Sirunyan V. Khachatryan et al. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *JINST* 12 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014.

[67] *CMS Muon results*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsBTV.

[68] Emil Bols et al. "Jet Flavour Classification Using DeepJet". In: *JINST* 15 (2020), P12012. DOI: 10.1088/1748-0221/15/12/P12012.

[69] A.M. Sirunyan et al. "Performance of missing transverse momentum reconstruction in proton-proton collisions at √ s = 13 TeV using the CMS detector". In: *Journal of Instrumentation* 14 (July 2019), P07004–P07004. DOI: 10.1088/1748-0221/14/07/P07004.

[70] "Measurement of associated W + charm production in pp collisions at ¯ = 7 TeV". In: *Journal of High Energy Physics* 2014.2 (Feb. 2014). ISSN: 1029-8479. DOI: 10.1007/jhep02(2014)013.

[71] "Measurement of associated Z + charm production in proton–proton collisions at ¯ TeV". In: *The European Physical Journal C* 78.4 (Apr. 2018). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-018-5752-x.

[72] "Measurements of the associated production of a W boson and a charm quark in proton–proton collisions at ¯ TeV". In: *The European Physical Journal C* 82.12 (Dec. 2022). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-022-10897-7.

[73] "Measurement of the production cross section for a W boson in association with a charm quark in proton–proton collisions at ¯ TeV". In: *The European Physical Journal C* 84.1 (Jan. 2024). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-023-12258-4.

[74] CMS Collaboration. *CMS Luminosity Measurements for the 2016 Data Taking Period*. https://cds.cern.ch/record/2257069. Geneva, 2017.

[75] CMS Collaboration. *CMS luminosity measurement for the 2017 data-taking period at* . https://cds.cern.ch/record/2621960. Geneva, 2018.

[76] CMS Collaboration. *CMS luminosity measurement for the 2018 data-taking period at* . https://cds.cern.ch/record/2676164. Geneva, 2019.

[77] A. M. Sirunyan et al. "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements". In: *The European Physical Journal C* 80.1 (Jan. 2020). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-7499-4.

[78] J. Alwall et al. "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions". In: *The European Physical Journal C* 53.3 (Dec. 2007), pp. 473–500. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-007-0490-5.

[79] *CMS B-tagging and Vertexing results*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsMUO.

[80] *CMS Egamma results*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEGM.

[81] Nikolaos Kidonakis. "Soft-gluon corrections for production at ". In: *Phys. Rev. D* 96 (3 Aug. 2017), p. 034014. DOI: 10.1103/PhysRevD.96.034014.

[82] Michal Czakon, David Heymes and Alexander Mitov. "High-Precision Differential Predictions for Top-Quark Pairs at the LHC". In: *Phys. Rev. Lett.* 116 (8 Feb. 2016), p. 082003. DOI: 10.1103/PhysRevLett.116.082003.

[83] Michał Czakon et al. "Top-pair production at the LHC through NNLO QCD and NLO EW". In: *Journal of High Energy Physics* 2017.10 (Oct. 2017). ISSN: 1029-8479. DOI: 10.1007/jhep10(2017)186.

[84] Stefano Catani et al. "Top-quark pair production at the LHC: fully differential QCD predictions at NNLO". In: *Journal of High Energy Physics* 2019.7 (July 2019). ISSN: 1029-8479. DOI: 10.1007/jhep07(2019)100.

[85] A. Tumasyan et al. "Measurement of differential production cross sections in the full kinematic range using jets events from proton-proton collisions at ". In: *Phys. Rev. D* 104 (9 Nov. 2021), p. 092013. DOI: 10.1103/PhysRevD.104.092013.

[86] A. Tumasyan et al. "Measurement of the top quark mass using a profile likelihood approach with the lepton+jets final states in proton–proton collisions at at = 13 TeV". In: *Eur. Phys. J. C* 83 (2023), p. 963. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-023-12050-4.

[87] *Jet Flavour Identification (MC Truth)*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagMCTools. 2016.

[88] Mykhailo Lisovyi, Andrii Verbytskyi and Oleksandr Zenaiev. "Combined analysis of charm-quark fragmentation-fraction measurements". In: *The European Physical Journal C* 76.7 (July 2016). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-016-4246-y.

[89] Michal Czakon and Alexander Mitov. "Top++: A program for the calculation of the top-pair cross-section at hadron colliders". In: *Comput. Phys. Commun.* 185 (2014), p. 2930. DOI: 10.1016/j.cpc.2014.06.021.

[90] John Campbell, Tobias Neumann and Zack Sullivan. "Single-top-quark production in the -channel at NNLO". In: *JHEP* 02 (2021), p. 040. DOI: 10.1007/JHEP02(2021)040.

[91] Richard D. Ball et al. "The PDF4LHC21 combination of global PDF fits for the LHC Run III". In: *J. Phys. G* 49.8 (2022), p. 080501. DOI: 10.1088/1361-6471/ac7216.

[92] Nikolaos Kidonakis and Nodoka Yamanaka. "Higher-order corrections for production at high-energy hadron colliders". In: *JHEP* 05 (2021), p. 278. DOI: 10.1007/JHEP05(2021)278.

[93] *Measurement of W and Z boson inclusive cross sections in pp collisions at 5.02 and 13 TeV*. https://cds.cern.ch/record/2868090. Geneva, 2023.

[94] A. M. Sirunyan et al. " boson pair production in proton-proton collisions at ". In: *Phys. Rev. D* 102 (9 Nov. 2020), p. 092001. DOI: 10.1103/PhysRevD.102.092001.

[95] A. M. Sirunyan et al. "Measurements of production cross sections and constraints on anomalous triple gauge couplings at TeV". In: *Eur. Phys. J. C* 81 (2021), p. 200. DOI: 10.1140/epjc/s10052-020-08817-8.

[96] A. Tumasyan et al. "Measurement of the inclusive and differential WZ production cross sections, polarization angles, and triple gauge couplings in pp collisions at = 13 TeV". In: *JHEP* 2022 (2022), p. 32. DOI: 10.1007/JHEP07(2022)032.

[97] A. M. Sirunyan et al. "Measurement of the inelastic proton-proton cross section at TeV". In: *JHEP* 07 (2018), p. 161. DOI: 10.1007/JHEP07(2018)161.

[98] *CMS Jet and Missing Energy Results*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsJME.

[99] R. Barate et al. "A Direct measurement of |V(cs)| in hadronic W decays using a charm tag". In: *Phys. Lett. B* 465 (1999), pp. 349–362. DOI: 10.1016/S0370-2693(99)01088-6.

[100] G. Abbiendi et al. "A Measurement of the Rate of Charm Production in W Decays". In: *Phys. Lett. B* 490 (2000), pp. 71–86. DOI: 10.1016/S0370-2693(00)00971-0.

[101] Aram Hayrapetyan et al. "The CMS statistical analysis and combination tool: COMBINE". https://cds.cern.ch/record/2895097?ln=es. 2024.

[102] A. Tumasyan et al. "Precision measurement of the W boson decay branching fractions in proton-proton collisions at ". In: *Phys. Rev. D* 105 (2022), p. 072008. DOI: 10.1103/PhysRevD.105.072008.

[103] Johannes Albrecht et al. "A Roadmap for HEP Software and Computing R&D for the 2020s". In: *Computing and Software for Big Science* 3.1 (Mar. 2019). ISSN: 2510-2044. DOI: 10.1007/s41781-018-0018-8.

[104] *IML Machine Learning Workshop*. https://indico.cern.ch/event/595059/. 2017.

[105] *20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research*. https://indico.cern.ch/event/855454/. 2021.

[106] *European AI for Fundamental Physics Conference*. https://indico.nikhef.nl/event/4875/. 2024.

[107] A. Aurisano et al. "A convolutional neural network neutrino event classifier". In: *JNIST* 11.09 (Sept. 2016), p. 23. ISSN: 1748-0221. DOI: 10.1088/1748-0221/11/09/p09001.

[108] Evan Racah et al. "Revealing Fundamental Physics from the Daya Bay Neutrino Experiment Using Deep Neural Networks". In: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (Dec. 2016), pp. 892–897. DOI: 10.1109/icmla.2016.0160.

[109] J. Renner et al. "Background rejection in NEXT using deep neural networks". In: *JNIST* 12.01 (Jan. 2017), p. 22. ISSN: 1748-0221. DOI: 10.1088/1748-0221/12/01/t01004.

[110] Julia Vázquez-Escobar, J.M. Hernández and Miguel Cárdenas-Montes. "Estimation of Machine Learning model uncertainty in particle physics event classifiers". In: *Computer Physics Communications* 268 (2021), p. 108100. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2021.108100.

[111] Radford M. Neal. "Bayesian learning for neural networks". PhD thesis. University of Toronto, 1995.

[112] Itamar Reis, Dalya Baron and Sahar Shahaf. "Probabilistic Random Forest: A machine learning algorithm for noisy datasets". In: `arXiv:abs/1811.05994` (Nov. 2018), p. 17.

[113] David Madras, James Atwood and Alex D'Amour. "Detecting Extrapolation with Local Ensembles". In: *ICLR 2020 International Conference on Learning Representations* (2019).

[114] Yarin Gal. "Bayesian Deep Learning". In: *Uncertainty in Deep Learning*. 2016. Chap. 3, pp. 29–61.

[115] Ian J. Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. Cambridge, MA, USA: MIT Press, 2016.

[116] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *J. Mach. Learn. Res.* 15 (June 2014). http://jmlr.org/papers/v15/srivastava14a.html, pp. 1929–1958.

[117] Vinod Nair and Geoffrey E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. https://www.cs.toronto.edu/~fritz/absps/reluICML.pdf. Haifa, Israel: Omnipress, 2010, pp. 807–814. ISBN: 9781605589077.

[118] Thomas M. Mitchell. "Artificial Neural Networks". In: *Machine Learning*. McGraw-Hill, Inc., 1997. Chap. 4, pp. 81–126. ISBN: 0070428077.

[119] J. Vázquez-Escobar, J.M. Hernández and M. Cárdenas-Montes. *Bayesian approximation with STR techniques*. https://github.com/juliavazquez3/github-upload.

[120] Leo Breiman. "Random Forests". In: *Mach. Learn.* 45.1 (Oct. 2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324.

[121] Schmidt A. Sander C. *CMS data analysis tutorial*. http://ippog.org/resources/2012/cms-hep-tutorial. International Particle Physics Outreach Group, Sept. 2014.

[122] Itamar Reis, Dalya Baron and Sahar Shahaf. *Code for Probabilistic Random Forest*. https://github.com/ireis/PRF.

[123] David Madras, James Atwood and Alex D'Amour. *Code for Local Ensembles*. https://github.com/dmadras/local-ensembles.

[124] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010.

[125] A.M. Sirunyan et al. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV." In: *JINST* 13.05 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011.

APPENDICES

Appendix A

Plots for baseline selection differentiating muon and electron channels

Distributions for the muon channel, events where the high- isolated lepton is a muon, correspond to Fig. A.1, A.2, A.3, A.4. Distributions for the muon channel are Fig. A.5, A.6, A.7, A.8. There are no meaningful differences that could indicate a systematic effect of using the muon or electron channel. Both sets of plots present the same description, data-simulation consistency is equally good.

