Search for new particles decaying to top quark pairs with the CMS experiment

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

vorgelegt von

Ksenia de Leo

Hamburg

2025

Gutachter/innen der Dissertation:	Prof. Dr. Johannes Haller Prof. Dr. Peter Schleper
Zusammensetzung der Prüfungskommission:	Prof. Dr. Johannes Haller Prof. Dr. Peter Schleper Prof. Dr. Jochen Liske PD Dr. Roman Kogler Dr. Jürgen Reuter
Vorsitzende/r der Prüfungskommission:	Prof. Dr. Jochen Liske
Datum der Disputation:	11.04.2025
Vorsitzender des Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Wolfgang J. Parak
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Markus Drescher
Dekan der Fakultät MIN:	Prof. Dr.-Ing. Norbert Ritter

Abstract

In this thesis a search for new particles decaying to top quark pairs is presented. The analysis is based on proton-proton collisions recorded with the CMS experiment at 13 TeV in the years 2016-2018, corresponding to an integrated luminosity of 138 fb^{-1}. Many theories of physics Beyond the Standard Model predict the existence of new particles that modify the $\text{t}\bar{\text{t}}$ mass spectrum and could explain some of the shortcomings of the Standard Model, connected for example to the hierarchy problem and the electroweak symmetry breaking. The models considered in this thesis include spin-1 particles, i.e. Z' bosons and g_{KK} gluons at the multi-TeV scale, as well as spin-0 particles, scalar (H) or pseudoscalar (A) heavy Higgs bosons at masses up to 1 TeV. The heavy Higgs signals present interference with the $\text{t}\bar{\text{t}}$ background, resulting in a peak-dip structure in the $\text{t}\bar{\text{t}}$ invariant mass spectrum, while spin-1 particles manifest themselves as peaks. The search is performed in the final state with a muon or an electron, jets and missing transverse momentum. Both the resolved and the boosted final state topologies are probed. In particular, novel machine-learning algorithms are used to identify the hadronic decay of the top quark in the highly Lorentz-boosted regime, where its decay products are collimated. Furthermore, a deep neural network for event classification is applied to categorize the events in the main backgrounds. Upper limits are placed on the production cross section of new spin-1 particles: masses up to 4.3 TeV, 5.3 TeV and 6.7 TeV are excluded for Z' bosons with 1%, 10% and 30% relative widths, respectively, and up to 4.7 TeV for g_{KK} gluons. Moreover, exclusion limits are placed on the coupling strength modifiers of H and A bosons for masses in the range 365-1000 GeV and 2.5% relative width.

The high instantaneous luminosity reached by the LHC in Run 2 leads to a high number of additional pp interactions in the same bunch crossing (pileup). It is fundamental to identify the interaction of interest in each event and to mitigate the effects of pileup on the object reconstruction. The PUPPI algorithm, used in CMS since Run 2, shows the best performance and it is the default algorithm in CMS for Run 3 and beyond. The new version of the algorithm for the re-reconstruction of Run 2 data, the Ultra Legacy (UL) reconstruction, is presented in this thesis. The new tune, PUPPI v15, features an improved track-vertex association that leads to an improved jet energy and p_T^{miss} resolution. PUPPI v15 is used in all the CMS analysis based on UL Run 2 data that use large-radius jets.

Zusammenfassung

In dieser Arbeit wird eine Suche nach neuen Teilchen, die in Top-Quark-Paare zerfallen, präsentiert. Die Analyse wird mit Proton-Proton-Kollisionen, die mit dem CMS-Experiment bei 13 TeV in den Jahren 2016 − 2018 durchgeführt. Die Daten entsprechen einer integrierten Luminosität von 138 fb^{-1}. Viele Theorien der Physik jenseits des Standardmodells sagen die Existenz neuer Teilchen voraus, die das $\text{t}\bar{\text{t}}$-Massenspektrum verändern, und könnten einige Effekte, die das Standardmodell nicht beschreibt, enklären, die beispielsweise mit dem Hierarchieproblem und der elektroschwachen Symmetriebrechung zusammenhängen. Die in dieser Arbeit berücksichtigten Modelle umfassen Spin-1-Teilchen, d. h. Z′-Bosonen und g_{KK}-Gluonen auf der Multi-TeV-Skala, sowie Spin-0-Teilchen, skalare (H) oder pseudoskalare (A) schwere Higgs-Bosonen mit Massen bis zu 1 TeV. Im Falle schwerer Higgs-Bosonen tritt Interferenz zwischen dem Signal und den $\text{t}\bar{\text{t}}$-Untergrundprozessen auf, was zu einer "Peak-Dip"-Struktur im $\text{t}\bar{\text{t}}$-invarianten Massenspektrum führt, während sich Spin-1-Partikel als Peaks manifestieren. Die Suche wird im Endzustand mit einem Myon oder einem Elektron, Jets und fehlendem Transversalimpuls durchgeführt. Es werden sowohl das Regime aufgelöster Jets als auch mit hohem Lorentz-Boost untersucht. Insbesondere werden neuartige Algorithmen des maschinellen Lernens verwendet, um den hadronischen Zerfall des Top-Quarks im Regime mit hohem Lorentz-Boost zu identifizieren, wo seine Zerfallsprodukte kollimiert sind. Darüber hinaus wird ein neuronales Netz zur Klassifizierung der Ereignisse in verschiedene Untergrundprozesse eingesetzt. Es werden obere Ausschlussgrenzen auf den Produktionswirkungsquerschnitt neuer Spin-1-Teilchen bestimmt. Massen bis zu 4.3 TeV, 5.3 TeV und 6.7 TeV für Z′-Bosonen mit 1%, 10% und 30% relativer Breite und bis zu 4.7 TeV für g_{KK}-Gluonen werden auf dem 95% Konfidenzlevel ausgeschlossen. Darüber hinaus werden obere Ausschlussgrenzen für die Kopplungsstärkemodifikatoren von H- und A-Bosonen für Massen im Bereich von $365 - 1000$ GeV und 2.5% relative Breite bestimmt.

Die hohe instantane Luminosität, die der LHC in Run 2 erreicht, führte zu einer hohen Zahl von zusätzlichen Proton-Proton-Interaktionen pro Bunch-Crossing (Pileup). Für die Analyse der Daten ist es von herausragender Bedeutung, in jedem Kollisionseregnis den harten Interaktionsprozess zu identifizieren und den Einfluss von Pileup auf die Objektrekonstruktion zu minimieren. Der PUPPI-Algorithmus, der seit Run 2 in CMS verwendet wird, zeigt hierbei die beste Performance, und ist zum Standardalgorithmus in CMS für Run 3 und darüber hinaus geworden. Die neue Version des Algorithmus für die Re-Rekonstruktion der Run 2-Daten, "Ultra Legacy" (UL) Rekonstruktion, wird in dieser Arbeit vorgestellt. Diese neue Version, PUPPI v15, verfügt über eine verbesserte Spur-Vertex-Zuordnung, die zu einer verbesserten Jetenergie- und p_T^{miss}-Auflösung führt. PUPPI v15 wird in allen CMS-Analysen verwendet, die auf UL Run 2-Daten basieren und Jets mit großem Radius verwenden.

List of own contributions

Search for new particles in the $t\bar{t}$ final state

I am the main analyzer of the CMS analysis searching for new particles decaying to top quark pairs in the lepton+jets final state. My own main contributions to the analysis are:

- Optimization of the analysis strategy and improvement of the sensitivity at low masses.
- Inclusion of a new signal interpretation with interference.
- Production of the simulation samples for the spin-1 signals.
- Studies and comparison of different top-tagging techniques.
- Derivation of data-to-simulation correction factors.
- Development and implementation of a deep neural network for event classification.
- Statistical interpretation of the results for the spin-1 and spin-0 signals.

As contact person, I am responsible for the CMS-internal review of the analysis, which includes:

- Regular presentations in CMS working group meetings.
- Documentation of the analysis strategy in a CMS-internal analysis note.
- Writing and editing the paper draft.
- Pre-approval presentation of the analysis.

The analysis is under the CMS-internal review and the publication is foreseen for the near future. The work was performed under the supervision of Prof. Dr. Johannes Haller and Dr. Roman Kogler, and in collaboration with Dr. Andrea Malara and Henrik Jabusch, whose contributions include the studies and statistical interpretation of non-resonant signals, see Ref. [1] for details.

Pileup mitigation techniques

I worked on the improvement and validation of the PUPPI algorithm for pileup mitigation in CMS. A new tune of the algorithm, PUPPI v15, was developed with an improved track-vertex association, resulting in better jet energy and p_T^{miss} resolution. The tune is implemented in the Ultra Legacy reconstruction of Run 2 data and it is used in the analysis presented in this thesis. The results have been published in a Detector Performance Note [2].

My own contributions consist in analyzing the new tunes produced by Dr. Anna Benecke and comparing them in terms of jet energy resolution, jet reconstruction efficiency and purity and jet substructure variables. Moreover I performed validation studies of the new tunes in Run 2 and Run 3 simulation. The work was performed under the supervision of Dr. Anna Benecke and Dr. Andreas Hinzmann.

Contents

Chapter 1

Introduction

The Standard Model (SM) of particle physics describes the elementary particles that constitute the Universe and three of the four fundamental interactions they experience. The SM is one of the most successful theories in history, thanks to the extremely precise tests of its parameters and the discovery of all the particles it predicts. Nevertheless, many open questions remain and various experimental observations can not be explained by the SM. For example, the gravitational force is not included in the theory and there is no viable candidate for a dark matter particle. To go beyond the known theory is one of the driving forces of particle physics research: many new theories are predicted to solve one or more of the SM shortcomings, and experiments search for new particles and forces that may be hidden at higher and higher energies.

One of the most important portals to new physics is the top quark: being the most massive elementary particle, it is expected to couple to new heavy particles predicted by many Beyond the Standard Model (BSM) theories, which could explain the *electroweak symmetry breaking* or the *hierarchy problem*. Such theories include heavy spin-1 particles at the TeV scale. e.g. Z' bosons or Kaluza-Klein excitations of the gluons (g_{KK}), or spin-0 particles, e.g. additional heavy Higgs bosons in the Two-Higgs-Doublet Models.

In this thesis a search for new massive particles that decay to top quark pairs in the lepton+jets final state is presented. The search is performed using proton-proton collisions at the Large Hadron Collider (LHC) at CERN, the most powerful particle accelerator and collider in the world. The experimental data are collected with the Compact Muon Solenoid (CMS) detector at the center-of-mass energy of 13 TeV during 2016-2018, corresponding to an integrated luminosity of 138 fb^{-1}. Searches for new particles decaying to top quark pairs have been already performed at the LHC by the ATLAS and CMS Collaborations at various center-of-mass energies, considering all possible final states of the top quark pair decay. To date, no discovery of such particles has been claimed.

The analysis presented in this thesis extends the previous CMS results by analyzing for the first time the full Run 2 dataset of the LHC and interpreting the results for different signal processes, including both spin-0 and spin-1 particles, in a model-independent approach. The spin-0 signals present an interference pattern with the SM $t\bar{t}$ background. The low mass as well as the high mass regimes are explored, that correspond to different final state topologies, and new techniques are used to identify the decay products of the top quarks. Furthermore, the sensitivity is improved with the use of a deep neural network event classifier.

The high instantaneous luminosity reached during Run 2 of the LHC results in a large number of simultaneous pp interactions for each bunch crossing. The identification of the main interaction and the mitigation of the effects of additional interactions (pileup) is of great importance for any physics analysis at the LHC. One of the pileup mitigation techniques used in CMS is the Pile-Up Per Particle Identification (PUPPI) algorithm. In this thesis, the optimization of PUPPI for the Ultra Legacy reconstruction of Run 2 data is presented. Given the great performance, PUPPI has become the official algorithm used in CMS in Run 3 and beyond.

The thesis is organized as follows: in Chapter 2 an overview of the theory of the SM is given and the physics of hadronic collisions is described. The open questions of the SM are presented in Chapter 3, together with the theories of new physics BSM connected to the top quark. The experimental setup is described in Chapter 4, with an overview of the LHC collider and the CMS experiment. In Chapter 5 the reconstruction of the objects in the CMS detector is described, while in Chapter 6 the pileup mitigation techniques used in CMS and the optimization of the PUPPI algorithm are presented. The search for new particles decaying to top quark pairs in the lepton+jets final state is described in Chapter 7. The conclusions of the thesis are discussed in Chapter 8.

Chapter 2

The Standard Model of Particle Physics

The Standard Model provides the theoretical description of the elementary particles and their interactions. The great success of the theory has been granted by numerous experimental confirmations of its predictions over the years and the discovery of all the postulated particles. In this Chapter the elementary particles and the fundamental interactions of the Standard Model will be briefly described (Sec. 2.1). Afterwards, the key elements of the physics of proton-proton collisions will be introduced in Sec. 2.2, and the event simulation with Monte Carlo generators will be presented in Sec. 2.3.

2.1 Particles and interactions of the Standard Model

The Standard Model (SM) of particle physics is the theory that describes the elementary particles and how they interact. It is a relativistic quantum field theory (QFT) and it successfully incorporates three of the four fundamental forces of nature: the electromagnetic, the weak and the strong interactions. Gravity, the fourth force, is not included in the SM, as currently there is not a quantum field theory formulation of this force. The electromagnetic and the weak forces are unified in the electroweak interaction at a scale above 100 GeV, the *electroweak scale*. At this energies the effects of gravity on elementary particles can be neglected. The elementary particles of the SM are the *fermions*, the building blocks of matter, and the *bosons*, the carrier particles associated to the interactions. The elementary particles and their properties are listed in Table 2.1.

Fermions are half-integer spin particles and they are grouped into quarks and leptons,

characterized by different quantum numbers, which indicate how they interact. Quarks and leptons are further categorized into three families, or *generations*, that differ only in the mass, increasing from the first to the third generation. Fermions obey the Dirac equation, which implies that for each particle there is an *anti-particle*, with the same mass and opposite charges. Anti-particles are conventionally represented with the symbol of the corresponding particle with a bar on the top (e.g. $\mathrm{q} \to \bar{\mathrm{q}}$). Depending on the chirality, fermions can be left-handed (negative chirality) or right-handed (positive chirality). Left-handed fermions are grouped into doublets and have weak isospin $T = 1/2$, while right-handed fermions are singlets and have $T = 0$. There are six types of quark flavours: up (u), down (d), charm (c), strange (s), top (t) and bottom (b). They can be grouped into up-type quarks (u, c, t) and down-type quarks (d, s, b). The up-type quarks have electric charge of $Q = +2/3\ e$, expressed in terms of the elementary electric charge e, and the down-type quarks have $Q = -1/3\ e$. The left-handed quark doublets are:

$$\begin{pmatrix} \mathrm{u} \\ \mathrm{d} \end{pmatrix}_L, \begin{pmatrix} \mathrm{c} \\ \mathrm{s} \end{pmatrix}_L, \begin{pmatrix} \mathrm{t} \\ \mathrm{b} \end{pmatrix}_L. \tag{2.1}$$

For up-type quarks the third component of the weak isospin is $T_3 = +1/2$, while down-type quarks have $T_3 = -1/2$. The right-handed quark singlets are:

$$\mathrm{u}_R,\ \mathrm{d}_R,\ \mathrm{c}_R,\ \mathrm{s}_R,\ \mathrm{t}_R,\ \mathrm{b}_R \tag{2.2}$$

and have $T_3 = 0$. Moreover, all quarks carry a color charge, the charge of the strong interaction.

The leptons consist of negatively charged leptons, the electron (e), the muon (μ) and the tau (τ), and the corresponding neutral leptons, the neutrinos ($\nu_e,\ \nu_\mu$ and ν_τ). Charge leptons carry an electric charge of $1e$. Similarly to quarks, leptons can be represented as left-handed doublets:

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L, \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L, \tag{2.3}$$

where the charged leptons have $T_3 = -1/2$ and the neutrinos have $T_3 = +1/2$. The right-handed singlets are:

$$e_R,\ \mu_R,\ \tau_R. \tag{2.4}$$

There are no right-handed neutrinos, as they are treated as massless particles in the SM. However, the observation of neutrino oscillations [3, 4], predicted by Pontecorvo in 1957 [5], is a demonstration that they are massive.

Bosons are integer spin particles that mediate the interactions among particles. The electromagnetic interaction is mediated by the photon (γ), the gluon (g) is the mediator

	generation	particle	spin	charge	mass
quarks	I	u	1/2	2/3	2.2 MeV
		d	1/2	-1/3	4.7 MeV
	II	c	1/2	2/3	1.3 GeV
		s	1/2	-1/3	93.4 MeV
	III	t	1/2	2/3	172.7 GeV
		b	1/2	-1/3	4.2 GeV
leptons	I	e	1/2	-1	511 keV
		ν_e	1/2	0	< 0.8 eV
	II	μ	1/2	-1	105.7 MeV
		ν_μ	1/2	0	< 0.8 eV
	III	τ	1/2	-1	1.8 GeV
		ν_τ	1/2	0	< 0.8 eV
bosons	-	γ	1	0	0
	-	g	1	0	0
	-	$W^\pm$	1	± 1	80.4 GeV
	-	Z	1	0	91.2 GeV
	-	H	0	0	125.3 GeV

Table 2.1: Table of the particles of the SM and their properties. Values from [11].

of the strong interaction and the $W^\pm$ and Z bosons are the carriers of the charged and neutral weak interaction, respectively. The photon and the gluon are massless, while the $W^\pm$ and Z bosons have mass. The $W^\pm$ has $Q = \pm 1e$ and $T_3 = \pm 1$, while the other bosons have $Q = 0$ and $T_3 = 0$. Among the bosons, the gluon is the only one carrying the color charge. The Higgs boson, a spin-0 *scalar boson*, completes the list of particles of the SM. It has been predicted in the 1960s by Higgs, Englert and Brout [6, 7] and discovered in 2012 by the CMS [8] and ATLAS [9] Collaborations at CERN. Through the interaction with the Higgs, all the particles acquire mass.

The following sections describe the fundamental interactions and are based on Refs. [10] and [11], unless stated otherwise. Natural units are used: $\hbar = c = 1$.

2.1.1 The electromagnetic interaction

The electromagnetic (EM) force is described by the QFT of Quantum Electrodynamics (QED) and it is based on the $U(1)_{EM}$ group symmetry. The charge which is conserved in QED is the electric charge Q. The photon γ is the gauge boson in the interaction: it is a massless, spin-1 particle and has no electric charge, which implies that no self-interaction of

the photon in QED is allowed. The Lagrangian density of the electromagnetic interaction is:

$$\mathcal{L}_{\mathrm{EM}} = \bar{\psi}(i\gamma_\mu D^\mu - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{2.5}$$

where ψ is the fermion field with mass m, γ_μ are the Dirac matrices and $F_{\mu\nu}$ is the EM field tensor. The D^μ is the covariant derivative: $D^\mu = \partial^\mu + iqA^\mu$, where A^μ is the field that can be identified as the photon. The coupling strength α_{EM} of QED, called *fine structure constant*, is given by:

$$\alpha_{\mathrm{EM}} = \frac{e^2}{4\pi\epsilon_0} \approx \frac{1}{137} \tag{2.6}$$

where ϵ_0 is the vacuum permittivity. The coupling strength increases with the momentum transfer q at which the interactions occur. Given that the mediator of the force is massless, the electromagnetic force has infinite range and decreases as $1/r^2$, where r is the distance among the interacting particles.

2.1.2 The strong interaction

The strong interaction is described by the Quantum Chromodynamics (QCD), a gauge theory based on the gauge group $\mathrm{SU(3)_C}$. The conserved charge in QCD is the color (C) charge. The Lagrangian of the QCD can be expressed similarly to the EM Lagrangian:

$$\mathcal{L}_{\mathrm{QCD}} = \bar{\psi}_q(i\gamma^\mu D_\mu - m)\psi_q - \frac{1}{4}G^a_{\mu\nu}G^{a\mu\nu} \tag{2.7}$$

and it acts on the quark fields ψ_q. The $G^a_{\mu\nu}$ is the field strength tensor and D_μ the covariant derivative, defined as:

$$D_\mu = \partial_\mu + ig_S t^a A^a_\mu \tag{2.8}$$

where g_S is the strong coupling constant, A^a_μ are the gluon fields and t^a are the eight generators of $\mathrm{SU(3)_C}$ and they are proportional to the Gell-Mann matrices $t^a = \frac{1}{2}\lambda^a$. Since the generators are 3×3 matrices, it follows that the quarks will have three additional degrees of freedom, the three color charges: red, blue and green. The mediators of QCD are thus eight massless gluons, which carry color charge themselves, therefore self-interaction of gluons is possible. While gluons carry a color and an anti-color, quarks carry one color charge.

A peculiar property of QCD is that the strong coupling, which can be expressed as $\alpha_S = g_S^2/4\pi$, behaves differently depending on the energy scale q of the interaction (Fig. 2.1). This is why it is referred to as *running coupling*. In particular, at low energies ($q \sim 1$ GeV) - or large distances - the coupling has values $\alpha_S \sim \mathcal{O}(1)$ and QCD processes can not be calculated with perturbation theory. In this regime, quarks can not exist freely

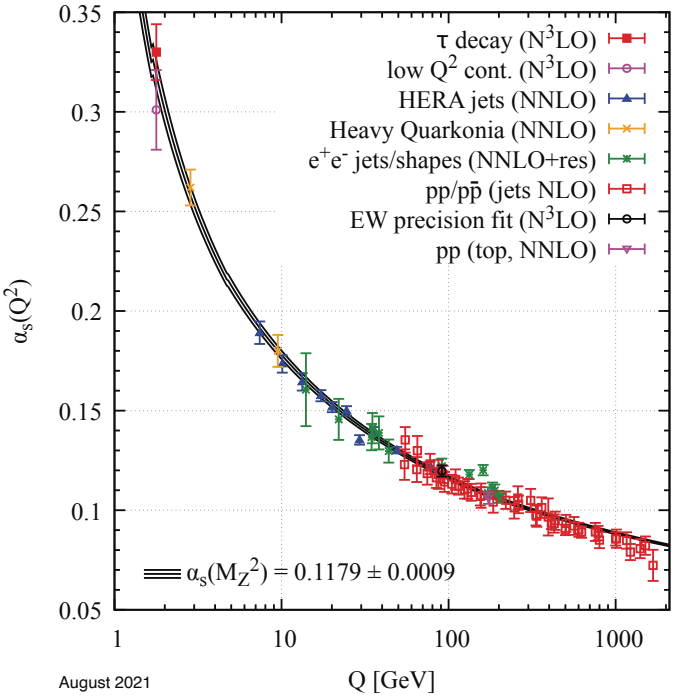

Figure 2.1: The strong coupling α_S as a function of the energy scale Q from different experimental measurements. Figure from [11].

and are forced to be in colourless bound states. This phenomenon is known as *confinement* and arises from the gluon self-interaction in QCD. Quarks are thus grouped together into *hadrons*: they can be formed by a quark-antiquark pair (mesons) or by three quarks (baryons). At higher energies ($q \sim 100$ GeV) the strong coupling decreases ($\alpha_S \sim 0.1$) and perturbation theory can be used. In this regime quarks can be treated as quasi-free particles inside the hadrons, a property known as *asymptotic freedom*.

2.1.3 The weak interaction

The weak interaction is described by the QFT based on the $SU(2)_L$ symmetry. The mediators of the force are the two charged W^+ and W^- bosons, with mass of 80.377 ± 0.012 GeV, and the neutral Z boson, with mass 91.1876 ± 0.0021 GeV. The charge related to the weak interaction is the weak isospin T and in particular its third component T_3 is conserved in the interaction. The charged current (CC) weak interaction is mediated by the $W^{\pm}$ bosons and it is the only interaction in the SM that violates parity: as a consequence only left-handed particles and right-handed anti-particles participate in the CC interaction. The Z boson, that mediates the neutral current (NC) interaction, should couple as well to left-handed particles and right-handed anti-particles, but experimentally it has been

observed that the Z couples both to left- and right-handed particles (and anti-particles), even though with different couplings. This behaviour can be explained by the electroweak unification, described in Sec. 2.1.4. Another difference between weak neutral and charge interactions is that the NC occurs between quarks of the same flavour, while the CC can occur between quarks of different generations. In the lepton sector, both the W and Z couple to leptons of the same generation. The reason why the CC can occur between quarks of different flavours is that the mass eigenstates of the quarks do not coincide with the weak eigenstates. The mixing mechanism is described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix [12, 13]. The CKM matrix ($\mathrm{V_{CKM}}$) provides the relation between the quark mass eigenstates q and the flavour ones q′. The relation is given by:

$$\begin{pmatrix} \mathrm{d}' \\ \mathrm{s}' \\ \mathrm{b}' \end{pmatrix} = \mathrm{V_{CKM}} \begin{pmatrix} \mathrm{d} \\ \mathrm{s} \\ \mathrm{b} \end{pmatrix} = \begin{pmatrix} \mathrm{V_{ud}} & \mathrm{V_{us}} & \mathrm{V_{ub}} \\ \mathrm{V_{cd}} & \mathrm{V_{cs}} & \mathrm{V_{cb}} \\ \mathrm{V_{td}} & \mathrm{V_{ts}} & \mathrm{V_{tb}} \end{pmatrix} \begin{pmatrix} \mathrm{d} \\ \mathrm{s} \\ \mathrm{b} \end{pmatrix}. \tag{2.9}$$

The $\mathrm{V_{CKM}}$ is a unitary matrix with four free parameters: three mixing angles and one complex phase. The probability of a transition from an up-type quark i to a down-type quark j is given by $|\mathrm{V}_{ij}|^2$; the values obtained experimentally [11] are:

$$\begin{pmatrix} |\mathrm{V_{ud}}| & |\mathrm{V_{us}}| & |\mathrm{V_{ub}}| \\ |\mathrm{V_{cd}}| & |\mathrm{V_{cs}}| & |\mathrm{V_{cb}}| \\ |\mathrm{V_{td}}| & |\mathrm{V_{ts}}| & |\mathrm{V_{tb}}| \end{pmatrix} = \begin{pmatrix} 0.97373 & 0.2243 & 0.00382 \\ 0.221 & 0.975 & 0.0408 \\ 0.0086 & 0.0415 & 1.014 \end{pmatrix} \tag{2.10}$$

from which it is clear that the transition probability is highest for the quarks of the same generation, with small off-diagonal values.

2.1.4 The electroweak unification

The electromagnetic and the weak force are unified in the *electroweak interaction* (EW), a model proposed by Glashow, Weinberg and Salam [14–16]. The idea behind the unification is that the two forces have many similarities and could originate from the same fundamental interaction. In particular the photon and the Z boson mediate the same interaction, with the difference that the γ is massless and the Z is a massive particle. The EW interaction is based on the gauge group $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$, with the new conserved charge being the *weak hypercharge* $Y = 2(Q - T_3)$. The gauge bosons are a triplet $W^a_{\mu\nu}$, with two charged components ($W^{(1/2)}_\mu$) and a neutral one ($W^{(3)}_\mu$), and a singlet $B_{\mu\nu}$, electrically neutral. The Lagrangian density of the EW interaction is:

$$\mathcal{L}_{\mathrm{EW}} = i\bar{\psi}\gamma^\mu D_\mu \psi - \frac{1}{4} W^a_{\mu\nu} W^{a\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}. \tag{2.11}$$

The fermion fields ψ are left-handed doublets and right-handed singlets under $\mathrm{SU(2)_L}$, as previously introduced. The covariant derivative D_μ is defined as:

$$D_\mu = \partial_\mu + ig_W T^a W_\mu^a + ig_{W'} \frac{Y}{2} B_\mu \tag{2.12}$$

where the g_W and $g_{W'}$ are the coupling constants of the $\mathrm{SU(2)_L}$ and $\mathrm{U(1)_Y}$ gauge groups, respectively. The physical $\mathrm{W^\pm}$, Z and γ bosons derive from the linear combinations of the $W_{\mu\nu}^a$, $B_{\mu\nu}$ and A_μ fields:

$$W^\pm = \frac{1}{\sqrt{2}}\left(W^{(1)} \mp iW^{(2)}\right) \tag{2.13}$$

and

$$\begin{aligned} A_\mu &= +B_\mu \cos\theta_W + W_\mu^{(3)} \sin\theta_W \\ Z_\mu &= -B_\mu \sin\theta_W + W_\mu^{(3)} \cos\theta_W \end{aligned} \tag{2.14}$$

where θ_W is the weak mixing angle or Weinberg angle.

2.1.5 Electroweak symmetry breaking and the Higgs boson

In the EW Lagrangian the gauge bosons and the fermions are treated as massless particles; if they were massive, the local $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ gauge invariance would be violated. Experimentally, it is clearly established that the fermions and the $\mathrm{W^\pm}$ and Z boson are massive. The *electroweak symmetry breaking* (EWSB) is the theory that allows the SM particles to acquire mass. The theory is also known as the *Higgs mechanism* and it was predicted independently by Higgs [6] and by Englert and Brout [7] in 1964. In the model, a new complex scalar field Φ is introduced, which spontaneously breaks the $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ symmetry. The new field is the *Higgs field*:

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{2.15}$$

with the related potential:

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2 \tag{2.16}$$

where μ and λ are two new parameters. The requirement $\lambda > 0$ assures a finite minimum in the potential. For $\mu^2 > 0$ the potential has a parabolic shape with one minimum at 0. For $\mu^2 < 0$ the potential assumes the so-called *sombrero-hat* shape, with a set of degenerate minima at:

$$\Phi^\dagger \Phi = \frac{v^2}{2} = -\frac{\mu^2}{2\lambda}. \tag{2.17}$$

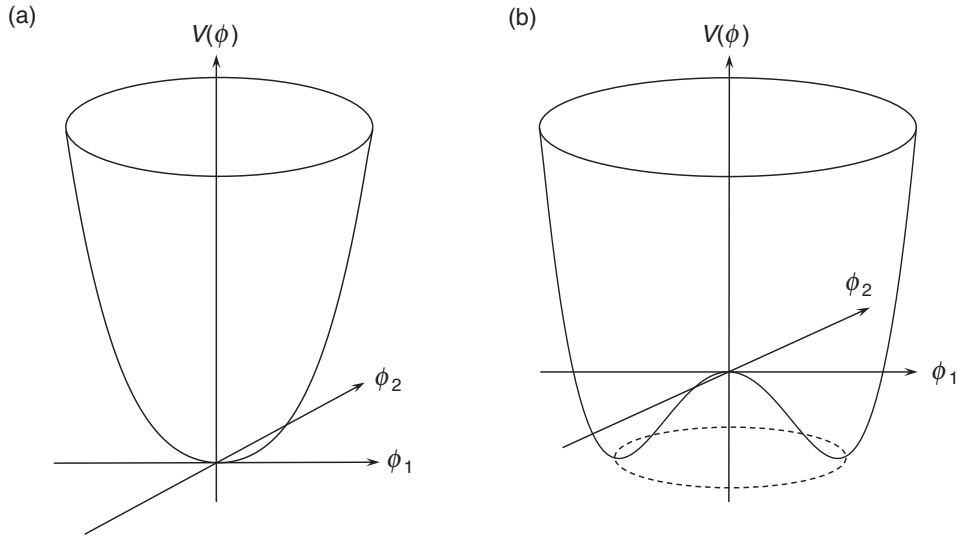

Figure 2.2: The potential $V(\Phi)$ for (a) $\mu^2 > 0$ and (b) $\mu^2 < 0$. Figure from [10].

The v is the *vacuum expectation value* (vev). Figure 2.2 shows the shape of the potential $V(\Phi)$ for a complex scalar field Φ as a function of its components ϕ_1 and ϕ_2. The two scenarios for $\mu^2 > 0$ and $\mu^2 < 0$ are represented. In the Higgs mechanism the values $\lambda > 0$ and $\mu^2 < 0$ are used. The choice of the vev leads to the *spontaneous symmetry breaking* of the $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ Lagrangian. After the symmetry breaking, the Higgs field can be expanded about the vacuum as:

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \tag{2.18}$$

where $h(x)$ is the new massive scalar particle introduced by the mechanism: the Higgs boson. Another consequence of the spontaneous symmetry breaking is the addition of new terms in the Lagrangian that provide the mass to the $\mathrm{W}^{\pm}$ and Z bosons, leaving the γ massless. The mass of the Higgs boson is:

$$m_{\mathrm{H}} = \sqrt{2\lambda} v \tag{2.19}$$

where the vacuum expectation value of the Higgs is $v \simeq 246$ GeV. The values of the masses of the W and Z bosons are given by:

$$\begin{aligned} m_{\mathrm{W}} &= \frac{1}{2} v g_W \\ m_{\mathrm{Z}} &= \frac{1}{2} v \sqrt{g_W^2 + g_{W'}^2} \end{aligned} \tag{2.20}$$

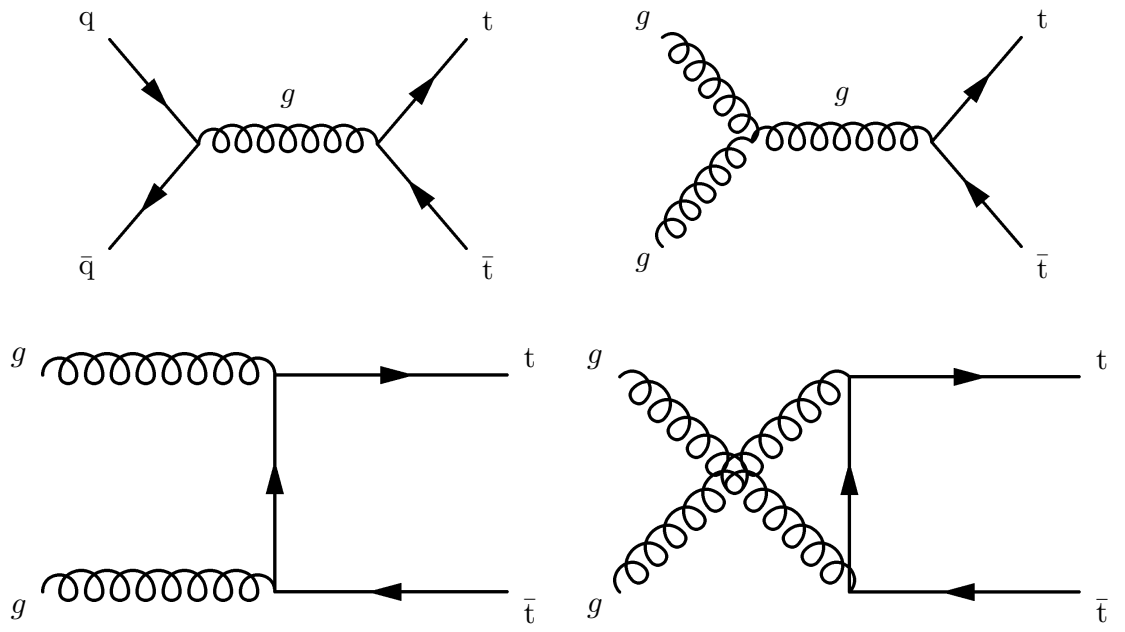

Figure 2.3: The Feynman diagrams of the $t\bar{t}$ production in pp collisions, via $q\bar{q}$ annihilation (upper left) and via gluon-gluon fusion in the s-channel (upper right), t-channel (lower left) and u-channel (lower right).

where g_W and $g_{W'}$ are the coupling constants of $SU(2)_L$ and $U(1)_Y$ introduced in the previous section. These masses can be expressed as a function of the Weinberg angle as:

$$\frac{m_W}{m_Z} = cos\theta_W. \tag{2.21}$$

Furthermore, it is possible to add mass terms to the fermions through the interaction with the Higgs field. The fermion masses are given by:

$$m_f = \frac{1}{\sqrt{2}} v y_f \tag{2.22}$$

with y_f the Yukawa coupling of the fermion f to the Higgs.

The Higgs boson has been the last missing piece of the SM for decades. Postulated in the 1960s, particle physics experiments have been searching for it as the final confirmation of the theory. Finally, on the 4^{th} of July 2012 the ATLAS [9] and CMS [8] experiments at CERN claimed the discovery of a new neutral boson compatible with the Higgs. Afterwards, precise measurements of the Higgs boson properties have been made and resulted to be all compatible with the SM predictions.

2.1.6 The top quark physics

The top quark is the most massive particle of the SM, with a mass of 172.69 ± 0.30 GeV [11]. It has been discovered in 1995 by the D0 and CDF Collaborations [17, 18] in $p\bar{p}$ collisions at the Tevatron. Due to its large mass, the t quark has a shorter lifetime compared to other quarks, of about $0.5 \cdot 10^{-24}$ s, which prevents it from forming bound states. Thus, the t decays immediately after production. Moreover, it has a large Yukawa coupling to the Higgs boson, giving it a important role in the SM and in new physics theories, as described in Chapter 3.