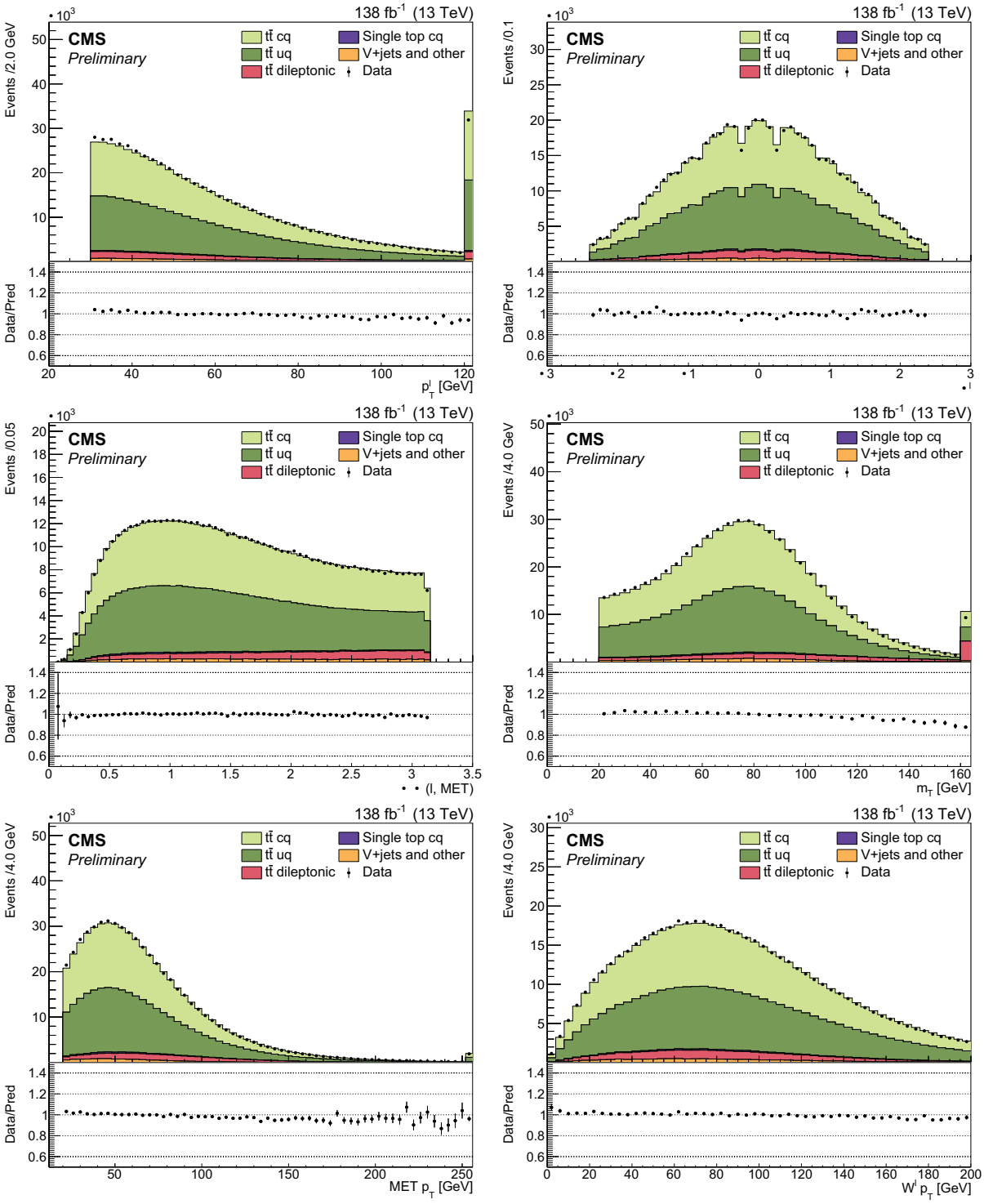


Figure A.1: Distributions for the high-p_T isolated lepton. The top left image depicts the transverse momentum of the prompt lepton and the top right image its pseudorapidity. The center left image is the azymuthal angle difference between the prompt lepton and the missing transverse momentum. The center right image shows the leptonic W transverse mass. Bottom left image is the missing transverse momentum and bottom right image displays the leptonic W transverse momentum.

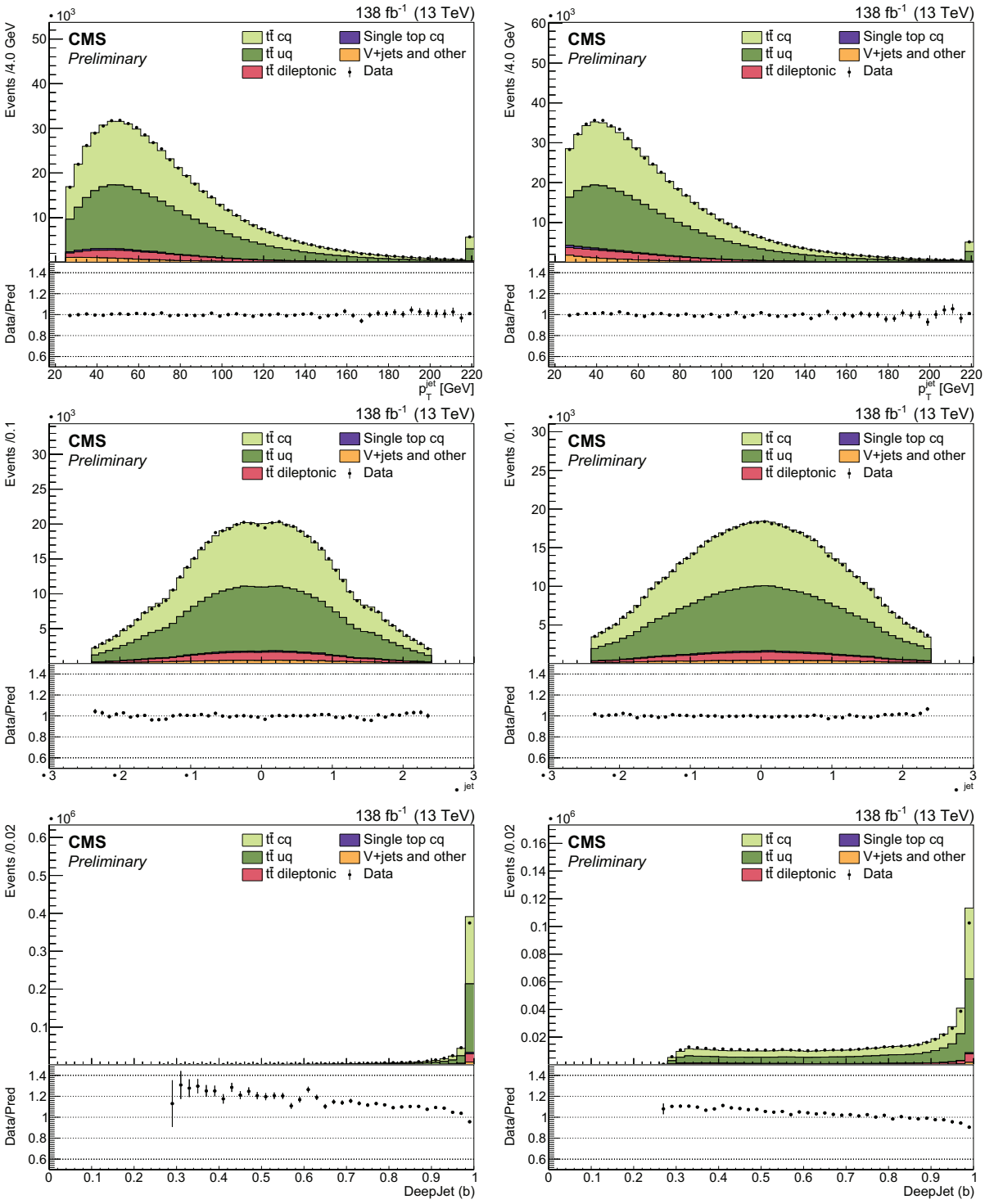


Figure A.2: B-tagged jet distributions. The left column corresponds to b1-jet, that with the highest b-tagging discriminant, and the right column to the other, b2-jet. The distributions are, from top to bottom, p_T, η and b-tagging discriminant score.

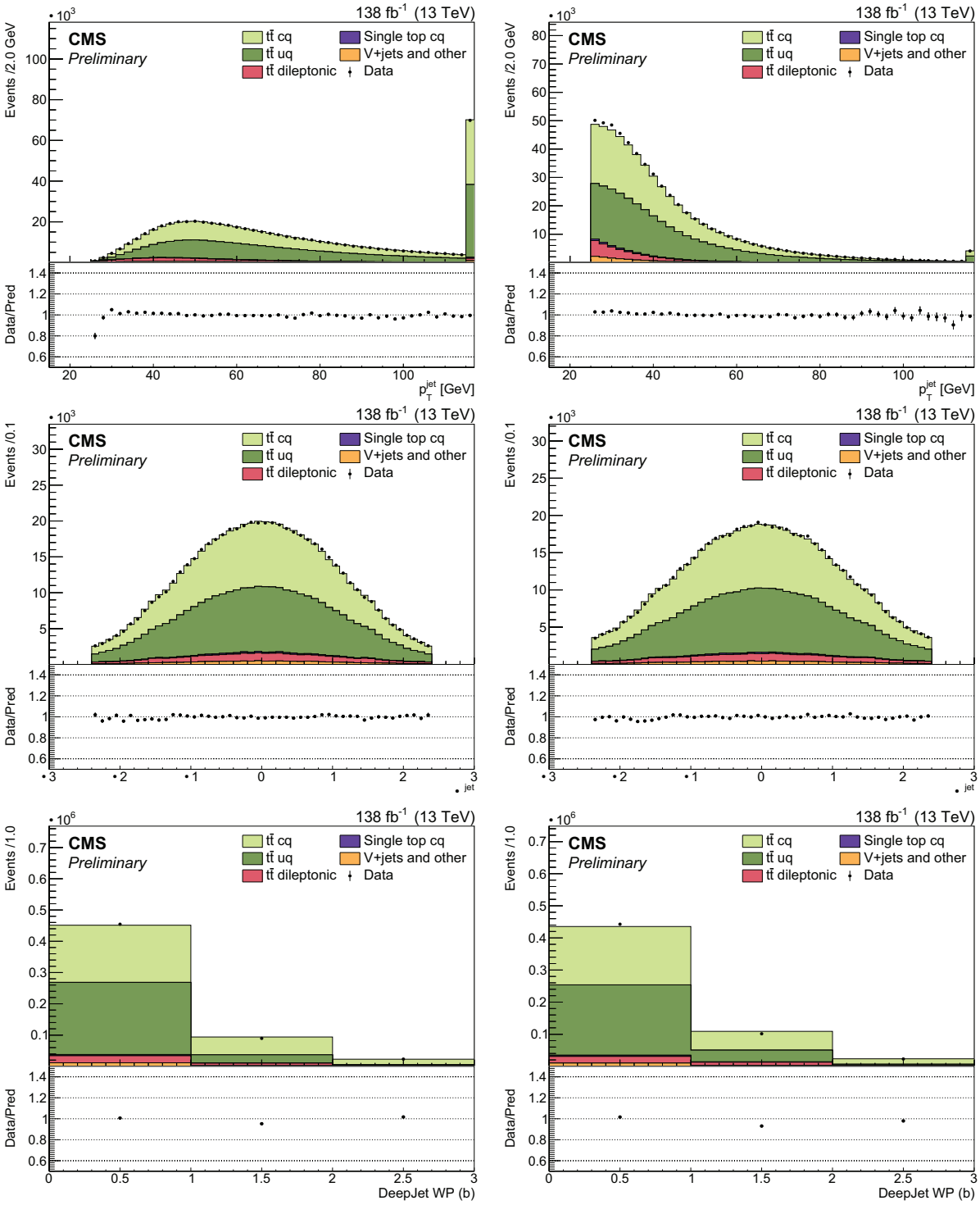


Figure A.3: Distributions of the two jets associated to the W boson decaying hadronically. The distributions of the left column correspond to the leading-p_T jet, W jet 1, and the right column to the subleading, W jet 2. The distributions are, from top to bottom, the transverse momentum, the pseudorapidity and the b-tag discriminant binned in working points (the first bin corresponds to jets not passing the loose or medium WPs, the center bin for jets satisfying the loose WP but not medium, and the last bin for jets satisfying the medium WP).

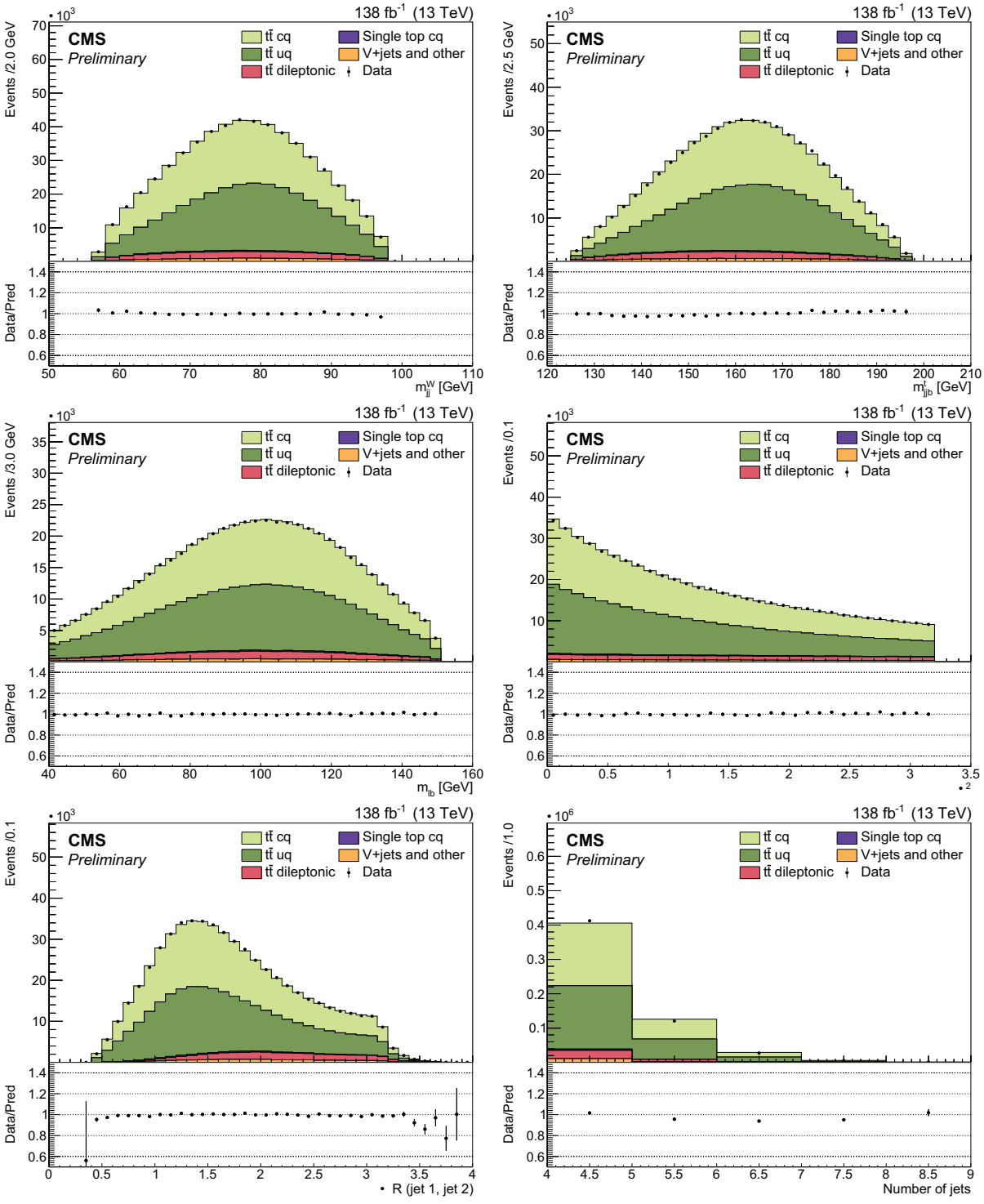


Figure A.4: Some relevant kinematic distributions. The top left image corresponds to the invariant mass of the dijet reconstructing the hadronic W boson. Top right image displays the invariant mass of this dijet plus the corresponding b-jet reconstructing the top quark mass. The center left image is the invariant mass of the high-p_T isolated lepton and the other b-tagged jet associated to the W boson decaying leptonically. The center right image is the distribution of the χ^2 test value for the kinematic constraints. The bottom left image is the ΔR between the jets reconstructing the hadronic W boson and the bottom right image is the number of jets distribution.

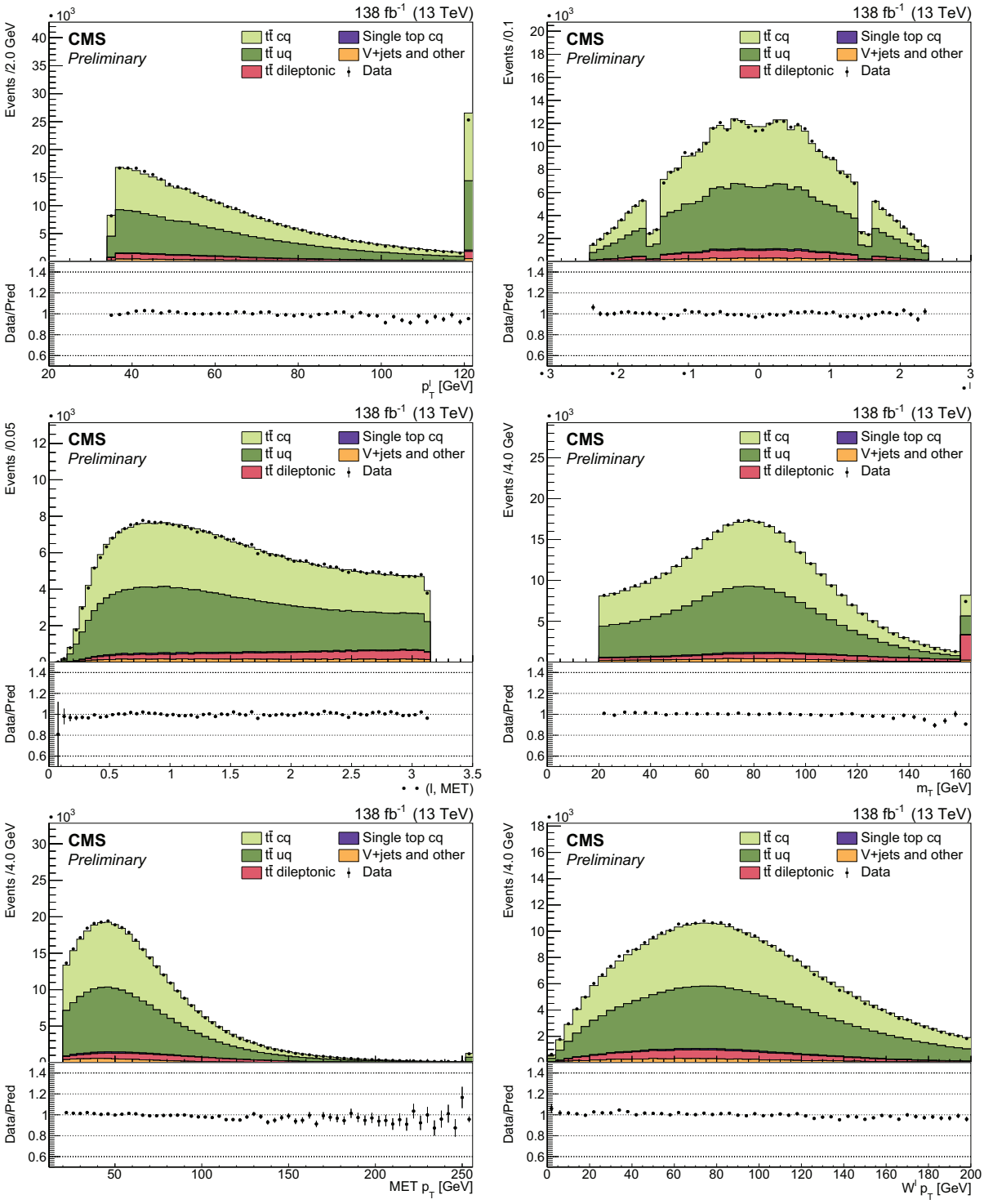


Figure A.5: Distributions for the high-p_T isolated lepton. The top left image depicts the transverse momentum of the prompt lepton and the top right image its pseudorapidity. The center left image is the azymuthal angle difference between the prompt lepton and the missing transverse momentum. The center right image shows the leptonic W transverse mass. Bottom left image is the missing transverse momentum and bottom right image displays the leptonic W transverse momentum.

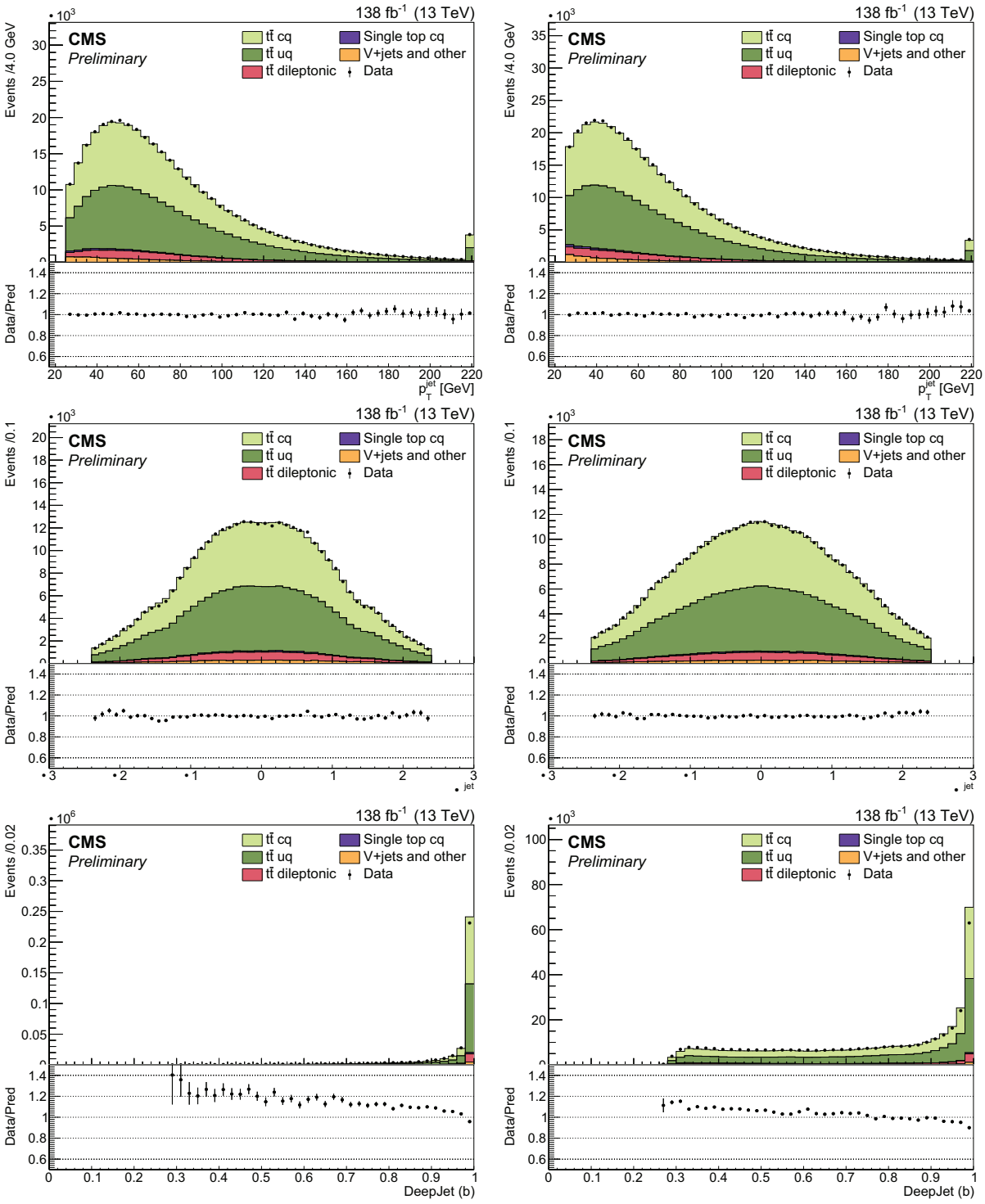


Figure A.6: B-tagged jet distributions. The left column corresponds to b1-jet, that with the highest b-tagging discriminant, and the right column to the other, b2-jet. The distributions are, from top to bottom, p_T, η and b-tagging discriminant score.