At hadron colliders, top quarks are produced mostly in top-antitop pairs ($t\bar{t}$) via strong interaction. The different leading-order production mechanisms of $t\bar{t}$ are shown in Fig. 2.3: the gluon-gluon fusion ($gg \to t\bar{t}$) and the quark-antiquark annihilation ($q\bar{q} \to t\bar{t}$). The gluon-gluon fusion production dominates at increasing collision energies $\sqrt{s}$ over the quark-antiquark annihilation because of the larger density of gluons inside the protons with respect to anti-quarks. In particular, at the LHC at the $\sqrt{s} = 13$ TeV about 90% of the $t\bar{t}$ pairs are produced via gluon-gluon fusion. On the other hand, at the Tevatron $p\bar{p}$ collider, $t\bar{t}$ pairs were produced via $q\bar{q}$ annihilation $\sim 85\%$ of the times at $\sqrt{s} = 1.96$ TeV. The total $t\bar{t}$ production cross section [19] is:

$$\sigma = 833.9^{+29.3}_{-36.6} \text{ pb}. \tag{2.23}$$

It has been calculated at NNLO in QCD with Top++2.0 [20] at $\sqrt{s} = 13$ TeV and assuming a top quark mass of 172.5 GeV. It is also possible to produce single top quarks in weak interactions, even though the probability is suppressed with respect to the pair production, because of the smaller coupling strength. The single top production mechanisms at LO are shown in Fig. 2.4: the s-channel, the t-channel and the production in association with a W boson.

Given the large value of $|V_{tb}|$ in the CKM matrix (Eq. 2.10), the top decays via weak interaction dominantly as $t \to Wb$. The subsequent decay of the W boson determines the final state of the t decay: leptonic decay $W \to l\nu$ (33%), and hadronic decay $W \to q\bar{q}'$ (67%). The two decay modes of the t quark are shown in Fig. 2.5. For the $t\bar{t}$ pairs, three different final states are possible:

- the lepton+jets final state, where one t decays leptonically and the other hadronically,
- the dilepton final state, where both t quarks decay leptonically,
- the all hadronic final state, where both t quarks decay hadronically.

The branching fractions of the different decays of $t\bar{t}$ are represented in Fig. 2.6. In the

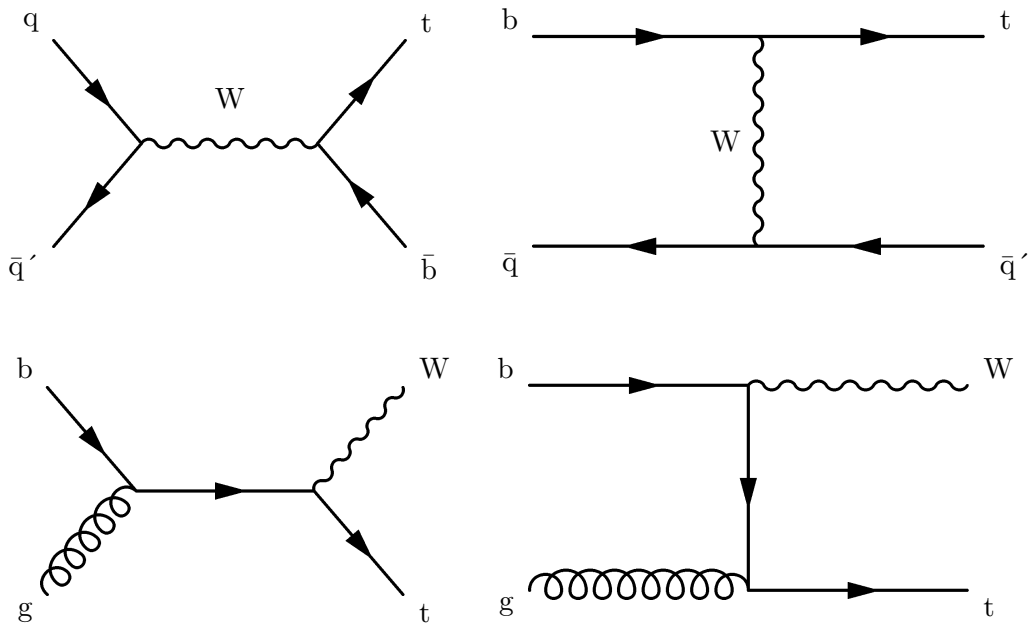

Figure 2.4: The Feynman diagrams of the single top production in pp collisions in the *s*-channel (upper left) and *t*-channel (upper right), and the t production in association with a W boson in the *s*-channel (lower left) and *t*-channel (lower right).

search presented in this thesis, the $t\bar{t}$ pairs decaying in the lepton+jets final states are analyzed, focusing on events with one e or one μ and jets.

2.2 Proton-proton collisions

High energetic proton-proton collisions allow to test the SM and to precisely measure its parameters and they are fundamental in the search for new physics. The world's most powerful hadron collider is the Large Hadron Collider (LHC) at CERN, where proton beams are accelerated and collide at unprecedented energies. To understand the experimental data of the LHC and make comparisons to theoretical predictions, the physics of proton-proton collisions has to be introduced.

2.2.1 The parton model

The proton is not an elementary particle, but it is made of constituents particles called *partons*. The three *valence quarks* are two up and one down quark (uud), they are the primary constituents and carry the electric charge and quantum numbers of the proton. They interact through exchange of gluons, which interact also among each other, forming a *sea* of gluons and quark-antiquark pairs, that are created and annihilated from vacuum

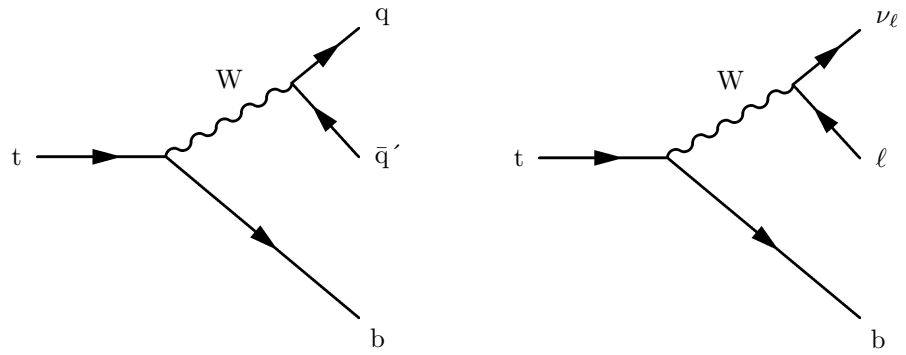

Figure 2.5: The Feynman diagrams of the hadronic (left) and leptonic (right) decay the top quark.

fluctuations. The fraction of the proton momentum carried by the parton is the *Bjorken variable* x. The parton distribution function (PDF) $f_{i/h}(x, Q^2)$ is the probability density function of a parton of type i in the hadron h, probed at the scale Q, with momentum fraction x. The PDFs are universal functions and can be extracted from experimental data at different energies. As an example, in Figure 2.7 the MSHT PDFs [22] at NNLO are shown, derived from a combination of LHC, HERA, Tevatron and fixed target data. Given a scale Q_0^2, it is possible to obtain the PDFs value at any scale $Q^2 > Q_0^2$ with the perturbative differential equation of Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) [23, 24].

2.2.2 The cross section and factorization

As a consequence of the composite nature of protons, in hadronic collisions the interactions occur between the partons inside the protons. A high energetic collision can be described by the *hard scattering*, i.e. the hard interaction of two partons, and the *underlying event* (UE), the particles that result from the break-up of the incoming protons (beam remnants), from the initial- and final-state radiation (ISR and FSR), and the soft interactions among the other partons. The cross section of a process can be calculated using the *factorization theorem* [25], where short- and long-distance contributions to the hard scattering are identified: they are the partonic cross section, which is calculated perturbatively, and the non-perturbative terms, e.g. hadronization, which are described by phenomenological models. Let us consider an interaction between two protons h_1 and h_2 with four-momenta P_1 and P_2, respectively. The center-of-mass energy of the collision $\sqrt{s}$ is defined as:

$$\sqrt{s} = (P_1 + P_2). \tag{2.24}$$

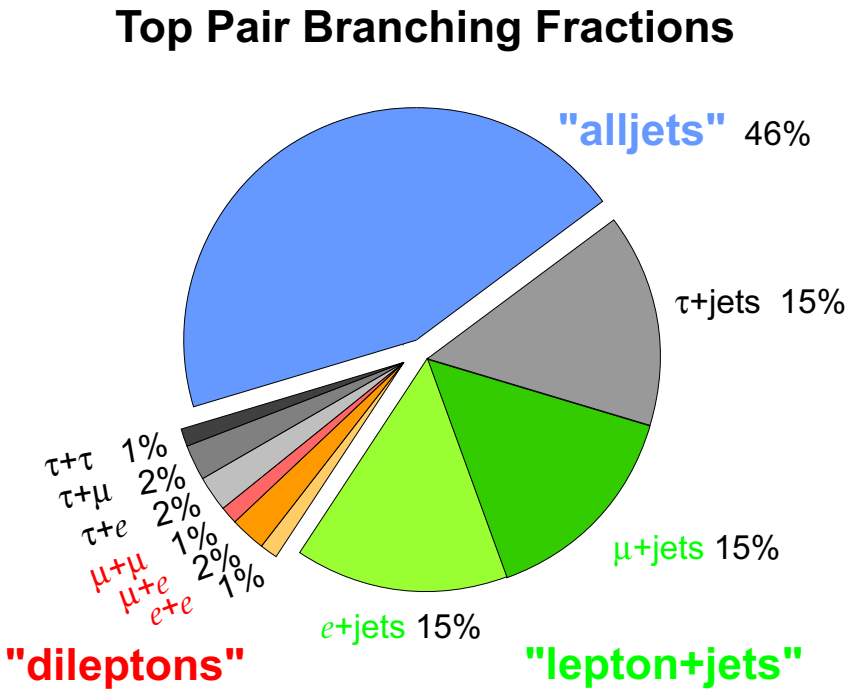

Figure 2.6: The branching fractions of the $t\bar{t}$ decay channels. Figure from [21].

The center-of-mass energy $\sqrt{\hat{s}}$ of the partonic interaction can be expressed as:

$$\sqrt{\hat{s}} = (p_1 + p_2) = \sqrt{x_1 x_2 s} = Q \tag{2.25}$$

where p_1 and p_2 are the partons' four-momenta, x_1 and x_2 their Bjorken-x variables, and Q is the energy scale of the interaction. Given that the values of x_1 and x_2 are not necessary equal, the interaction can be Lorentz-boosted along the z-direction. However, in the collinear approximation the partons do not carry transverse momenta, thus for momentum conservation the sum of the transverse momenta of the final state particles must vanish.

The cross section σ of a process $h_1 h_2 \to X$ can be factorized into the partonic cross section $\hat{\sigma}$ and the PDFs of the interacting partons:

$$\sigma_{h_1 h_2 \to X} = \sum_{i,j} \int\int \mathrm{d}x_1 \mathrm{d}x_2 f_{i/h_1}(x_1, \mu_F^2) f_{j/h_2}(x_2, \mu_F^2) \hat{\sigma}_{ij \to X}(\hat{s}, \mu_F^2, \mu_R^2). \tag{2.26}$$

The sum runs over the possible types i, j of the initial partons, f_{i/h_1} and f_{j/h_2} are the PDFs of the partons i and j inside the hadrons h_1 and h_2, respectively, and μ_F^2 is the *factorization scale*. The partonic cross section can be calculated in perturbation theory and depends on the factorization and renormalization (μ_R^2) scales. The values of μ_F^2 and μ_R^2 are arbitrary and usually take the value of the interaction scale Q. The Matrix Element

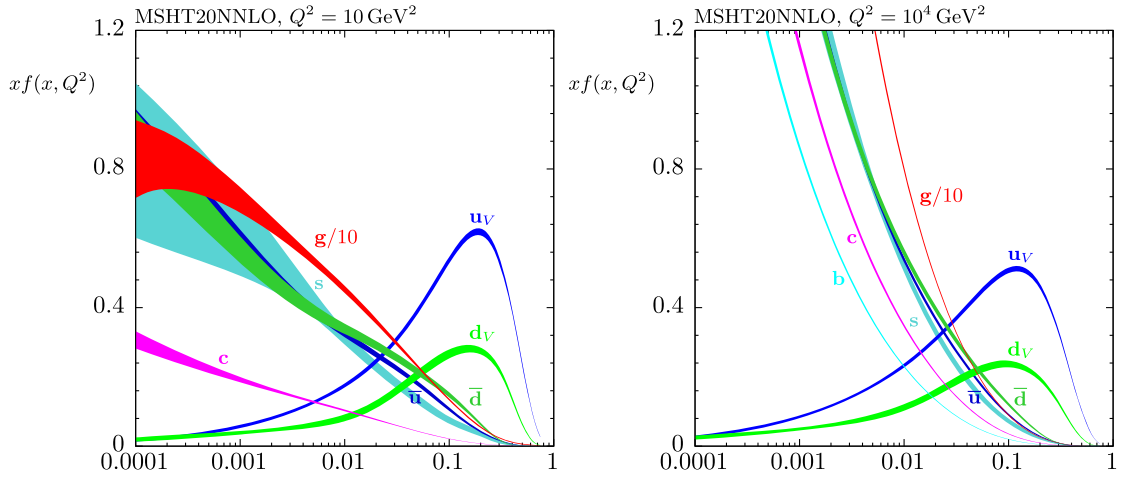

Figure 2.7: The NNLO MSHT PDFs at the scales 10 GeV^2 (left) and 10^4 GeV^2 (right). Figure from [22].

(ME) is the mathematical formulation of the partonic hard scattering.

2.2.3 Hadronization

The quarks and gluons originating in pp collisions do not propagate freely due to colour confinement, but are observed as colorless hadrons, a phenomenon called hadronization. It is not possible to describe this process in perturbative QCD, but phenomenological models are required. One of the models that is mostly used in event simulation is the "string" or "Lund" model [26]. An example of hadronization is shown in Fig. 2.8. Given a quark and an antiquark that move apart at high velocity, a color field is established between them. As the particles move apart, the color potential increases with their distance r as $V(r) = kr$, with $k \simeq 1$ GeV/fm. When the distance, or energy, is large enough, the color flux breaks and creates another quark-antiquark pair. The new $q\bar{q}$ pair connects to the previous $q\bar{q}$ with color fluxes, and the process continues. When the energy is sufficiently small, quarks and antiquarks combine into colorless hadrons. The particles in the final state follow the direction of the initiating parton and can be grouped into one single object called *jet*.

2.3 Event generators

Event simulations are indispensable tools to compare experimental data to theory predictions, and can be used to design future detectors and study new experimental techniques. A general overview on event generators can be found in Ref. [27].

When considering proton-proton collision physics, it is necessary to link the particles

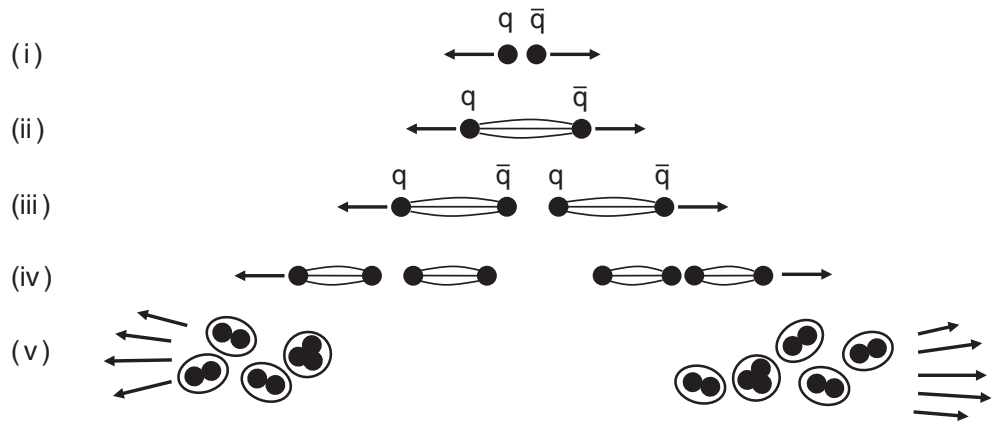

Figure 2.8: The hadronization process in the Lund model. Figure from [10].

detected by the experiments to the hard scattering processes that generated them. To the complexity of the final state, the theoretical difficulties have to be added. In fact, there are many steps in the hadronic collision physics that can not be calculated in perturbation theory, but rely on phenomenological models. Moreover, the perturbative QCD calculations themselves can be very complex. Event generators are based on the Monte Carlo (MC) method and simulate events in different steps, starting from the hard scattering, then adding the hadronization and the UE simulation. The event generators most commonly used in LHC experiments are Pythia8 [28], Powheg [29,30], Herwig [31] and MadGraph5_aMC@NLO [32].

The event simulation starts with the ME convoluted with the PDFs. All general-purpose generators provide LO matrix elements for the $2 \to 1$, $2 \to 2$ and $2 \to 3$ processes. For higher-order simulation, dedicated generators are used (e.g. MadGraph). The second step is the Parton Shower (PS), which describes the emission of quarks and gluons from the initial- and final-state partons. The PS covers a wide energy range, from high scales, comparable to the hard process, down to low scales of $\mathcal{O}(1\text{ GeV})$, where partons hadronize. The PS simulation is typically done with Pythia8 or Herwig; for the generators that do not include PS, the final-state particles from the hard process simulated with ME generators have to be matched to the PS. The algorithms used for the matching are MLM [33] and FxFx [34].

Finally, the UE is simulated, that includes the beam remnants, the soft interactions among partons and the ISR and FSR. Furthermore, the additional pp interactions that happen concurrently to the interaction of interest have to be taken into account. Such interactions, called pileup (PU) are simulated at this stage as well, and the generator usually used is Pythia8. The hadronization and the UE rely on phenomenological models

and depend on many unknown parameters. The choice of the parameters can vary from one generator to another. A defined set of values of the parameters of a given model is known as *tune*. The tunes can be optimized by comparing the simulation to data in variables that are sensitive to such parameters.

The particles originating from the proton-proton collision travel through and interact with the detector. The simulation of the interaction of the particles with the detector material is implemented with GEANT4 [35], the most commonly used tool in physics experiments for the purpose. In particular, the program includes the full simulation of all the detector components of the CMS experiment.

Chapter 3

Physics Beyond the Standard Model

The SM is the most successful theory describing the elementary particles and their interactions, as was presented in the previous Chapter. Nevertheless, there are some open questions that the SM can not answer. They are, on one hand, limitations of the theory, as there are properties within the SM that do not have a clear explanation. On the other hand, there are experimental observations that are in contradiction with the SM predictions. These open questions hint to the existence of new theories Beyond the Standard Model (BSM), that extend the SM and try to solve some of its problems. The top quark plays an important role in many BSM models, being the heaviest elementary particle and thus a perfect portal to new massive particles.

In this Chapter some of the open questions of the SM will be presented in Sec. 3.1 and the theories of physics BSM will be described in Sec. 3.2. The models that predict new heavy resonances that couple to top quarks, which include spin-1 particles, as Kaluza-Klein gluons g_{KK} and heavy Z' bosons, and spin-0 particles, as heavy Higgs bosons, are presented in Sec. 3.3. Finally, the results of previous searches for such particles will be summarized.

3.1 The open questions of the Standard Model

Despite the great success of the SM, there are still some key questions that can not be answered and experimental observations that are not described by this theory. Some of the open questions of the SM are summarized in the following.

Gravity

Gravity is one of the four fundamental forces of particle physics, but it is the only one that is not included in the SM. Nevertheless, gravity is a very weak force with respect to the others, and the SM is considered valid up until the Planck scale $\Lambda_{Pl} \sim 10^{19}$ GeV, where gravity is non-negligible anymore. On the other hand gravity is well described by the theory of General Relativity. The impossibility to unify the gravitational force in the SM is one of the shortcomings of the theory.

The hierarchy problem and fine-tuning

A problem directly related to gravity is the observation of a large difference between the weak interaction and the gravitational force, which is of the order of 10^{24}. This large discrepancy is known as the *hierarchy problem*, which seems unnatural and is not explained by the SM. A consequence of this large difference in energy scales is the *fine-tuning* of the Higgs mass. The measured Higgs mass is given by:

$$m_{\mathrm{H}}^2 \approx m_{bare}^2 + \Delta m_{\mathrm{H}}^2 \tag{3.1}$$

where m_{bare} is the bare mass of the Higgs and Δm_{H}^2 are the quantum corrections, given by the virtual contributions of all the particles that couple to the Higgs boson. The quantum corrections depend on the energy scale Λ and the largest contribution is the term $\frac{3}{8\pi^2} y_t^2 \Lambda^2$, where y_t is the Yukawa coupling of the top quark, the heaviest particle of the SM. If no physics BSM is present up to the Planck scale, then the quantum loop corrections to the Higgs mass should be ~ 30 orders of magnitude larger than the measured Higgs mass, unless there is a fine tuning of the bare mass parameter that precisely cancels out these corrections.

Fermion generations

An interesting feature of the SM is the presence of three generations of fermions and the similarities among quarks and leptons. They are ordered by mass and have electric charge which is a multiple of $e/3$. There is no explanation in the SM for the number of fermion generations and no prediction for their masses.

Flavour anomalies

Lepton flavour universality (LFU) assumes that the gauge couplings to the three generations of leptons are the same. Nevertheless, there are experimental observations, e.g. from the BaBar [36] and Belle [37] experiments, that hint to violation of LFU in the b sector, known

as flavour anomalies. The recent results from the LHCb collaboration [38], on the other hand, are in agreement with the SM predictions, reducing the tension.

Matter-antimatter asymmetry

In our Universe today a matter-antimatter asymmetry is observed, with almost complete lack of antimatter. The imbalance between matter and antimatter can not be explained by the SM nor by Cosmological models. The only mechanism that can break the matter-antimatter asymmetry is the CP-violation. In the SM, CP is violated in the CMK matrix, but it is not sufficiently large to explain the observed excess of matter. This hints to the possibility of CP-violation in other sectors of the SM.

Neutrino masses

In the SM neutrinos are predicted to be massless particles, but neutrino oscillations, predicted in 1957 [5] and observed by the SuperKamiokande [3] and SNO [4] experiments, are possible only if neutrinos have mass. From the measurements, we know that at least two of the three neutrinos have mass and they differ by $\mathcal{O}(1)$ eV.

Muon anomalous magnetic moment

The measurements of the muon anomalous magnetic moment a_μ show deviations from the SM prediction at more than 5 standard deviations [39], hinting to the presence of new physics not described by the SM.

Dark matter and dark energy

From the measurements of the mass-energy content of the Universe we know that only the $\sim 5\%$ is in the form of visible matter. Another $\sim 27\%$ is made of *dark matter*, a type of matter that does not interact with the electromagnetic and strong force, but is massive. The evidence for its existence comes from the experimental measurements of the rotation curves of galaxies [40], gravitational lensing [41] and the cosmic microwave background [42, 43]. Moreover, the accelerating expansion rate of the Universe indicates the presence of *dark energy*, that makes up $\sim 68\%$ of the Universe. So far, no viable candidate for dark matter is included in the SM, nor there is an explanation for dark energy.

3.2 Theories Beyond the Standard Model

There are many theories of physics BSM that are designed to answer one or more of the open questions presented in the previous section. These theories can extend the SM with new

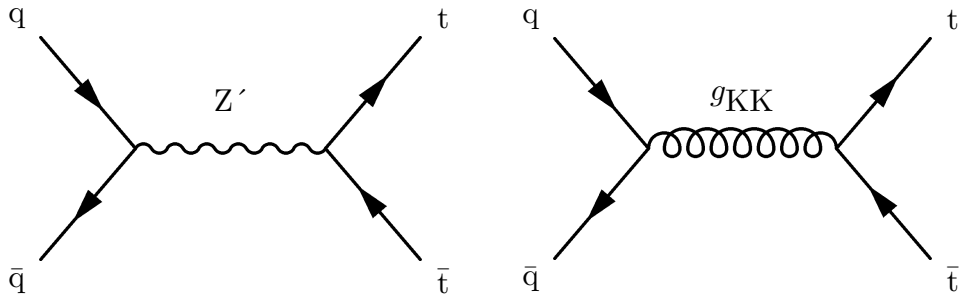

Figure 3.1: The Feynman diagrams of the production of a Z′ boson (left) and a g_{KK} gluon (right) decaying to $t\bar{t}$.

particles, new forces or extra dimensions that can manifest at the TeV scale, and usually target only part of the limitations of the SM. Other theories are more complex, like Grand Unified Theories (GUT) [44] or Supersymmetry (SUSY) [45]. Grand Unified Theories aim at unifying all the three fundamental forces of the SM, in the same way the electromagnetic and weak interactions have been unified in the EW interaction. At a high energy scale ($\Lambda_{GUT} \sim 10^{16}$ GeV) the gauge groups of the SM are embedded in a higher symmetry group, that is broken below this scale, giving the fundamental interactions as they are described by the SM. GUTs are also known as *theories of everything*. Supersymmetry attempts to answer almost all of the open questions of the theory of particle physics. In SUSY an extra symmetry is included between fermions and bosons, called *supersymmetry*, which turns bosonic states into fermionic states and vice-versa. For each particle of the SM there is a super-partner (*s-particle*) that differs only in the spin by half a unit. The minimal extension of the SM that includes SUSY in the Minimal Supersymmetryc Standard Model (MSSM) [46]. SUSY could solve the hierarchy problem and provide a candidate for the dark matter, the *neutralino*, but at date no evidence of this theory has been found.

Theories BSM that predict the existence of new heavy particles that decay to top quark pairs are presented in detail in the following.

3.3 New particles decaying to top quark pairs

3.3.1 Spin-1 particles

The first category of models considered includes heavy spin-1 resonances: Kaluza-Klein gluons g_{KK} and Z′ bosons. The Feynman diagram of the production of such new particles and the decay to a top quark pair is shown in Fig. 3.1.

Kaluza-Klein gluons

A solution to the hierarchy problem is presented by the Randall-Sundrum I framework (RS1) [47] of a warped extra dimension. Theories of extra dimensions postulate that the observable Universe resides on a 4D hypersurface (*brane*) embedded in the *bulk*, the higher dimensional space [48]. The warping of the extra dimension causes a large ratio of energy scales, so that the energy scale at one end of the extra dimension is much larger than the one at the other end. In the RS1 theory, there is a warped 5D bulk with two branes: the Planck brane, where gravity is strong, and the TeV brane, where the SM particles live, and gravity penetrates into the extra dimension. It is possible to write the 5D metric as:

$$ds^2 = e^{-2\beta(y)}\eta_{\mu\nu}dx^\mu dx^\nu - dy^2 \tag{3.2}$$

where $e^{-2\beta(y)}$ is the warp factor, the fifth dimension has radius r and coordinate y in $[0, \pi r]$. The Planck and TeV branes live at $y = 0$ and $y = \pi$, respectively.

In the RS1 theory, however, there are contributions to flavour changing neutral current processes (FCNC) and to SM electroweak precision test observables (EWPT) that are too large compared to experimental measurements. A proposed solution [49] postulates that not only gravity, but also the SM fields can propagate in the extra dimension. If the first and second generation fermions are placed near the Planck brane, while the Higgs and the third generation fermions near the TeV brane, then the contributions to FCNC and to EWPT are suppressed. In this way the hierarchies in the SM Yukawa couplings can also be explained.

In this extended RS1 model, Kaluza-Klein (KK) parters of SM particles are predicted. The KK gauge bosons are localized near the TeV brane and thus are expected to be massive and to couple mostly to third generation fermions, given the higher Yukawa couplings. The relevant couplings of the KK gauge states with respect to the SM couplings are:

$$\begin{aligned} \frac{g_{\mathrm{RS}}^{\mathrm{q\bar{q}},l\bar{l}G^1}}{g_{\mathrm{SM}}} &\simeq \xi^{-1} \approx \frac{1}{5}, \quad \frac{g_{\mathrm{RS}}^{Q^3\bar{Q}^3,l\bar{l}G^1}}{g_{\mathrm{SM}}} \approx 1, \\ \frac{g_{\mathrm{RS}}^{\mathrm{t}_R\bar{\mathrm{t}}_R G^1}}{g_{\mathrm{SM}}} &\simeq \xi \approx 5, \quad \frac{g_{\mathrm{RS}}^{GGG^1}}{g_{\mathrm{SM}}} \approx 0 \end{aligned} \tag{3.3}$$

where l = leptons, q=u,d,c,s,b$_R$, $Q^3 = (\mathrm{t,b})_L$, G and G^1 are the SM and first KK states of gauge fields, and g_{SM} and g_{RS} are the SM and RS1 gauge couplings. The factor ξ is equal to $\sqrt{\log(M_{Pl}/\mathrm{TeV})}$, where $M_{Pl} = 2 \times 10^{18}$ GeV is the Planck mass. Among the predicted KK gauge particles, the Kaluza-Klein partner of the gluon g_{KK} has the highest expected production rate at the LHC. The main production mechanism is via u$\bar{\mathrm{u}}$ and d$\bar{\mathrm{d}}$ annihilation

and its production cross section is shown in Fig. 3.2 (top) for 14 TeV pp interactions. The branching fraction is shown in Fig. 3.2 (center): g_{KK} decays to $\mathrm{t\bar{t}}$ pairs 94% of the times. The g_{KK} is a broad resonance: for masses above 1 TeV the width is about $m_{g_{\mathrm{KK}}}/6$, as depicted in Fig. 3.2 (bottom).

Z' bosons

Many theories of new physics predict the existence of a heavy, neutral boson Z' that is associated with a new gauge group. There are two classes of models predicting Z' resonances: in the first class the new boson couples weakly, as in the Sequential Standard Model (SSM) [50], where the Z' couples to the SM particles like the SM Z boson. In the second class of models, the Z' couples strongly and preferentially to top quarks, as in topcolor models [51]. Such theories could explain the large mass of the top quark through the formation of a top condensate and the mechanism of electroweak symmetry breaking.

In this dissertation, models where the Z' couples preferentially to top quarks are considered. Such models [52] predict that one or more of the SM SU(N) gauge groups can be extended into SU(N) × SU(N). Generally, first and second generation fermions transform under one of the SU(N), and the third generation fermions under the other. When SU(N) × SU(N) spontaneously breaks, massive gauge bosons arise that couple differently to different generation fermions. Based on the choice of the couplings, different variants of the model are defined.

The model used in the search presented in this thesis is the leptophobic topcolor model, denoted as Model IV [53], where the Z' couples only to the first and third generations of quarks and has no significant couplings to leptons. In this framework, the $\mathrm{SU(3)}_C$ gauge group is embedded in $\mathrm{SU(3)}_1 \times \mathrm{SU(3)}_2$, and the breaking $\mathrm{SU(3)}_1 \times \mathrm{SU(3)}_2 \rightarrow \mathrm{SU(3)}_C$ produces $\mathrm{t\bar{t}}$ and $\mathrm{b\bar{b}}$ condensates, resulting in top and bottom quarks with the same mass of around 600 GeV. To remove this degeneracy, a new component has to be added to the model, a *tilting* to increase the formation of top condensates over bottom condensates. A simple tilting mechanism is given by the embedding of $\mathrm{U(1)}_Y$ into $\mathrm{U(1)}_1 \times \mathrm{U(1)}_2$. This gives rise to a Z' boson from $\mathrm{U(1)}_2$.

The Lagrangian for Model IV is:

$$\mathcal{L}_{IV} = \left(\frac{1}{2} g_1 \cot\theta_H\right) \mathrm{Z}'^{\mu} \left(\bar{\mathrm{t}}_L \gamma_\mu \mathrm{t}_L + \bar{\mathrm{b}}_L \gamma_\mu \mathrm{b}_L + f_1 \bar{\mathrm{t}}_R \gamma_\mu \mathrm{t}_R + f_2 \bar{\mathrm{b}}_R \gamma_\mu \mathrm{b}_R - \bar{\mathrm{u}}_L \gamma_\mu \mathrm{u}_L - \bar{\mathrm{d}}_L \gamma_\mu \mathrm{d}_L - f_1 \bar{\mathrm{u}}_R \gamma_\mu \mathrm{u}_R - f_2 \bar{\mathrm{d}}_R \gamma_\mu \mathrm{d}_R\right) \quad (3.4)$$

where g_1 is the SM coupling constant of $\mathrm{U(1)}_Y$, $\cot\theta_H$ is the ratio of the two $\mathrm{U(1)}_i$ coupling constants and f_1 and f_2 are the relative strengths of the couplings of right-handed up- and

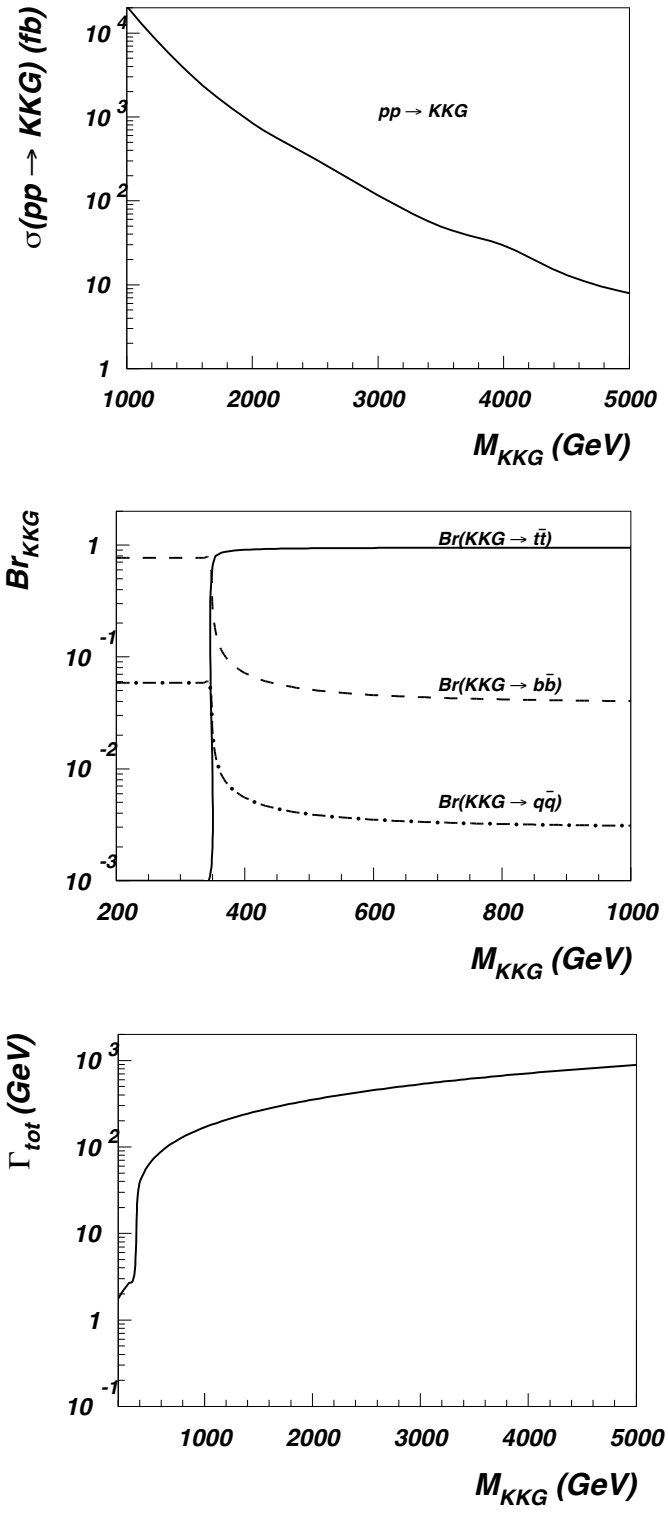

Figure 3.2: The production cross section of g_{KK} for 14 TeV pp interactions as a function of its mass (upper). The branching fractions (middle) and the total decay width (lower) of the g_{KK} as a function of its mass. Figures from [49].

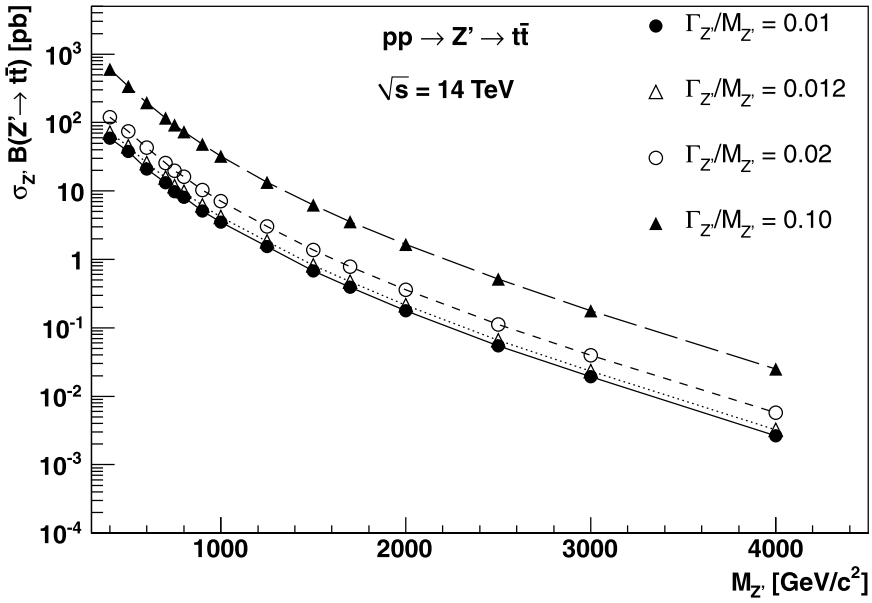

Figure 3.3: The production cross section times branching fraction of Z′ decaying to $t\bar{t}$ for 14 TeV pp interactions. Different relative widths are shown with different markers. Figure from [54].

down-type quarks with respect to those of the left-handed quarks. The following conditions are used: $f_1 > 0$ to be $t\bar{t}$ attractive and/or $f_2 < 0$ to be $b\bar{b}$ repulsive, $\cot\theta_H >> 1$ to avoid the fine-tuning.