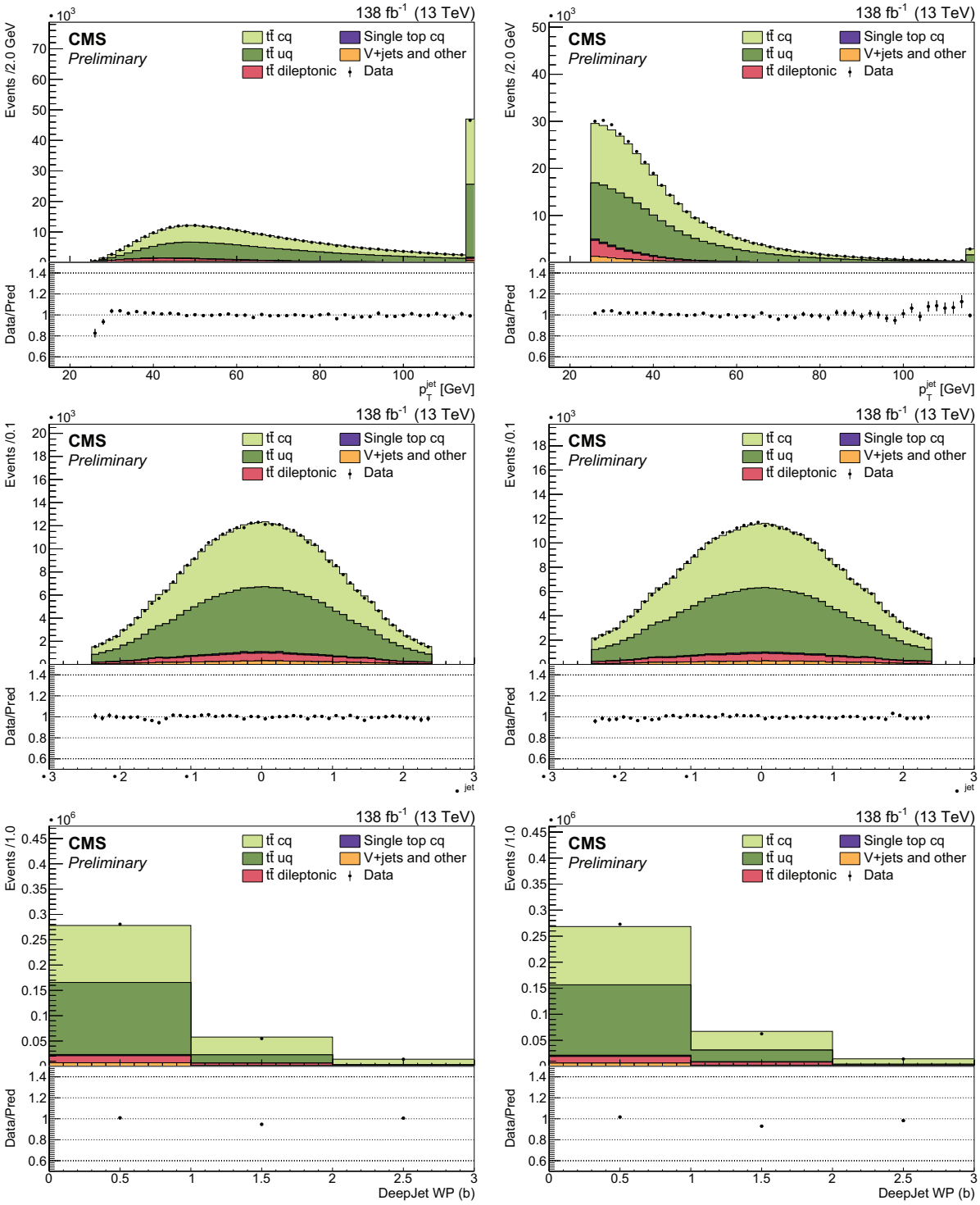


Figure A.7: Distributions of the two jets associated to the W boson decaying hadronically. The distributions of the left column correspond to the leading-p_T jet, W jet 1, and the right column to the subleading, W jet 2. The distributions are, from top to bottom, the transverse momentum, the pseudorapidity and the b-tag discriminant binned in working points (the first bin corresponds to jets not passing the loose or medium WPs, the center bin for jets satisfying the loose WP but not medium, and the last bin for jets satisfying the medium WP).

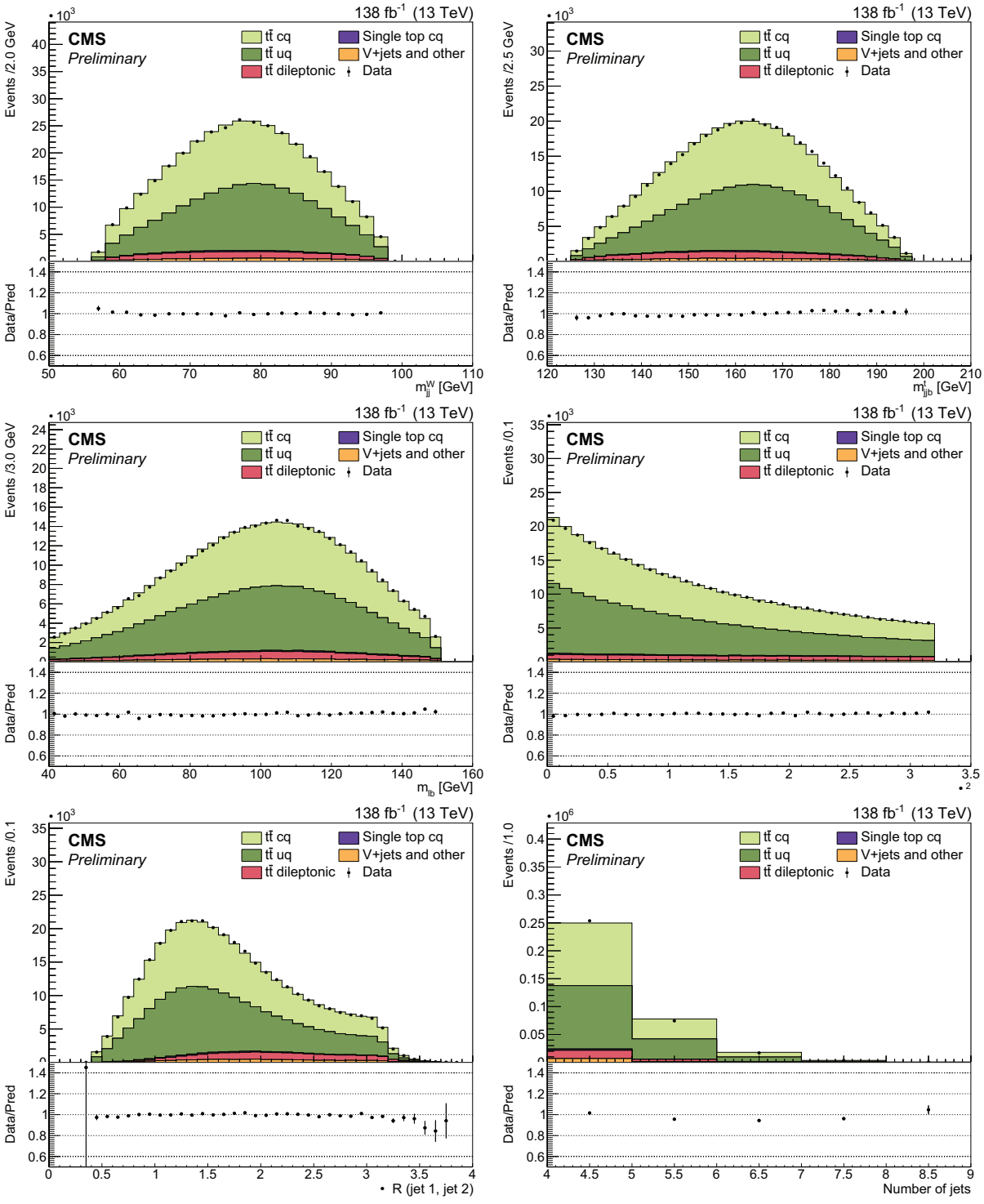


Figure A.8: Some relevant kinematic distributions. The top left image corresponds to the invariant mass of the dijet reconstructing the hadronic W boson. Top right image displays the invariant mass of this dijet plus the corresponding b-jet reconstructing the top quark mass. The center left image is the invariant mass of the high-p_T isolated lepton and the other b-tagged jet associated to the W boson decaying leptonically. The center right image is the distribution of the χ^2 test value for the kinematic constraints. The bottom left image is the ΔR between the jets reconstructing the hadronic W boson and the bottom right image is the number of jets distribution.

Appendix B

Cross check using DeepJet c-tagging

We have evaluated the sensitivity in the measurement of $_{c}^{W}$ that can be achieved using the DeepJet heavy flavour tagger [125, 68]. We have used the charm tagging Tight operation working point on the cvl and cvb discriminants in order to keep the mistag rate at the minimum. This WP selects from the semileptonic $t\bar{t}$ baseline sample, where there are equal contributions of W cq and W uq events, about 30% of the W cq events and 10% of the W uq events.

The available c tagging scale factors for tag and mistag rates are applied, and the corresponding yields are calculated and used in the datacard to perform the fit with combine described in Sec. 5, using the same 4 event categories (muon and electron prompt lepton, charm tagged and non-tagged events).

As recommended for applying the tagging scale factors to the simulation, we compute the tagging efficiency for jets satisfying our selection requirements. We fit a function dependent on the jet for each data-taking year and jet flavour and use those, in combination with the provided scale factors, to correct the sample.

The charm tagging systematic effect is estimated by using the up and down scale factors provided by the BTV group. The resulting systematic uncertainty is 10% for the c-tagged event categories and 5% for the non-c-tagged categories. The systematic uncertainties resulting from the charm mistag rates are 15% for c-tagged event categories and 1% for the non-c-tagged categories.

The results of the combine fit are shown in Fig. B.1, and the impact of the systematic effects on the uncertainty of the fit parameters is displayed in Fig. B.2. The sensitivity on the POI, which provides the expected uncertainty in the $_{c}^{W}$ measurement, is around 14%, much worse than the expected sensitivity achieved by the muon-based charm tagging method. The uncertainty is dominated by the systematic effects related to charm tagging (uncertainties in charm tagging and mistagging). The increase in the number of events, and the corresponding reduction of the statistical uncertainty, achieved with the DeepJet c-tagging (30% efficiency vs 2% for the DeepJet and muon-based charm tagging methods), is irrelevant given that the measurement is dominated by the systematic uncertainty. The DeepJet tagger does not allow to separate c-tagged events in OS and SS events, and

therefore, the charm mistag rates must be evaluated with the simulation, suffering from the large uncertainty in the mistag scale factors (∼15%). The large uncertainty in the c tagging scale factors (5-10%) also contributes significantly to the deteriorated sensitivity to the R_c^W measurement.

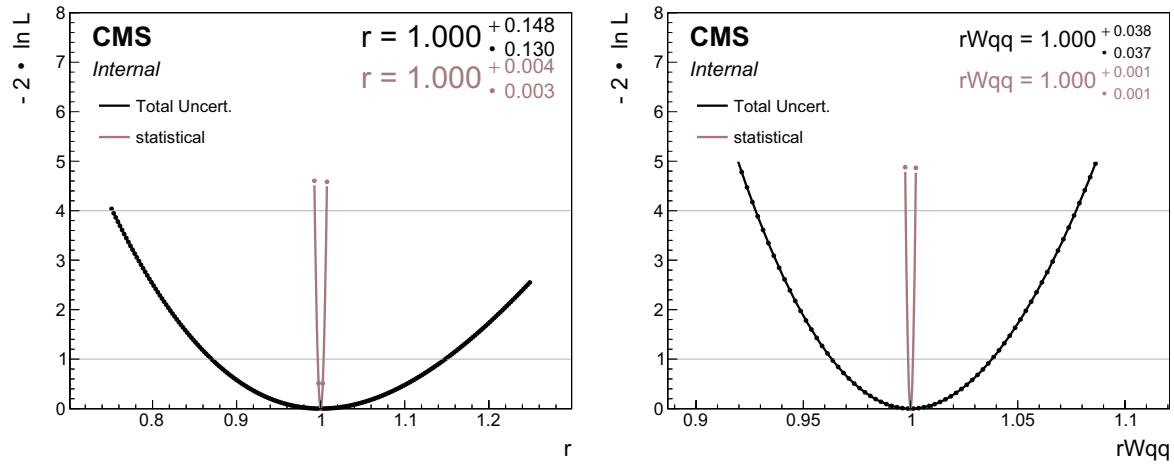


Figure B.1: Scan of the likelihood function of the r POI (top) and the $r_{W_{qq}}$ POI (bottom) in the fit.

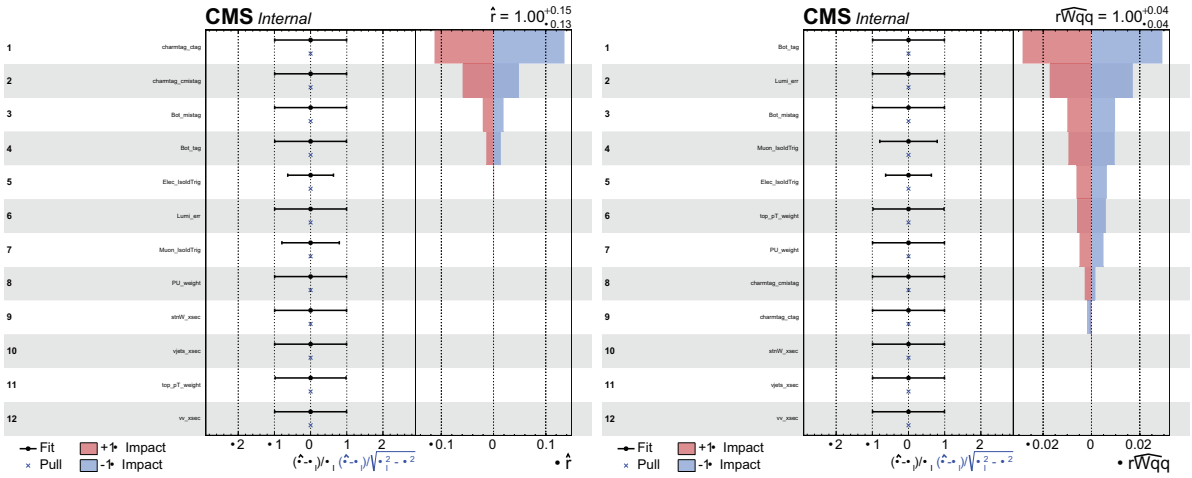


Figure B.2: Impact of the various systematic uncertainties in the determination of the r and $r_{W_{qq}}$ POIs.

Figures B.3 and B.4 show some kinematic distributions for the charm tagged and not-tagged samples, respectively, including the postfit systematic uncertainty band.

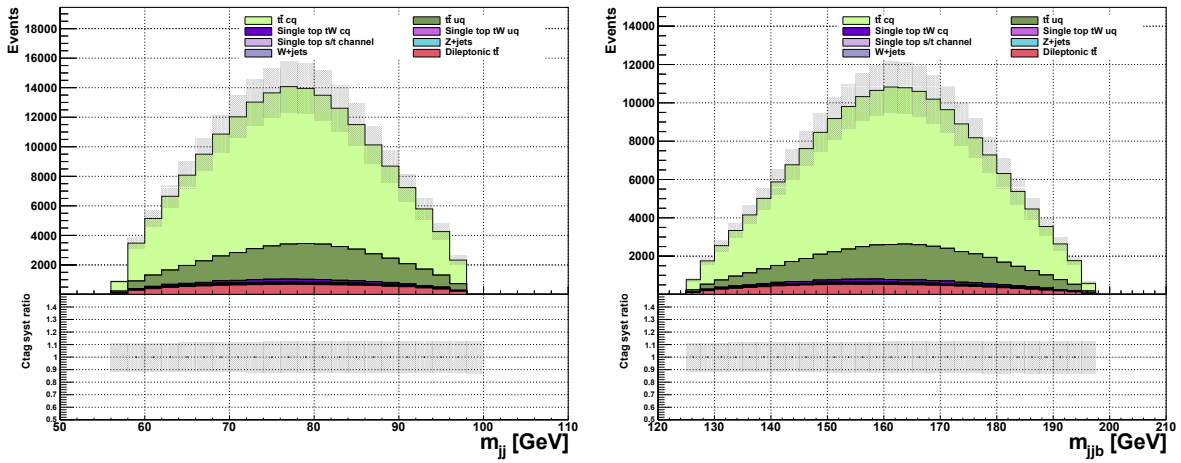


Figure B.3: Kinematic distributions for the sample of semileptonic $t\bar{t}$ events selected by the DeepJet charm tagging Tight requirement: invariant mass of the dijet reconstructing the hadronic W boson (left) and invariant mass of the two jets associated with the hadronic W boson plus the corresponding bottom jet forming the trijet that reconstructs the top quark that decays into the hadronic W boson.

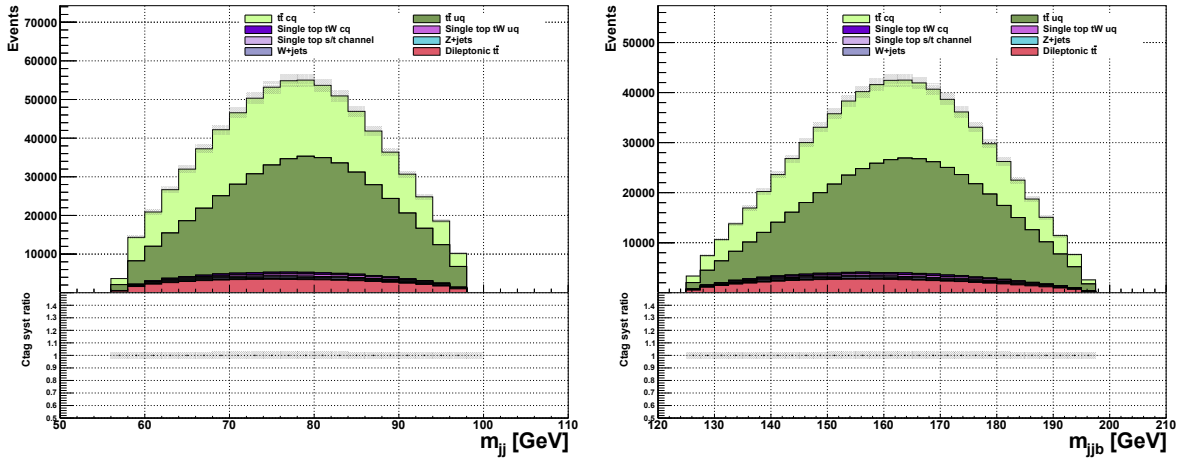


Figure B.4: Kinematic distributions for the sample of semileptonic $t\bar{t}$ events not selected by the DeepJet charm tagging Tight requirement: invariant mass of the dijet reconstructing the hadronic W boson (left) and invariant mass of the two jets associated with the hadronic W boson plus the corresponding bottom jet forming the trijet that reconstructs the top quark that decays into the hadronic W boson.

Appendix C

Plots differentiating muon and electron channels for OS-SS subtracted sample

Distributions for the muon channel, events where the high- isolated lepton is a muon, correspond to Fig. C.1, C.2, C.3, C.4, C.5. Distributions for the muon channel are Fig. C.6, C.7, C.8, C.9, C.10.

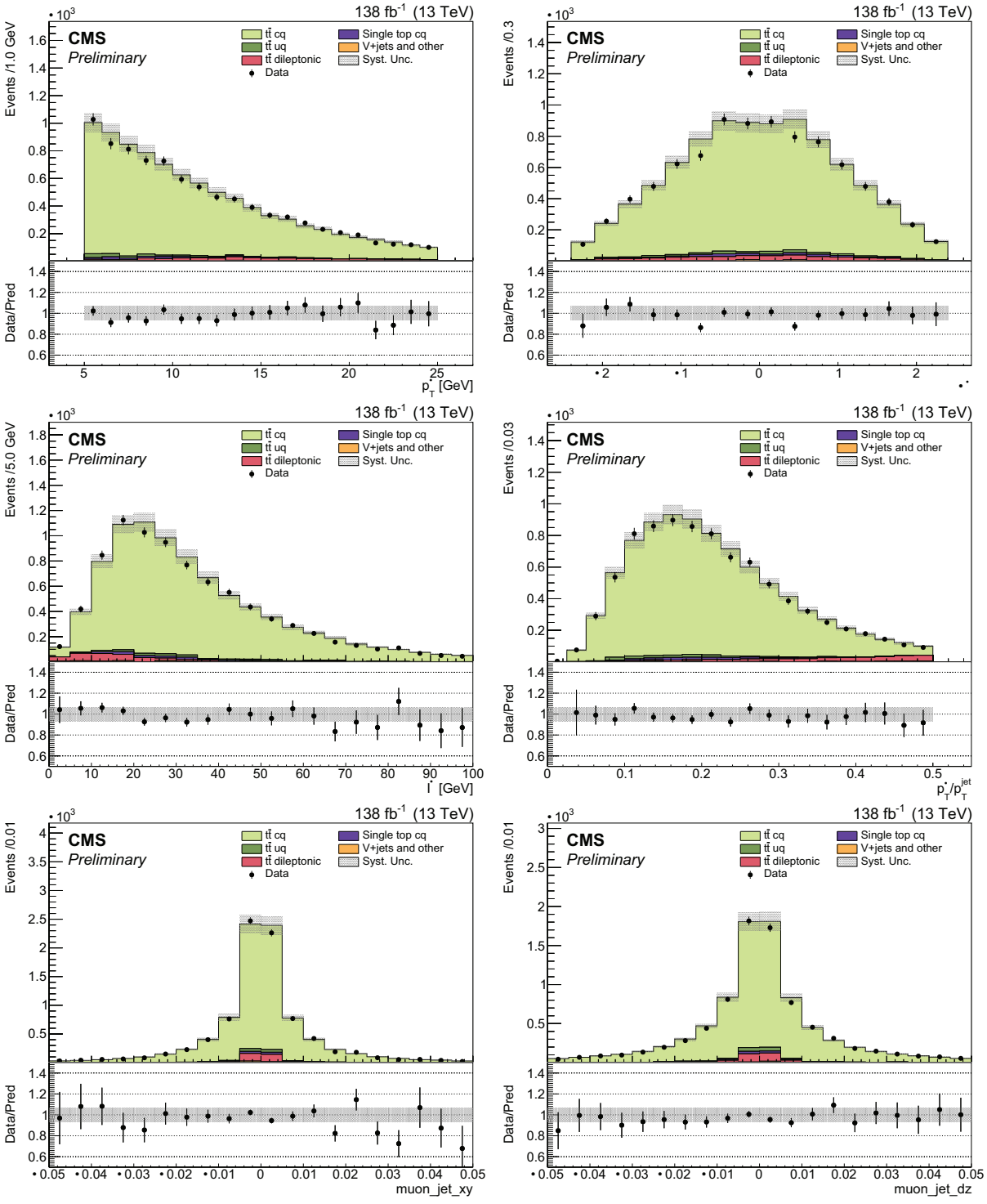


Figure C.1: Charmed-tagged selected sample, exclusive muon channel, OS − SS distributions related to the muon inside the tagged c jet: Muon p_T (top-left), η (top-right), isolation (middle-left), $z = p_T^{\mu}/p_T^{jet}$ (middle-right), transverse impact parameter d_{xy} (bottom-left), and longitudinal impact parameter d_z (bottom-right).