The LO cross section is then controlled by the three parameters $\cot\theta_H$, f_1 and f_2. In the leptophobic, top-phyllic scheme the values are set to $f_1 = 1$ and $f_2 = 0$ and $\cot\theta_H$, proportional to the total decay width, is the only free parameter. The production cross section of $Z' \to t\bar{t}$ is shown in Fig. 3.3 for 14 TeV pp interactions. Three relative widths are considered in this thesis: 1%, 10% and 30%.

For both Z′ and g_{KK} resonances, the interference with the SM $t\bar{t}$ production is negligible: at the LHC $t\bar{t}$ pairs are produced mostly via gluon-gluon fusion, while both Z′ and g_{KK} resonances are produced via $q\bar{q}$ annihilation.

3.3.2 Spin-0 particles

New spin-0 resonances are predicted in extensions of the SM that include additional Higgs bosons, like the Two-Higgs-Doublet Models (2HDM) [55] or SUSY. In these theories a spectrum of Higgs bosons is expected: two neutral scalars h and H, a neutral pseudoscalar A and two charged scalars $H^{\pm}$. In particular, in the search presented in this thesis the focus is on the type-II 2HDM, which can be considered as a generalization of the MSSM, because

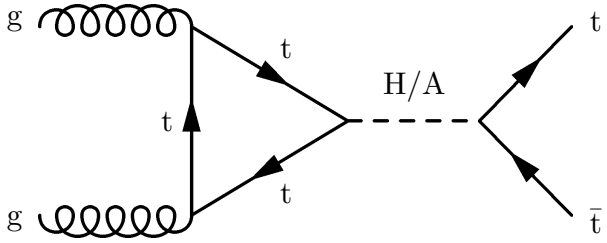

Figure 3.4: The Feynman diagram of a heavy Higgs boson H or A decaying to $t\bar{t}$.

in this model the decays to top quark pairs are enhanced for a large part of the parameter space. The Feynman diagram of a neutral heavy Higgs boson H or A that decays to $t\bar{t}$ is presented in Fig. 3.4. An interesting feature is that such signals interfere with the SM $t\bar{t}$ production, resulting in the characteristic *peak-dip* structure in the $t\bar{t}$ mass spectrum.

Type-II 2HDM

In the 2HDMs [55] a second Higgs doublet is introduced in the SM. Given the two Higgs doublets denoted as Φ_1 and Φ_2, it is possible to write the scalar potential as:

$$V = m_{11}^2\Phi_1^\dagger\Phi_1 + m_{22}^2\Phi_2^\dagger\Phi_2 - m_{12}^2(\Phi_1^\dagger\Phi_2 + \Phi_2^\dagger\Phi_1) + \frac{\lambda_1}{2}(\Phi_1^\dagger\Phi_1)^2 + \frac{\lambda_2}{2}(\Phi_2^\dagger\Phi_2)^2 + \lambda_3\Phi_1^\dagger\Phi_1\Phi_2^\dagger\Phi_2 + \lambda_4\Phi_1^\dagger\Phi_2\Phi_2^\dagger\Phi_1 + \frac{\lambda_5}{2}[(\Phi_1^\dagger\Phi_2)^2 + (\Phi_2^\dagger\Phi_1)^2] \quad (3.5)$$

where m_{ij} and λ_k are free parameters of the model. To assure that the CP symmetry is conserved, all the parameters are assumed to be real. The minimization of the potential V gives the basis:

$$\langle\Phi_1\rangle_0 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v_1\end{pmatrix}, \quad \langle\Phi_2\rangle_0 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v_2\end{pmatrix} \quad (3.6)$$

where v_1 and v_2 are the vevs of Φ_1 and Φ_2, respectively. The resulting scalar fields are:

$$\Phi_a = \begin{pmatrix}\phi_a^+\\(v_a + \rho_a + i\eta_a)/\sqrt{2}\end{pmatrix}, \quad a = 1, 2. \quad (3.7)$$

Of the eight fields, three give mass to the $W^\pm$ and Z bosons, and the remaining five are the scalar Higgs fields: two neutral scalars (CP-even) h and H, one neutral pseudoscalar (CP-odd) A and two charged scalars (CP-even) $H^\pm$. Since no new scalar lighter that the SM Higgs has been discovered, $m_h < m_H$ is assumed and h is identified with the SM Higgs. This assumption is know as the *alignment limit*.

There are different types of 2HDMs depending on the way the Higgs bosons couple to

the fermions. They are summarized in Table 3.1.

Model type	u_i	d_i	l_i
Type-I	Φ_2	Φ_2	Φ_2
Type-II	Φ_2	Φ_1	Φ_1
Lepton-specific	Φ_2	Φ_2	Φ_1
Flipped	Φ_2	Φ_1	Φ_2

Table 3.1: The types of 2HDMs depending on the coupling of the Higgs to the fermions, where u_i are up-type quarks, d_i are down-type quarks and l_i are charged leptons.

In this dissertation, the focus is on type-II 2HDM, of which the MSSM is a subset. In type-II 2HDM the CP is conserved and FCNC are absent. To avoid FCNC at tree-level, each of the SM fermions couples only to one of the doublets Φ_1 and Φ_2. With this assumption, there is a Z_2 symmetry that is softly broken [56]. Under the Z_2 symmetry, the doublets transform as $(\Phi_1, \Phi_2) \to (-\Phi_1, \Phi_2)$. Considering no CP violation in the vevs, the values v_1 and v_2 are real and non-negative. The parameters of the model are:

- the masses of the Higgs bosons m_{h}, m_{H}, m_{A}, $m_{\mathrm{H}^\pm}$
- the vevs v_1 and v_2 with the relation $v_1^2 + v_2^2 = v^2 = (246\ \mathrm{GeV})^2$, or $v_2/v_1 = \tan\beta$
- the mixing angle α between h and H.

The couplings of the neutral Higgs bosons to the other SM particles are given in Table 3.2 for type-II 2HDM. The h and H couple to bosons and fermions, while A couples only to fermions. In the alignment limit the couplings of h match the SM couplings for $\sin(\beta - \alpha) \to 1$.

	VV	u_i	d_i,l_i
g_{h}	$\sin(\beta-\alpha)$	$\cos\alpha/\sin\beta$	$-\sin\alpha/\cos\beta$
g_{H}	$\cos(\beta-\alpha)$	$\sin\alpha/\sin\beta$	$\cos\alpha/\cos\beta$
g_{A}	0	$\cot\beta$	$\tan\beta$

Table 3.2: The couplings at tree-level of the neutral Higgs bosons to vector bosons V, up-type quarks u_i, down-type quarks d_i and charged leptons l_i. The couplings are divided by the corresponding coupling of the SM Higgs boson.

The Yukawa Lagrangian is:

$$\mathcal{L}_{Yukawa} = -\sum_{f=\mathrm{u,d},l} \frac{m_f}{v}\left(g^f_{\mathrm{h}}\bar{f}f\mathrm{h} + g^f_{\mathrm{H}}\bar{f}f\mathrm{H} - ig^f_{\mathrm{A}}\bar{f}\gamma_5 f\mathrm{A}\right) - \left\{\frac{\sqrt{2}\mathrm{V_{ud}}}{v}\bar{\mathrm{u}}(m_{\mathrm{u}}g^{\mathrm{u}}_{\mathrm{A}}P_L + m_{\mathrm{d}}g^{\mathrm{d}}_{\mathrm{A}}P_R)\mathrm{dH}^+ + \frac{\sqrt{2}m_l g^l_{\mathrm{A}}}{v}\bar{\nu}_L l_R \mathrm{H}^+ + H.c.\right\} \quad (3.8)$$

where g_x are the parameters given in Table 3.2 and P_L and P_R are the projection operators for left- and right-handed fermions, respectively. In particular, the terms in the Yukawa Lagrangian for H and A coupling to top quarks are:

$$\mathcal{L}_{Yukawa,\mathrm{H}} = -\frac{m_{\mathrm{t}}}{v}g^{\mathrm{t}}_{\mathrm{H}}\bar{\mathrm{t}}\mathrm{t}\mathrm{H}, \quad \text{and} \quad L_{Yukawa,\mathrm{A}} = \frac{m_{\mathrm{A}}}{v}ig^{\mathrm{t}}_{\mathrm{A}}\bar{\mathrm{t}}\gamma_5\mathrm{t}\mathrm{A}. \quad (3.9)$$

The A and H bosons can be produced at the LHC via gluon-gluon fusion (ggF) with top quarks in the loop (see Fig. 3.4). The SM Higgs-like vector-boson fusion (VBF) mode is not possible, as in the alignment limit the couplings to vector bosons are suppressed. This means that the interference with SM $gg \to \mathrm{t\bar{t}}$ has to be taken into account. The interference can be constructive or destructive and it manifests as a peak-dip structure in the mass spectrum of the new particles. The exact interference pattern, meaning a more enhanced peak, a more enhanced dip, or a peak-dip, depends on the specific parameters of the signal, e.g. the mass and relative widths of the particle. The cross section for $gg \to \mathrm{t\bar{t}}$ is presented in Figure 3.5 considering the additional heavy H or A bosons.

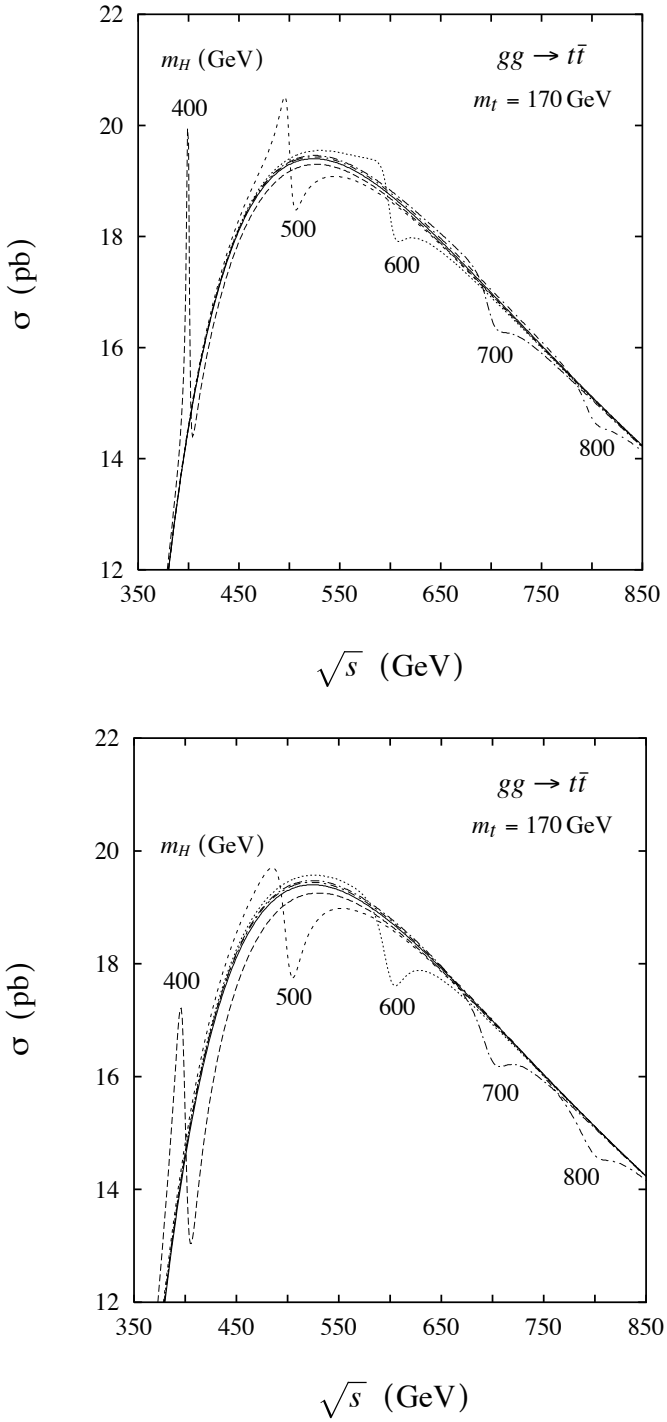

Figure 3.5: The cross section for $gg \to t\bar{t}$ as a function of the $t\bar{t}$ invariant mass. The effects of the inclusion of a heavy scalar (upper) or pseudoscalar (lower) Higgs boson are shown for different masses of the new particle. Figures from [57].

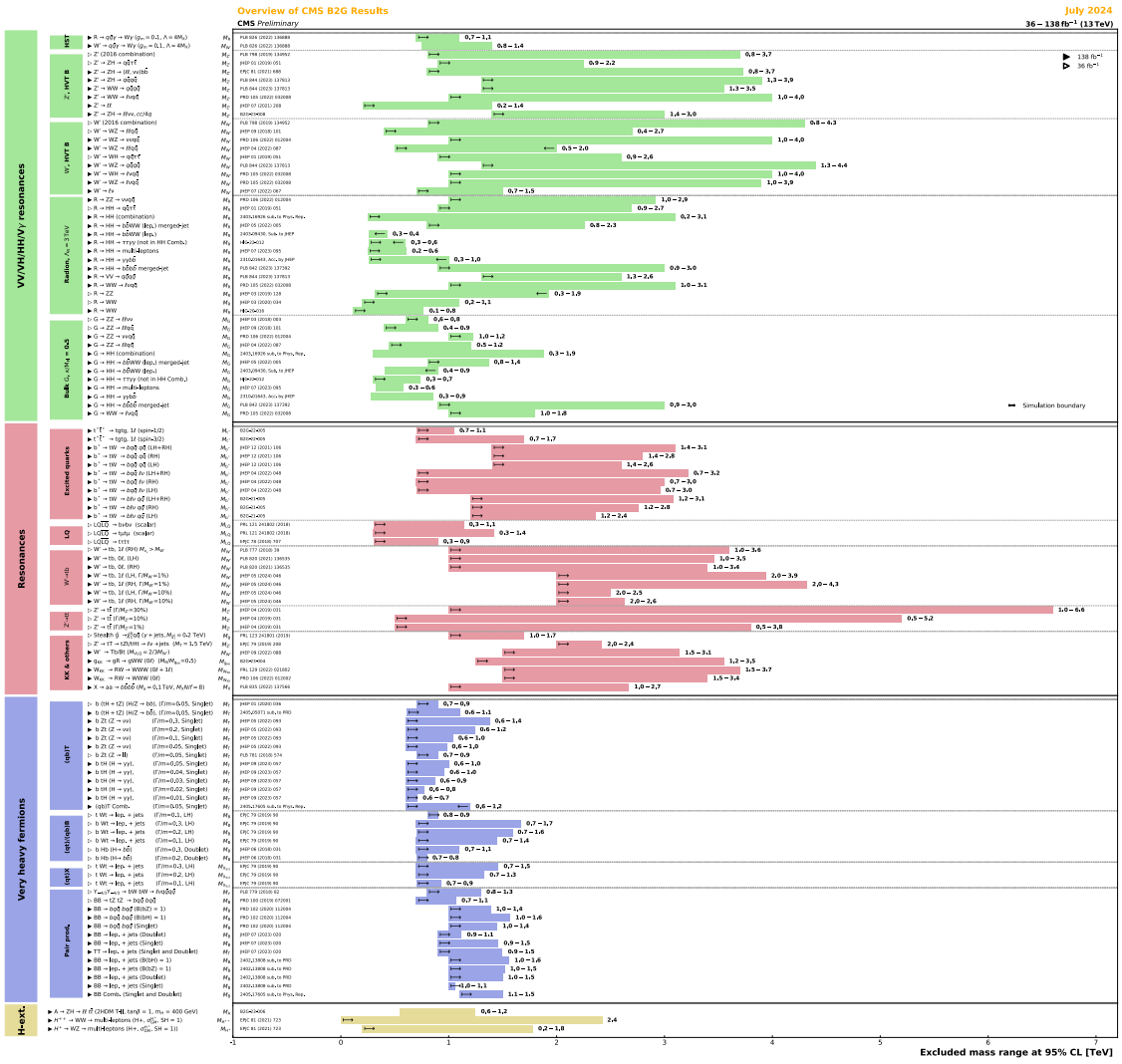

Figure 3.6: Summary of CMS results on searches for new particles coupling to third generation quarks or SM bosons. Figure from [58].

3.3.3 LHC results

As presented previously, many theories predict the existence of new particles that couple to top quarks. At the LHC, a plethora of searches for such particles is carried out using data recorded with the ATLAS and CMS experiments. A summary of searches for heavy resonances coupling to third generation quarks or SM bosons preformed by the CMS Collaboration is presented in Fig. 3.6. So far, no new particles have been discovered. In the following, the previous searches for spin-1 and spin-0 particles decaying to top quark pairs will be briefly presented, and the latest results from the CMS Collaboration will be discussed.

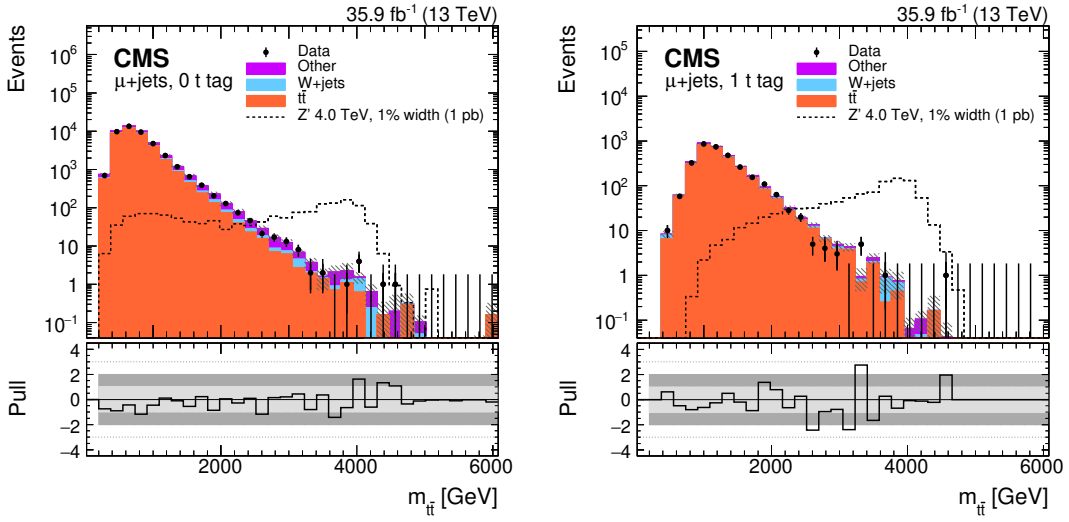

Figure 3.7: The $m_{t\bar{t}}$ distributions for the lepton+jets channel with a muon in the final state in the 0 t tag SR (left) and in the 1 t tag SR (right). The expected contribution of a Z′ boson with mass of 4 TeV and relative width of 1% is shown. Figures from [75].

Spin-1 resonances

Searches for Z′ resonances and g_{KK} gluons have been performed already at the Tevatron and then at the LHC at various center-of-mass energies. No new physics has been found to date, and upper limits have been placed on the production cross section of $Z'/g_{KK} \to t\bar{t}$. The first searches for leptophobic Z′ decaying to top quark pairs have been performed by the CDF [59–62] and D0 Collaborations [63, 64] at the Tevatron at $\sqrt{s} = 1.96$ TeV. Masses up to 900 GeV were excluded at 95% confidence level (CL). At the LHC, searches for Z′ and g_{KK} have been performed by CMS and ATLAS at 7 TeV [65–69], at 8 TeV [70–72] and at 13 TeV [72–76]. The most stringent limits to date on g_{KK} have been derived by the CMS Collaboration at $\sqrt{s} = 13$ TeV using 35.9 fb^{-1} of data and g_{KK} masses up to 4.55 TeV are excluded [75]. For Z′ resonances, different relative widths have been probed. For relative widths of 1%, 10% and 30% the excluded masses are 3.80, 5.25 and 6.65 TeV, respectively, and they have been obtained by the CMS Collaboration at $\sqrt{s} = 13$ TeV using 35.9 fb^{-1} of data [75]. For the 1.2% relative width, the best limit is obtained by the ATLAS Collaboration at 13 TeV analyzing 139 fb^{-1} and it corresponds to 4.1 TeV [76].

The most recent published CMS result [75], of which the analysis presented in this thesis is an extension, is discussed in the following. Heavy Z′ bosons and g_{KK} gluons are searched for in all the three possible $t\bar{t}$ decay modes: the lepton+jets, the dileptonic and the all hadronic final states, and the results are combined to enhance the final sensitivity. The analysis looks at deviations in the invariant $t\bar{t}$ mass spectrum: a possible signal

would appear as a peak over a falling background. The focus is on the boosted final state: the particles produced in the decay of a heavy particle can obtain a large Lorentz-boost, which consequently causes their decay products to be collimated. Different techniques are developed to reconstruct and identify top quarks with large Lorentz boost (see Sec. 5.6), referred to as *t tagging*. On the other hand, the decay products of particles that decay at rest are well separated.

The dileptonic channel selects events which contain exactly two oppositely charged leptons (ℓ), either muons or electrons, on which no isolation requirement is placed, to allow the reconstruction of boosted t quarks. Moreover, at least two small-radius jets (cf. Sec. 5.5) are selected, one of which has to be identified as originating from the fragmentation of a b quark (c.f. Sec. 5.6.1), and finally missing transverse energy p_T^{miss} (see Sec. 5.7) is required, which accounts for the presence of a neutrino. The main irreducible background is the SM $\mathrm{t\bar{t}}$ process, like for the other two analysis channels, while the main reducible background for the dilepton channel arises from the Z+jets process. All the backgrounds are estimated from simulation. An angular variable is used to define the signal region (SR) and the control regions (CRs) of the analysis. The sensitive variable is S_T, defined as:

$$S_T = \sum_{i=1}^{N_{\text{jet}}} p_{T_i}^{\text{jet}} + \sum_{i=1}^{2} p_{T_i}^{\ell} + p_T^{\text{miss}}, \tag{3.10}$$

where $p_T^{\text{jet}(\ell)}$ is transverse momentum of the jet (lepton).

The lepton+jets channel selects events with exactly one non-isolated electron or muon with high p_T, at least two jets and p_T^{miss}. Large-radius jets are considered as well, which are t tagged with a dedicated selection using the jet substructure (Sec. 5.6.2). The main reducible background is W+jets. The sensitive variable is the $\mathrm{t\bar{t}}$ invariant mass $m_{\mathrm{t\bar{t}}}$, which is reconstructed using a χ^2 approach. All the backgrounds shapes are estimated from simulation. A boosted decision tree (BDT) is used to separate the W+jets process and define the SR and CRs. The regions are further divided based on the presence of a t tagged large-radius jet and finally a selection on the χ^2 value is applied to further reduce background contributions.

Finally, the all hadronic channel targets events with two large-radius, t tagged jets. The main reducible background is the QCD multijet process, which is estimated from data. The number of sub-jets identified as originating from a b quark is used to define the SRs and CRs, together with the difference in rapidity between the two large-radius jets. The sensitive variable is $m_{\mathrm{t\bar{t}}}$ in this channel as well.

The results in the lepton+jets channel are shown in Fig. 3.7 for events with a muon in the final state and separated for events with 0 or 1 t tagged jets. The expected contribution

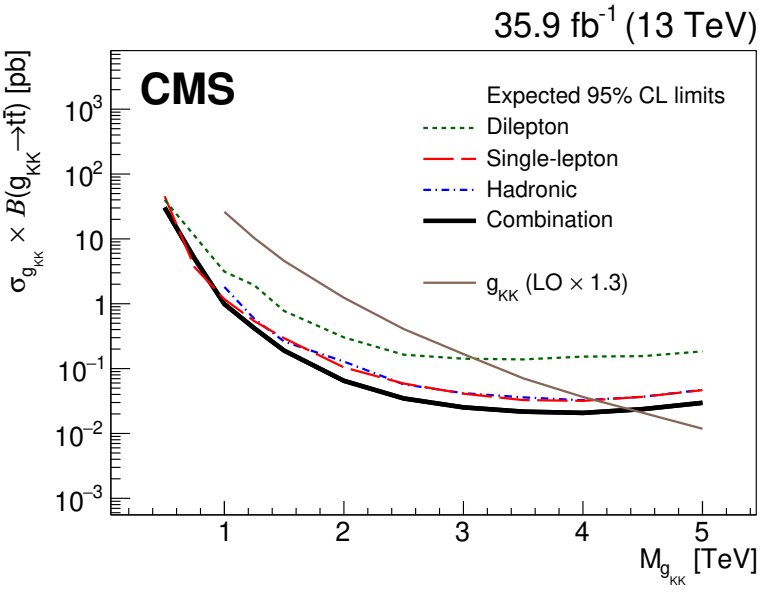

Figure 3.8: The expected exclusion limits at 95% CL on the production cross section of the g_{KK} gluon as a function of the gluon mass, showing the sensitivity of each channel and the combination. Figure from [75].

from a signal corresponding to a Z′ boson with mass of 4 TeV is also shown. Using the combined results of all three channels, the exclusion limits on the product of cross section and branching fraction are obtained for the four models considers. In Fig. 3.8 the expected contribution of each channel is shown separately, together with the combination, for the g_{KK} signal. The dilepton channel is sensitive below 1 TeV, while the lepton+jets and all hadronic channels, which have similar performance, lead the sensitivity at higher masses. The expected and observed exclusion limits are shown in Fig. 3.9 for the combination of the three channels for the Z′ and g_{KK} signals.

Spin-0 resonances

Searches for heavy Higgs bosons A/H decaying to top quark pairs have been performed by ATLAS at $\sqrt{s} = 8$ TeV [77] and $\sqrt{s} = 13$ TeV [78] and by CMS at $\sqrt{s} = 13$ TeV using 35.9 fb^{-1} [79] and 138 fb^{-1} [80] of data. The most stringent constrains to date on the coupling strength modifiers g_H and g_A are reported by the CMS Collaboration [80] for relative widths from 0.5 to 25% and masses in the range $365 - 1000$ GeV. An excess has been observed close to the $\mathrm{t\bar{t}}$ production threshold with a significance above 5 standard deviations, more compatible with the pseudoscalar than the scalar hypothesis. The excess is compatible with a $\mathrm{t\bar{t}}$ bound state ($\eta_{t\bar{t}}$) with a cross section of 7.1 pb.

The latest CMS result [80] is summarized in the following. The analysis targets the lepton+jets and dileptonic final states in the resolved regime, which is characterized by

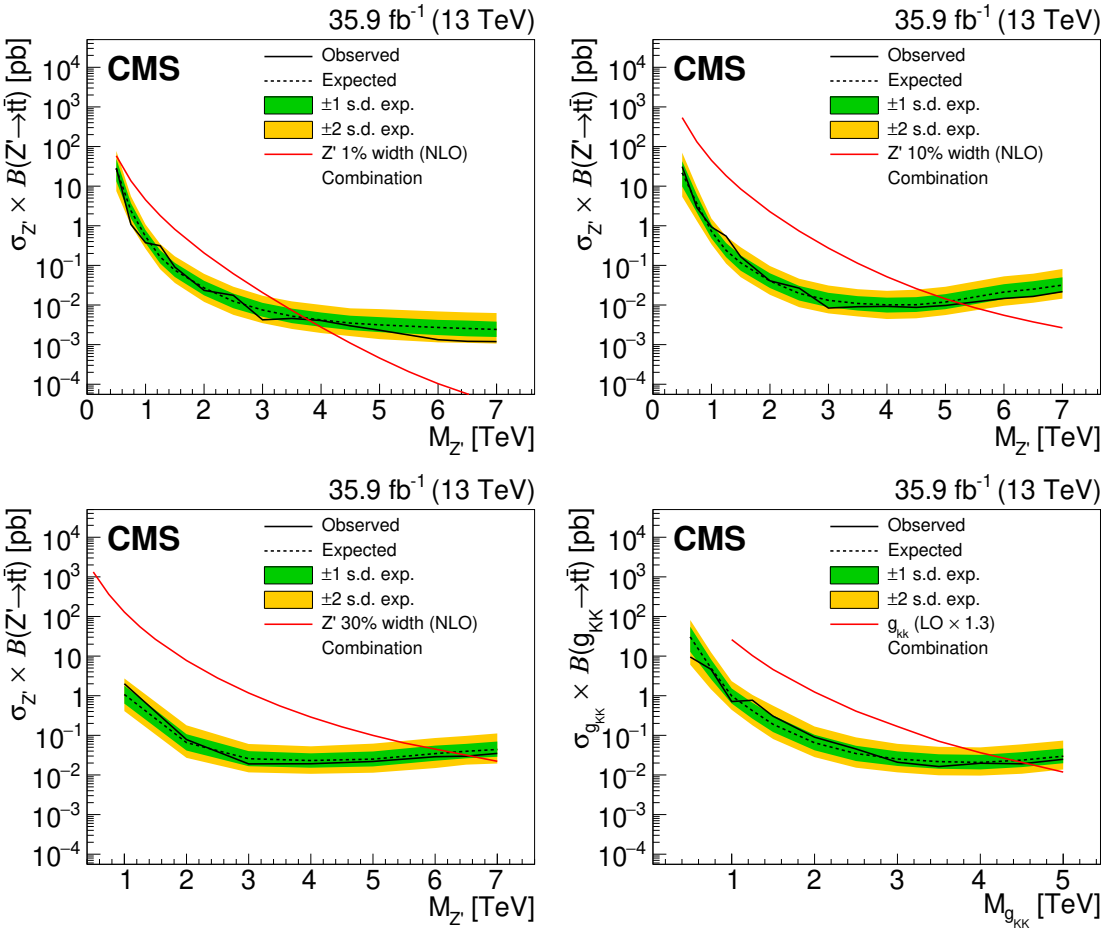

Figure 3.9: The observed and expected exclusion limits at 95% CL on the production cross section of the new particle as a function of the particle mass for the Z′ bosons with 1%, 10% and 30% relative widths and for the g_{KK} gluon. Figures from [75].

the presence of isolated leptons and small-radius jets. The H/A signals interfere with the SM $\mathrm{t\bar{t}}$, resulting in a peak-dip structure in the $m_{\mathrm{t\bar{t}}}$ spectrum. Moreover, the signal and the backgrounds show different angular properties. The lepton+jets channel selects events with exactly one isolated electron or muon, at least three jets, of which at least two b tagged, and p_T^{miss}. The main irreducible background is $\mathrm{t\bar{t}}$, common to both final states, which is estimated from simulation, while the QCD background is estimated from data. The sensitive variables of the search are two: $m_{\mathrm{t\bar{t}}}$ and the cosine of θ^*, an angular variable sensitive to the spin of the decaying particle. The dilepton channel selects two oppositely charged leptons, at least two jets, of which at least one b tagged, and p_T^{miss}. All the backgrounds are estimated from simulation. For the reducible $\mathrm{Z}/\gamma + \mathrm{jets}$ background, the total yield in simulation is corrected from data. Three are the sensitive variables used: $m_{\mathrm{t\bar{t}}}$ and two spin correlation variables c_{hel} and c_{han}.

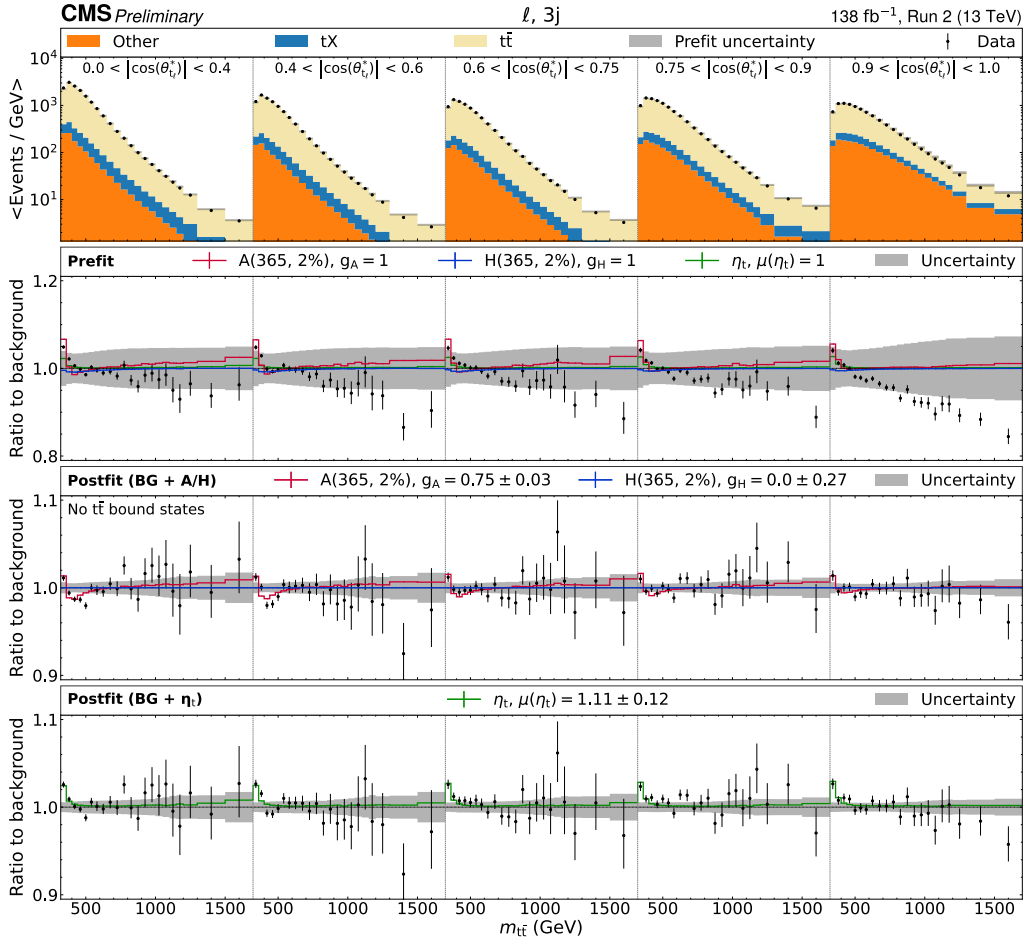

Figure 3.10: The $m_{\mathrm{t\bar{t}}}$ distributions in bins of $|\cos(\theta^*)|$ in the lepton+jets channel for events with exactly 3 jets. The prefit as well as the postfit ratios of data to backgrounds are shown. In the postfit panels, the contribution of a A/H signal or of the $\eta_{t\bar{t}}$ bound state is included. Figure from [80].

The results from the lepton+jets final state are shown in Fig. 3.10. The expected contribution from three interpretations are shown: a signal corresponding to a pseudoscalar boson A or scalar boson H with mass of 365 GeV, or a $\eta_{t\bar{t}}$ bound state. The results from the two channels are combined to increase the sensitivity of the search and expected and observed exclusion limits on the coupling strength modifiers are obtained. The limits are shown in Fig. 3.11 for the pseudoscalar scenario, without the inclusion of the $\eta_{t\bar{t}}$ bound state in the background prediction. A deviation at low $m_{\mathrm{t\bar{t}}}$ values, close to the $\mathrm{t\bar{t}}$ production threshold, is observed. The results with the inclusion of the $\eta_{t\bar{t}}$ contribution to the background are shown in Fig. 3.12. In this case, the observed constrains agree with the expectation.

The search presented in this thesis extends the previous CMS results by analyzing

the full Run 2 dataset, corresponding to 138 fb^{-1} of pp data, and targets the lepton+jets final state for both the resolved and the boosted regimes. The analysis is carried out in a model-independent approach and the result interpretation is performed for spin-1 and spin-0 signals, taking into account interference effects. The analysis targets non-resonant effects as well, described in Ref. [1], which are not presented in this thesis.