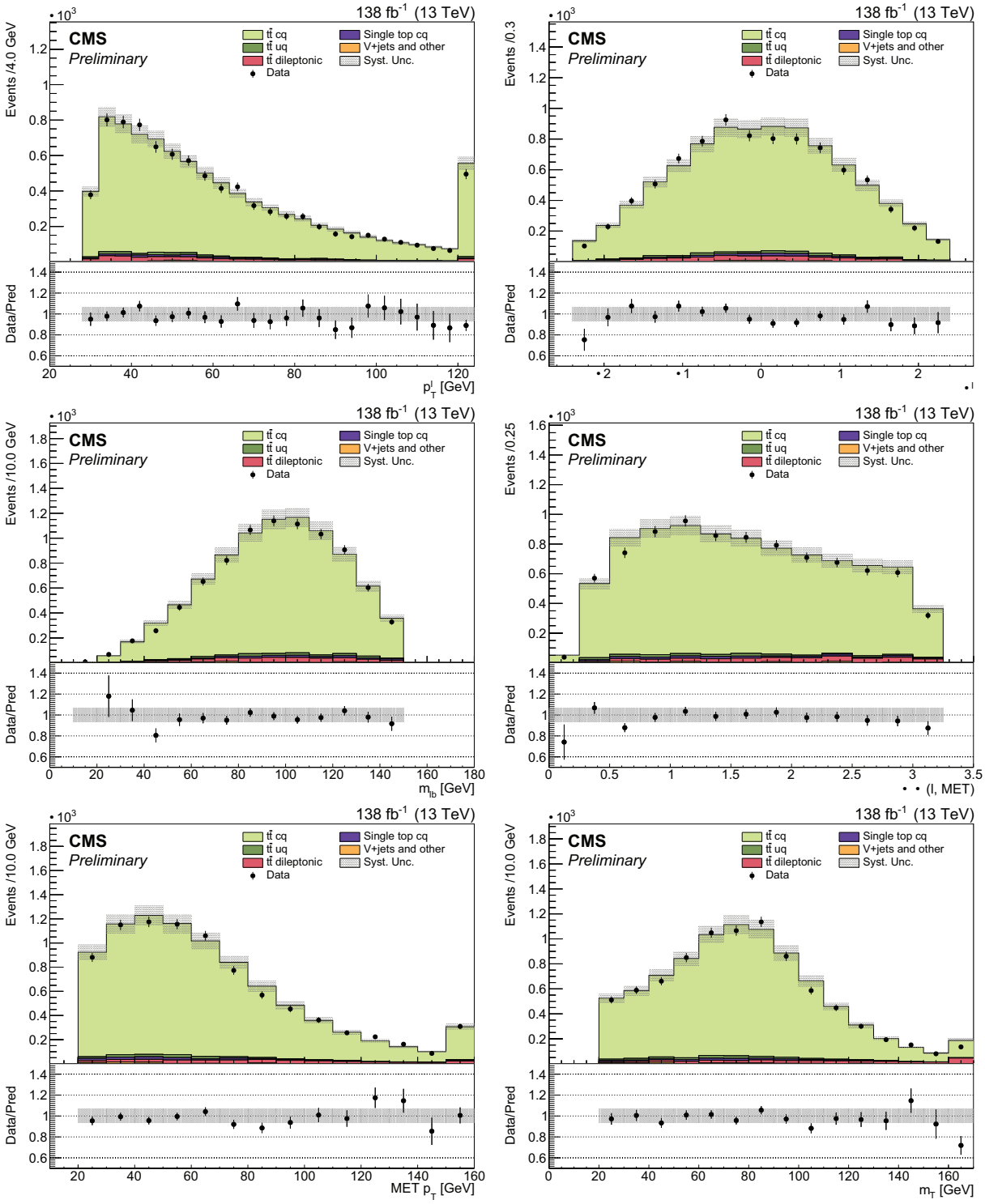


Figure C.2: Charmed-tagged selected sample, exclusive muon channel, OS − SS distributions related to the prompt high-p_T lepton: lepton p_T (top-left), lepton η (top-right), invariant mass of the lepton and the b jet (middle-left), difference in azymuthal angle of the lepton and $\vec{p}_T^{\,miss}$ (middle-right), missing momentum p_T^{miss} (bottom-left), and transverse mass (bottom-right).

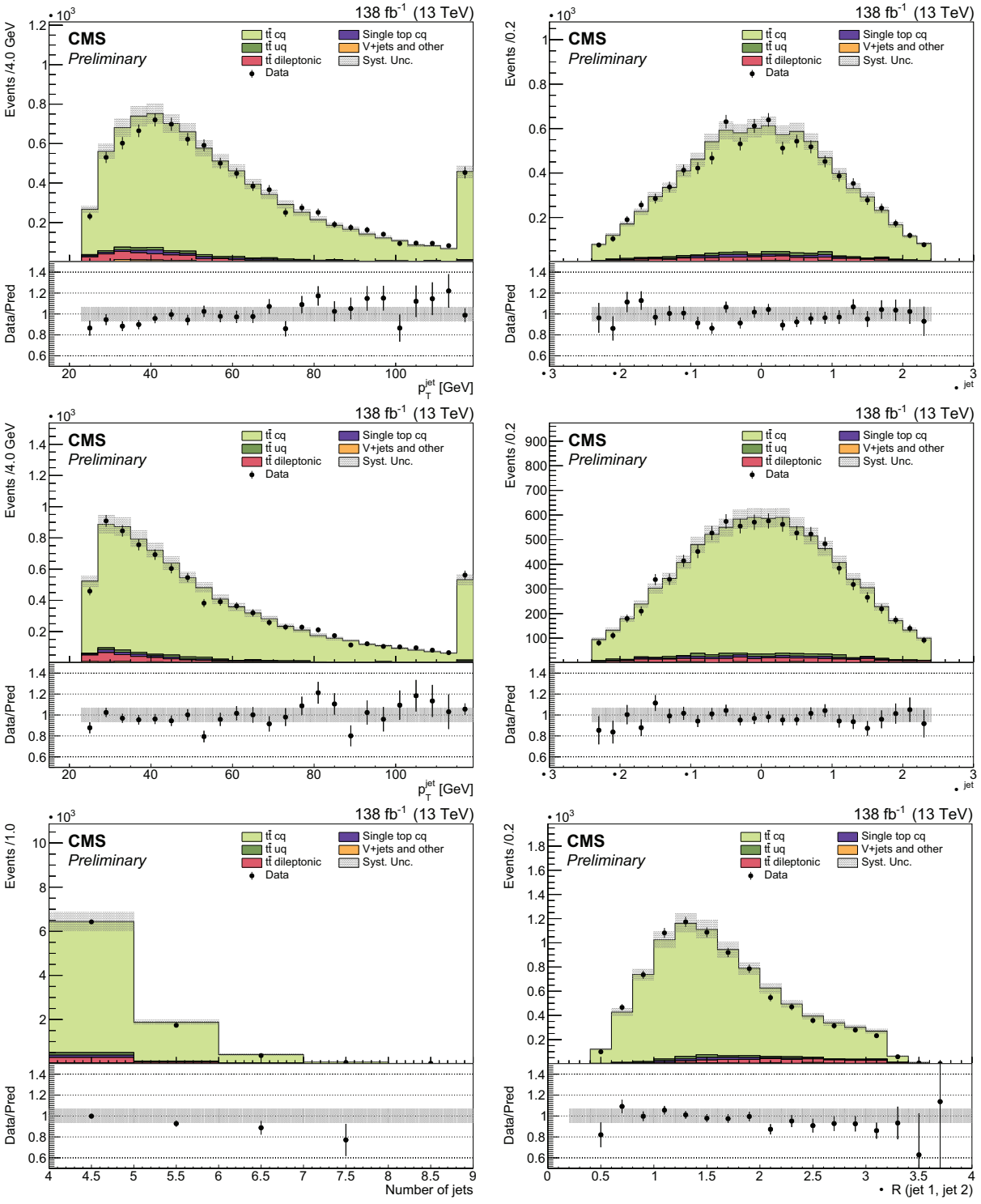


Figure C.3: Charmed-tagged selected sample, exclusive muon channel, OS − SS distributions related to the jets associated with the W boson: jet p_T (top-left) and jet η (top-right) for the jet tagged as c-jet, jet p_T (middle-left) and jet η (middle-right) for the other jet, event jet multiplicity (bottom-left) and ΔR between the two jets (bottom-right).

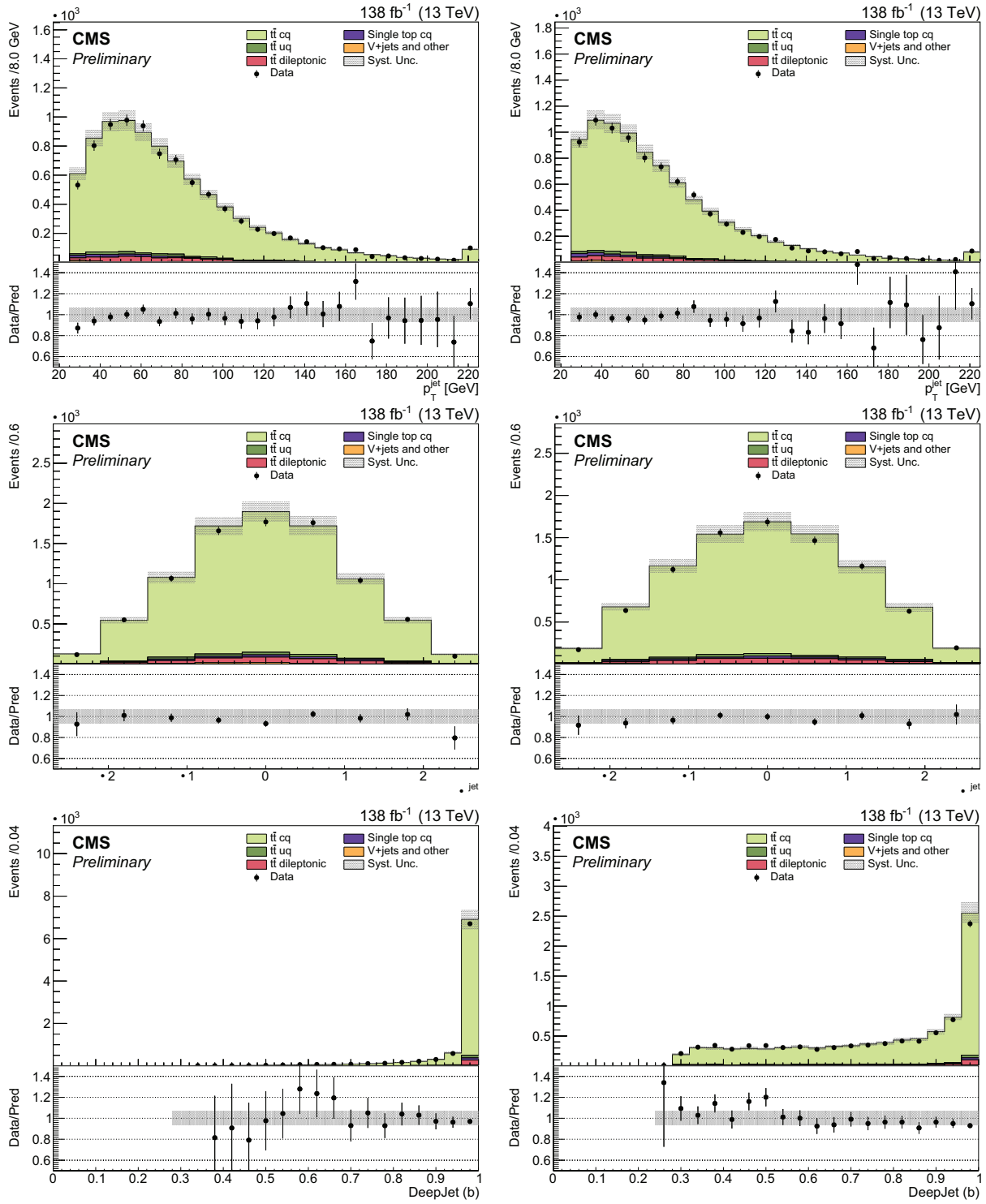


Figure C.4: Charmed-tagged selected sample, exclusive muon channel, OS − SS distributions related to the b-tagged jets: jet p_T, η, and b-tag score for the jet with the highest b-tag score(left); jet p_T, η, and b-tag score for the other b-tagged jet (right).

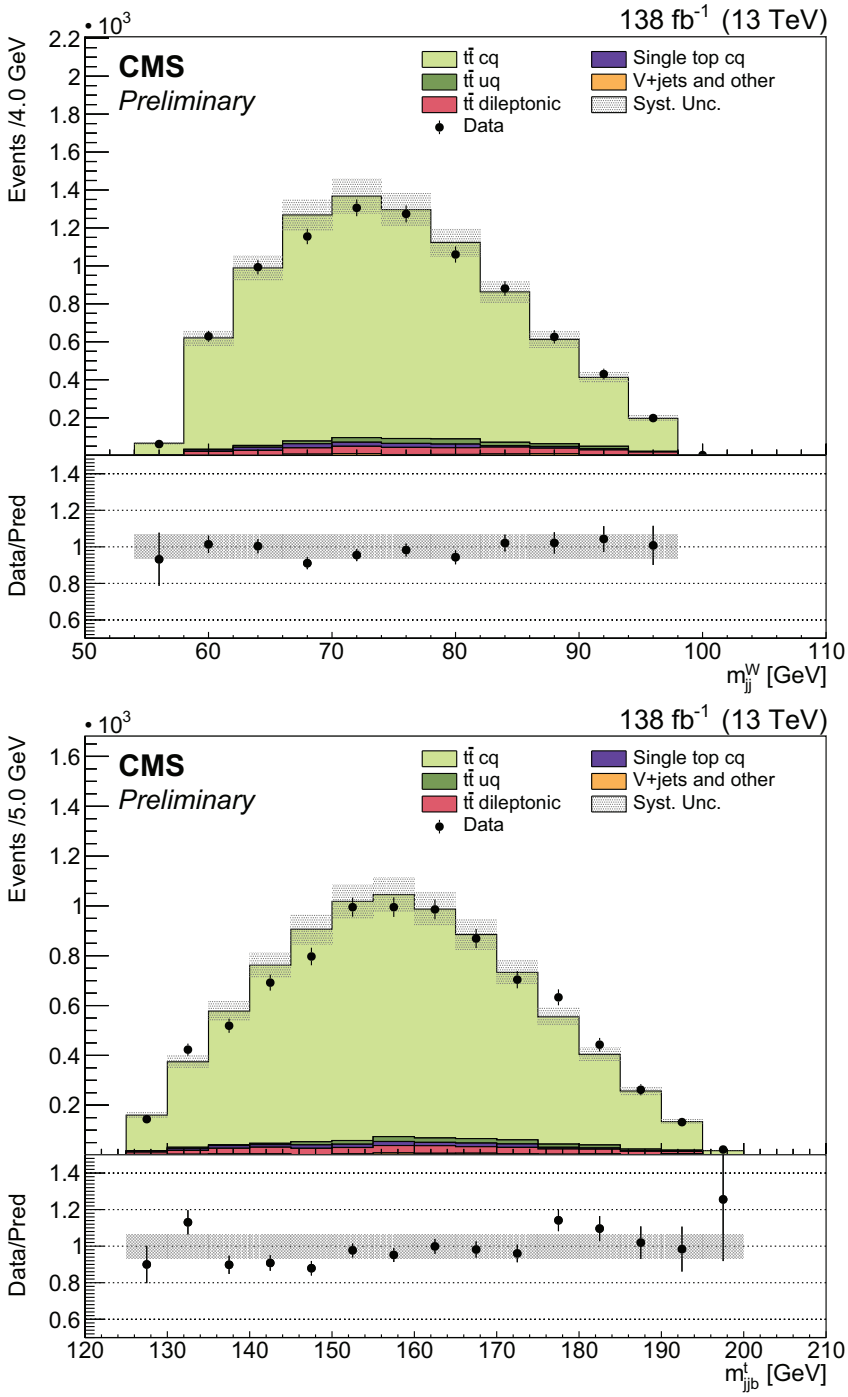


Figure C.5: Charmed-tagged selected sample, exclusive muon channel, OS − SS distributions for the invariant mass of the two jets associated with the W boson (top) and invariant mass of the three jets associated with the top quark (bottom).

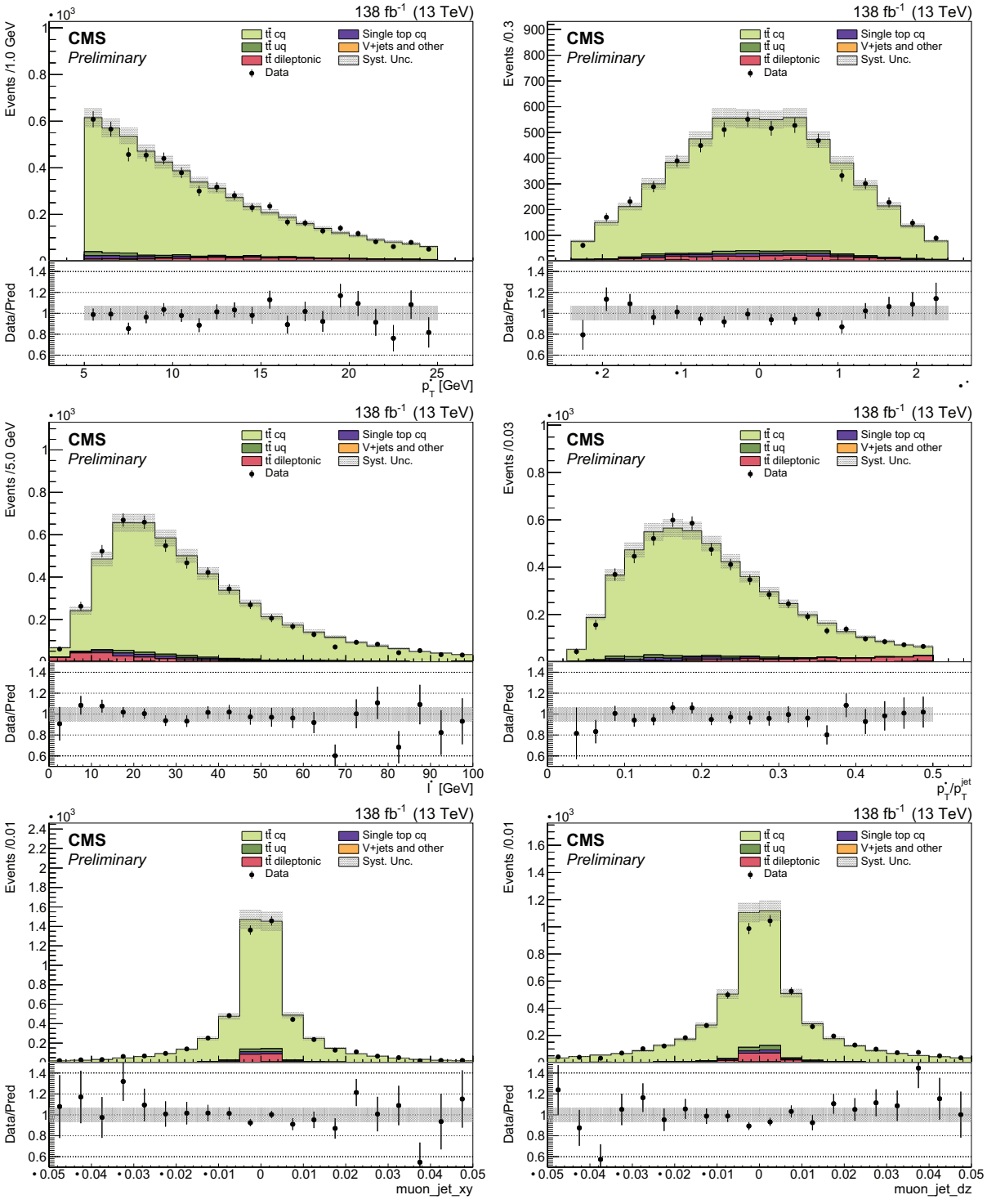


Figure C.6: Charmed-tagged selected sample, exclusive electron channel, OS − SS distributions related to the muon inside the tagged c jet: Muon p_T (top-left), η (top-right), isolation (middle-left), $z = p_T^{\mu}/p_T^{jet}$ (middle-right), transverse impact parameter d_{xy} (bottom-left), and longitudinal impact parameter d_z (bottom-right).

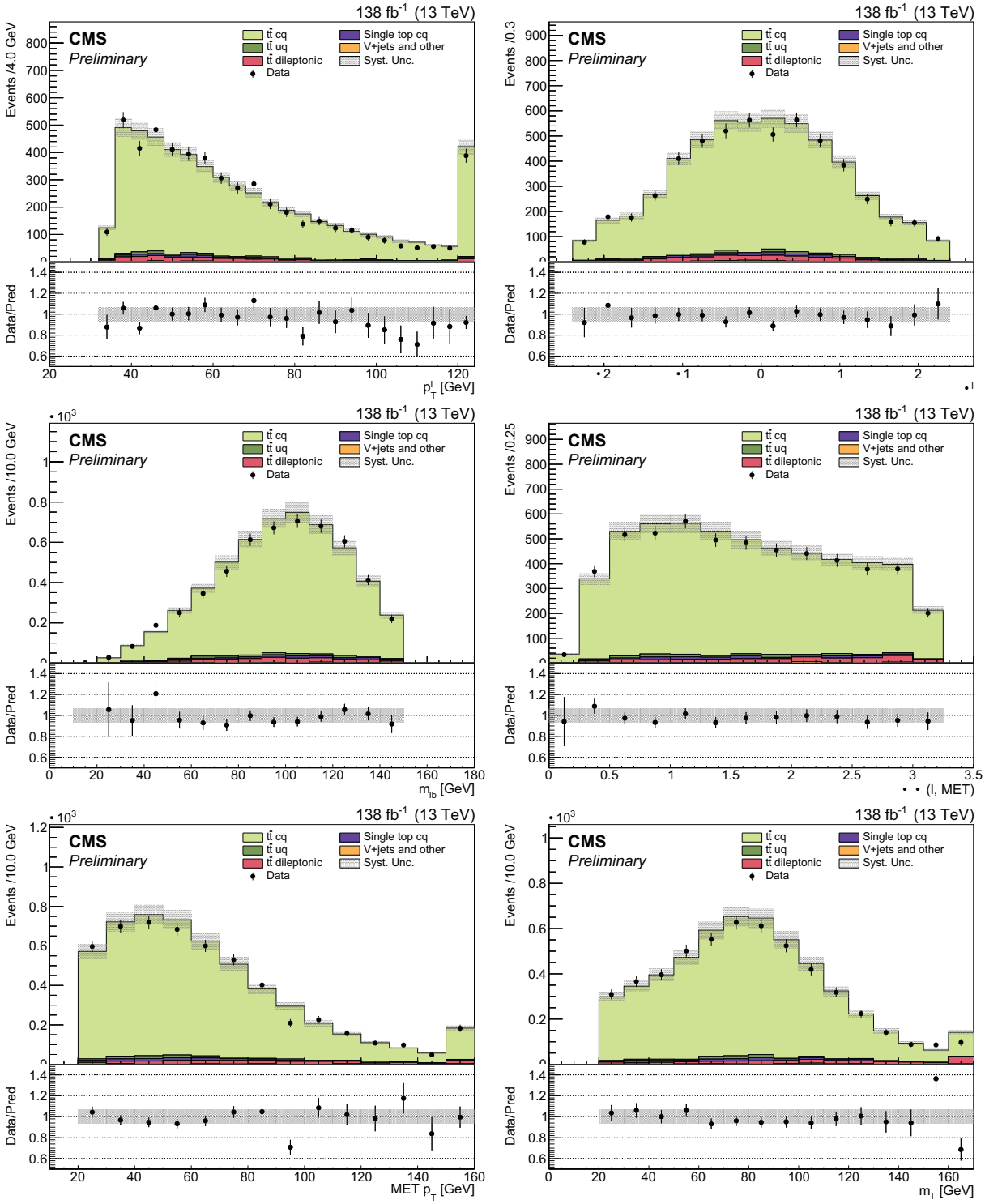


Figure C.7: Charmed-tagged selected sample, exclusive electron channel, OS − SS distributions related to the prompt high-p_T lepton: lepton p_T (top-left), lepton η (top-right), invariant mass of the lepton and the b jet (middle-left), difference in azymuthal angle of the lepton and $\vec{p}_T^{\,miss}$ (middle-right), missing momentum p_T^{miss} (bottom-left), and transverse mass (bottom-right).