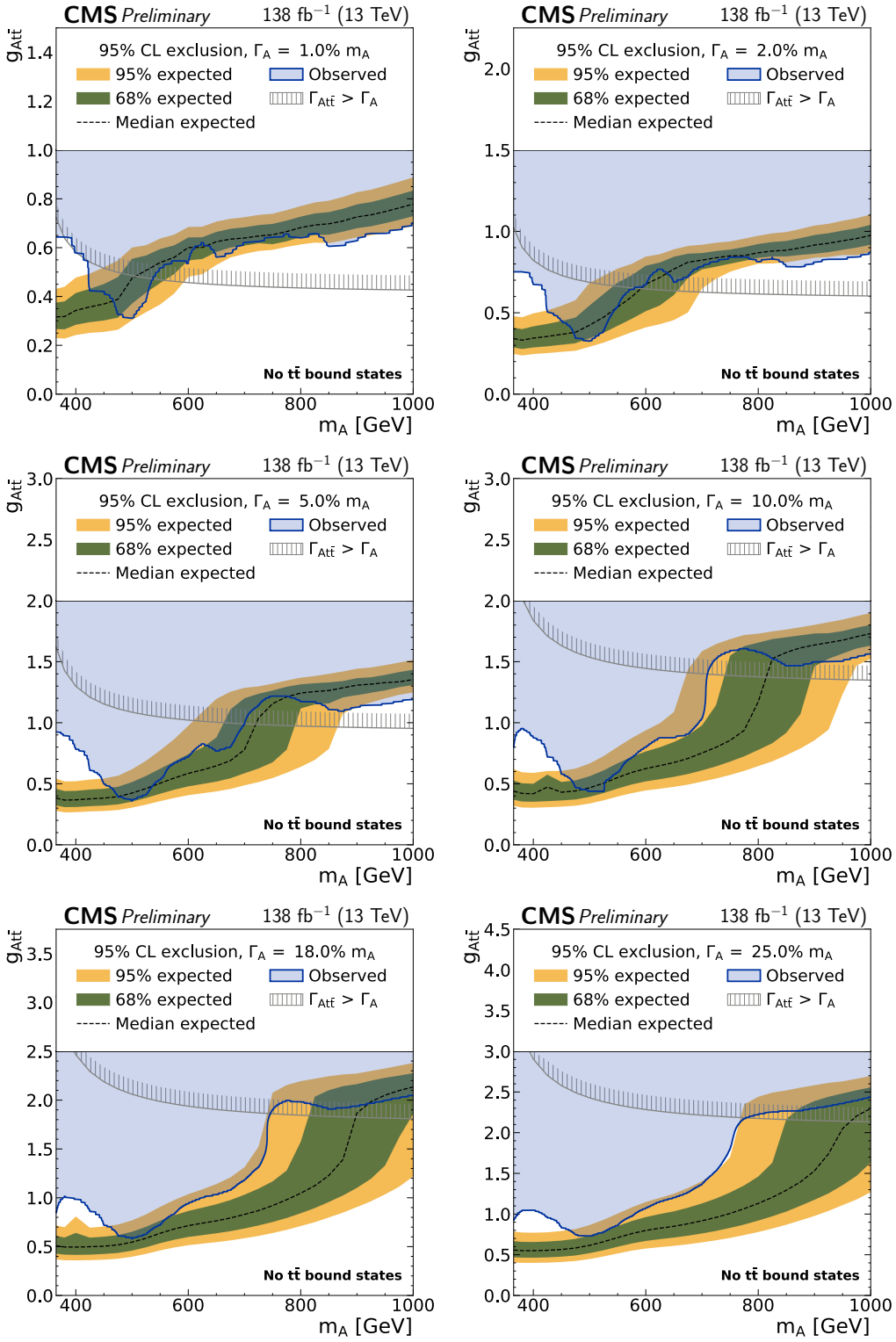

Figure 3.11: The observed and expected exclusion limits at 95% CL on the coupling strength modifier for a pseudoscalar boson A as a function of the boson mass, for the 1, 2, 5, 10, 18 and 25% relative widths. Figure from [80].

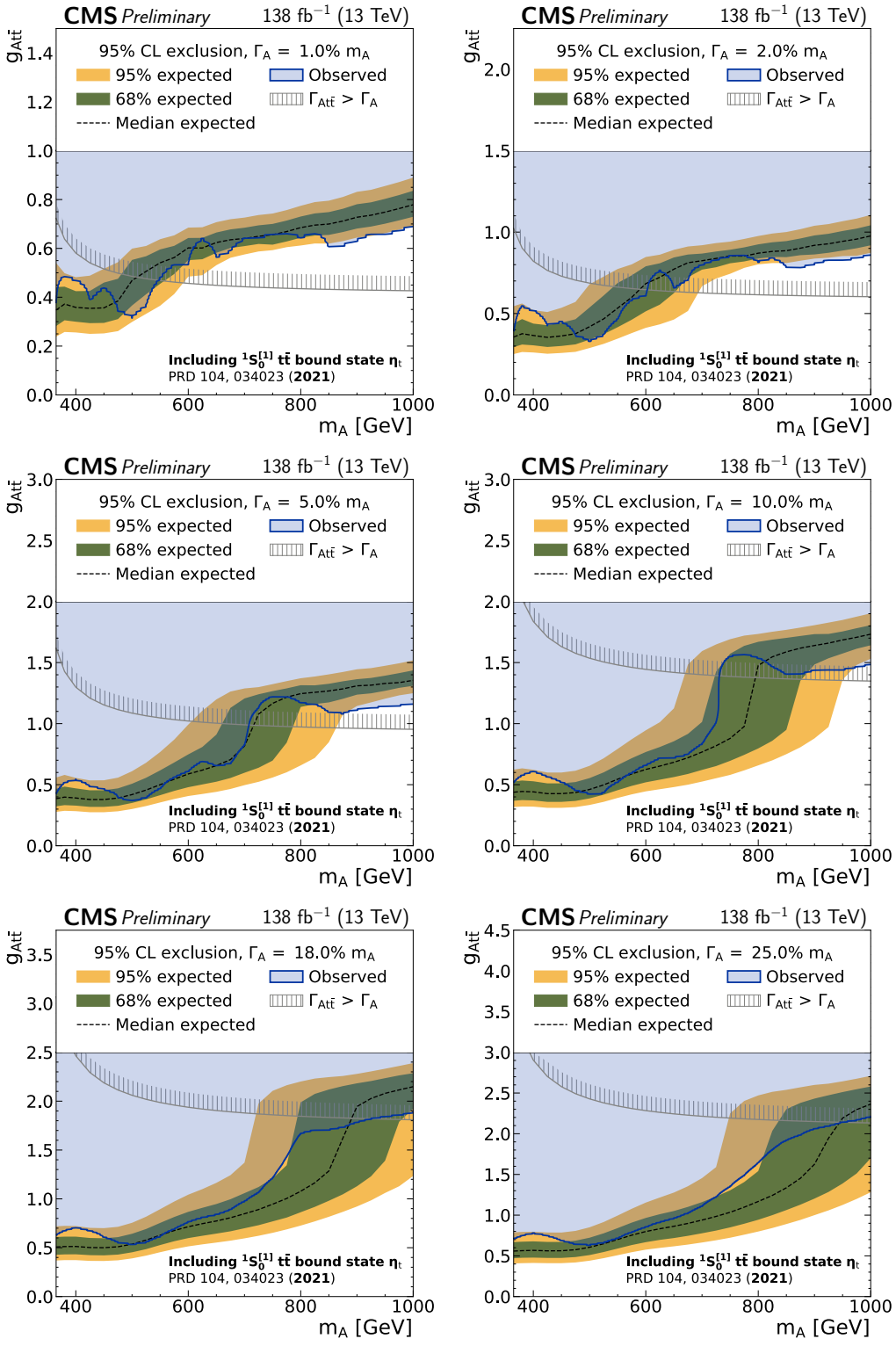

Figure 3.12: The observed and expected exclusion limits at 95% CL on the coupling strength modifier for a pseudoscalar boson A as a function of the boson mass, for the 1, 2, 5, 10, 18 and 25% relative widths. The contribution of $\eta_{t\bar{t}}$ to the background is included. Figure from [80].

Chapter 4

Experimental setup

The analysis presented in this thesis uses 13 TeV pp collision data recorded with the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC). The data have been collected during 2016-2018 and correspond to an integrated luminosity of 138 fb^{-1}. The experimental setup will be discussed in this Chapter: in Section 4.1 the LHC collider will be presented, followed by the description of the CMS detector and its components in Section 4.2.

4.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [81] is the largest and most powerful particle collider in the world operated at the *Conseil européen pour la recherche nucléaire* (CERN), an international organization for particle physics research, and it is located on the France–Switzerland border near Geneva. The LHC lies in the tunnel previously constructed for the Large Electron-Positron (LEP) collider, located between 45 and 170 meters underground, with a circumference of 26.7 km and a slope of 1.4%. It accelerates and collides beams of protons and heavy ions, which interact in four interaction points (IPs), where the main experiments are located. In the following only the proton operation mode will be described, being the one relevant for this thesis.

A strong magnetic field of 8.3 T is used to bend the protons in the beamlines. The magnetic field is produced by 1232 superconductive niobium-titanium (NbTi) dipole magnets of 14.3 m length, while 392 quadrupole magnets focus the beams.

The aim of the LHC physics programme is to perform precise measurements of the SM, including the Higgs boson discovered in 2012, as well as to discover new phenomena, which could be possible thanks to the high energies at which it operates. The four main experiments at the LHC are: CMS (Compact Muon Solenoid), ATLAS (A Toroidal LHC

ApparatuS), LHCb (Large Hadron Collider beauty) and ALICE (A Large Ion Collider Experiment). ATLAS and CMS are general-purpose detectors that study high-energy pp and heavy ion collisions. They aim to precisely measure SM physics processes and to search for new physics phenomena. The CMS detector will be described in detail in the following Section. The LHCb experiment is designed to study rare decays of b and c hadrons and to measure the parameters of CP violation. It is a single-arm spectrometer, as b hadrons are mostly produced in the same forward direction. Finally, ALICE is specialized on heavy ion physics with the aim of studying the quark-gluon plasma.

A complex injection chain that is shown in Fig. 4.1 collects and accelerates the protons before they enter the LHC. The protons are obtained from gaseous hydrogen via ionisation and then accelerated in different steps. First they are sent to the Linear Accelerator (LINAC2) that accelerates them up to 50 MeV. Then they are injected in three circular accelerators: the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS) where they reach an energy of 450 GeV. Finally they are grouped into beams and injected in the LHC ring. The beams enter in two counterrotating beamlines and are further accelerated in the main ring via radio frequency cavities at 400 MHz, acquiring 0.5 MeV per revolution. The designed energy of each beam is 7 TeV, corresponding to a center-of-mass energy of $\sqrt{s} = 14$ TeV. The LHC operated at center-of-mass energy of $\sqrt{s} = 7$ and 8 TeV during 2010-2011 and 2012, respectively. During Run 2 (2015-2018) the center-of-mass energy was 13 TeV, while at the start of Run 3 in 2022 a value of $\sqrt{s} = 13.6$ TeV has been reached, setting a new world record. It has not been possible to arrive at the target energy of 14 TeV in Run 3, because the magnet training to reach 7 TeV beam energies has not been achieved, obtaining a stable performance at 6.8 TeV.

The protons travelling in the two circular beamlines are grouped in 2808 bunches of $1.15 \cdot 10^{11}$ protons each. The bunches are separated by a distance of 7.5 m, which corresponds to a collision every 25 ns. The collision rate is thus 40 MHz.

An important parameter of the accelerator is the instantaneous luminosity $\mathcal{L}$, that is proportional to the number of events produced in the collider. The number of events N for a given process is:

$$N = \sigma \int \mathcal{L} dt \tag{4.1}$$

where σ is the cross section of the process. Therefore, with high values of instantaneous luminosity, the expected number of rare processes events increases. The luminosity in the collider can be described as:

$$\mathcal{L} = \frac{n_b N_b^1 N_b^2 f}{4\pi\sigma_x\sigma_y} F \tag{4.2}$$

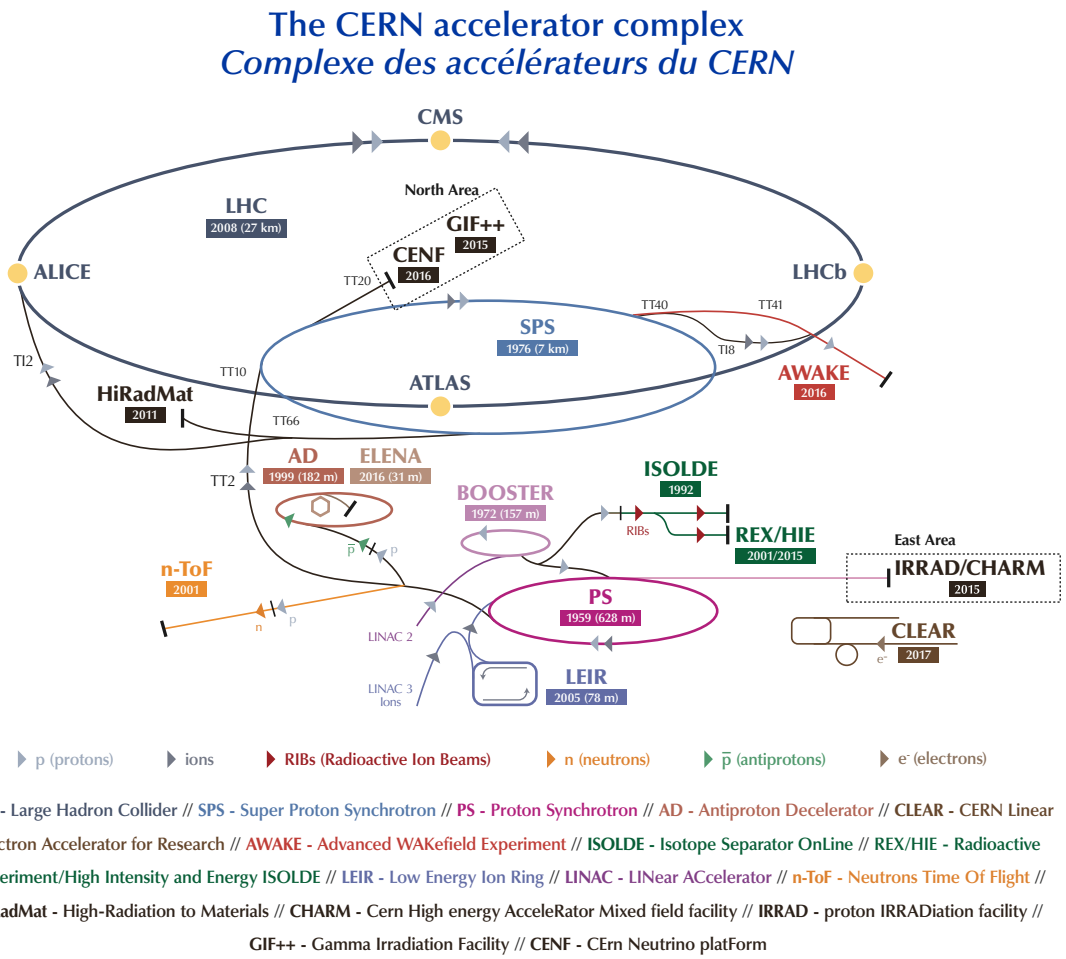

Figure 4.1: Sketch of the CERN accelerator complex. Figure from [82].

where n_b is the number of bunches per beam, $N_b^{1/2}$ is the number of particles per bunch in each of the two beams, and f the revolution frequency. The parameters $\sigma_{x,y}$ are the transverse beam sizes in the x and y direction at the IP and F is the geometric luminosity reduction factor due to the crossing angle. The design peak instantaneous luminosity of the LHC is $\mathcal{L} = 10^{34}$ cm^{-2}s^{-1}, which was already achieved in 2016 and more than doubled in 2018. The total integrated luminosity delivered by LHC to CMS as a function of time is shown in Fig. 4.2.

While the high instantaneous luminosity increases the rate of rare and potentially new processes, it comes with a high number of pp collisions happening in the same bunch crossing. It is extremely important for physics analysis to identify the main pp interaction for each bunch crossing, which is the one of interest, and to reduce the effects of the additional pp collisions, called pileup (PU). Chapter 6 is focused on pileup mitigation and on the techniques most commonly used in CMS.

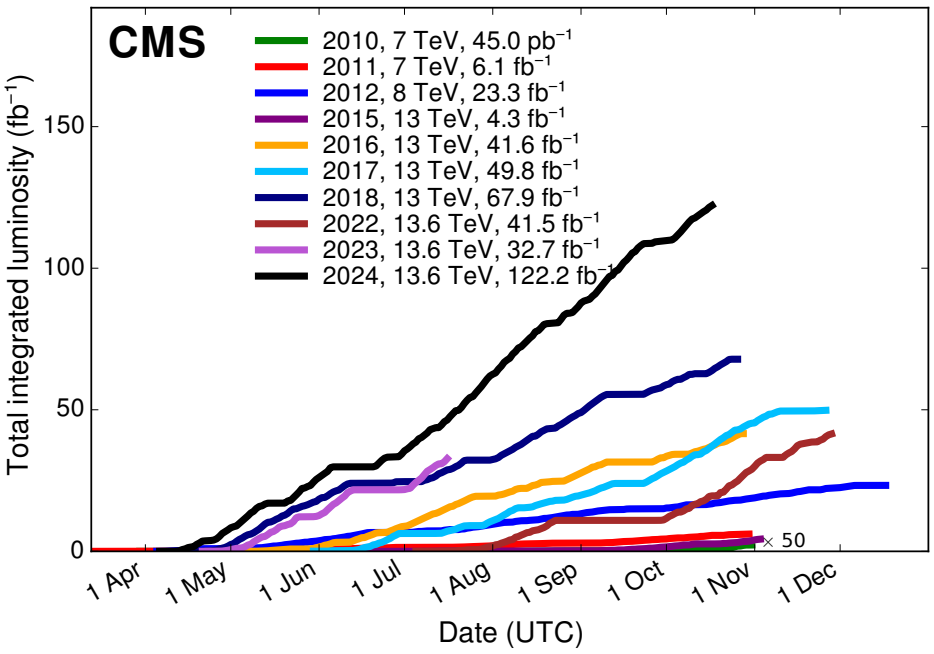

Figure 4.2: The total integrated luminosity delivered by LHC to the CMS experiment as a function of time for pp collisions, for the period 2010-2024. Figure from [83].

4.1.1 The coordinate system

The LHC coordinate system has its origin at the IP, with the x-axis pointing radially towards the LHC center, the y-axis pointing vertically upwards and the z-axis lying along the beam axis in counterclockwise direction. The polar coordinates (r, θ, ϕ) are more commonly used in CMS, given the cylindrical symmetry of the experiment. Given that in high-energy pp collisions the interactions occur between partons, with an unknown fraction of the proton momentum in the z direction, the collisions are boosted along the z-axis (see Sec. 2.2). Thus the coordinates have to be Lorentz-invariant under boosts along the beam axis. While it is already the case for ϕ and r, the polar angle θ is not Lorentz-invariant. Instead, the *pseudorapidity* η is used:

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]. \tag{4.3}$$

Differences in η are invariant under Lorentz-boost along the z-axis. Another quantity, used in experiments as CMS, that is by construction Lorentz-invariant is $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ which is a measure of angular separation.

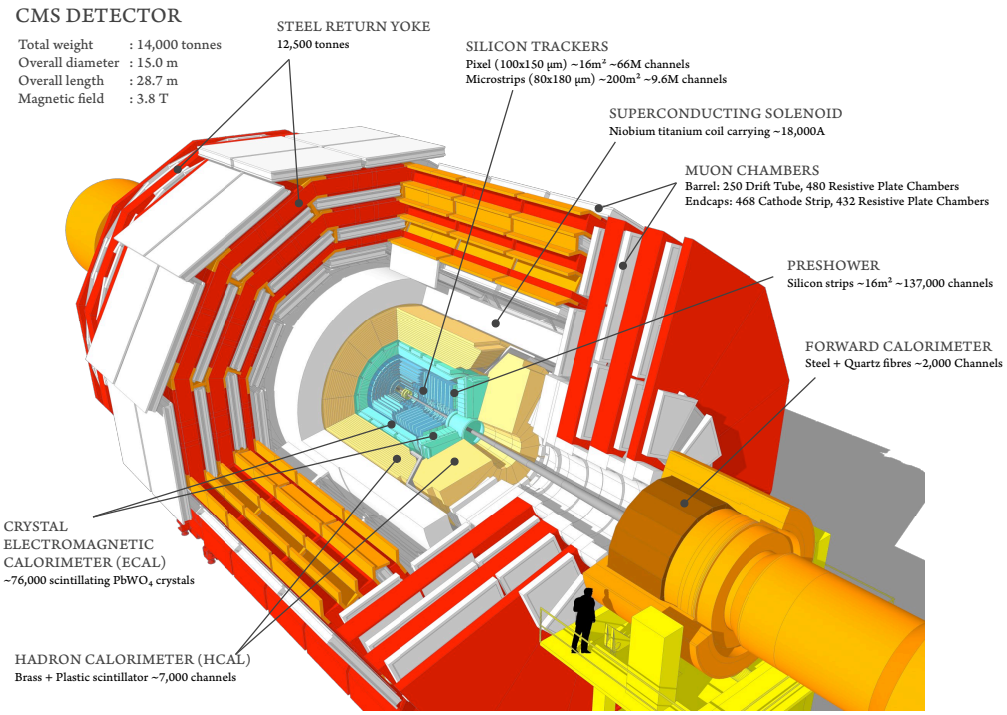

Figure 4.3: A schematic view of the CMS experiment, showing the sub-detectors and the superconducting solenoid. Figure from [84].

4.2 The Compact Muon Solenoid detector

The CMS experiment [85, 86] is a multi-purpose detector situated at IP 5, about 100 m underground. The detector was designed as a discovery machine: its main goals are the discovery of the Higgs boson and the measurements of its properties, the search for new physics as well as the precise measurements of the SM parameters. In particular, with the aim of finding the Higgs boson in the *golden* channels, $H \to \gamma\gamma$ and $H \to 4\ell$, the measurement of muons and photons with extremely high resolution was essential for the detector design.

The key feature of the CMS experiment is the powerful superconducting solenoid magnet which bends the trajectories of charged particles and allows to measure their properties. The solenoid provides a magnetic field of 3.8 T, it is 13 m long and has an inner diameter of 5.9 m. The cylindrical structure, symmetrical around the beamline, is divided in two regions: the *barrel*, the central part that is coaxial with the beamline, and two *endcaps*, one on the forward and one on the backward side, for a coverage of almost 4π. A sketch of the CMS detector is shown in Fig. 4.3. The detector is composed of a series of sub-detectors in a layered structure, each dedicated to the measurement of a particular type of particles. Starting from the collision point and moving outwards the sub-detectors

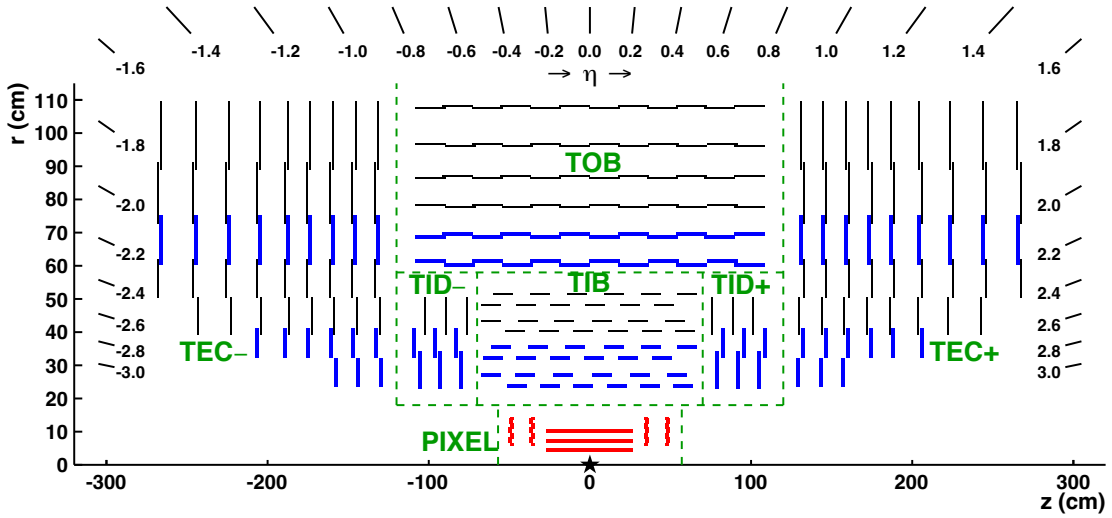

Figure 4.4: A schematic view of the CMS tracking system in the $r - z$ plane. Figure from [87].

are: the tracker, the electromagnetic calorimeter, the hadronic calorimeter and the muon system, that is placed in the return yoke of the superconducting solenoid. The detector is 21.6 m long with a diameter of 14.6 m. The different sub-detectors of the CMS experiment and the solenoid magnet will be briefly described in the following Sections.

4.2.1 Inner tracking system

The tracking system, or tracker, is the innermost sub-detector of the CMS experiment and it is used to measure the trajectory and the charge sign of the charged particles with very high precision and to reconstruct primary and secondary vertices. It lies completely within the magnetic field produced by the solenoid, needed to bend the charged particles and enabling the measurement of their sign and momentum, and it is made entirely by silicon detectors. The tracker has a length of 5.8 m and a diameter of 2.5 m and covers the pseudorapidity range up to $|\eta| = 2.5$. The tracking system is made by two parts: a pixel detector, closest to the IP, and a strip tracker, with a total active area of about 200 m^2. A schematic representation of the CMS tracking system is shown in Fig. 4.4.

The pixel detector is composed of three barrel layers with radii of 4.4, 7.3 and 10.2 cm and two endcap disks per side, at 34.5 and 46.5 cm from the IP. The pixels have a size of 100×150 μm^2. During the long shut-down between the 2016 and 2017 data-taking periods, the pixel detector has been upgraded (Phase-I upgrade) to recover the performance in the high instantaneous luminosity regime and cope with radiation damage. In the new layout an additional layer has been added to the barrel and to each endcap. The Phase-I detector is made of four barrel layers at $r = 3.0$, 6.8, 10.2 and 16.0 cm and three endcap disks at

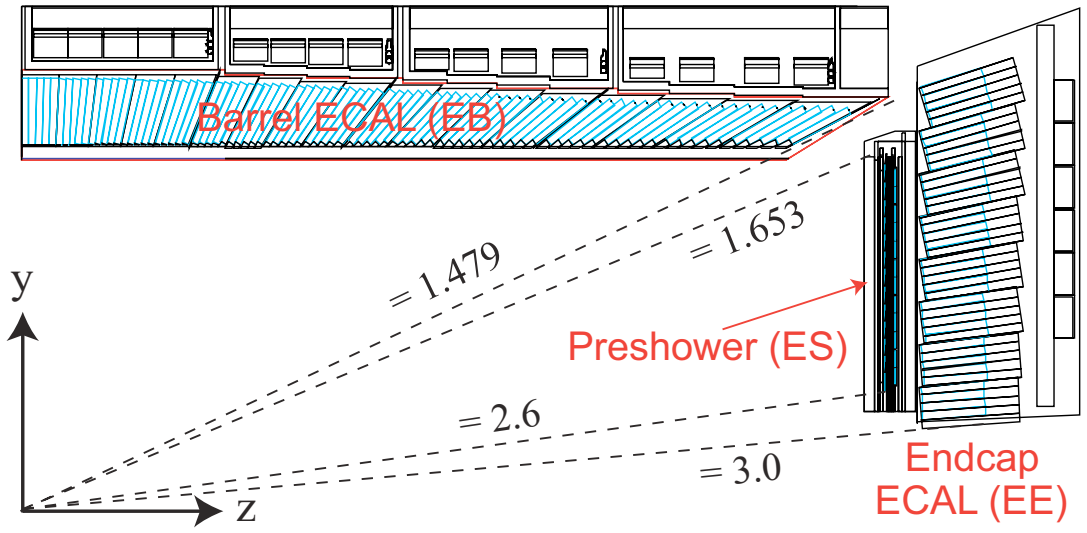

Figure 4.5: A schematic view of the ECAL calorimeter. Figure from [86].

$|z| = 29.1$, 39.6 and 51.6 cm.

The innermost detector is surrounded by the silicon micro-strip tracker, which covers a ragion between $r = 20$ and 110 cm. The strip tracker is made of different sub-modules: in the central region there are the tracker inner barrel (TIB) and tracker outer barrel (TOB). In the forward region there are the tracker inner disk (TID) and the tracker endcaps (TEC). Given the lower particle flux in this region with respect to the region of pixel detector, the size of the strips can be larger: in the inner region the strips have a surface of 10 cm $\times$ 80 μm, while in the outer region their surface is 25 cm $\times$ 180 μm. The TIB is made of 4 layers and covers the region up to $r = 55$ cm, while the TOB has 6 layers and reaches up to $r = 110$ cm and $|z| = 118$ cm. The TECs, one at each end, are located in the region $124 < |z| < 282$ cm and $22.5 < |r| < 113.5$ cm and are made of 9 disks each. Finally the TID covers the gap between the TIB and TEC and is composed of 3 disks per side.

4.2.2 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) is placed outside of the tracking system and it is used to measure the energy of electrons and photons. When these particles travel through the calorimeter, they deposit progressively their energy until they stop, by means of electromagnetic shower production. The CMS ECAL is a hermetic, homogeneous, scintillating crystal calorimeter. The choice of the material has to satisfy the requirements of high granularity and fast response of the detector, and the limited space available inside the solenoid magnet. The requirements are fulfilled by lead tungstate ($PbWO_4$), which has high density ($\rho = 8.28$ g/cm^3), short radiation length ($X_0 = 0.89$ cm) and small

Molière radius ($R_M = 2.2$ cm). Moreover it has a short scintillation decay time: 80% of the scintillation light is emitted in the time of a bunch crossing (25 ns).

The calorimeter is made of a barrel part (EB), covering the region $|\eta| < 1.479$, and two endcaps (EE), extending the coverage to $1.479 < |\eta| < 3.0$. In EB there are 61200 crystals, each 23 cm long, corresponding to 25.8 X_0. The crystals in the EE are 7324 per side, with a length of 22 cm, corresponding to 24.7 X_0.

A pre-shower sampling calorimeter (ES) is placed in front of EE and it is used to identify the photons originating from the decays of neutral pions and to determine more precisely the position of electrons and photons. The ES is located in the region $1.653 < |\eta| < 2.6$ and is made of two layers: lead radiators and silicon strips. The photodetectors used to detect the scintillation light are silicon avalanche photodiodes (APDs) in the barrel and vacuum phototriodes (VPTs) in the endcaps. The layout of ECAL can be seen in Fig. 4.5. The energy resolution of ECAL [88], measured in test beams, can be parametrized as:

$$\frac{\sigma_E}{E} = \frac{S}{\sqrt{E}} \oplus \frac{N}{E} \oplus C = \frac{2.8\%}{\sqrt{E/\mathrm{GeV}}} \oplus \frac{12\%}{E/\mathrm{GeV}} \oplus 0.3\% \tag{4.4}$$

where S is the stochastic term, N the electronic noise term and C a constant related to calibration errors and inhomogeneities.

4.2.3 Hadronic calorimeter

The hadronic calorimeter (HCAL) is used to measure the energy of hadrons, which have a larger interaction length compared to electrons and photons, and travel through ECAL without being absorbed. HCAL is an hermetic sampling calorimeter, made of alternating layers of absorber and active medium. The detector is divided in four parts: the hadron barrel (HB), the endcap (HE), the outer hadron calorimeter (HO) and the forward hadron calorimeter (HF). The first three parts of the calorimeter are made of brass absorber layers alternated with plastic scintillators. The HF, being in the region with the highest particle flux, is made of steel plates with quartz fibres as active material, which makes it able to detect both hadronic and electromagnetic showers.

The HB is located between the ECAL and the solenoid coil and extends up to $|\eta| < 1.3$, while the HO lies between the solenoid and the muon system, covering the region up to $|\eta| < 1.26$. For the outer calorimeter the solenoid acts as an absorber. The HE lies in the range $1.3 < |\eta| < 3$. Finally, the HF is placed outside the magnet yoke, at $|z| = 11.2$ m from the IP, in the range up to $|\eta| < 5$. A longitudinal view of the HCAL can be seen in Fig. 4.6. The energy resolution for single neutral pions, determined in test beams [89], can

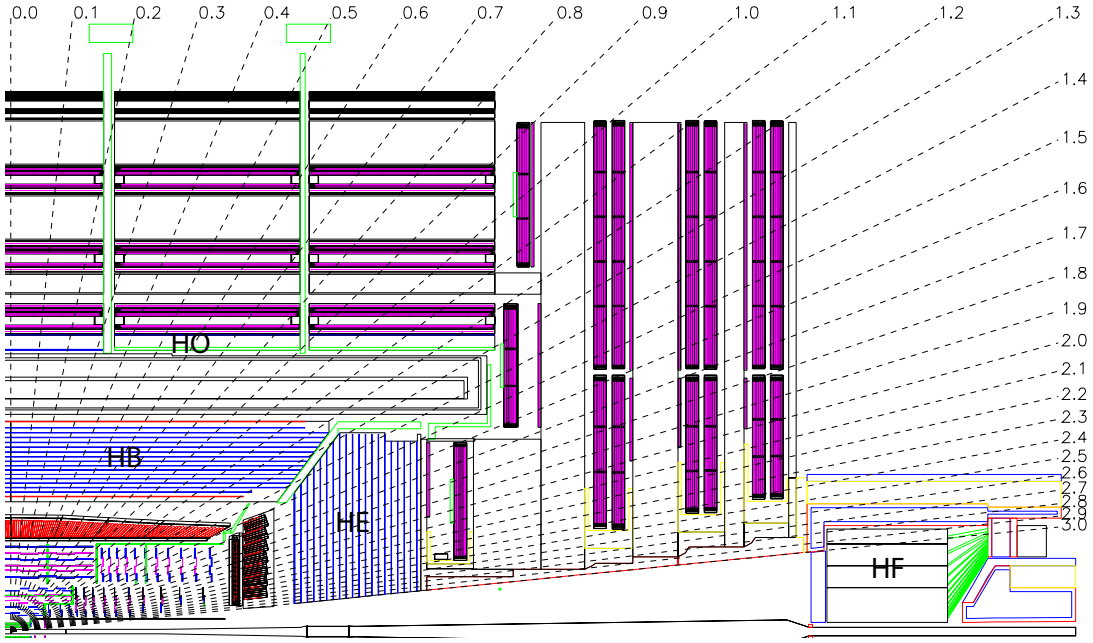

Figure 4.6: A schematic view of the HCAL calorimeter. Figure from [85].

be expressed as:

$$\frac{\sigma_E}{E} = \frac{111.5\%}{\sqrt{E/\text{GeV}}} \oplus 8.6\% \tag{4.5}$$

where the same parametrization as in Eq. 4.4 is used.

4.2.4 Superconducting solenoid

The superconducting solenoid magnet is a central part of the CMS experiment, providing the magnetic field needed to bend the trajectories of the charged particles and thus measure their momentum. The solenoid encloses the inner tracker and the calorimeters, it is 12.9 m long and has a diameter of 5.9 m. It is made of NbTi in 4 layers and provides a homogeneous magnetic field of 3.8 T in the inner part. The magnetic field is closed by an iron yoke, which hosts the muon detector system. The magnetic filed on the outside is of 2 T and bends the trajectories of the muons in the opposite direction with respect to the inner filed, providing an even more precise measurement of their momentum.

4.2.5 Muon system

The muon system is used to identify and measure muons with the highest precision and constitutes an important part of the CMS detector. It is the outermost part of the experiment: the reason is that muons can travel through all the CMS sub-detectors

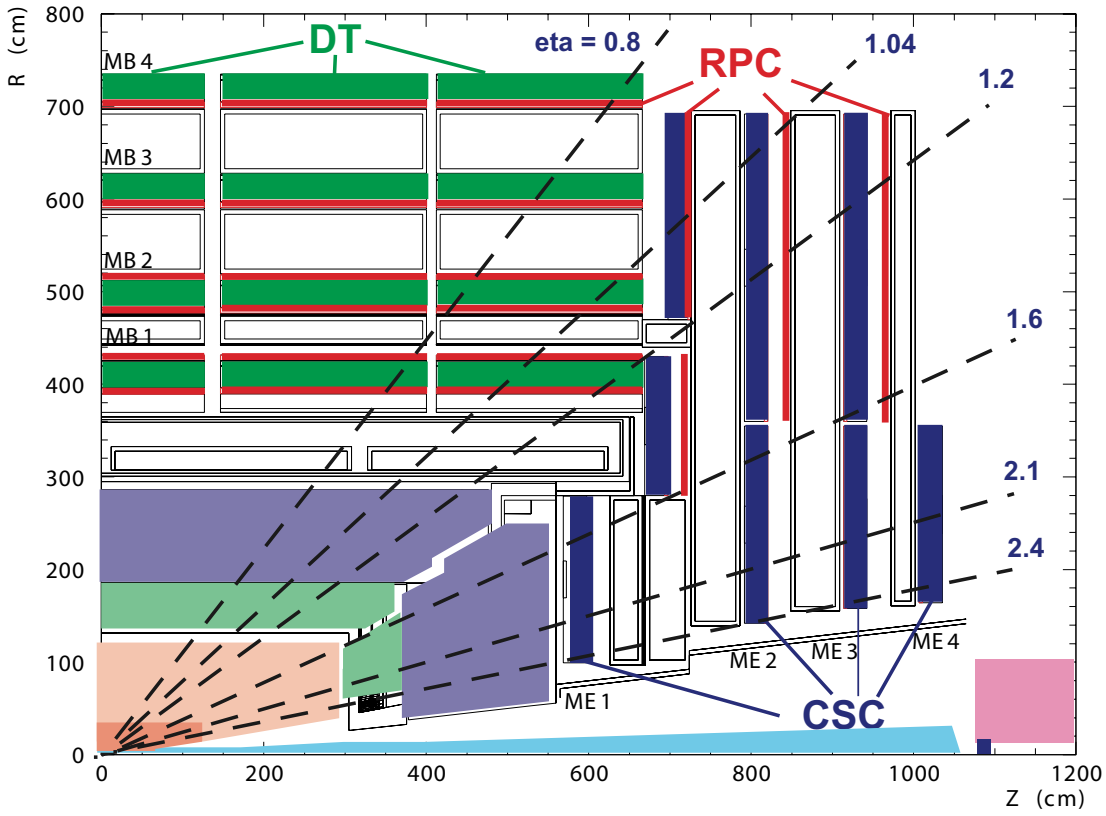

Figure 4.7: A schematic view of the CMS muon system. Figure from [86].

with a minimum loss of energy and are the only particles, except for neutrinos, that are not absorbed in the calorimeters. The muon system consists of three gaseous detectors: drift tubes (DT) in the barrel ($|\eta| < 1.2$), cathode strip chambers (CSC) in the endcaps ($0.9 < |\eta| < 2.4$) and resistive plate chambers (RPC) in both the regions, covering the range up to $|\eta| < 1.6$.

The choice of the material was driven by the different conditions in the different pseudorapidity regions and by the large surface to cover. The DT are placed in the barrel region, where the magnetic field is low and uniform and the muon rate is low. They are made of 4 stations, alternate with the iron return yoke. In the forward region, that is characterized by a high background rate and high, non-uniform magnetic field, the CSC are used, because of the fast response and high radiation tolerance. They consist of 4 stations in each endcap. The RPC are made of 6 stations in the barrel region and 3 in each of the endcaps. They provide a fast response and an independent muon trigger system. A representation of the muon detectors is shown in Fig. 4.7.

4.2.6 Trigger system

The high luminosity of the LHC and the high bunch crossing rates of 40 MHz result in 10^9 events produces per second. It is technically impossible to store all the events produced, nor it is needed, as the interesting events are very rare compared to well known processes, like QCD multijet production. In order to select and store the interesting events, in CMS a two-level trigger system is used. First the Level-1 trigger (L1), based on hardware, is employed, followed by the High-Level trigger (HLT), which is software-based. The L1 uses the information of the calorimeters and muon system to keep or reject events in 3.2 μs. The event rate is reduced to 100 kHz. The HLT processes the events from the L1 employing complex algorithms and uses the information of all the sub-detectors. The decision to keep an event is made in 50 ms and the rate is further reduced to 100 Hz. In CMS, the combination of selections and filters applied in the HLT is referred to as *Path*. The commonly used HLT Paths require one or more final state objects above a certain transverse momentum or energy threshold.

Chapter 5

Object reconstruction in CMS

The events produced in pp collisions at the LHC have to be reconstructed starting from the raw electronic signals left by particles in the CMS detector. Exploiting the different signatures the particles leave in each detector layer, it is possible to combine the information and reconstruct each physics object in the most precise way. The algorithm used in CMS is called Particle Flow: starting from the tracks and energy clusters, it reconstructs and identifies muons, electrons, photons and hadrons. The Particle Flow algorithm will be described in Sec. 5.1, followed by the reconstruction of the primary vertices in Sec. 5.2. The objects important for the analysis presented in this thesis will be described in detail: muons (Sec. 5.3), electrons (Sec. 5.4), jets (Sec. 5.5) and missing transverse momentum (Sec. 5.7). A particular emphasis is put on jet reconstruction and on the identification of the jets originating from the decays of b and t quarks (Sec. 5.6).

5.1 Particle Flow

The Particle Flow (PF) algorithm [88] aims at reconstructing particles starting from the signatures they leave in each sub-detector of CMS and merging together the elements into final objects called PF candidates. The elements that are reconstructed are the trajectories of charged particles in the inner tracker and in the muon system (*tracks*), the energy deposits in the electromagnetic and hadronic calorimeters (ECAL and HCAL *clusters*), and the primary vertices (PV) and secondary vertices (SV). Tracks and clusters can be linked together to obtain the PF blocks, which finally can be identified as neutral and charged hadrons, photons, electrons and muons (PF candidates). The PF approach for particle identification and reconstruction shows excellent performance and its success is permitted by the fine granularity of the CMS detector components. In Fig. 5.1 an example of how different particles interact with the CMS sub-detectors in shown.

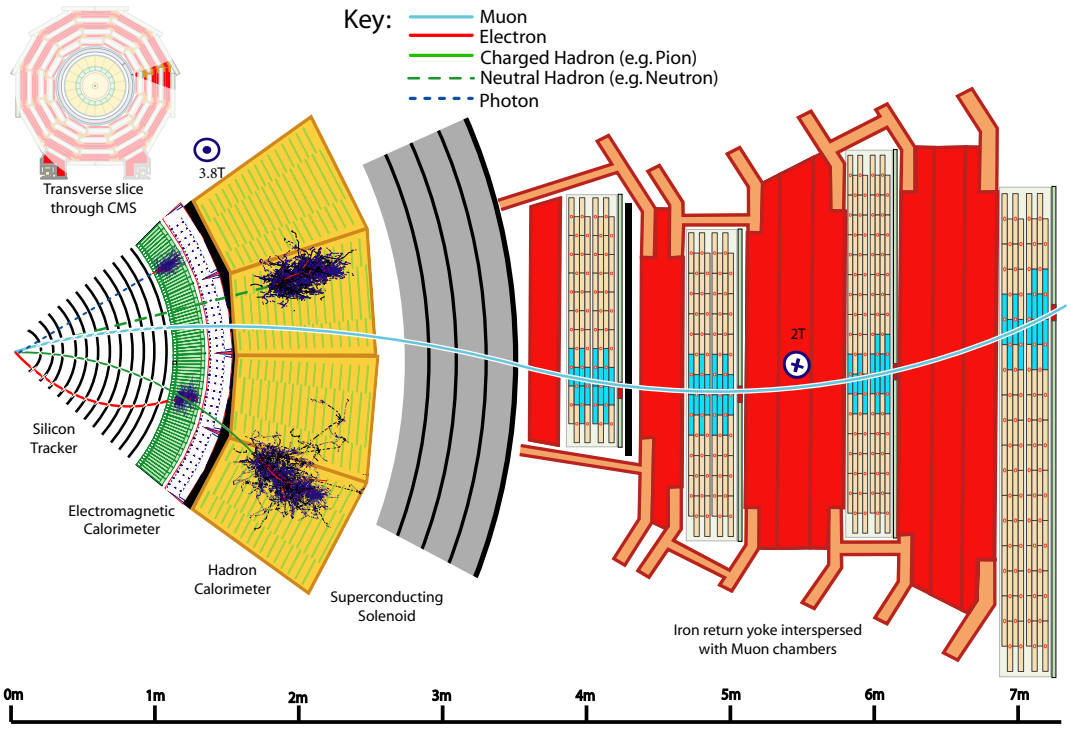

Figure 5.1: Representation of a transversal section of CMS, showing the sub-detectors and how different particles interact with them. Figure from [88].

5.1.1 Tracks

The starting point of the PF algorithm is the reconstruction of charged tracks in the tracking system, which allows to determine the direction and momentum of charged particles that travel through the tracker and bend in the magnetic field provided by the solenoid. The track finding algorithm is based on the Kalman Filter [90] and it consists of the following steps: first a seed is generated from hits compatible with the trajectory of a charged particle, then the trajectory is built using information from all the tracker layers and finally a fit is performed to determine the properties of the charged particle, e.g. its origin, direction and p_T. To reduce the inefficiency in the track reconstruction while keeping the misreconstruction rate low, the track finder is applied iteratively, with selection criteria loosened in the each step. In each iteration, the hits associated to tracks are masked to reduce random association in the following iterations.

As can be been in Fig. 5.2, the iterative tracking increases the efficiency while keeping a smaller misreconstruction rate compared to the single iteration. Moreover, it extends the acceptance down from 1 GeV to 200 MeV in the tracks p_T. Even though the efficiency worsens at high transverse momentum, the energy and angular resolutions of the reconstructed charged hadrons can maintain a small value thanks to the excellent calorimeter

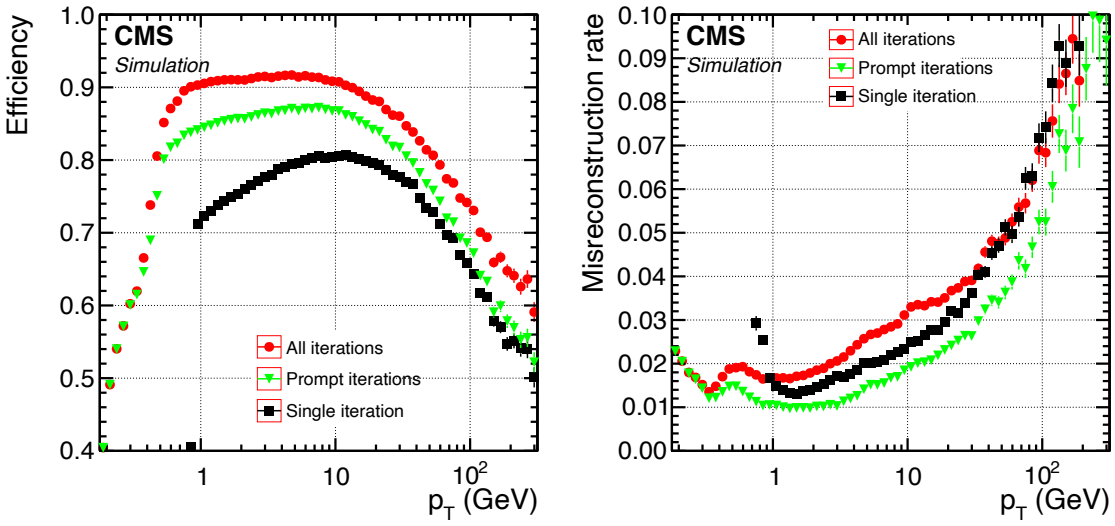

Figure 5.2: Efficiency (left) and misreconstruction rate (right) of the track finder algorithm as a function of the track p_T, for charged hadrons in multijet events without pileup interactions. The black squares indicate the single iteration algorithm, in green and red the iterative method, with the green triangles corresponding to prompt iterations based on seeds with at least one hit in the pixel detector and the red circles to all iterations, including those with displaced seeds. Only tracks with $|\eta| < 2.5$ are taken into account. Figure taken from [88].

resolution achieved at high p_T.

5.1.2 Calorimeter clusters

The reconstruction of calorimeter clusters is useful to find the energy and direction of neutral particles, to separate neutral particles from charged particles, to identify electrons through their energy deposits and the ones from bremsstrahlung photons, and to measure the energy of charged hadrons when tracker information is not sufficient. The clustering algorithm has been developed for PF reconstruction and it is applied separately on ECAL, HCAL − each divided into barrel and endcaps − and on preshower layers. In the HF, no clustering is needed as each cell gives rise to an HF EM or HF HAD cluster. The first step of the clustering is the seed identification: only cells above a given seed threshold are considered, and with energy larger than the energy in the neighbours cells. Then the topological clusters are built from the seeds, by aggregating cells with common sides or corners and with energy larger then the cell threshold, which corresponds to twice the noise level. With an algorithm based on a Gaussian-mixture model, clusters within topological clusters are reconstructed and their position and energy are extracted. Photons and neutral hadrons are reconstructed from calorimeter clusters that are separated from the position

of the tracks of charged particles. Finally, the calibration of the calorimeter response is performed to maximize the probability of reconstructing photons and neutral hadrons, while keeping low the misreconstructed energy excesses, that appears when neutral clusters overlap with charged clusters.

5.1.3 The link algorithm

A particle that travels through the detector can leave signatures in different sub-components and thus can create different PF elements in each of them. The next part of the PF algorithm addresses the linking, which connects the PF elements that originate from the same particle to create the so called PF blocks. The link algorithm tests all pairs of elements, restricted to nearest neighbours in the (η, ϕ) plane, in a given event. In the following, the link between different elements is described. To link calorimeter clusters and tracks, the track's last hit is extrapolated to the calorimeters and its position has to be within the cluster area. A link distance is defined as the distance between the track and the cluster in the (η, ϕ) plane. The cluster-to-cluster link can be established between HCAL and ECAL or between ECAL and preshower clusters. The link is possible if the position of the cluster in the more granular calorimeter is consistent with the envelope of the cluster in the less granular calorimeter. Again, a link distance is defined as the distance between the clusters in the (η, ϕ) plane for HCAL-ECAL or in the (x, y) plane for ECAL-preshower links. Charged tracks originating from a common SV can be linked together. The tracks from the inner tracker can be linked to the tracks in the muon detector. If multiple clusters/tracks are linked together, the pair with smallest distance is chosen.

After all the PF blocks are reconstructed, the particle identification is performed to obtain the final PF candidates. The identification proceeds as follow: first muons are reconstructed and the corresponding PF elements are removed from the list of PF blocks. Then electrons are reconstructed, together with their bremsstrahlung photons and in the same step isolated photons are identified. The tracks and ECAL/preshower clusters are removed from the remaining PF blocks. Finally, the cross-identification of the remaining PF elements is performed to reconstruct charged hadrons, neutral hadrons and non-isolated photons. After the identification of all the particles, a post-processing step is applied to account for possible particle misidentification and misreconstruction. The most important case in which the misidentification plays a role is when a high p_T muon is misreconstructed, leading to an artificially large p_T^{miss} in the event.

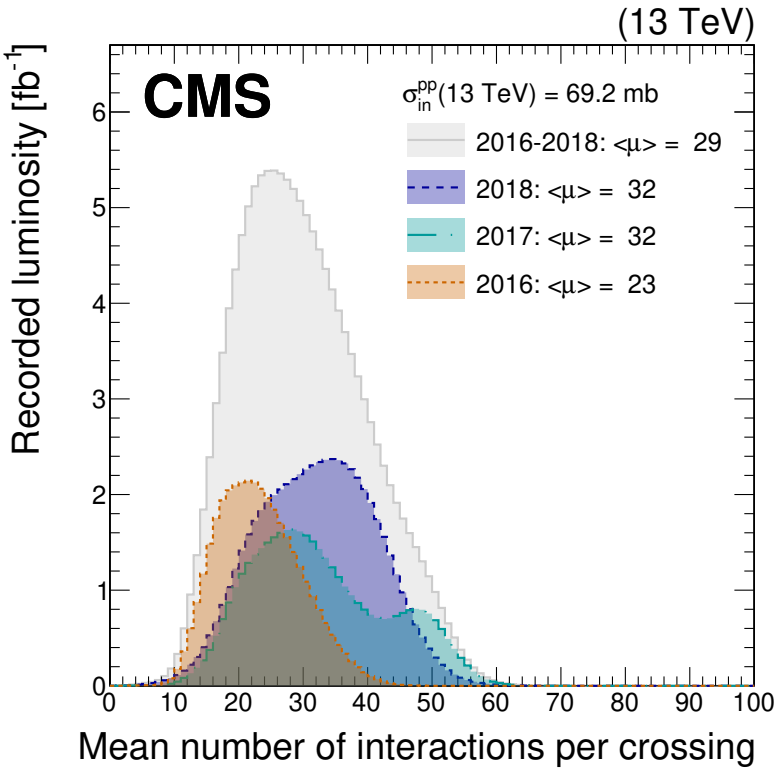

Figure 5.3: The mean number of pp inelastic interactions per bunch crossing in data during the Run 2 of the LHC. Figure taken from [91].

5.2 Vertex Reconstruction

One of the main challenges of the object reconstruction is given by the large number of pp interactions that happen in each bunch crossing, which is a consequence of the high instantaneous luminosity reached by LHC. It is important to identify all the interaction vertices and their position and to distinguish the leading vertex (LV) from the pileup vertices (PU), the additional interactions happening in same bunch crossing (in-time pileup) or in the nearby bunch crossings (out-of-time pileup). The mean number of interactions during Run 2 of the LHC was 29, as is depicted in Fig. 5.3. This number is increasing with increasing luminosity: it is around 60 in Run 3 and will be 140-200 in the High-Luminosity LHC (HL-LHC).

The CMS collaboration has developed an algorithm for vertex reconstruction [92], with the goal of finding the PV in each event and distinguish the leading one from PU vertices. The algorithm is performed in three steps: first the tracks are selected, then the tracks originating from a common vertex are clustered together, and finally the position of each vertex is found by fitting the associated tracks. The tracks that enter the algorithm have to fulfill stringent quality criteria, e.g. minimum number of hits associated to the track

and small distance to the beam spot. The track clustering is done with a *deterministic annealing* (DA) algorithm [93]. With this algorithm it is possible not only to find the track-vertex assignment and the vertex position, but also the number of vertices per event.

The resulting vertices with at least two tracks assigned are fitted with the *adaptive vertex fitter* [94], to find the vertex parameters, as the position x, y, z and covariance matrix, and the fit parameters, like the degrees of freedom. In the fit, to each track a weight w_i between 0 and 1 is assigned, indicating the probability of the track to originate from the given vertex. The number of degrees of freedom of the fit is given by:

$$n_{\text{dof}} = -3 + 2 \sum_{i=1}^{\#\text{tracks}} w_i \tag{5.1}$$

where w_i is the weight of the i-track and the sum is performed on all tracks associated to the vertex. The value of n_{dof} can be used to select the true pp interactions. Each vertex has to satisfy quality criteria in order to be kept for the analysis described in this thesis. The criteria are: $n_{\text{dof}} > 4$, $\sqrt{x^2 + y^2} < 2$ cm and $|z| < 24$ cm. In order to identify the LV in each event, the vertex with highest sum of physics objects p_T is chosen. All the other reconstructed vertices in the event are considered PU vertices.

The object reconstruction in CMS suffers from the effects of particles originating from PU. The two algorithms that are used in CMS for pileup suppression are the Charged Hadron Subtraction (CHS) algorithm [88] and the Pile Up Per Particle Identification (PUPPI) algorithm [95], which are applied on the PF candidates. In CHS, the charged particles that are used in the fit of PU vertices are removed. In PUPPI, on the other hand, it is possible to remove the contribution of charged as well as neutral PU. A detailed description of PU mitigation techniques is presented in Chapter 6.

5.3 Muons

Muons are reconstructed with the PF algorithm using information from the inner tracker and the muon detector system. In particular, the muon spectrometer provides very high identification efficiency, while with the inner tracker it is possible to measure the momentum precisely. Three different types of muons can be reconstructed:

- *standalone muons*: reconstructed from hits in the muon detector. The fitting of the trajectory results in the *standalone muon tracks*.
- *tracker muons*: tracks from inner tracker are extrapolated to the muon system. If there is at least one segments matching the extrapolated trajectory, then the track is identified as a *tracker muon track*.

- *global muons*: they are reconstructed by matching *standalone muon tracks* to the tracks in the inner tracker. All the hits are combined and fitted to obtain a *global muon track*. At $p_T > 200$ GeV, the momentum resolution is improved with respect to *tracker muons*.

Different types of identification (ID) criteria [96] are defined in CMS, with different levels of efficiency and purity in the muon reconstruction. The *Loose muon ID* selects prompt muons from the LV and muons from light and heavy flavour decays. It has the highest efficiency ($> 99\%$), while keeping a low misidentification rate. The *Medium muon ID* aims at identifying prompt muons and muons from heavy flavour decays. It is equivalent to the *Loose muon ID*, but with additional quality criteria on the tracks. Its efficiency is 98%. The *Tight muon ID* is optimized for suppression of muons from in-flight decays and from hadronic punch-through. It has extra muon-quality requirements and its efficiency is $96 - 97\%$. The *Soft muon ID* aims at reconstructing low p_T muons and it has been developed for B physics analysis. Last, the *High momentum muon ID* is optimized for muons with $p_T > 200$ GeV. The requirements are similar to the *Tight muon ID*, with an additional requirement on the relative p_T error, for a proper momentum measurement, and without the fit $\chi^2 < 10$ condition, in order to recover inefficiencies when high p_T muons radiate and produce EM showers and give rise to hits in the muon chambers. The efficiency of this ID is $96 - 98\%$.

The IDs chosen for the muons used in the analysis described in this thesis are presented in detail in the following. The *Tight muon ID* is applied on muons at low transverse momentum ($p_T < 55$ GeV). The corresponding selection cuts are summarized in the following:

- The muon candidate is reconstructed as a *global muon*.
- The normalized χ^2 of the muon global track fit is less than 10.
- At least one muon-chamber hit is included in the global muon track fit.
- Require muon segments in at least two muon stations.
- Its tracker track has transverse impact parameter $d_{xy} < 2$ mm with respect to the LV and a longitudinal distance $d_z < 5$ mm.
- The number of pixel hits is greater than 0.
- Number of tracker layers with hits is greater than 5.

For muons at high transverse momentum ($p_T > 55$ GeV), the *High momentum muon ID* is used. The following selection cuts are applied:

- The muon candidate is reconstructed as a *global muon*.
- At least one muon-chamber hit is included in the global muon track fit or in the TuneP fit.
- Require muon segments in at least two muon stations. If there is only one matched station it must be a tracker muon and satisfy at least one of the following conditions: 0 or 1 expected matched station based on the extrapolation of the inner track, the matched station should not be the first one, or has at least two matched RPC layers.
- The p_T relative error of the muon best track is less than 30%.
- Its tracker track has transverse impact parameter $d_{xy} < 2$ mm with respect to the LV and a longitudinal distance $d_z < 5$ mm.
- The number of pixel hits should be greater than 0.
- Number of tracker layers with hits greater than 5.

Additionally, the low p_T muons have to satisfy the *PFIso* criteria for isolation, that correspond to a requirement on the relative isolation I_{rel}, defined as:

$$I_{rel=}\frac{\sum p_T^{\text{CH from LV}} + \max(0, \sum E_T^{\text{NH}} + \sum E_T^{\gamma} - 0.5 \sum p_T^{\text{CH from PU}})}{p_T^{\mu}} \tag{5.2}$$

where CH refers to charged hadrons and NH to neutral hadrons. The sum runs over the particles in a cone of radius $\Delta R < 0.4$ around the muon. The *Tight* working point of the isolation ID has been chosen, defined by a cut of $I_{rel} < 0.15$, which corresponds to a selection efficiency of 95%.

The efficiencies of ID and isolation criteria are different between data and simulation, for this reason scale factors provided by the CMS collaboration are applied on simulated events. The scale factors are applied on muons as a function of their p_T and η.

5.4 Electrons

Electrons are reconstructed using the energy deposits in the EM calorimeter and the tracks in the inner tracker with the PF algorithm. ECAL clusters are used as electron seeds for isolated, energetic electrons. Since electrons can emit bremsstrahlung photons while traveling through the detector material, and photons can convert into electron-positron pairs, the final result is a shower of particles depositing energy in more that one ECAL cluster. Superclusters (SCs) are then built to gather all the energy of the original electron

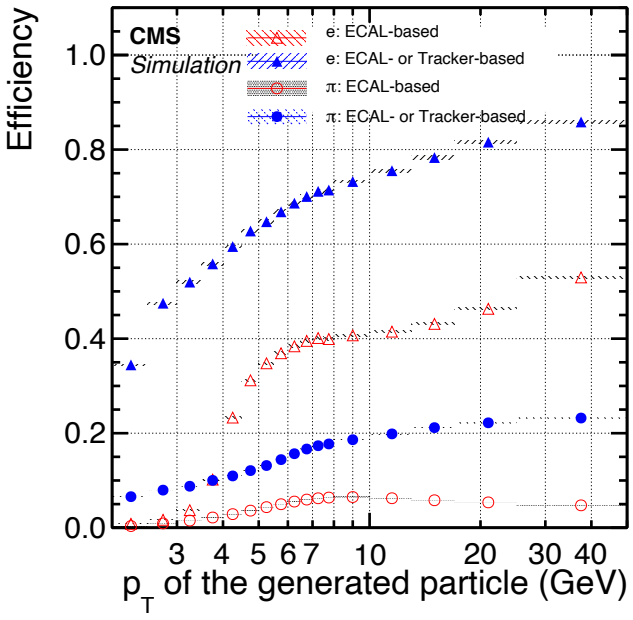

Figure 5.4: Efficiency of electron seeding for electrons (triangles) and pions (circles), as a function of the particle p_T. The ECAL based seeding is shown in red, while the combination of ECAL and track based seeding is shown in blue. Figure taken from [88].

and the bremsstrahlung photons. On the other hand, for electrons in jets, ECAL-based seeding is not optimal, since the presence of other particles in the jet overlapping with the electron would worsen the efficiency. Moreover, the propagation of such ECAL clusters to the tracker hits can cause misreconstruction, as they can be compatible with multiple hits in the inner tracker. For this reason track-based electron seeding is used to identify electrons that are missed by the ECAL-based seeding. In Fig. 5.4 the efficiency of the electron seeding in shown for ECAL based seeds only and for the addition of tracking seeds. With the latter, the efficiency is improved by almost a factor of two and the electron reconstruction is extended down to 2 GeV. Finally, in the case of photon emission, the electron momentum can change, together with the its curvature in the tracker. A dedicated tracking algorithm, the Gaussian-sum filter (GSF), is used to extract the tracking parameters. ECAL clusters, SCs and tracks are given as input to the link algorithm of Particle Flow to reconstruct electrons and distinguish them from photons.

After the identification of electrons, a set of quality criteria is applied in order to use them in analyses. One of the strategies for electron identification is a sequential cut-based selection [97], which includes requirements on different variables. Different working points are defined, based on the values of the selection cuts, and correspond to different levels of efficiency.

For the electrons used in this thesis, a multivariate technique (MVA) [97] developed

by the CMS collaboration is used. The MVA-based identification makes use of a set of variables that are combined to produce a single discriminating output. The discrimination is achieved with a series of BDTs, trained on electrons from DY+jets simulation, in different $|\eta|$ and E_T bins. Moreover, the BDTs are trained with and without isolation variables. In general, the MVA approach has better background rejection for a given signal efficiency compared to the cut-based approach. Three different working points (WPs) are provided for the MVA-based ID: *wp90*, *wp80*, corresponding to 90% and 80% efficiency respectively, and the *wpLoose*, with 98% efficiency, generally used used for vetoing or for multilepton analyses. In this thesis, the MVA ID with *wp80* working point has been chosen. For electrons at low transverse momentum ($p_T < 120$ GeV) the MVA version with isolation is used, while the ID without isolation is used for electrons at high p_T. As for the muons, the differences in the efficiencies of the electron reconstruction and identification in data and simulation have to be taken into account. Dedicated scale factors provided by the CMS collaboration are applied in simulation to electrons as a function of their p_T and $|\eta_{\mathrm{SC}}|$.

5.5 Jets

Quarks and gluons produced in high energy pp collisions do not propagate freely due to colour confinement, but they hadronize, producing sprays of colourless particles (see Sec. 2.2.3). These bunches of particles are grouped together and reconstructed in single objects called *jets*. One of the main challenges in particle reconstruction at the LHC is the correct reconstruction of jets and the identification of the particles from which they originate and their kinematic properties.

5.5.1 Jet clustering algorithms

Among the different jet clustering algorithms that have been developed for collider experiments, the most important and used ones are infrared and collinear (IRC) safe. This requirement imposes the jet to be invariant under the emission of soft radiation or splitting in the jet direction. If a jet is IRC, then it is possible to make comparisons to predictions from perturbative QCD.

One approach to jet clustering is given by *sequential recombination algorithms*, like the k_T [98, 99], the anti-k_T [100] and the Cambridge-Aachen (CA) [101, 102] algorithms. They are all based on the same procedure and differ only on the parameters chosen in the clustering. Given a list of entities, or particles, one starts by defining the distance d_{ij} between the entities i and j, and the distance d_{iB} between the entity i and the beam B. The clustering proceeds as follows:

- the smallest among the distances d_{ij} and d_{iB} is identified
- if the smallest is d_{ij}, then the particles i and j are combined into a new entity
- if the smallest is d_{iB}, then i is defined as *jet* and removed from the list of entities.

The distances are calculated again and the procedure is repeated until there are no particles left. The distances are defined as:

$$d_{ij} = \min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2} \quad \text{and} \quad d_{iB} = k_{ti}^{2p} \tag{5.3}$$

where k_{ti} is the transverse momentum of the entity i, R is the parameter corresponding to the maximum radius of the cone used in the clustering and Δ_{ij}^2 is the distance in the (y, ϕ) plane between the particles i and j:

$$\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2 \tag{5.4}$$

where y_i is the rapidity and ϕ_i the azimuthal angle of i.

The value of the parameter p defines the type of the algorithm: $p = 0$ for the CA algorithm, $p = 1$ for k_T and $p = -1$ for anti-k_T. The difference in p can be interpreted as the different way in which the clustering proceeds. For CA, the particles are clustered based on the geometrical distance: first the particles close in (y, ϕ) are clustered. The k_T algorithm combines soft and collinear particles first, while the anti-k_T the hardest particles first. The latter algorithm produces jets with regular shapes, with a radius of size R centered on the hardest particle, while the other algorithms produce irregular jet shapes. This result can be observed in Fig. 5.5, where the same particles are combined using different clustering parameters.

Another approach to jet clustering has been developed where the jet radius R is not a fixed parameter, but it depends on the p_T of the decaying particle. This method is particularly useful for the identification of the hadronic decays of highly energetic particles. The higher the p_T of the decaying particle, the more collimated its decay products. The jet radius should be large enough to catch all the decay products in a single jet, but it should not be too large, in order not to cluster additional radiation in the jet. In Fig. 5.6 the maximum distance between the decay products of a top quark is shown as a function of the quark p_T.

The *variable R* (VR) algorithm [104] replaces the parameter R of Eq. 5.3 with a radius that adapts dynamically to the p_T of the decaying particle:

$$R_{\text{eff}}(p_T) = \frac{\rho}{p_T} \tag{5.5}$$

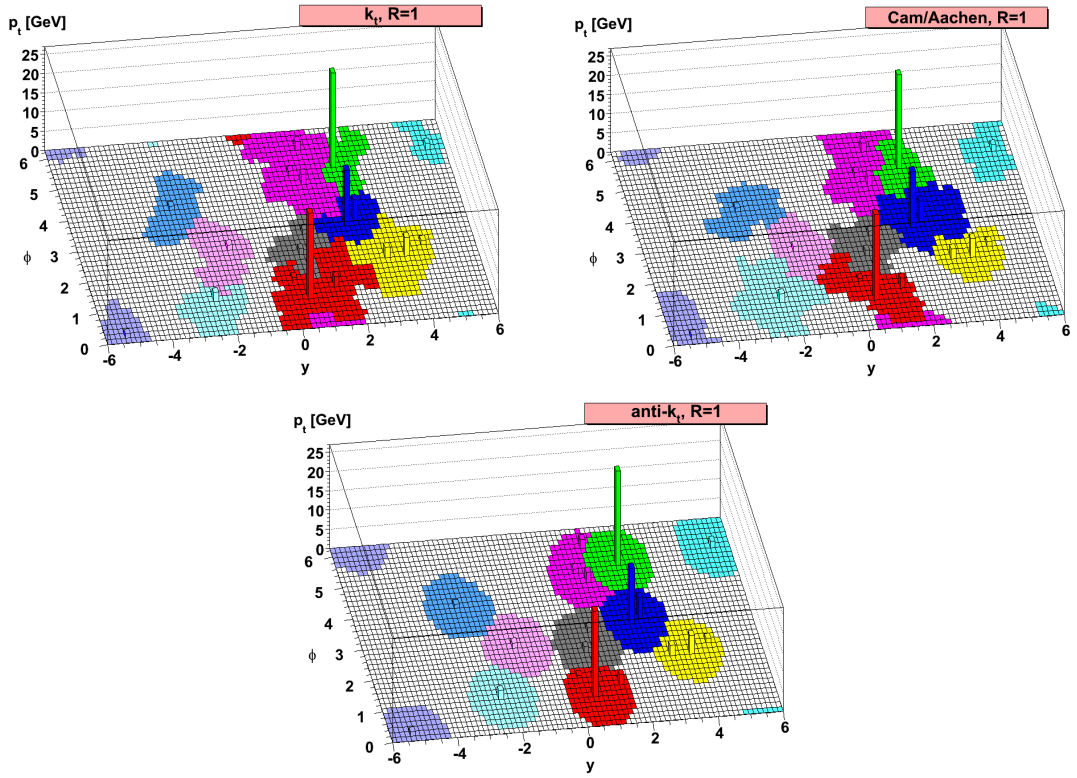

Figure 5.5: The jet shapes for the k_t (upper left), the CA (upper right) and the anti-k_T (lower) clustering algorithms in the (y, ϕ) plane. Figure taken from [100].

where ρ is the scale that determines the slope of R_{eff}. The boundaries on R_{eff} are introduced:

$$\begin{cases} R_{\text{min}} & \text{for } \rho/p_T < R_{\text{min}} \\ R_{\text{max}} & \text{for } \rho/p_T > R_{\text{max}} \\ \rho/p_T & \text{elsewhere.} \end{cases} \tag{5.6}$$

The value of the parameter p determines the clustering procedure of the VR algorithm.

The *Heavy Object Tagger with Variable R* (HOTVR) [105] algorithm is based on VR, but it modifies the clustering procedure by adding a mass jump veto [106]. This criterion assures that no additional radiation is clustered in the jet and simultaneously allows the identification of subjets.

Jet clustering is implemented in CMS with the FastJet [107] framework. The particles that enter the clustering procedure are the PF candidates. Moreover, based on the pileup mitigation technique that is used, different type of jets are reconstructed: CHS jets and PUPPI jets. The standard jet clustering algorithm in CMS is anti-k_T. The radius parameters used are $R = 0.4$ for small-radius jets and $R = 0.8$ for large-radius jets and the corresponding jets are referred to as AK4 and AK8 jets.

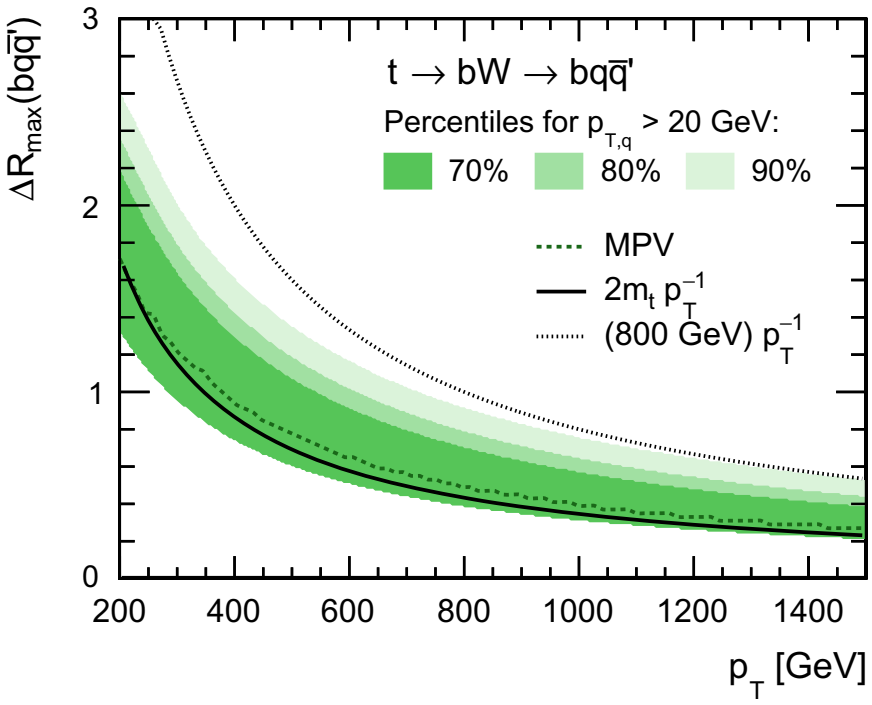

Figure 5.6: The maximum distance between the quarks from the hadronic decay of a top quark as a function of the top quark p_T. Figure from [103].

5.5.2 Jet calibration

The energy of jets at detector-level needs to be calibrated to match the true value of particle-level jets, which are clustered from stable (lifetime $c\tau > 1$ cm) and visible (non-neutrinos) particles. The differences between detector-level and particle-level are due to the effects of pileup, detector response and noise. In CMS jet energy corrections (JECs) are derived and applied in a factorized approach [108–110], as is represented in Fig. 5.7, to calibrate the jet energy scale (JES) and the jet energy resolution (JER). In each step different techniques are used to mitigate different effects and they can be derived from simulated events or from data.

The first corrections are the L1 offsets and they correct the jet energy scale due to the contribution of pileup. They are applied both on data and simulation and are parametrized as a function of the jet p_T, η, area A and energy density ρ in the event. These corrections are applied on CHS jets, while they are not needed for PUPPI jets, since the PU contribution in PUPPI jets is already removed by the PUPPI algorithm itself.