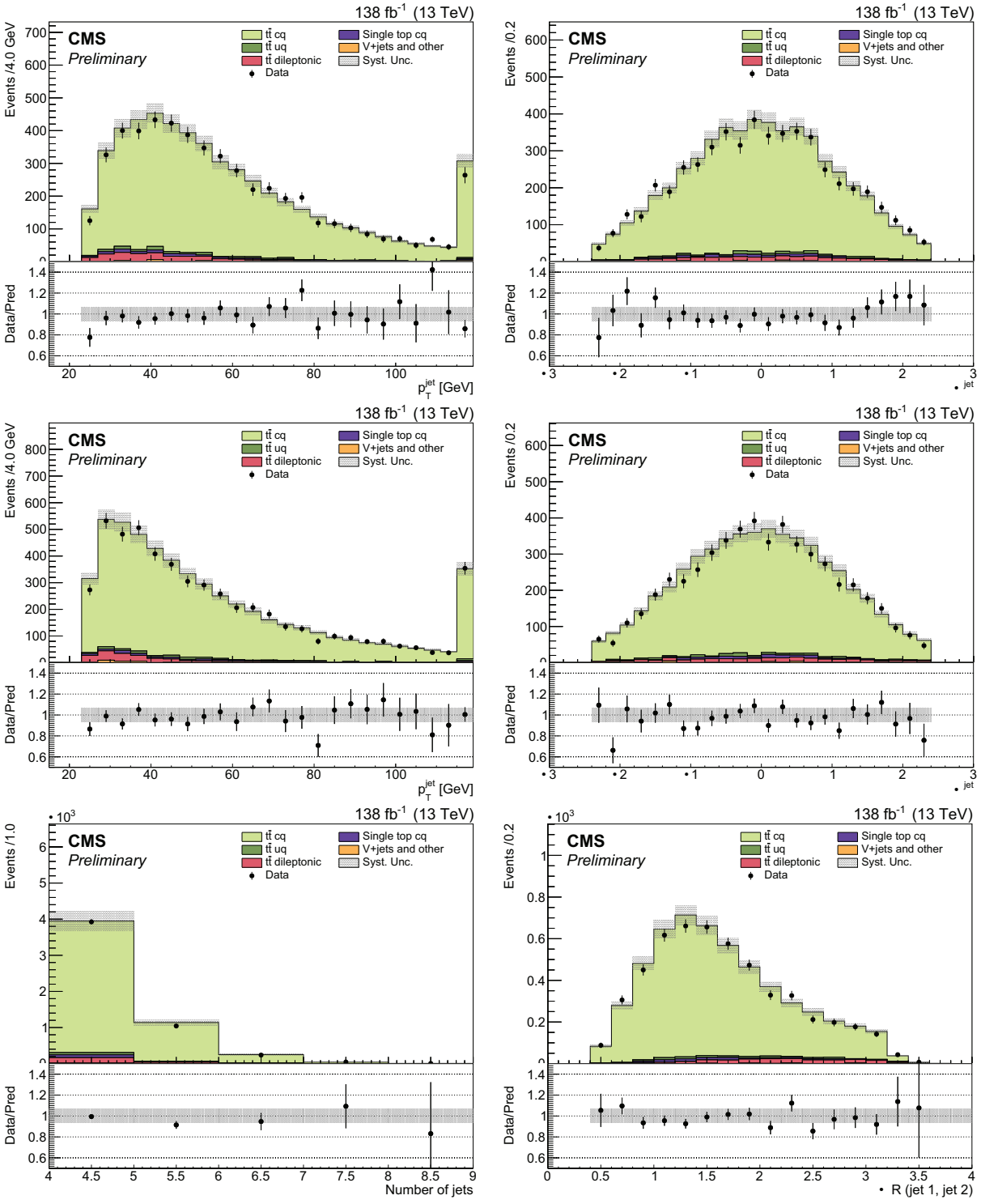


Figure C.8: Charmed-tagged selected sample, exclusive electron channel, OS − SS distributions related to the jets associated with the W boson: jet p_T (top-left) and jet η (top-right) for the jet tagged as c-jet, jet p_T (middle-left) and jet η (middle-right) for the other jet, event jet multiplicity (bottom-left) and ΔR between the two jets (bottom-right).

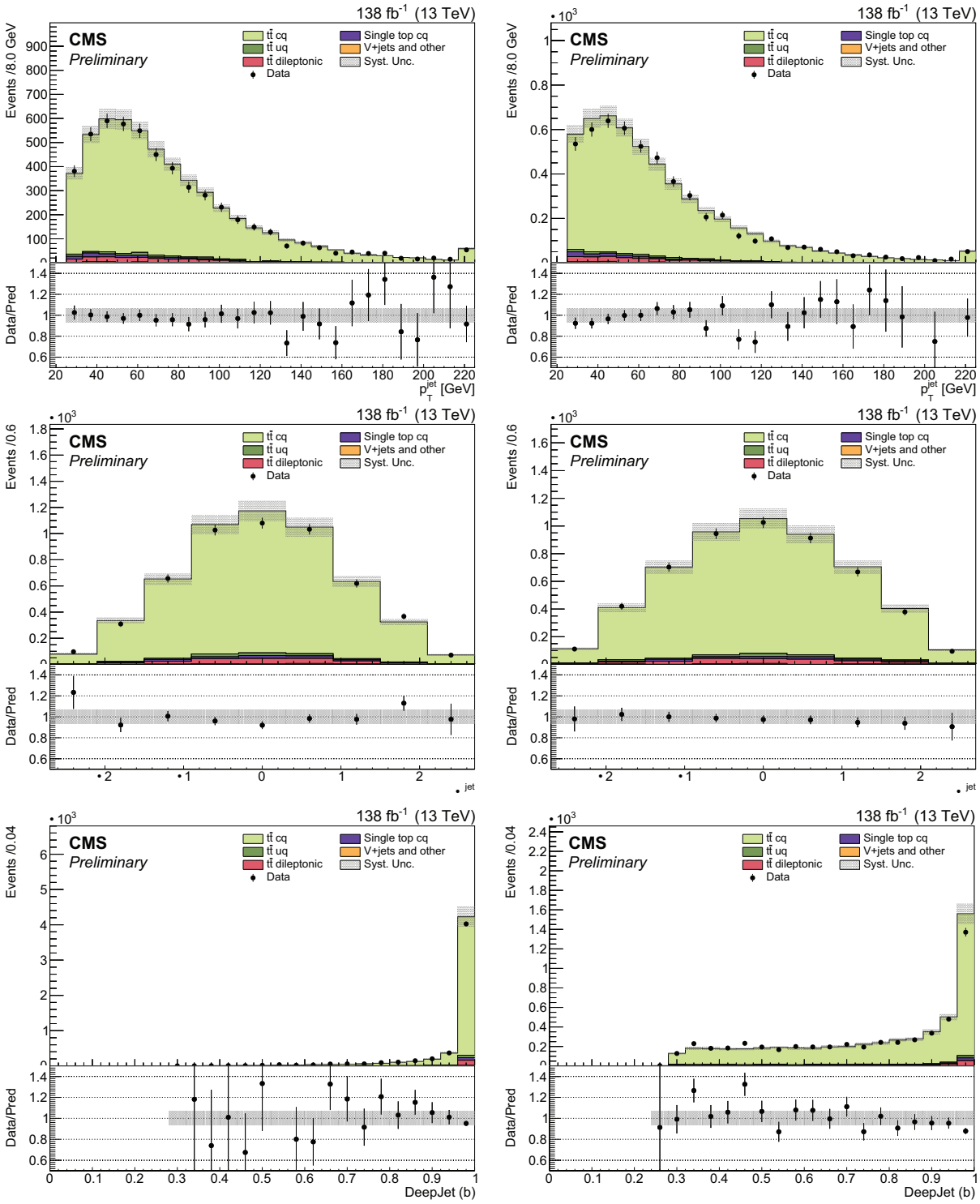


Figure C.9: Charmed-tagged selected sample, exclusive electron channel, OS – SS distributions related to the b-tagged jets: jet p_T, η, and b-tag score for the jet with the highest b-tag score(left); jet p_T, η, and b-tag score for the other b-tagged jet (right).

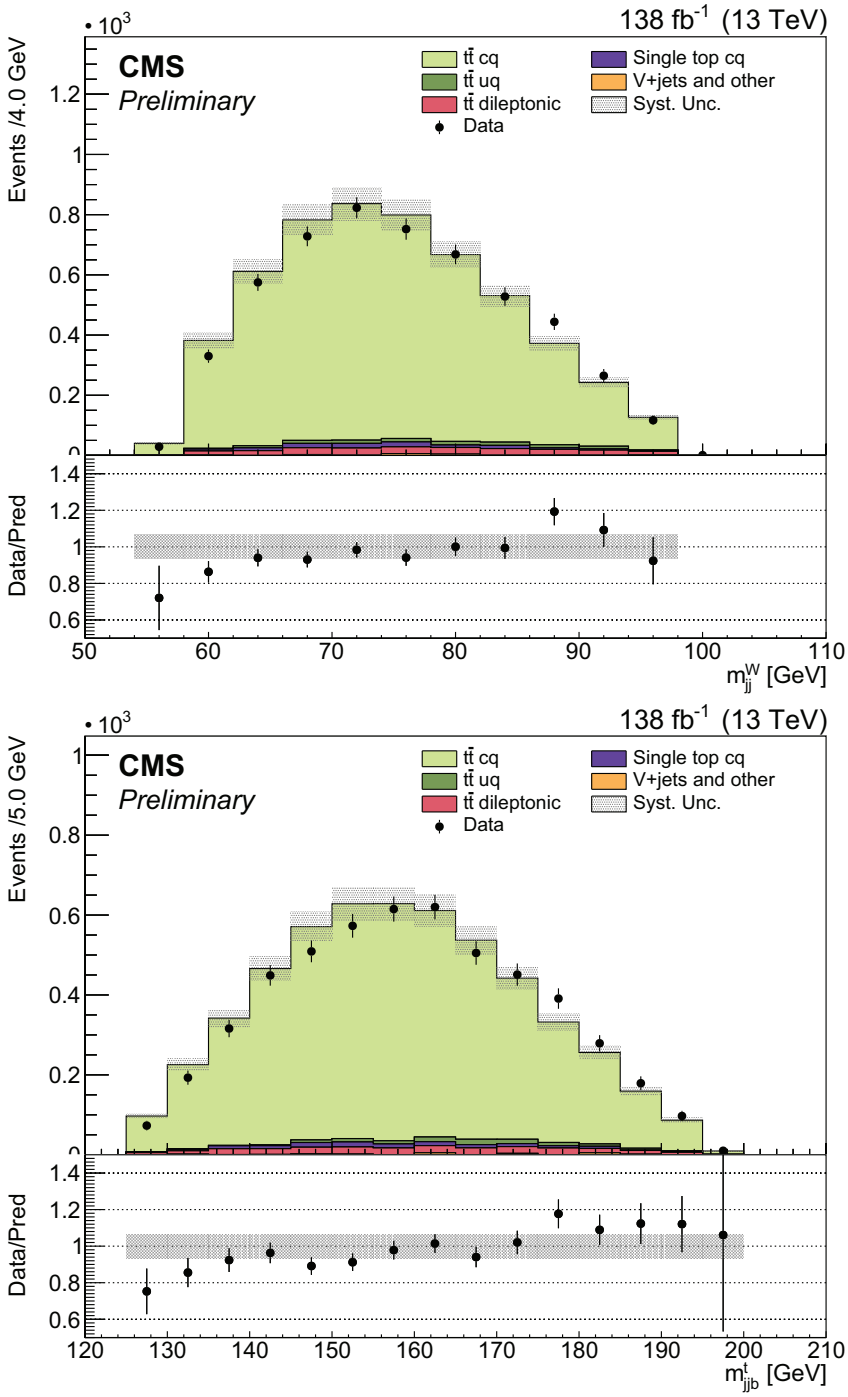


Figure C.10: Charmed-tagged selected sample, exclusive electron channel, OS−SS distributions for the invariant mass of the two jets associated with the W boson (top) and invariant mass of the three jets associated with the top quark (bottom).

Appendix D

Computation of results for uncertainty estimation

Computation of the classifier distribution

In order to assess the discrimination capability of the classification algorithms, we construct with the test sample the distribution of the values of the classification parameter, separately for true signal and background events. The resulting histogram, normalized to unit area, can be regarded as the probability density function of the classification parameter.

It is assumed that the prediction for an event follows a normal distribution, being the mean of the distribution the average prediction and the standard deviation the square root of the prediction variance. It is assumed then that $^{-}$ Var where , having events to be labeled.

We divide the space of prediction, which is , into a desired number of bins . The fact that the prediction for an event falls into a bin can be viewed as if it followed a Bernouilli distribution , where is the probability of an event to fall into a bin (delimited by) calculated as given $^{-}$ Var , .

The number of events with predicted value in a bin will follow, therefore, a sum of Bernouilli distributions. We estimate the expected value of each bin and its variance. The weights of each event is taken into account when computing these quantities.

following the properties , where and are random variables and and are constants, and if and are independent from each other.

Calculation of classification performance metrics

Expected values and variance for TPR and FPR are calculated following a similar procedure than that outlined in appendix A. The probability of an event to be classified as signal, given a cut classification parameter is calculated as given $^{-}$ Var . If an event is classified as signal or not is modeled as a Bernouilli distribution . As a reminder, the TPR value for a cut point is the ratio between the true signal events predicted as such and all the condition signal events. Therefore, we can model the TPR as a sum of Bernouillis multiplied by a constant. Likewise, the FPR, ratio between true background events classified as signal and condition background, can be obtained. The expected values and the variance are estimated.

TPR —— TPR ——

FPR —— FPR ——

where is the set of condition signal events and the set of condition background. are the event weights that are used to combine events in the test sample according to their real proportion in nature.

List of Figures

List of Tables