The second step is given by the L2 relative and L3 absolute corrections which correct for detector effects. In this step the jet response, which is the ratio of the jet p_T at detector-level over particle-level, is corrected to be close to unity. Then L2L3 residual corrections are applied on data as a function of η and p_T to correct for residual differences in the detector

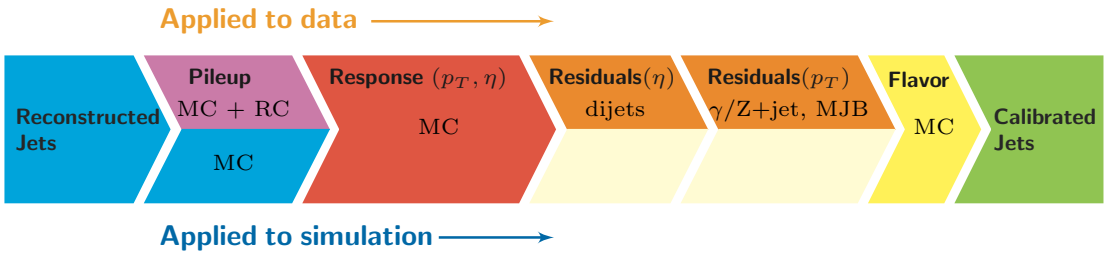

Figure 5.7: Representation of the factorized JECs that are applied to simulation and to data. Figure from [108].

response between data and simulation. Finally, optional L5 flavour corrections are derived to mitigate the difference in response of jets originating from particles of different flavour.

After the calibration of the JES, the JER is corrected. The JER is defined as the ratio of the width of the jet response over the mean, and it is typically broader in data than in simulation. Correction factors are derived to smear the JER in simulation and match it to the one obtained in data.

5.6 Jet tagging

In the analysis presented in this thesis searches for new heavy particles decaying to top quark pairs are performed. The top quark decays as $t \to Wb$, with $W \to q\bar{q}'$ or $W \to l\nu$. If the new decaying particle is at the TeV scale, the top quarks acquire a large Lorentz boost and their decay products are collimated and can be reconstructed in single, large-radius jets. The identification (tagging) of the particle from which a jet originates is then a crucial part of the analysis. In the analysis presented in this thesis, both the tagging of jets originating from b quarks and from the hadronic decay of top quarks are performed. An important challenge is given by the high rate of QCD multijet production at the LHC. The most important tool for jet identification is jet substructure: a complete overview of jet substructure can be found in Ref. [103].

5.6.1 Identification of b quark jets

The identification of b-initiated quarks (b-tagging) is of particular interest for this thesis. Jets originating from heavy-flavour quarks (b and c quarks) present particular characteristics that can be used to distinguish them from the jets initiated from light-flavour quarks and gluons. The most important feature is the large lifetime of B hadrons, that is of the order of ~ 1.5 ps. The B hadrons can thus travel for a few mm*s* before decaying and this leads to a displacement of the decay products of the B hadrons with respect to the PV. The tracks

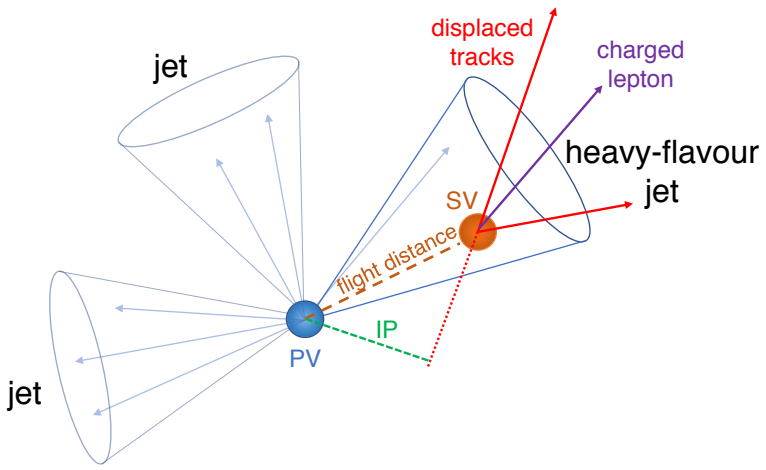

Figure 5.8: Representation of a heavy-flavour jet with a SV that is displaced from the PV. Figure from [111].

from their decays can be used to reconstruct a SV [111] thanks to the high resolution of the CMS inner tracker. An illustration of a heavy-flavour jet with a SV is shown in Fig. 5.8.

In the search presented in this thesis the algorithm used to identify b jets is DeepJet (or DeepFlavour) [112], that has been developed for Run 2. DeepJet is a multi-classifier that can distinguish jets originating from b, c, light-flavour quarks and gluons. The algorithm is based on a neural network that uses as input the information of charged and neutral jet constituents, SVs and event observables. The medium working point of the algorithm has been chosen, that corresponds to 1% misidentification rate for light-flavour quarks.

5.6.2 Identification of t quark jets

Jet substructure techniques are exploited to tag the hadronic decays of top quarks and distinguish them from the QCD multijet background. The most important substructure variables for jet tagging are presented in the following. Based on them, more sophisticated algorithms have been developed. The machine-learning based top tagger used in this thesis, DeepAK8, is then described in detail.

Substructure observables

A sensitive variable to identify t quark jets is the invariant jet mass m_{jet}, which is defined as the invariant mass of all the jet constituents. The jet mass is sensitive to the mass of the decaying particle, providing good discrimination between top jets and QCD jet production. However, m_{jet} can include the additional contributions from pileup, underlying event and initial state radiation, that affect in particular large-radius jets. To remove soft and wide

angle radiation and thus obtain a more precise prediction of m_{jet}, *grooming* techniques are applied. The most used grooming algorithm is *soft-drop* (SD) [113]. Given a jet j with radius R_0, soft-drop reclusters the jet constituents using the CA algorithm and then declusters the jet by un-doing the last step of CA. This results in two subjets j_1 and j_2 for which the following condition is checked:

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{cut} \left(\frac{\Delta R_{12}}{R_0} \right)^{\beta} \tag{5.7}$$

where p_{Ti} are the transverse momenta of the subjets, R_{12} is their distance in the (y, ϕ) plane and the parameters z_{cut} and β are the soft threshold and the angular exponent, respectively. If the condition is satisfied, then j is identified as the soft-drop jet. Otherwise, the subjet with larger p_T is defined as j and the procedure is repeated. If j can not be declustered anymore, then either j is removed (*tagging mode*) or j is identified as the final SD jet (*grooming mode*).

In CMS the *grooming mode* is used and the two soft-drop parameters are set to $z_{cut} = 0.1$ and $\beta = 0$. The mass of a jet after the application of the algorithm is referred to as soft-drop mass m_{SD}.

Another powerful tool to discriminate jets is based on the jet energy distribution. In jets originating from the hadronic decays of top quarks, the energy is deposited in three regions (three-prong structure). On the other hand, in jets originating from QCD, one expects the jet to deposit the energy in one region (one-prong structure). The N-subjettiness [114] algorithm is designed to predict how many subjets N are (at most) in a jet. The N-subjettiness variable τ_N is defined as:

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \, \min(\Delta R_{1,k}, \Delta R_{2,k}, ..., \Delta R_{N,k}) \tag{5.8}$$

where the sum runs over the k jet constituent particles, $p_{T,k}$ are their transverse momenta and $\Delta R_{J,k}$ is the distance between a subjet J and a constituent particle k. The factor d_0 is given by:

$$d_0 = \sum_k p_{T,k} R_0 \tag{5.9}$$

where R_0 is jet radius. From these definitions, it is clear that a low value of τ_N is obtained for jets consistent with N subjets. Even more discrimination power is given by the ratios of N-subjettinesses: $\tau_{MN} = \tau_M / \tau_N$, where $M > N$. In particular, to distinguish top jets from QCD jets the ratio $\tau_{32} = \tau_3 / \tau_2$ is used.

HOTVR

While jet substructure variables are powerful tools to tag top-initiated jets, more sophisticated techniques have been developed for the identification of hadronic decays of heavy objects. An example is HOTVR, a jet clustering and identification algorithm introduced in Sec. 5.5.1. The HOTVR algorithm not only clusters jets with a variable radius, but it allows to identify the hadronic decays of boosted particles. The HOTVR parameters can be set to tag the jets originating from top quarks. Namely, the requirements are:

- the number of subjets $N_{subjet} \geq 3$
- the fractional p_T of the leading subjet with respect to the jet $f(p_T) < 0.8$
- the jet mass $140 < m_{jet} < 220$ GeV
- the N-subjettiness ratio $\tau_{32} < 0.56$
- the minimum pairwise mass of two subjets $m_{ij}^{min} > 50$ GeV.

Thanks to the variable jet radius, it is possible to achieve a good tagging performance both at low and high transverse momentum.

DeepAK8

Novel top tagging techniques based on machine learning have become more and more used in CMS. The tagging algorithm used in this thesis is the DeepAK8 tagger [115]. DeepAK8 is a multi-classifier able to identify the hadronic decays of t quarks, Higgs and vector bosons and their various decay channels (e.g. $\mathrm{Z} \to \mathrm{b\bar{b}}$, $\mathrm{Z} \to \mathrm{c\bar{c}}$, $\mathrm{Z} \to \mathrm{q\bar{q}}$). It takes as inputs a "particle list": up to 100 jet constituents per jet ordered by p_T, each with 42 particle properties, like p_T, change, energy, angular separation. Moreover, a "secondary vertex list" is included and it consists of up to 7 SVs ordered by 2D impact parameter significance (S_{IP2D}), each with 15 properties, like displacement, kinematics and quality criteria. A one-dimensional convolutional neural network (CNN) is applied to each of the lists and the outputs are combined in a fully connected neural network. In this way it is possible to exploit not only all the input information, but also the correlations. The architecture of DeepAK8 is shown in Fig. 5.9. In order to reduce the p_T bias, the jet distributions are reweighted to flat distributions.

Mass decorrelation

In many machine learning based taggers the mass of a jet is learned by the algorithm, even if it is not used as input. This can lead to *mass-sculpting* effects: after the application of

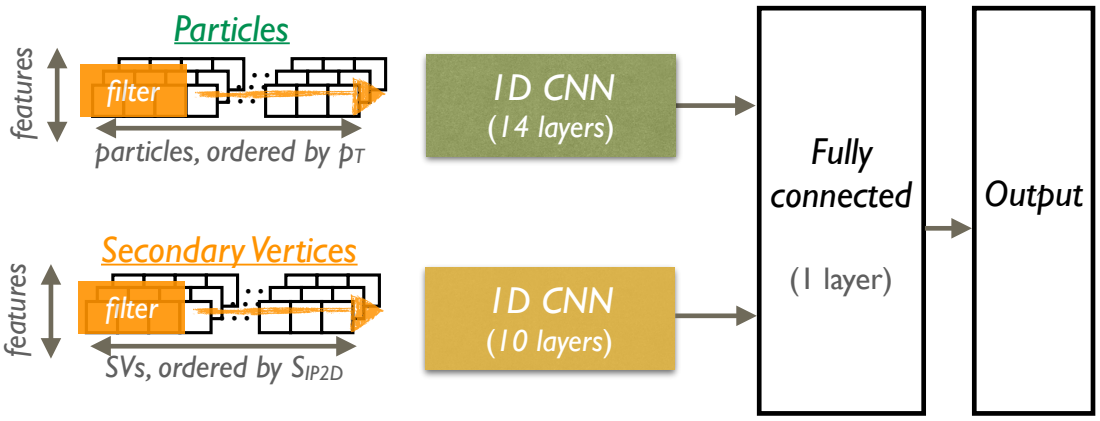

Figure 5.9: Representation of the architecture of the DeepAK8 tagger. Figure from [115].

the tagger, the jet mass distribution of background samples becomes similar to the jet mass distribution of the signal. This is an undesirable feature, especially if the jet mass itself is used to separate signal from background, or if the mass of the signal is unknown. To avoid mass-sculpting, *mass decorrelated* (MD) versions of the taggers have been developed. The mass-sculpting effect can be seen in Fig. 5.10 for various taggers used in CMS.

The mass decorrelated version of the DeepAK8 tagger, DeepAK8-MD, has been developed using an adversarial training approach. The jet distributions are reweighted to be flat in m_{SD}. The architecture is modified by adding a mass prediction network, that predicts the mass of jets from the features extracted by the CNNs. The accuracy of the mass prediction in then included as a penalty term to prevent the algorithm from being correlated with the mass. The architecture of DeepAK8-MD is shown in Fig. 5.11.

Tagging performance

A common tool to evaluate the performance of tagging algorithms is to calculate signal and background efficiencies, ϵ_S and ϵ_B, in simulation. They are defined as:

$$\epsilon_S = \frac{N_S^{tagged}}{N_S^{tot}} \quad \text{and} \quad \epsilon_B = \frac{N_B^{tagged}}{N_B^{tot}} \tag{5.10}$$

where N_S^{tot} (N_B^{tot}) is the total number of signal (background) events and N_S^{tagged} (N_B^{tagged}) represents the number of signal (background) events after the application of the tagger. In the case of top tagging, the signal consists of hadronically decaying top quarks and the background is the QCD multijet process. The efficiency is evaluated in terms of the receiver operating characteristic (ROC) curve. The performance of different taggers for the benchmark top tagging is presented in Fig. 5.12 (top). The best discriminating

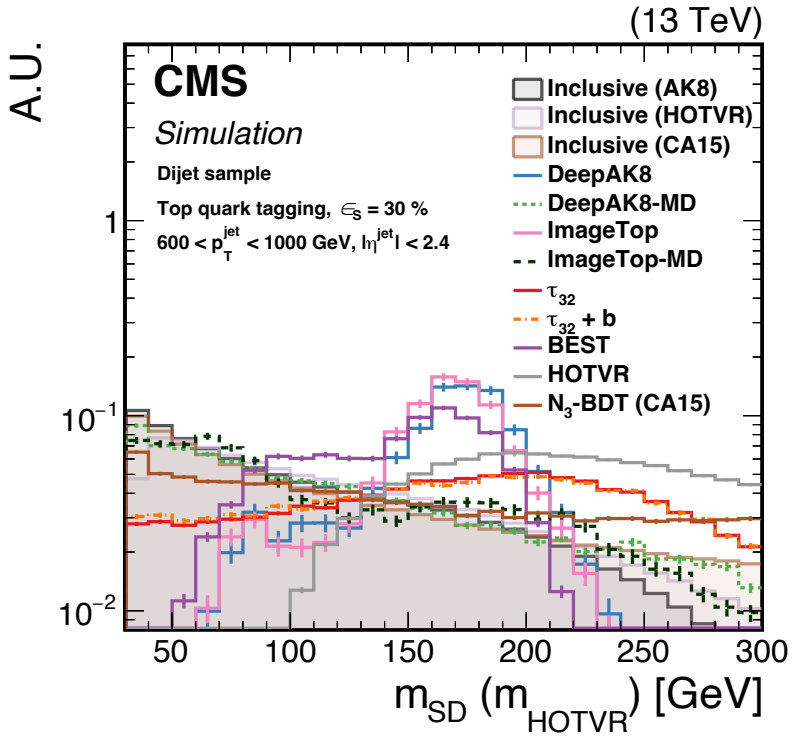

Figure 5.10: The m_{SD} distribution for a QCD sample before and after the application of different top taggers. Figure from [115].

power at high transverse momentum is obtained with the DeepAK8 algorithm, while the mass-decorrelated version yields only a slight loss in performance.

It is also important to check the robustness of the taggers to changes in jet kinematics, for example the efficiency ϵ_S or the misidentification rate ϵ_B as a function of the transverse momentum of the generated particle. In Fig. 5.12 (bottom) the efficiency and the misidentification rate are shown for various top tagging algorithms. The efficiency as a function of p_T shows that the HOTVR tagger has a stable efficiency for the whole p_T range, as expected, while other taggers that use fixed radius jets, like DeepAK8, have a lower efficiency at low p_T, that increases until ~ 600 GeV, where it becomes the highest. The same is true for the misidentification rate, which is constant for HOTVR. However, a lower misidentification rate is obtained with DeepAK8.

5.7 Missing transverse energy

Neutrinos are the only SM particles that can travel through the CMS detector without interacting with its materials, being thus undetected. If new particles, neutral and weakly interacting, were produced at the LHC, they would also be unobserved. Nevertheless, if such particles are produced together with other particles that leave a signature in the

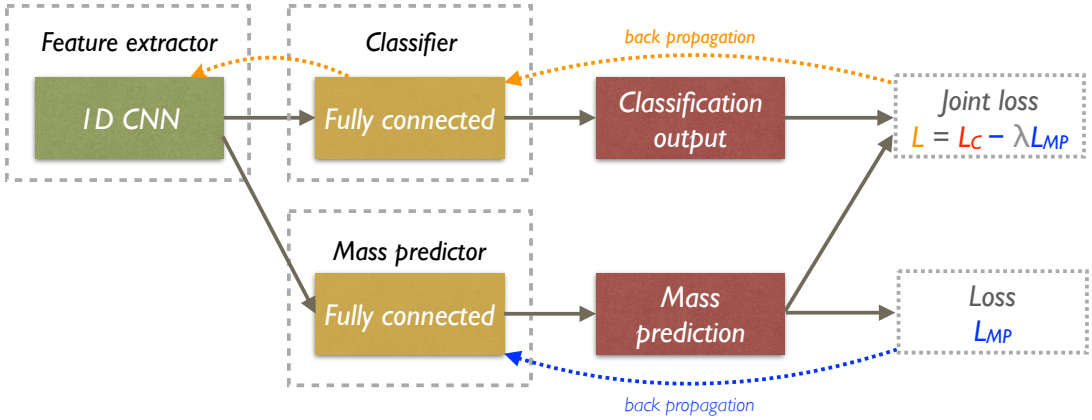

Figure 5.11: Representation of the architecture of the DeepAK8-MD tagger. Figure from [115].

detectors, the imbalance in the transverse momentum, denoted as $\vec{p}_T^{\,\text{miss}}$, with magnitude p_T^{miss}, can be used to infer their presence. Often it is referred to as missing transverse energy (MET) as well. The precise measurement of p_T^{miss} is therefore crucial for SM measurements and searches that target final states containing neutrinos or neutral, weakly interacting particles.

An important distinction has to be made on the reconstructed p_T^{miss}: there is a *genuine* p_T^{miss} that comes from the production of neutrinos, and there is a p_T^{miss} component that can arise from misreconstruction and miscalibration of physics objects, detector effects, noise or pileup.

In CMS the $\vec{p}_T^{\,\text{miss}}$ reconstruction is based on the PF algorithm and it is defined as the negative vector p_T sum of all the PF candidates in each event [116]. It is referred to as PF p_T^{miss} and used in most of the CMS analysis based on Run 2 data. A second algorithm has been developed based on the PUPPI algorithm. The PUPPI $\vec{p}_T^{\,\text{miss}}$ is defined as the negative vector p_T sum of all the PF candidates weighted with their PUPPI weight. In this thesis, the PUPPI p_T^{miss} is used. The advantage of using the PUPPI algorithm in the reconstruction of the missing transverse momentum is the reduction of the pileup dependence.

To correct for the object miscalibration, the jet energy corrections calculated for AK4 jets are propagated to p_T^{miss}. The result is the "Type-I" corrected p_T^{miss}. More details about the PF and PUPPI MET will be presented in Subsec. 6.2.5.

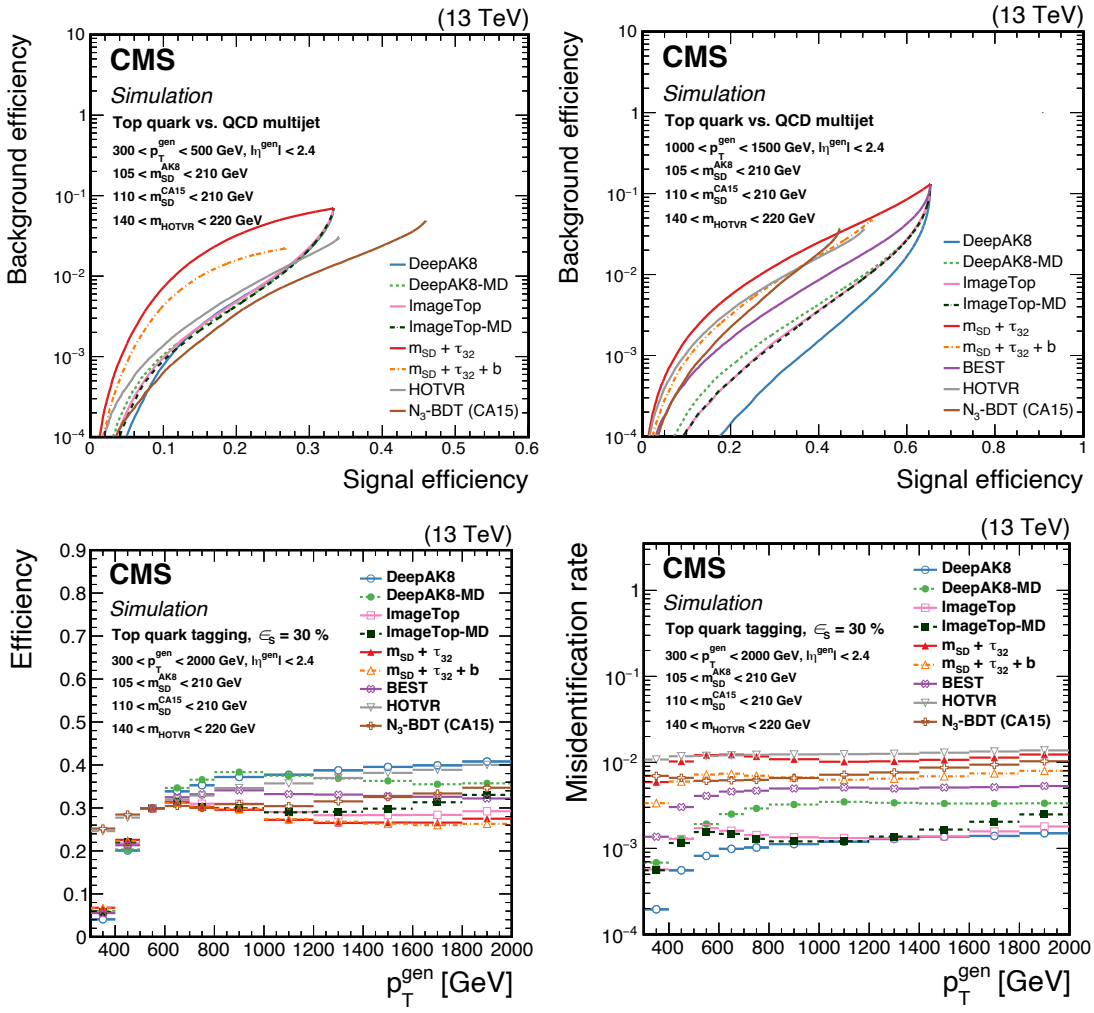

Figure 5.12: Upper: the top tagging performance comparison in terms of the ROC curve for generated particles with low p_T (right) and high p_T (left). Lower: the efficiency (left) and the misidentification rate (right) as a function of the generated particle p_T for top tagging. Figures from [115].

Chapter 6

Pileup mitigation

With the high instantaneous luminosity reached by the LHC, more and more data can be collected, giving access to rare SM phenomena and potentially to new physics. However, more instantaneous luminosity comes with an increase in the number of pp collisions that happen in each bunch crossing. It becomes crucial to correctly identify the main interaction in each bunch crossing and to mitigate the effects of the additional collisions, the pileup (PU). In this Chapter the main pileup mitigation techniques used in CMS will presented: the PUPPI algorithm and the CHS algorithm. The PUPPI algorithm is used in CMS since Run 2 and it has been the default PU mitigation technique for large-radius jets, while CHS was the standard algorithm for small-radius jets. However, given its excellent performance, PUPPI became the default algorithm in CMS for both small- and large-radius jets for Run 3 and beyond. After describing the PUPPI algorithm in detail, the new tune developed for the *Ultra Legacy* (UL) reconstruction of Run 2, PUPPI v15, will be introduced and its performance in comparison to the default version of PUPPI and to CHS will be presented. The results here presented have been published in the Detector Performance Summary [2].

6.1 Pileup mitigation in CMS

An important challenge at hadron colliders consists in the event reconstruction in a high pileup scenario. During Run 2 of the LHC the mean number of pp interactions per bunch crossing was 29, with peaks up to 60. In Run 3 there is a mean of 60 interactions per bunch crossing and it is expected to be in the range 140-200 in HL-LHC, requiring an excellent handle of the pileup contribution in order to perform physics analyses. The vertex reconstruction in CMS has been presented in Sec. 5.2.

Two main algorithms are used in CMS to mitigate the effects of pileup: the Charged Hadron Subtraction (CHS) algorithm [88] and the Pile Up Per Particle Identification

(PUPPI) algorithm [95]. Both algorithms are applied on the PF candidates before the jet clustering procedure to reduce the contribution of particles originating from pileup interactions in the reconstruction of jets and p_T^{miss}.

In CHS the tracking information is used to identify the vertices from which the charged particles originate. The charged hadrons that are associated to one of the PU vertices are removed from the list of PF candidates used to reconstruct the jets. While this procedure allows to remove the charge PU contribution, it presents some drawbacks. First, it only allows to remove charged PU, but nothing is done to mitigate the contribution of neutral PU. Second, the procedure is valid only within the tracker volume, where it is possible to identify the tracks of charged particles. Additional jet-area-based energy corrections have to be applied on the reconstructed jets to mitigate the contribution of neutral PU and of charged PU outside the tracker volume.

Another type of jets present in pp collisions are the *PU jets*. These are jets that are made entirely from PU: they can be QCD jets, i.e. jets originating from a soft pp interaction, or stochastic jets, which are formed when particles originating from various vertices are grouped together and reconstructed as a jet. The PU jet ID [117] is applied on top of CHS jets to remove PU jets. The PU jet ID is a BDT-based algorithm that is applied on low-p_T jets and discriminates jets made entirely of PU.

To overcome these issues and to mitigate the effects of PU for both charged and neutral particles, not only within the tracker volume, the PUPPI algorithm has been developed. In PUPPI each particle in an event is given a weight w, that corresponds to the probability of the particle to originate from the LV ($w = 1$) or from a PU vertex ($w = 0$). The weight is assigned considering the charged PU distribution in the event and the particles surrounding the particle of interest. The weight assignment is based on the properties of the particle:

- for charged particles from LV $w = 1$,
- for charged particles from PU $w = 0$,
- for charged particles without vertex association $w = 1$ if $|d_z| < 0.3$ cm, $w = 0$ otherwise, where d_Z is the distance of closest approach to the LV along the z-axis,
- for neutral particles $0 < w < 1$.

The particle's four-momentum is then scaled by the PUPPI weight w. A sketch of the LV and PU vertices and the different types of particles is shown in Fig. 6.1.

The algorithm used to calculate the weight for neutral particles is described in the following. First a variable α_i is defined for each neural particle i:

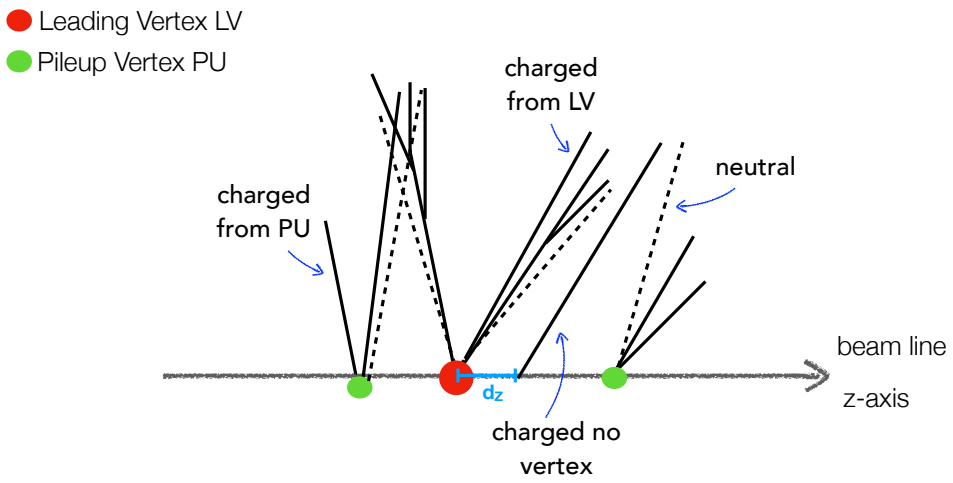

Figure 6.1: A sketch depicting the leading vertex (LV) and two PU vertices (PU) in a simplified event. The charged particles are represented with solid lines, while the neutral particles with dashed lines. A particle can be either associated to the LV, to one of the PU vertices or it can have no vertex association. The d_z is the distance of closest approach to the LV along the z-axis.

$$\alpha_i = \log \sum_{j \neq i, \Delta R_{ij} < R_0} \left(\frac{p_{T,j}}{\Delta R_{ij}} \right)^2 \tag{6.1}$$

where the sum runs over the particles j in a cone of radius $R_0 = 0.4$ around the particle i, $p_{T,j}$ is the transverse momentum of the particle j and ΔR_{ij} is the distance between particles i and j in the (η, ϕ) plane. Within the tracker volume $|\eta| < 2.5$, the particles j are the charged particles originating from the LV, while for $|\eta| > 2.5$ all the reconstructed particles are used. If there are no particles in the cone, the default value of $\alpha = 0$ is used. From this definition, it is clear that the value of α_i is high for high energetic and collinear particles, while it is low for soft, wide-angle particles. Then it is assumed that the neutral PU distribution in an event is the same as the charged PU distribution. The charged PU is used to calculate the α distribution of PU, from which the median $\bar{\alpha}_{PU}$ and the root-mean-square RMS_{PU} values are extracted. For the region of the detector with no tracking information ($|\eta| > 2.5$), it is not possible to calculate $\bar{\alpha}_{PU}$ and RMS_{PU}. Instead, the values from the tracker region are used and they are multiplied by a transfer factor, with values reported in Table 6.1. A signed χ^2_i is calculated for each neutral particle:

$$\chi^2_i = \frac{(\alpha_i - \bar{\alpha}_{PU})|\alpha_i - \bar{\alpha}_{PU}|}{\text{RMS}^2_{PU}}. \tag{6.2}$$

Finally a weight w_i is calculated using the cumulative distribution function of the χ^2 with

one degree of freedom:

$$w_i = F_{\chi^2,\mathrm{NDF}=1}(\chi_i^2). \tag{6.3}$$

The distributions of the α variable, the signed χ^2 and the PUPPI weight are shown in Fig. 6.2.

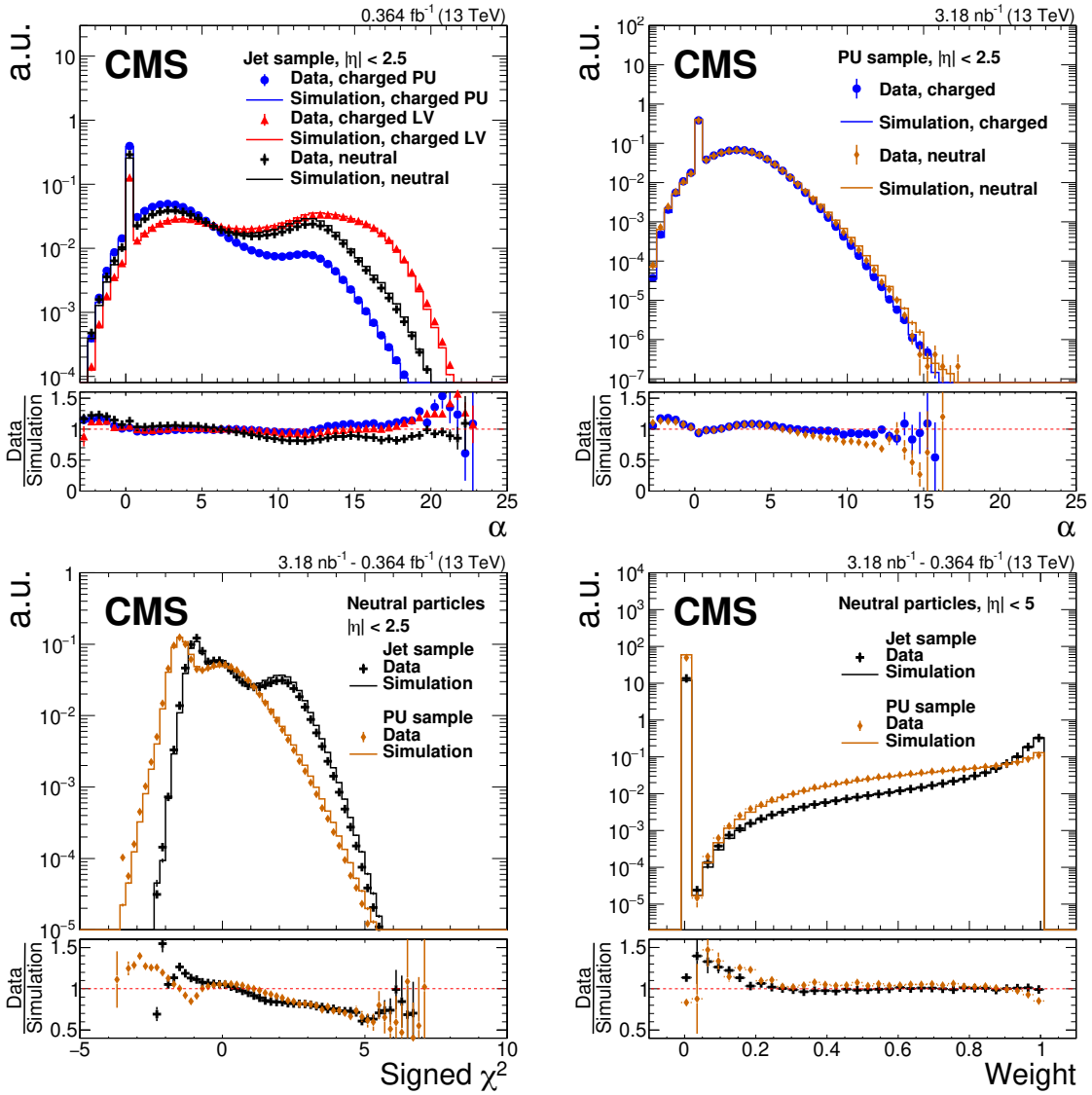

Figure 6.2: The distributions of the α variable in the jet sample (upper left) and in the PU sample (upper right) and the distribution of the signed χ^2 (lower left) and the PUPPI weight (lower right) for neutral particles. Data are represented with solid markers and simulation with solid lines. Figure from [91].

Only particles with a weighted p_T above a threshold are kept: $p_T \cdot w > (A + B \cdot N_{\mathrm{vertices}})$. The threshold is calculated as a function of the number of reconstructed vertices N_{vertices}. The parameters A and B have been optimized to increase the jet response and to improve

the $\vec{p}_T^{\,\text{miss}}$ resolution and their values are shown in Table 6.1. The performance of the

$\lvert\eta\rvert$	A [GeV]	B [GeV]	TF $\bar{\alpha}_{PU}$	TF RMS_{PU}
0-2.5	0.2	0.015	1	1
2.5-3	2.0	0.13	0.9	1.2
3-5	2.0	0.13	0.75	0.95

Table 6.1: The values of the tunable parameters A and B and the transfer factors (TF) used in the PUPPI algorithm in different η regions.

PUPPI algorithm and its comparison to CHS have been extensively studied and reported in Ref. [91] using 2016 data. Given the very good performance of PUPPI, it has been decided to make it the default pileup mitigation algorithm in CMS for Run 3 and beyond. Nonetheless, further improvements have been made to the PUPPI algorithm to obtain an even better performance, especially in terms of jet energy resolution at high p_T. The improved version of PUPPI, called PUPPI v15, is presented in Ref. [2]. It has been already implemented in the latest re-reconstruction of Run 2 data, the UL reconstruction, and it is used in the analysis presented in this thesis. The developments implemented in PUPPI v15 and its performance will be discussed in the following Section. The main difference with respect to the previous version of the algorithm is given by the improvement in the track-vertex association.

6.2 PUPPI v15

The new tune of PUPPI, denoted as PUPPI v15, has been implemented in the UL reconstruction of Run 2 data and it is the starting version of the algorithm in Run 3. The changes of PUPPI v15 with respect to the default version are:

- improved track-vertex association for charged particles that are not used in any vertex fit. The PUPPI weight is assigned based on the kinematic properties of each particle as:
  - for $p_T > 20$ GeV: $w = 1$,
  - for $p_T < 20$ GeV and $|\eta| > 2.4$: $w = 1$ if $|d_z| < 0.3$ cm, $w = 0$ otherwise,
  - for $p_T < 20$ GeV and $|\eta| < 2.4$: calculate w as for neutral particles.

    The separation of particles at $|\eta| = 2.4$ is motivated by the fact that in CMS only particles below this threshold are used in the vertex fitting procedure.

- Recover vertex splitting or track stealing by PU vertices. If the particle comes from the 1^{st} or 2^{nd} PU vertex and $|d_z| < 0.2$ cm: $w = 1$.

- Protection assuring high weights for high p_T neutrals. The protection works by checking if the weight w_i of a neutral particle lies below or above a threshold, if it is below, then it is increased to the corresponding value of the threshold. For particles with p_T between 20 and 200 GeV the threshold is: $w_i < p_T \cdot \frac{1}{200-20} - \frac{20}{200-20}$; for $p_T > 200$ GeV the weight is always set to 1.

- Tuning of the parameters A and B for the weighted p_T protection. The new parameters are shown in Table 6.2.

$\|\eta\|$	A [GeV]	B [GeV]
0-2.5	0.2	0.015
2.5-3	1.7	0.08
3-5	2.0	0.08

Table 6.2: The values of the tunable parameters A and B for the PUPPI v15 algorithm in different η regions.

The validation of PUPPI v15 is performed by comparing the new tune to the default version of PUPPI, called PUPPI v11a, and to the CHS algorithm using the UL reconstruction of the 2017 dataset, for which the mean number of interactions is $< \mu >= 32$. It has been studied in terms of jet energy resolution, jet reconstruction efficiency and purity, substructure variables and missing transverse momentum.

6.2.1 Jet energy resolution

The jet energy resolution (JER) is obtained from the jet energy response, which is the ratio of the reconstruction-level jet p_T over the particle-level jet p_T. The response distribution can be considered gaussian to a good approximation, and it is fitted with a gaussian function in an iterative procedure. The fit is repeated three times, each time setting the range to $[\mu - 1.5\sigma,\ \mu + 1.5\sigma]$, where μ and σ are the mean and the width of the previous fit, respectively. The JER is defined as the ratio σ/μ. The JER as a function of the particle-level jet p_T is shown in Fig. 6.3 in six different η bins. The jets are clustered with the anti-k_T algorithm with a radius of 0.4 and the jet energy corrections are applied. For $|\eta| < 2.5$, PUPPI v15 outperforms PUPPI v11a and it is as good as, or better than, CHS in the whole p_T spectrum. For $|\eta| > 2.5$, all the algorithms reach the same level of JER, due to the lack of tracking information in this region.

In Fig. 6.4 the JER is presented as a function of the number of interactions, for two η regions and considering low and high p_T jets. The jets with PUPPI v15 show the best JER and the best stability as a function of PU.

6.2.2 Jet reconstruction efficiency and purity

Another important measure for pileup mitigation techniques is the jet reconstruction efficiency and purity, which tell us how good the algorithm is in reconstructing all the LV jets in an event, and if it is able to reconstructs only LV jets and not PU jets. The efficiency is defined as the fraction of particle-level jets matched to reconstruction-level jets, over particle-level jets. The matching is performed by checking that the distance in the (η, ϕ) plane is $\Delta R < 0.4$. The thresholds on the transverse momentum are $p_T > 20$ GeV for reconstruction-level jets and $p_T > 30$ GeV for particle-level jets. The purity is defined as the fraction of reconstruction-level jets matched to particle-level jets, over reconstruction-level jets. The thresholds on the transverse momentum for the purity are $p_T > 30$ GeV for reconstruction-level jets and $p_T > 20$ GeV for particle-level jets. Different p_T thresholds are set on reconstruction- and generator-level jets in order to be independent on the effects of the jet energy corrections. The efficiency and purity are presented in Fig. 6.5 in different η bins. The efficiency of PUPPI v15 is increased with respect to the previous version of PUPPI, while there is a slight loss in purity for $|\eta| > 2.5$. This is due to the lower requirement on the weighted p_T protection, which allows more PU particles to be clustered in the jets.

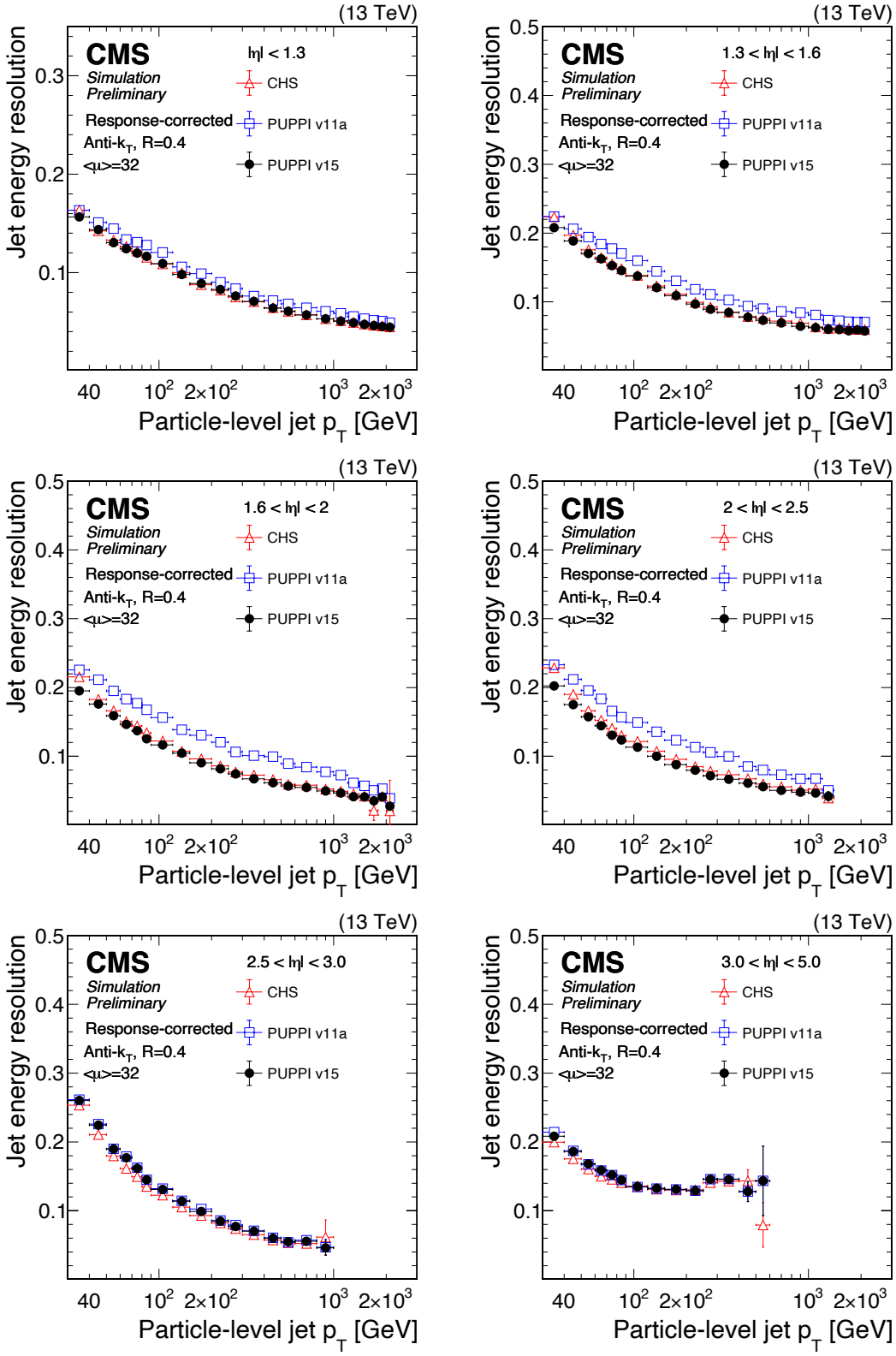

Figure 6.3: The JER as a function of the particle-level jet p_T for jets with PUPPI v15 (black circles), PUPPI v11a (blue squares) and CHS (red triangles) in six η regions. The QCD multijet simulated sample is used. The jets have a radius of 0.4 and the jet energy corrections are applied. Published in [2].

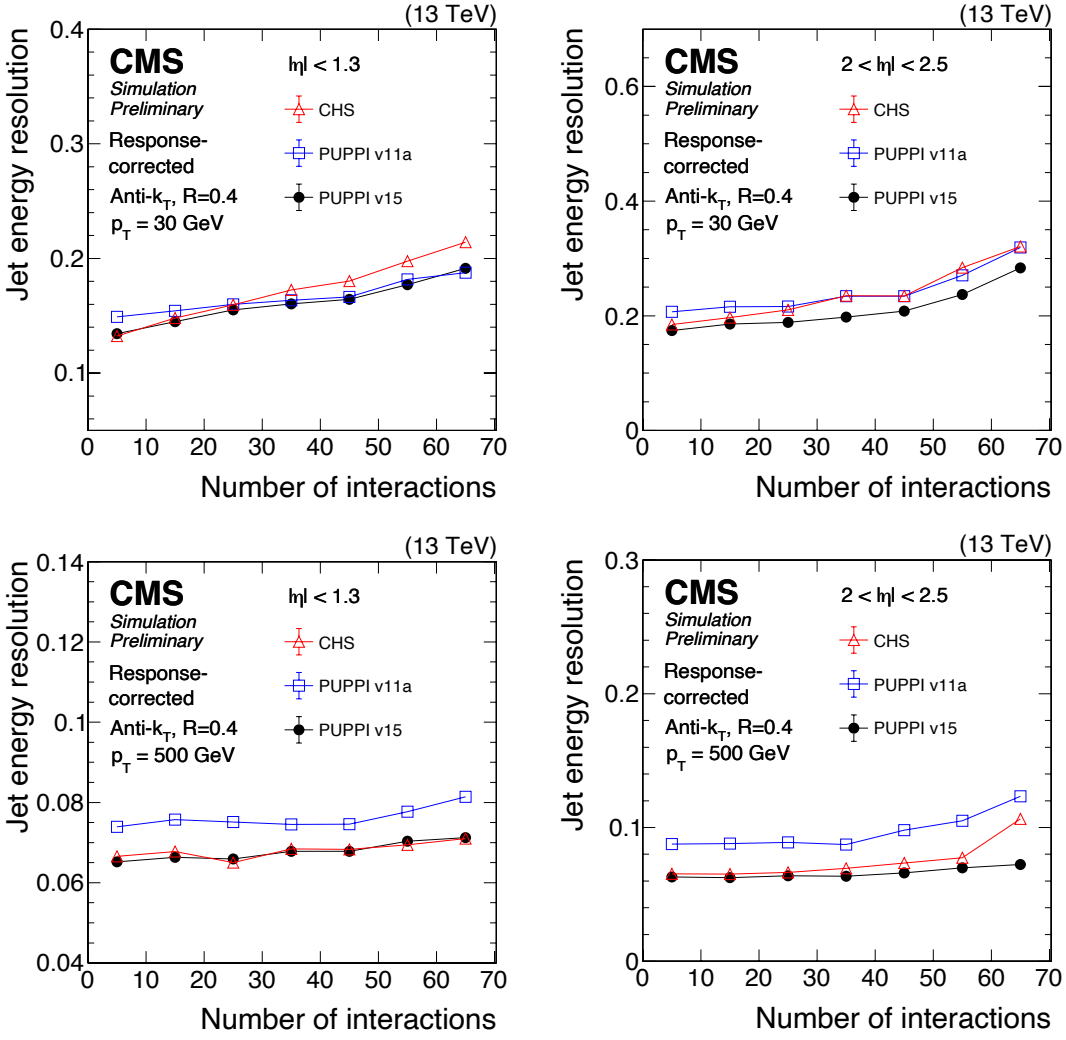

Figure 6.4: The JER as a function of the number of interactions for jets with PUPPI v15 (black circles), PUPPI v11a (blue squares) and CHS (red triangles), on the left-hand side for the region $|\eta| < 1.3$ and on the right-hand side for $2 < |\eta| < 2.5$. Two exemplary p_T of the reconstructed jets have been chosen: low p_T jets, with $p_T = 30$ GeV (upper) and high p_T jets, with $p_T = 500$ GeV (lower). The QCD multijet simulated sample is used. The jets have a radius of 0.4 and the jet energy corrections are applied. Published in [2].

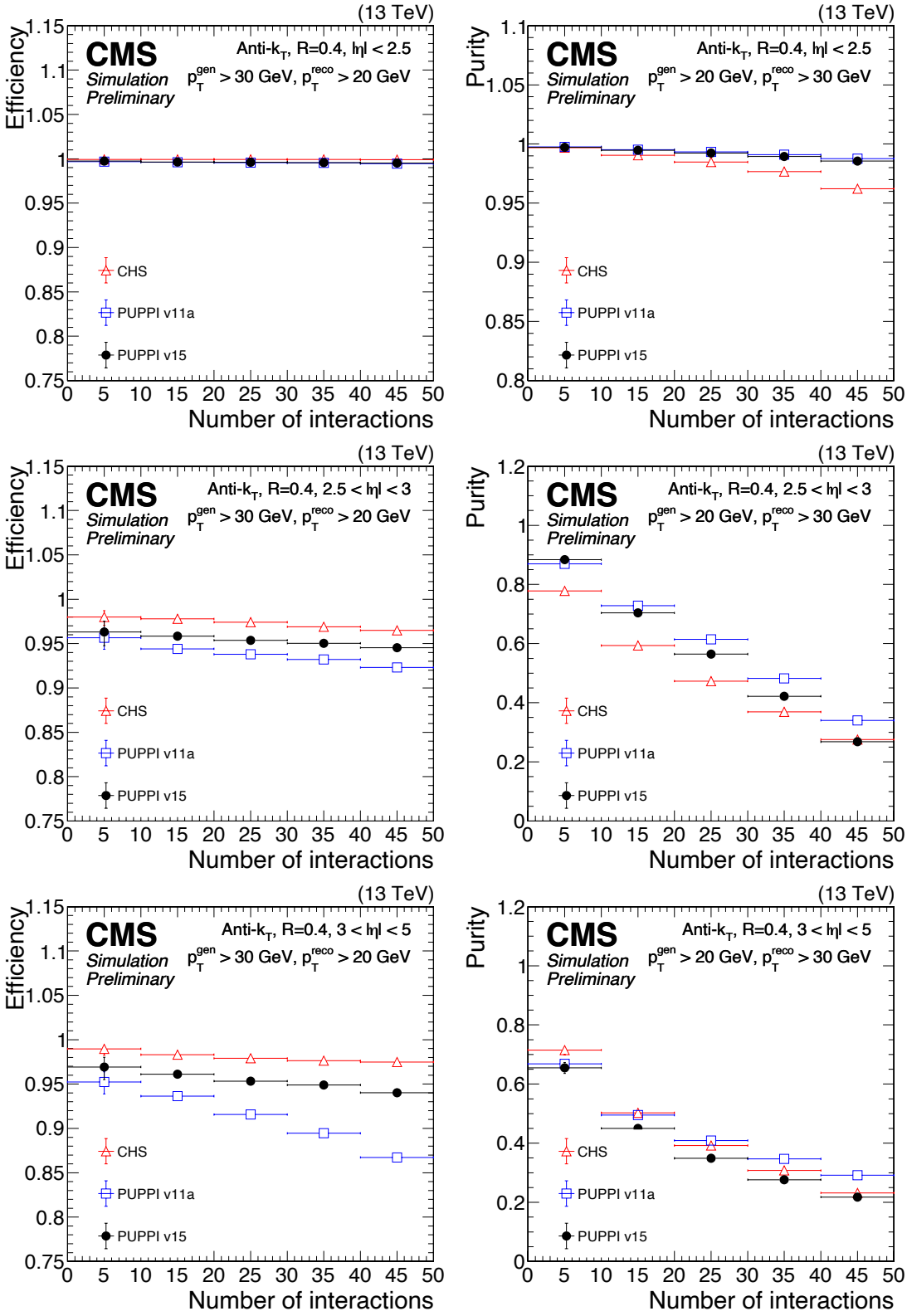

Figure 6.5: The jet reconstruction efficiency (right) and purity (left) as a function of the number of interactions in three different η bins. The Z+jets simulated sample is used. The jets are reconstructed with the anti-k_T algorithm with a radius of 0.4 and the jet energy corrections are applied. Published in [2].

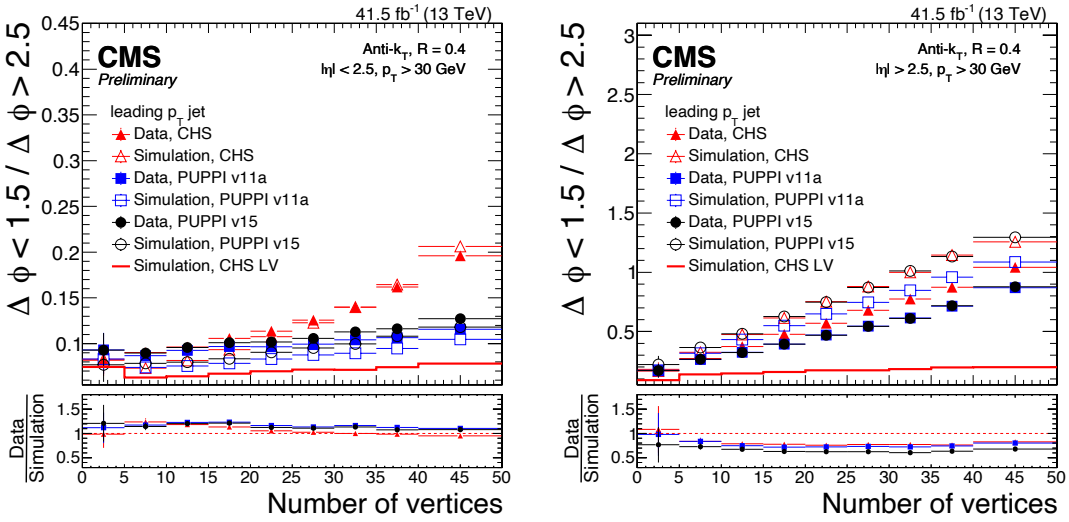

Figure 6.6: The jet rate in the PU-enriched region over the jet rate in the LV-enriched region as a function of the number of vertices. The PUPPI v15 jets are represented with black circles, PUPPI v11a with blue squares and CHS with red triangles, for simulation (open markers) and data (filled markers). The rate of CHS jets matched to a particle-level jet is shown with the red solid line. The Z+jets events are used. The jets are reconstructed with the anti-k_T algorithm with a radius of 0.4 and the jet energy corrections are applied. In the lower panel the data over simulation ratio is shown. Published in [2].

6.2.3 Pileup jet rate

In order to determine the efficiency in the rejection of PU jets, the PU jet rate is evaluated. For this study Z+jets events are used in both data and simulation, where the Z boson decays into a pair of muons. The jets that overlap with one of the muons ($\Delta R < 0.4$) are removed. It is possible to identify the jets originating from the LV as the ones recoiling against the Z boson, while all the additional jets originate most probably from PU. The separation is made based on the distance in the azimuth between the Z boson and the leading jet in p_T. If $\Delta\phi$(Z, jet) < 1.5 the event is PU-enriched, while if $\Delta\phi$(Z, jet) > 2.5 the event is LV-enriched. In Fig. 6.6 the rate of PU-enriched events over LV-enriched events is presented as a function of the number of vertices. As a reference, the rate of CHS jets matched to a particle-level jet is shown. For $|\eta| < 2.5$ the PUPPI algorithm shows a stable performance against PU, while for $|\eta| > 2.5$ all the algorithms have an increased PU jet rate with increasing PU.

6.2.4 Jet substructure variables

Jet substructure variables are crucial for analysis using boosted objets, as described in the previous chapter. The effects of PU in large-radius jets can degrade substructure observables, as the jet soft-drop mass or the N-subjettiness variables. The performance

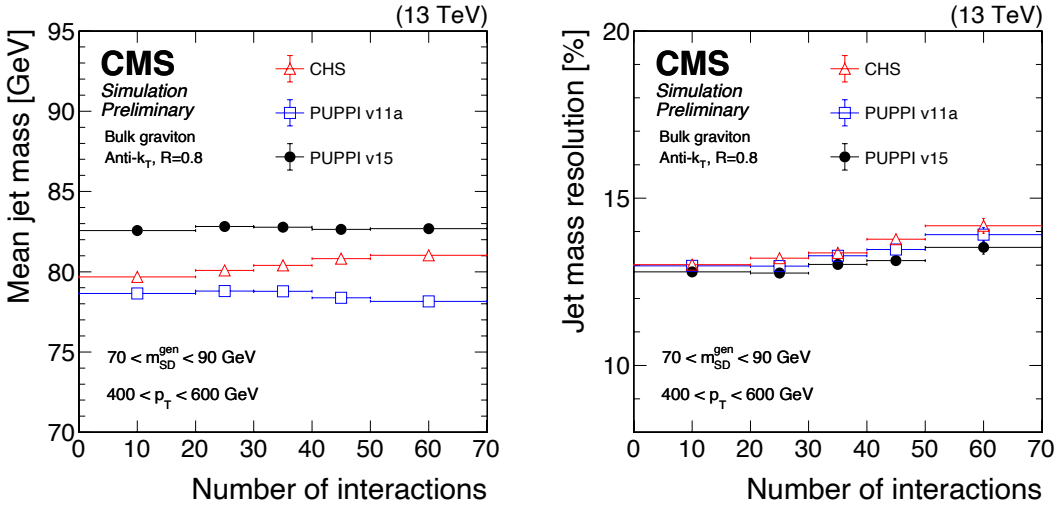

Figure 6.7: The mean jet mass (left) and the jet mass resolution (right) for AK8 jets as a function of the number of interactions. The jets are reconstructed with the PUPPI v15 (black circles), PUPPI v11a (blue squares) and CHS (red triangles) algorithms. The jets are selected with $70 < m_{\mathrm{SD}} < 90$ GeV and $400 < p_T < 600$ GeV in a MC sample of a bulk graviton decaying to a pair of scalar bosons, that subsequently decay to a pair of quarks. Published in [2].

of the pileup mitigation algorithms on jet substructure has been studied using simulated events of a bulk graviton decaying to a pair of scalar bosons, that subsequently decay into a pair of quarks. The scalar bosons are reconstructed with AK8 jets, with a selection on the jet soft-drop mass $70 < m_{\mathrm{SD}} < 90$ GeV, to mimic the W boson. The mean jet mass and the jet mass resolution as a function of the number of interactions are presented in Fig. 6.7. The mean and the width of the jet mass distribution are extracted with a gaussian fit in an iterative procedure, and the jet mass resolution is obtained as the ratio of the width over the mean. The mean jet mass for the PUPPI algorithms is stable against PU, while for CHS the mass increases with increasing PU. The absolute value of the mass is different among the algorithms because the AK8 jets in the MC sample used are not calibrated. In terms of the jet mass resolution the best performance is obtained for PUPPI v15. The median value of the N-subjettiness variable τ_{21} is shown in Fig. 6.8. Again, both the versions of PUPPI are stable with increasing PU, which is not the case for CHS.

6.2.5 Missing transverse energy performance

The $\vec{p}_T^{\,\mathrm{miss}}$ reconstruction in CMS is based on the PF algorithm and it is defined as the negative vector p_T sum of all the PF candidates in the event. Since the CHS algorithm works only on charged particles in the tracker volume, it is not applied in the calculation of p_T^{miss}. The PUPPI algorithm, on the other hand, can be used. The PUPPI $\vec{p}_T^{\,\mathrm{miss}}$ is

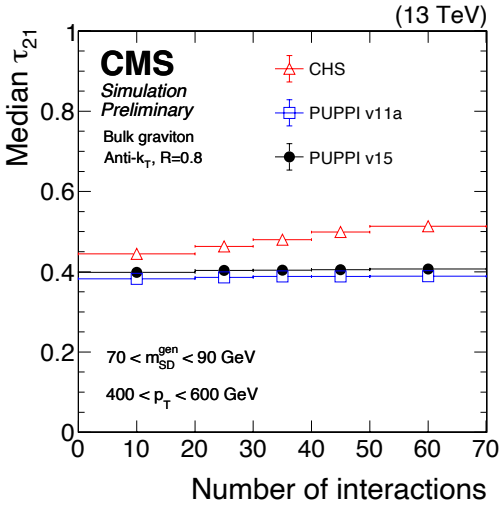

Figure 6.8: The median τ_{21} for AK8 jets as a function of the number of interactions. The jets are reconstructed with the PUPPI v15 (black circles), PUPPI v11a (blue squares) and CHS (red triangles) algorithms. The jets are selected with $70 < m_{\mathrm{SD}} < 90$ GeV and $400 < p_T < 600$ GeV in a MC sample of a bulk graviton decaying to a pair of scalar bosons, that subsequently decay to a pair of quarks. Published in [2].

defined as the negative vector p_T sum of all the PF candidates weighted with their PUPPI weight. The PUPPI algorithm is modified for the calculation of p_T^{miss}: all the charged leptons are considered prompt and excluded from the α calculation, and the photons with $p_T > 20$ GeV and $|\eta| < 2.5$ and all the leptons are given a weight of 1. This modification is required as PU particles surrounding prompt leptons would be given high weights, thus creating a PU dependence.

Since any miscalibration of objects has an impact on the p_T^{miss} reconstruction, it is important to calibrate p_T^{miss} properly. The *Type-I* correction is given by the propagation of the jet energy corrections in the following way:

$$\vec{p}_T^{\,\mathrm{miss}} = \vec{p}_T^{\,\mathrm{miss,\ raw}} - \sum_{jets} (\vec{p}_{T,jet}^{\,\mathrm{corr}} - \vec{p}_{T,jet}) \tag{6.4}$$

where $\vec{p}_T^{\,\mathrm{miss,\ raw}}$ is the uncorrected $\vec{p}_T^{\,\mathrm{miss}}$ and the sum runs on all AK4 jets with $p_T > 15$ GeV. It is applied both the PF and PUPPI p_T^{miss}, where for the latter PUPPI jets are used.

The p_T^{miss} performance is measured in events with no genuine p_T^{miss}, where all the momentum imbalance is given by object miscalibration. Events with Z bosons and jets, where the Z boson decays into a pair of muons, are used, since no genuine p_T^{miss} is expected. The hadronic recoil is defined as $\vec{u} = -\vec{p}_T^{\,\mathrm{miss}} - \vec{p}_T(\mathrm{Z})$ and is illustrated in Fig. 6.9.

The parallel $u_{\parallel}$ and perpendicular $u_{\perp}$ components of the hadronic recoil are used to check the resolution and response of p_T^{miss}. In particular, $u_{\parallel}$ is sensitive to the jet energy

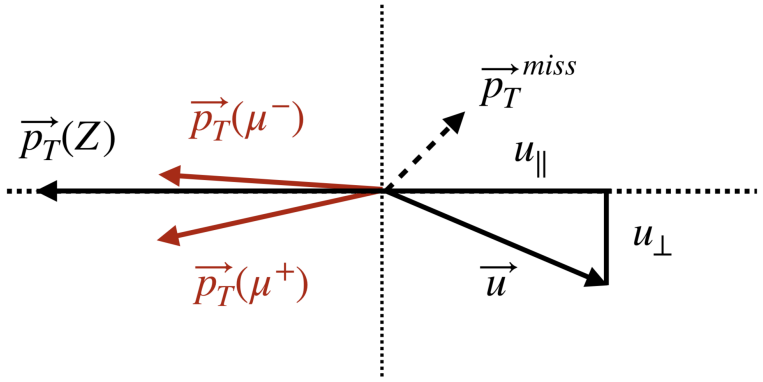

Figure 6.9: An illustration of the Z $\rightarrow \mu\mu$ kinematics. The $\vec{u}$ vector indicates the hadronic recoil, where $u_\parallel$ and $u_\perp$ are its parallel and perpendicular component, respectively.

resolution, while $u_\perp$ is sensitive to the PU contribution in the event. The performance is shown in Fig. 6.10 for PF and PUPPI p_T^{miss}. The PUPPI v15 p_T^{miss} shows the highest response and the resolution is improved with respect to PF p_T^{miss} and is more stable against pileup.

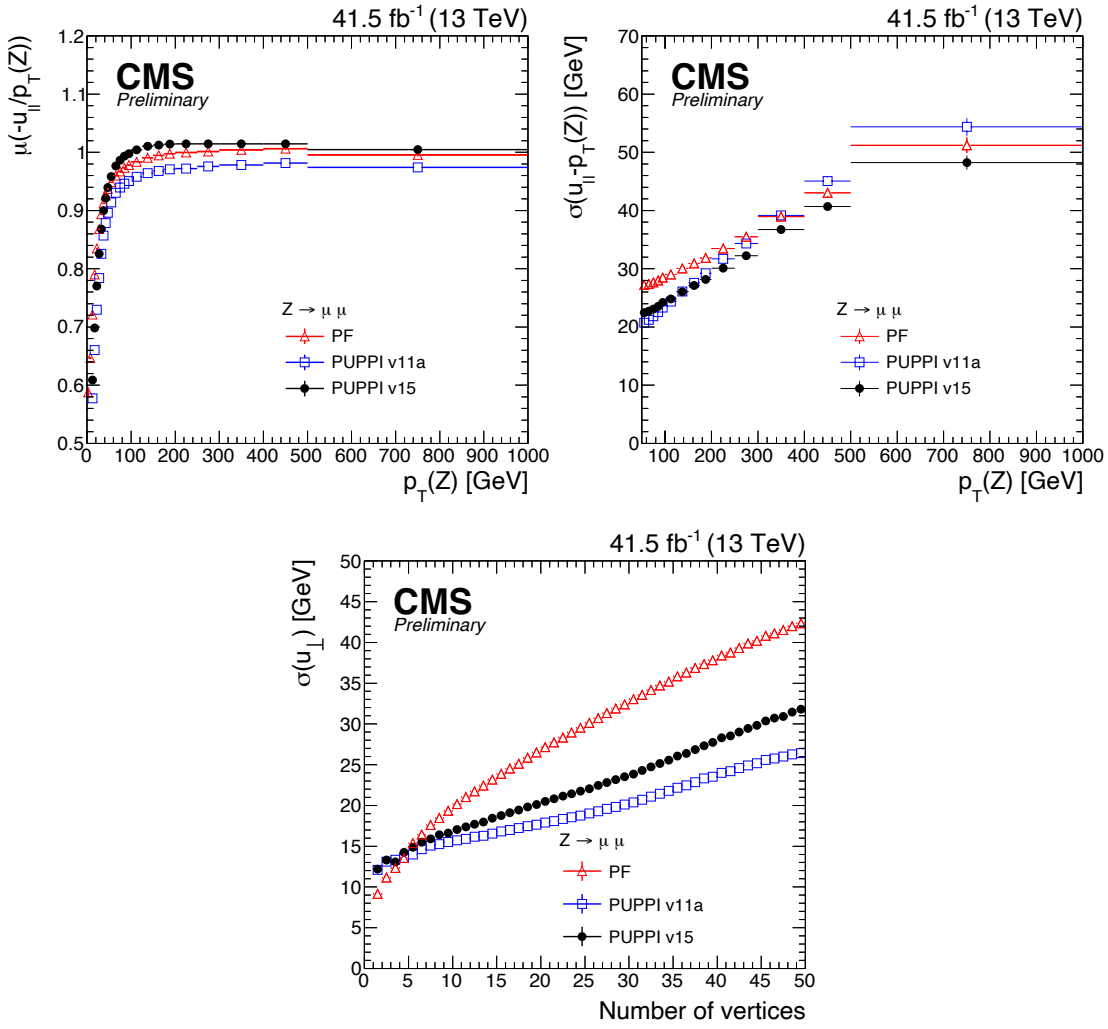

Figure 6.10: The hadronic recoil response (upper left) and resolution (upper right) for $u_{\parallel}$ as a function of the Z boson transverse momentum and the hadronic recoil resolution (lower) for $u_{\perp}$ as a function of the number of vertices. The performance is shown for PF (red triangles), PUPPI v11 (blue squares) and PUPPI v15 (black circles) p_T^{miss} using Z+jets data. Published in [2].

Chapter 7

Search for new particles decaying to top quark pairs

In this Chapter, a search for new resonances decaying to $t\bar{t}$ in the lepton+jets final state is presented. This search uses data collected at $\sqrt{s} = 13$ TeV during 2016-2018 with the CMS experiment at the LHC, corresponding to an integrated luminosity of 138 fb^{-1}. The analysis focuses on new resonances at the TeV scale, including models of Z' bosons, Kaluza-Klein gluons and additional heavy Higgs bosons, and targets both the resolved and the boosted regimes.

First, the analysis overview is presented (Sec. 7.1), followed by the datasets and simulated samples used in the search (Sec. 7.2). In Sec. 7.3 the event selection is described and in Sec. 7.4 the reconstruction of the $t\bar{t}$ system is presented. The neural network approach for event classification is introduced in Sec. 7.5, followed by the definition of the search variables and the final event categorization in Sec. 7.6. The systematic uncertainties and the statistical interpretation are described in Sections 7.7 and 7.8, respectively. Finally, the results are summarized in Sec. 7.9 and the conclusions close the Chapter.

7.1 Analysis overview

This analysis looks for new particles decaying to top quark-antiquark pairs ($X \to t\bar{t}$). In particular, the lepton+jets final state is studied, where one top decays hadronically ($t \to W^+b \to q\bar{q}'b$) and the other leptonically ($\bar{t} \to W^-\bar{b} \to l^-\bar{\nu}\bar{b}$)[1]. Final states with one electron or one muon, jets and missing transverse momentum are considered. The two complementary final states, the all hadronic and the dileptonic channels, are analyzed by different teams and the combination of the results is foreseen. The orthogonality of the

[1]The corresponding charge conjugated processes is included implicitly.

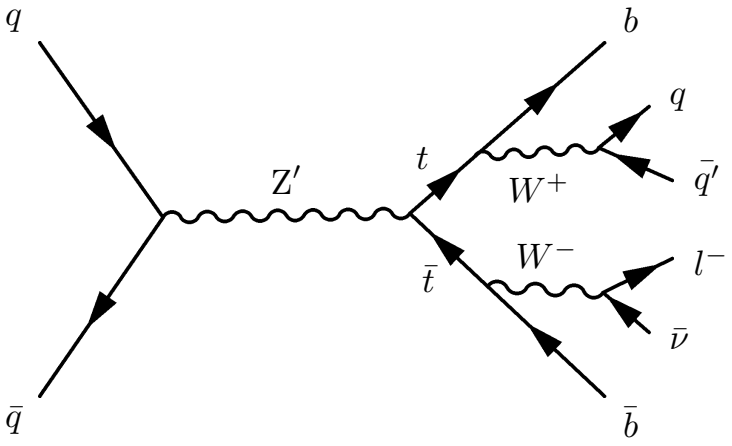

Figure 7.1: The leading order Feynman diagram of the production of a Z′ boson and its decay to $t\bar{t}$ in the lepton+jets final state.

searches is assured by the event selections applied in the analyses.

A model-independent search is performed, testing different models of physics BSM which include the coupling of new heavy particles to t quarks, as presented in Chapter 3. The new particles probed are Z′ bosons, Kaluza-Klein gluons g_{KK} and additional scalar H or pseudoscalar A Higgs bosons. As an example, the Feynman diagram of the production and consequent decay of a Z′ boson is depicted in Fig. 7.1. These hypothetical particles have masses that range from 365 GeV up to several TeV. The different masses of the decaying particles lead to very different kinematic regimes: at low energies - the *resolved regime* - all the final state particles are well separated from one another. On the other hand, at higher energies - the *boosted regime* - the t quarks are Lorentz-boosted and their decay products are collimated and can not be reconstructed as separate, isolated objects. Dedicated reconstruction algorithms and selections have to be deployed in the different cases. In particular, in the resolved regime the leptons are isolated and the jets are reconstructed with a radius of $R = 0.4$. In the boosted regime, the leptons are non-isolated and on the hadronic leg of the decay the jets are reconstructed with a large radius of $R = 0.8$ and are t-tagged using the DeepAK8 algorithm (cf. Sec. 5.6).

After the selection of the final state objects and the measurement and application of correction factors, the $t\bar{t}$ pairs are reconstructed. The main irreducible background is the SM $t\bar{t}$ production, while reducible backgrounds are given by the W+jets, single t and QCD processes, and all of them are predicted from simulation. Events are divided in signal regions (SRs) and control regions (CRs) to better constrain the cross section rates of the SM processes. The classification task is performed with a deep neural network (DNN), which improves the selection efficiency compared to traditional, cut-based selections. The

sensitive variables used in the search are the invariant mass of the reconstructed $\mathrm{t\bar{t}}$ pair, $m_{\mathrm{t\bar{t}}}$, and the angular variable $\cos(\theta^*)$, sensitive to the spin of the decaying particle. The final step consists in the statistical interpretation of the results. Stringent exclusion limits are placed on the production cross section times branching fraction of Z' bosons and g_{KK} gluons, or on the coupling strength modifiers of the H and A bosons, and the results are compared to previous CMS searches [75, 80]. The results for spin-1 particles correspond to the most stringent limits to date.

7.2 Datasets and simulated events

7.2.1 Datasets

This analysis uses pp collision data collected with the CMS detector at $\sqrt{s} = 13$ TeV. The data are recorded during the 2016-2018 years of the LHC Run 2, corresponding to an integrated luminosity of 138 fb^{-1}. The UL reconstruction of data is used and four run periods are considered: UL16preVFP (19.5 fb^{-1}), UL16postVFP (16.8 fb^{-1}), UL17 (41.5 fb^{-1}) and UL18 (59.8 fb^{-1})[2].

Given that the target final state contains one muon or electron, jets and missing transverse momentum, single lepton datasets are used, namely the SingleMuon and SingleElectron data streams. For 2018 the EGamma data stream is used, which contains both single electron and single photon datasets. Moreover, in 2016 and 2017 the SinglePhoton data stream is added to the SingleElectron one, in order to recover the trigger inefficiencies for high p_T electrons. Only good data taking periods, which pass data-quality certification of the CMS Java Script Object Notation (JSON) files, are selected for the analysis.

High-level trigger paths with one muon or one electron are used. Different triggers are used for low p_T and for high p_T leptons, to obtain the highest sensitivity in both the resolved and boosted regimes. Moreover, the single photon triggers are used to increase the efficiency for high p_T electrons. The triggers include a requirement on the online p_T of the lepton (or the energy E of the photon), and a logical *or* of various triggers is applied, which differ in the object reconstruction methods. In particular, the triggers for low p_T leptons include an isolation requirement, while it is not the case for leptons at high transverse momentum, given the boosted topology of the final state. The application of the low or high p_T triggers is based on the reconstructed (offline) p_T of the muon or electron. The trigger requirements are summarised in Tables 7.1 and 7.2 for the muon and electron

[2]The data collected in 2016 are split into two eras, pre- and post-VFP, which include a different track reconstruction. The reason is a dead-time in the detector read-out due to a high energy deposit in the strip sensors.

datasets, respectively, for the different years.

The difference in the trigger efficiency in data and simulation is corrected using scale factors (SFs). The SFs for the muon triggers are provided by the CMS Collaboration. The electron trigger SFs are measured in the analysis with the "orthogonal dataset" method. The procedure of the SFs extraction is presented in Sec. 7.3.2.

Year	low p_T regime	high p_T regime
UL16 pre&postVFP	$p_T^\mu > 24$ GeV	$p_T^\mu > 50$ GeV
UL17	$p_T^\mu > 27$ GeV	$p_T^\mu > 50$ GeV $\vee$ $p_T^\mu > 100$ GeV
UL18	$p_T^\mu > 24$ GeV	$p_T^\mu > 50$ GeV $\vee$ $p_T^\mu > 100$ GeV

Table 7.1: The trigger requirements for the SingleMuon datasets. The triggers in the low p_T regime include the isolation requirement.

Year	low p_T electrons	high p_T electrons
UL16 pre&postVFP	$p_T^e > 27$ GeV	$p_T^e > 115$ GeV $\vee$ $E^\gamma > 175$ GeV $\vee$ $p_T^e > 27$ GeV
UL17	$p_T^e > 35$ GeV	$p_T^e > 115$ GeV $\vee$ $E^\gamma > 200$ GeV $\vee$ $p_T^e > 35$ GeV
UL18	$p_T^e > 32$ GeV	$p_T^e > 115$ GeV $\vee$ $E^\gamma > 200$ GeV $\vee$ $p_T^e > 32$ GeV

Table 7.2: The trigger requirements for the SingleElectron, SinglePhoton and EGamma datasets. The triggers in the low p_T regime include the isolation requirement.

7.2.2 Simulated samples

New particles decaying to $t\bar{t}$ have been generated at leading order (LO) with different generators for each signal model considered. The complete list of signal samples, with the cross section and the number of weighted events generated for each sample, is given in Appendix A. Spin-1 Z' signals are generated with MadGraph5_aMC@NLO [32] in a model where the Z' boson has the same right- and left-handed couplings to fermions as the SM Z bosons. Masses ranging from 400 GeV to 9 TeV have been generated with three different relative widths: 1%, 10% and 30%. The Z' bosons decay to top quark pairs in all generated events.

Kaluza-Klein gluons g_{KK} are generated with Pythia8 [28] for masses between 500 GeV and 6 TeV, in a model where the branching fraction of the resonance to $t\bar{t}$ is about 94% and the width is $m_{g_{KK}}/6$.

Heavy Higgs bosons A/H are generated with MadGraph5_aMC@NLO for masses between 365 GeV and 1 TeV and three relative widths: 2.5%, 10% and 25%. For this model,

the interference with SM $t\bar{t}$ has to be considered, thus for each signal the resonant and the interference part are simulated separately. The decay to $t\bar{t}$ pairs, which subsequently decay into one lepton, one neutrino and jets, is implemented.

For illustration purposes, the production cross section values of the Z' and g_{KK} signal samples are assumed to be 1 pb and for the Z' and heavy Higgs bosons the 10% relative width is chosen, unless stated otherwise.

There are different SM processes that make up the background of the search. The full list of SM samples is presented in Table 7.3, together with the cross section values. In Appendix A, the Table with the weighted number of generated events for each era is shown. The main, irreducible background is SM $t\bar{t}$ production. It is generated with POWHEG [29, 30] at next-to-leading-order (NLO). The production of single top is generated at NLO as well. For the t-channel and for the production in association with a W boson (tW) POWHEG is used, while single top in the s-channel is generated with MadGraph5_aMC@NLO. Vector bosons produced in association with jets, W+jets and DY+jets (also denoted as V+jets altogether), are generated with MadGraph5_aMC@NLO at LO. These samples are binned in H_T, the sum of momenta of all the final state partons in the matrix element. The multijet QCD process is generated with Pythia8 and binned in H_T. Finally, diboson samples (WW, WZ and ZZ) are generated with Pythia8 at LO.

For all the samples, the parton shower and hadronization are simulated with Pythia8. The underlying event tune CP5 [118] has been used for all SM and signal samples, except for the Z' signal samples, that have the CP2 tune [118]. The NNPDF3.1 [119] PDF set is used for all simulated samples. All MC samples include the simulation of in-time and out-of-time pileup and are reweighted so that the pileup distribution in simulation matches the one observed in data. The reweighting is done using a minimum-bias cross section of 69.2 mb ($\pm 4.6\%$) [120]. Finally, in order to compare simulation with recorded data, simulated events are reweighted according to the integrated luminosity of $L = 138$ fb^{-1}, as:

$$w = \frac{\sigma L}{N} \tag{7.1}$$

where σ is the cross section of each process and N the number of generated events.

Experimental corrections

Additional corrections have to be applied to data and simulation to account for detector effects. In simulation, weights are applied to account for the L1 trigger prefiring issue [121]. The reason is a timing shift in ECAL in 2016 and 2017 that was not propagated to the L1 triggers, resulting in the wrong association of trigger objects to the previous bunch

Process	$\sigma \times$ BR [pb]
$t\bar{t}$ semileptonic	$3.64 \cdot 10^2$
$t\bar{t}$ all hadronic	$3.80 \cdot 10^2$
$t\bar{t}$ dileptonic	$8.73 \cdot 10^1$
W($\to l\nu$)+jets, $70 < H_T < 100$ GeV	$1.27 \cdot 10^3$
W($\to l\nu$)+jets, $100 < H_T < 200$ GeV	$1.25 \cdot 10^3$
W($\to l\nu$)+jets, $200 < H_T < 400$ GeV	$3.36 \cdot 10^2$
W($\to l\nu$)+jets, $400 < H_T < 600$ GeV	$4.52 \cdot 10^1$
W($\to l\nu$)+jets, $600 < H_T < 800$ GeV	$1.10 \cdot 10^1$
W($\to l\nu$)+jets, $800 < H_T < 1200$ GeV	$4.94 \cdot 10^0$
W($\to l\nu$)+jets, $1200 < H_T < 2500$ GeV	$1.16 \cdot 10^0$
W($\to l\nu$)+jets, $H_T > 2500$ GeV	$2.62 \cdot 10^{-2}$
DY($\to ll$)+jets, $70 < H_T < 100$ GeV	$1.40 \cdot 10^2$
DY($\to ll$)+jets, $100 < H_T < 200$ GeV	$1.40 \cdot 10^2$
DY($\to ll$)+jets, $200 < H_T < 400$ GeV	$3.84 \cdot 10^1$
DY($\to ll$)+jets, $400 < H_T < 600$ GeV	$5.21 \cdot 10^0$
DY($\to ll$)+jets, $600 < H_T < 800$ GeV	$1.27 \cdot 10^0$
DY($\to ll$)+jets, $800 < H_T < 1200$ GeV	$5.68 \cdot 10^{-1}$
DY($\to ll$)+jets, $1200 < H_T < 2500$ GeV	$1.33 \cdot 10^{-1}$
DY($\to ll$)+jets, $H_T > 2500$ GeV	$2.98 \cdot 10^{-3}$
WW	$7.59 \cdot 10^1$
WZ	$2.76 \cdot 10^1$
ZZ	$1.21 \cdot 10^1$
single t/$\bar{t}$ s-channel	$3.36 \cdot 10^0$
single t t-channel	$1.36 \cdot 10^2$
single $\bar{t}$ t-channel	$8.10 \cdot 10^1$
single t tW-channel	$1.95 \cdot 10^1$
single $\bar{t}$ tW-channel	$1.95 \cdot 10^1$
QCD, $50 < H_T < 100$ GeV	$1.86 \cdot 10^8$
QCD, $100 < H_T < 200$ GeV	$2.36 \cdot 10^7$
QCD, $200 < H_T < 300$ GeV	$1.55 \cdot 10^6$
QCD, $300 < H_T < 500$ GeV	$3.24 \cdot 10^5$
QCD, $500 < H_T < 700$ GeV	$3.03 \cdot 10^4$
QCD, $700 < H_T < 1000$ GeV	$6.44 \cdot 10^3$
QCD, $1000 < H_T < 1500$ GeV	$1.12 \cdot 10^3$
QCD, $1500 < H_T < 2000$ GeV	$1.08 \cdot 10^2$
QCD, $H_T > 2000$ GeV	$2.20 \cdot 10^1$

Table 7.3: List of SM simulated samples used in the analysis. The $\sigma \times$ BR of each process in given in pb.

crossings. As two bunch crossings can not fire L1 triggers consecutively, the events with high ECAL energy in $2 < |\eta| < 3$ are wrongly vetoed. Event weights are applied to simulation to account for this issue and they are defined as the product of the non-prefiring probability of all objects (photons, muons, jets):

$$w = 1 - P(\text{prefiring}) = \prod_i (1 - \varepsilon_i^{pref}(\eta, p_T)). \tag{7.2}$$

Another detector misfunction happened in 2018, when two HCAL modules stopped working during RunC and RunD. This issue had an impact on the measurement of jets energies in the region defined by $-3 < \eta < -1.3$ and $-1.57 < \phi < -0.87$, with the consequence of jet miscalibration and electron misidentification. To account for these effects, the events containing jets or leptons in this region have been vetoed. The veto has been applied to data and simulated events in UL18. Simulated events have been weighted to account for the affected luminosity fraction.

Theory corrections

Theory corrections are applied on simulated samples. The Kaluza-Klein gluon samples are generated at LO, and the LO cross section values are obtained from Pythia8. To account for higher order corrections, the cross sections are multiplied by a κ factor with value 1.3 [122]. The heavy Higgs cross sections are multiplied by κ factors as well, to take into account next-to-next-to-leading-order (NNLO) cross section calculations [123]. The resonant part of the signal is multiplied by $\kappa_{res} = \sigma_{\text{NNLO}}/\sigma_{\text{LO}}$, while the κ factor for the interference part is given by $\kappa_{int} = \sqrt{\kappa_{res} \times \kappa_B}$, where $\kappa_B = \sigma_{\text{NNLO}}/\sigma_{\text{LO}}$ of SM $\mathrm{t\bar{t}}$ production. The κ factors of scalar and pseudoscalar Higgs bosons as a function of the boson mass are depicted in Fig. 7.2.

The V+jets and diboson samples are simulated at LO. For the V+jets simulation, to account for missing higher-order QCD and electroweak (EWK) contributions, correction factors are applied as a function of the boson transverse momentum [124] . The NLO corrections are shown in Fig. 7.3. Additionally, the V+jets MC production campaigns differ in 2016 and in 2017/2018, resulting in a difference in the distribution of the vector boson p_T. The difference has been included in the NLO QCD correction factors. The cross sections of the diboson samples are multiplied by κ factors that take into account NLO (WZ) and NNLO (WW, ZZ) calculations.

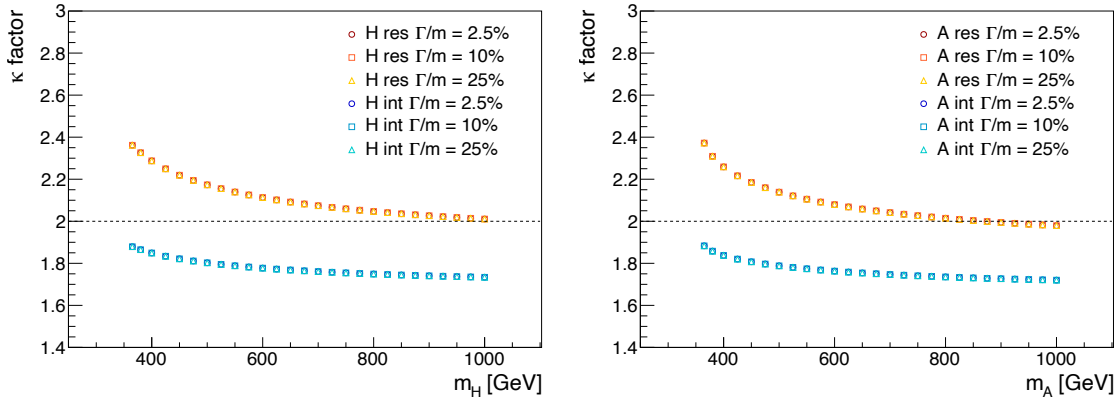

Figure 7.2: The κ factors for the scalar (left) and pseudoscalar (right) Higgs bosons as a function of the boson mass. The κ factors for the resonant and interference signals are shown for the three relative widths used in the analysis.

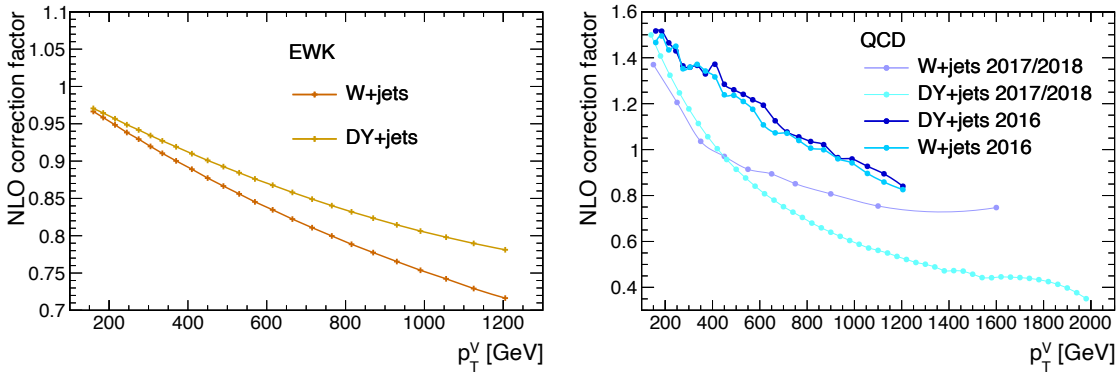

Figure 7.3: The NLO EWK (left) and QCD (right) correction factors applied on the W+jets and DY+jets simulated samples as a function of the vector boson p_T. Derived from [124].

7.3 Event selection

Events that contain two top quarks, one decaying leptonically and the other decaying hadronically, are selected. Two channels are identified, based on the flavour of the charged lepton originating from the leptonic top decay: the muon channel (μ+jets) and the electron channel (e+jets). Moreover, to be sensitive to both low and high mass resonances, two categories of events are considered: events in the resolved regime, for which all the decay products of the top quarks are well separated and reconstructed as different objects, and events in the boosted regime, where the leptons are not isolated and the hadronically decaying top quark is reconstructed in one single, large-radius jet. Different selections and reconstruction techniques are used based on the lepton flavour and kinematic regime, to ensure the best sensitivity in each scenario, and they are presented in the following. The

definition of the objects used in this search is presented in Chapter 5.

7.3.1 Baseline selection

Given the difference in the reconstruction, trigger and selection of events containing muons or electrons, two different event selections are used, one for the μ+jets channel and one for the e+jets channel.

The following selections are applied in the μ+jets channel:

- each event contains exactly one muon with $p_T^{\mu} > 30$ GeV and $|\eta^{\mu}| < 2.4$,
- muons with $30 < p_T^{\mu} < 55$ GeV fulfill the"CutBasedIdTight" and "PFIsoTight" ID,
- muons with $p_T^{\mu} > 55$ GeV fulfill the "CutBasedIdGlobalHighPt" and satisfy the isolation requirement:

$$\Delta R_{min}(l, \text{jet}) > 0.4 \ \vee \ p_{T,rel}(l, \text{jet}) > 25 \text{ GeV} \tag{7.3}$$

  where $\Delta R_{min}(l, \text{jet})$ is the minimum distance in ΔR of the muon with respect to all the AK4 jets with $p_T > 15$ GeV and $p_{T,rel}(l, \text{jet})$ is the transverse momentum of the muon with respect to the AK4 jet that is closest in ΔR,
- muons fulfill the trigger requirements described in Sec. 7.2.1,
- AK4 jets have $p_T > 30$ GeV and $|\eta| < 2.5$ and pass the tight WP of jet ID,
- at least two AK4 jets with $p_T > 50$ GeV are required,
- at least one AK4 jet is b-tagged,
- AK8 jets have $p_T > 200$ GeV and $|\eta| < 2.5$ and pass the tight WP of jet ID,
- a selection of $p_T^{\text{miss}} > 70$ GeV is applied.

The following selections are applied in the e+jets channel:

- each event contains exactly one electron with $p_T^{e} > 35/38/35$ GeV (for 2016/2017/2018) and $|\eta^{e}| < 2.5$,
- electrons with $35/38/35 < p_T^{e} < 120$ GeV (for 2016/2017/2018) fulfill the"mva-based electron ID wp80" with isolation,
- electrons with $p_T^{e} > 120$ GeV fulfill the "mva-based electron ID wp80" without isolation and satisfy the isolation requirement defined in Eq. 7.3,

- electrons fulfill the trigger requirements described in Sec. 7.2.1,
- AK4 jets have $p_T > 30$ GeV and $|\eta| < 2.5$ and pass the tight WP of jet ID,
- at least one AK4 jet with $p_T > 50$ GeV is required,
- a second AK4 jet with $p_T > 40$ GeV is required,
- at least one AK4 jet is b-tagged,
- AK8 jets have $p_T > 200$ GeV and $|\eta| < 2.5$ and pass the tight WP of jet ID,
- a selection of $p_T^{\text{miss}} > 60$ GeV is applied.

The AK4 and AK8 jets used in the event selection are reconstructed with the PUPPI algorithm and they are referred to as PUPPI jets. The tune PUPPI v15, presented in Chapter 6, is used. If AK8 jets are present, they are t-tagged with the DeepAK8-MD tagger. The b-tagging and t-tagging requirements will be described in Sections 7.3.3 and 7.3.4, respectively.

The differences in the μ+jets and e+jets channels arise mostly from the differences in the definitions of the leptons at low and high p_T. The p_T thresholds on the leptons that define the low- and high-p_T regions are chosen following the requirements of the HLT triggers. For muons, the low-p_T category goes from a p_T of 30 GeV to 55 GeV. In this regime, isolation in applied through the ID. For the electrons, on the other hand, the low-p_T regime goes from 35 up to 120 GeV and these electrons are isolated, while the high-p_T region for electrons start only above 120 GeV.

The cuts on the p_T of AK4 jets and p_T^{miss} have been optimized to reduce the background contribution and maximize the sensitivity of the search. In particular, the suppression of QCD is of high importance, as this background is not well modelled by MC simulation. First, QCD is reduced with the application of the isolation ID or custom lepton isolation in Eq. 7.3. The remaining multijet background contribution is reduced with the application of tight cuts on the sub-leading AK4 jet p_T and on p_T^{miss}. This difference in the isolation for the two lepton flavours makes the QCD contribution in the e+jets channel smaller than in the μ+jets channel, resulting in less stringent requirements for the jet and MET cuts. Finally, a $\Delta\eta(j_1, j_2) < 3$ selection is applied between the two leading AK4 jets, to further reduce the QCD multijet background contribution.

Events with additional leptons, namely muons with $p_T^{\mu} > 25$ GeV and $|\eta^{\mu}| < 2.4$ that pass the "CutBasedIdTight" ID and electrons with $p_T^{e} > 25$ GeV and $|\eta^{e}| < 2.5$ that satisfy the "cut-based electron ID" with tight WP, are discarded to ensure orthogonality with the dileptonic analysis. To be orthogonal with the all hadronic channel, events with more than one t-tagged AK8 jet are vetoed.

7.3.2 Electron trigger efficiency measurement

As described in Sec. 7.2.1, a combination of HLT paths is used in this analysis. SFs are applied to MC to account for the differences in the trigger efficiencies between data and simulation. For the μ+jets channel, official SFs are provided by the CMS Collaboration. For the electrons, instead, it is necessary to derive the correction factors for the specific trigger combination used in the analysis. The SFs are measured in a dataset orthogonal to the one used in the search and are applied as a function of the electron η and p_T. The orthogonal dataset is obtained selecting events with one muon and one electron in the final state and it is dominated by the dileptonic $t\bar{t}$ process. The $e\mu$ sample is selected as:

- the events are selected from the SingleMuon datasets and have to pass one of the single muon HLT described in Sec. 7.2.1,
- exactly one muon with $p_T^{\mu} > 30$ GeV and $|\eta^{\mu}| < 2.4$,
- exactly one electron with $p_T^{e} > 35/38/35$ GeV for 2016/2017/2018 and $|\eta^{e}| < 2.5$,
- high-p_T leptons have to satisfy the custom isolation requirement defined in Eq. 7.3,
- at least two AK4 jets with $p_T > 50$ GeV and $|\eta| < 2.5$ are required,
- at least one AK4 jet with $p_T > 30$ GeV and $|\eta| < 2.5$ has to be b-tagged,
- a selection on $p_T^{\text{miss}} > 70$ GeV is applied.

The kinematic distributions of leptons and jets after the dilepton selection are presented in Fig. 7.4. The UL17 period is chosen as example. As expected, the $t\bar{t}$ production is the dominant process in this region.

After this selection, the efficiency of electrons passing the combination of electron trigger paths is derived for data (ε_{DATA}) and simulation (ε_{MC}). The efficiencies are defined as:

$$\varepsilon = \frac{\text{N(selection+trigger)}}{\text{N(selection)}} \tag{7.4}$$

where the numerator contains the number of events that pass the dilepton event selection and the electron HLT, while the denominator contains the number of events that pass the selection. The efficiency is calculated as a function of the electron η in three different p_T bins: $p_T < 120$ GeV, $120 < p_T < 200$ GeV and $p_T > 200$ GeV. The values of 120 and 200 GeV follow the requirements of the electron and photon HLT, respectively. The electron trigger SFs are obtained as: $SF = \varepsilon_{DATA}/\varepsilon_{MC}$. The efficiencies for data and MC and the SFs are shown in Fig. 7.5 for the UL17 period.

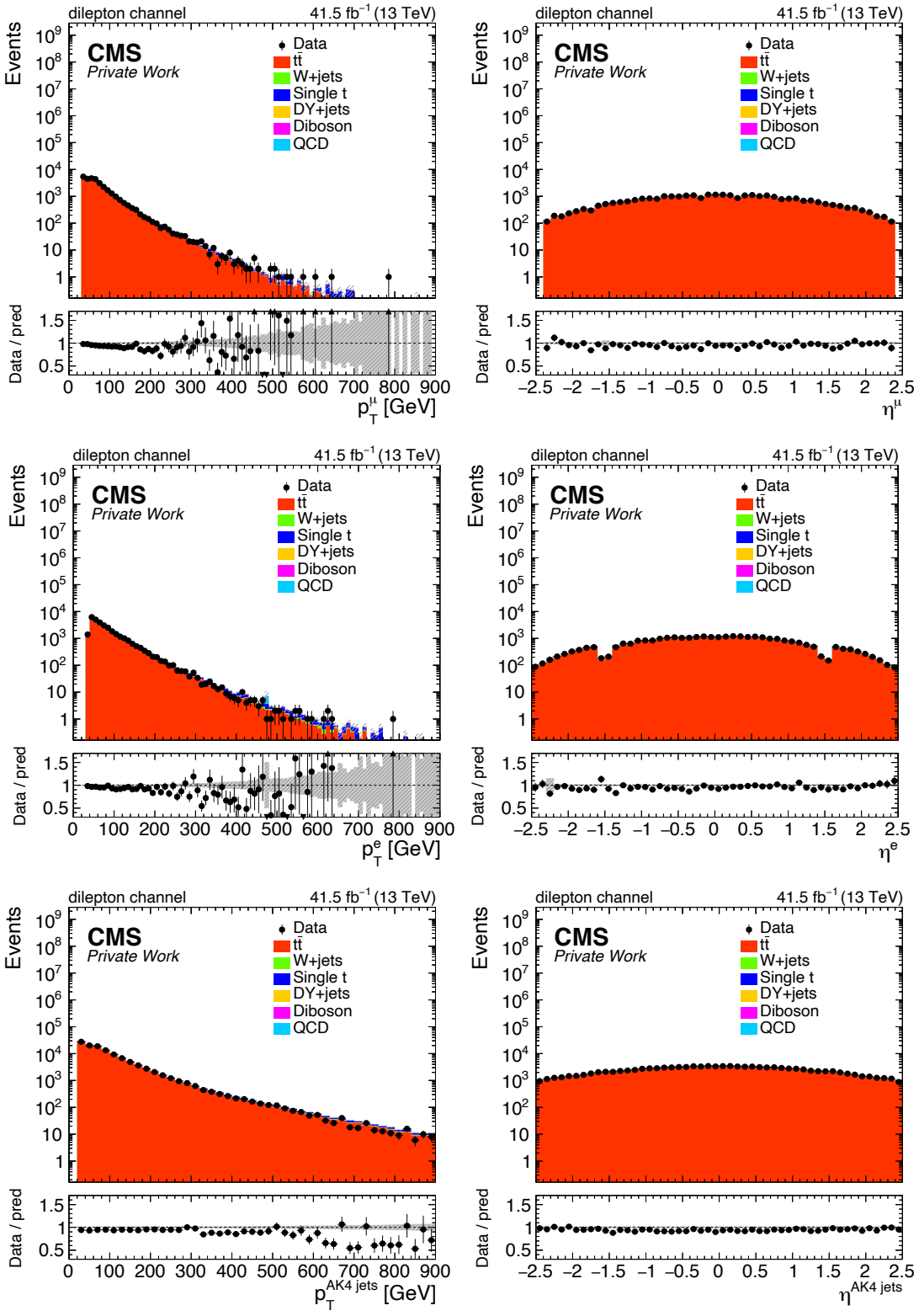

Figure 7.4: The p_T and η distributions of the muon (upper), electron (middle) and AK4 jets (lower) after the dilepton selection for the UL17 period. The grey band represents the statistical uncertainty.

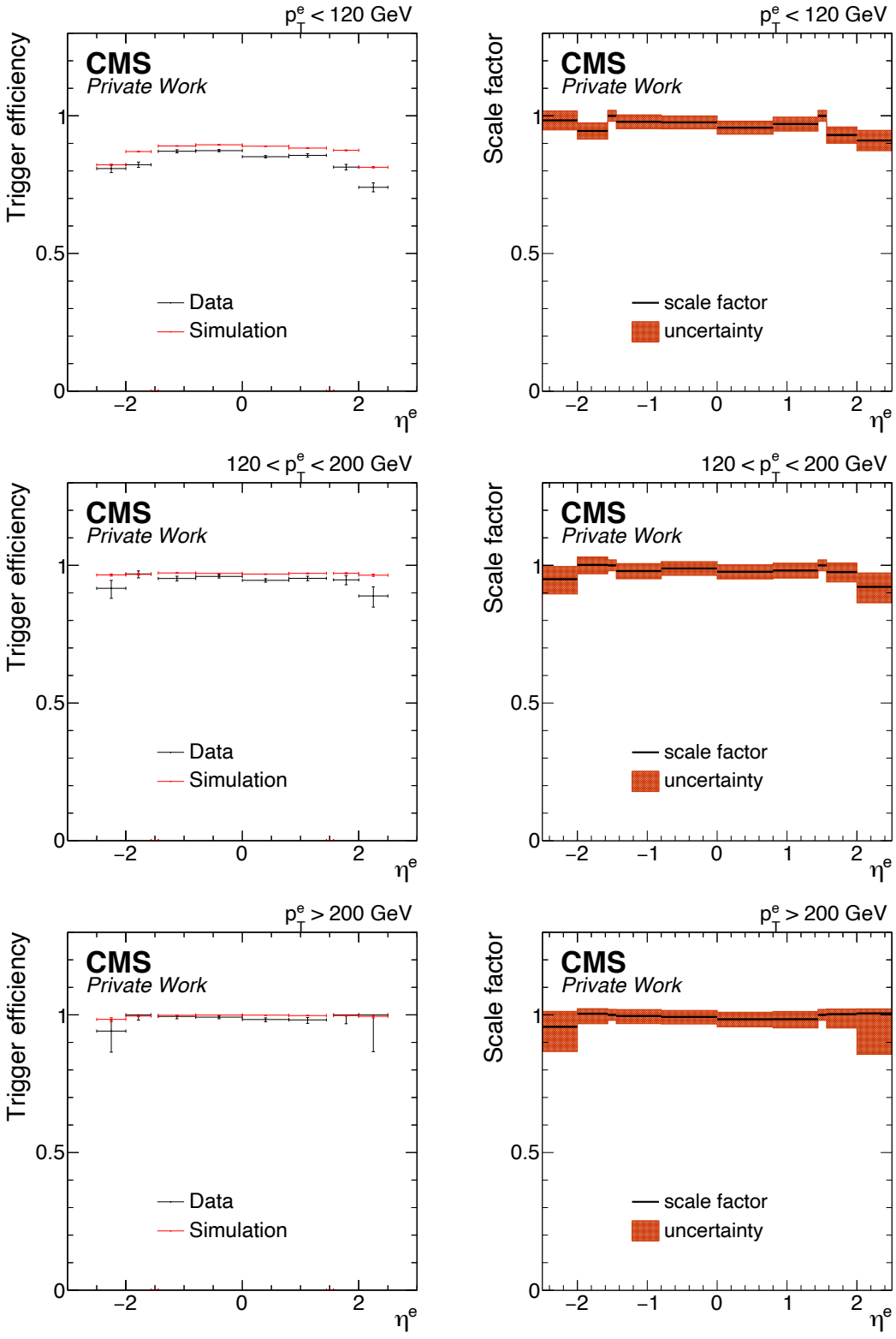

Figure 7.5: The trigger efficiencies for data and simulation (left) and the SFs (right) as a function of the electron η for the UL17 period. Three p_T bins are used: $p_T < 120$ GeV (upper), $120 < p_T < 200$ GeV (middle) and $p_T > 200$ GeV (lower).

7.3.3 b-tagging and correction measurement

The requirement of at least 1 b-tagged AK4 jet is applied in the baseline selection. The DeepJet algorithm [112] is used with the medium WP, which corresponds to an efficiency of 73.3%, 71.4%, 79.1% and 80.7% for the four eras, respectively. Since the training of the tagger and the SF derivation have been performed on CHS jets, the direct application on PUPPI jets is not possible. To overcome this, a matching between PUPPI and CHS jets is done, and the b-tagging criteria are applied on the CHS jets that are matched to the PUPPI jets present in the analysis. A dedicated study on the performance of the two pileup suppression algorithms is performed. The main difference between CHS and PUPPI jets arises because with the PUPPI algorithm more PU is removed from jets (both neutral and charged PU contributions) and more jets made entirely of PU are rejected. This results in kinematic differences between jets reconstructed with the two algorithms, as can be seen in Fig. 7.6. Here the baseline selection is applied on the $t\bar{t}$ sample, using either PUPPI or CHS AK4 jets. With PUPPI there are more events with lower jet multiplicity compared to CHS and the CHS jet p_T distribution shows more low-p_T jets than PUPPI. In the DeepJet score distribution there is a shape difference which prevents the use of DeepJet on PUPPI jets directly.

The matching between PUPPI and CHS is done in the following way:

- for each PUPPI jet, the closest CHS jet in ΔR is identified,
- the condition ΔR(PUPPI jet, CHS jet) < 0.2 is applied, to assure that the correct CHS jet is assigned to each PUPPI jet,
- if no CHS jet is matched to a PUPPI jet, the PUPPI jet is rejected.

The efficiency of the matching of PUPPI to CHS jets is greater than 99%. After the matching, it is possible to apply the b-tagging algorithm to the matched CHS jet. Events with at least 1 b-tagged jet are kept for the analysis.

The b-tagging SFs are applied to MC samples to correct for differences in the b-tagging efficiency between data and simulation. The SFs are provided by the CMS Collaboration and they are applied using the information of the matched CHS jets. The application of b-tagging SFs can cause shape-changing effects on kinematic distributions such as number of jets and jet p_T, independently of the PU mitigation algorithm chosen. This effect can be reduced using dedicated correction factors, which are derived for this analysis as a function of number of jets (N_{jets}) and jet H_T, where H_T is the sum of the p_T of all jets in the event. The 2D SFs are calculated by deriving the (N_{jets}, H_T) distribution in the analysis region - without any requirement on b-tagging - before and after the application

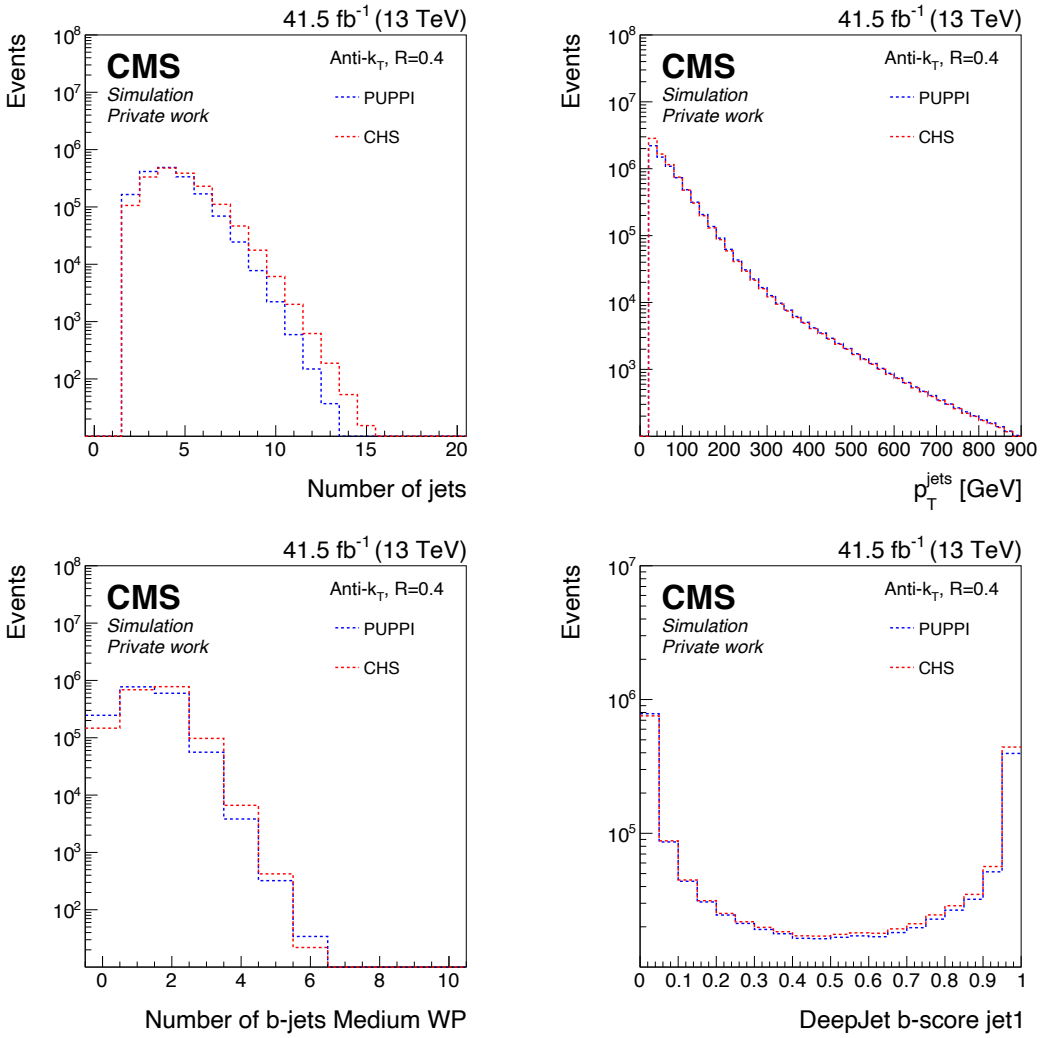

Figure 7.6: The number (upper left) and the p_T (upper right) of AK4 jets, the number of b-jets (lower left), and the DeepJet b-score of the leading AK4 jet (lower right) for the $t\bar{t}$ sample in the μ+jets channel UL17.

of the b-tagging SFs. The correction factors are calculated for each MC sample. In Fig. 7.7 the number of jets, leading jet p_T and DeepJet score distributions are shown for the $t\bar{t}$ simulation without any b-tagging SF, with the b-tagging SFs and with the b-tagging+2D SFs. It can be observed that the changes in the shape, caused by the application of b-tag SFs, are reduced using the 2D SFs in the jet p_T and number of jets distributions. The shape of the DeepJet score is not affected, as desired.

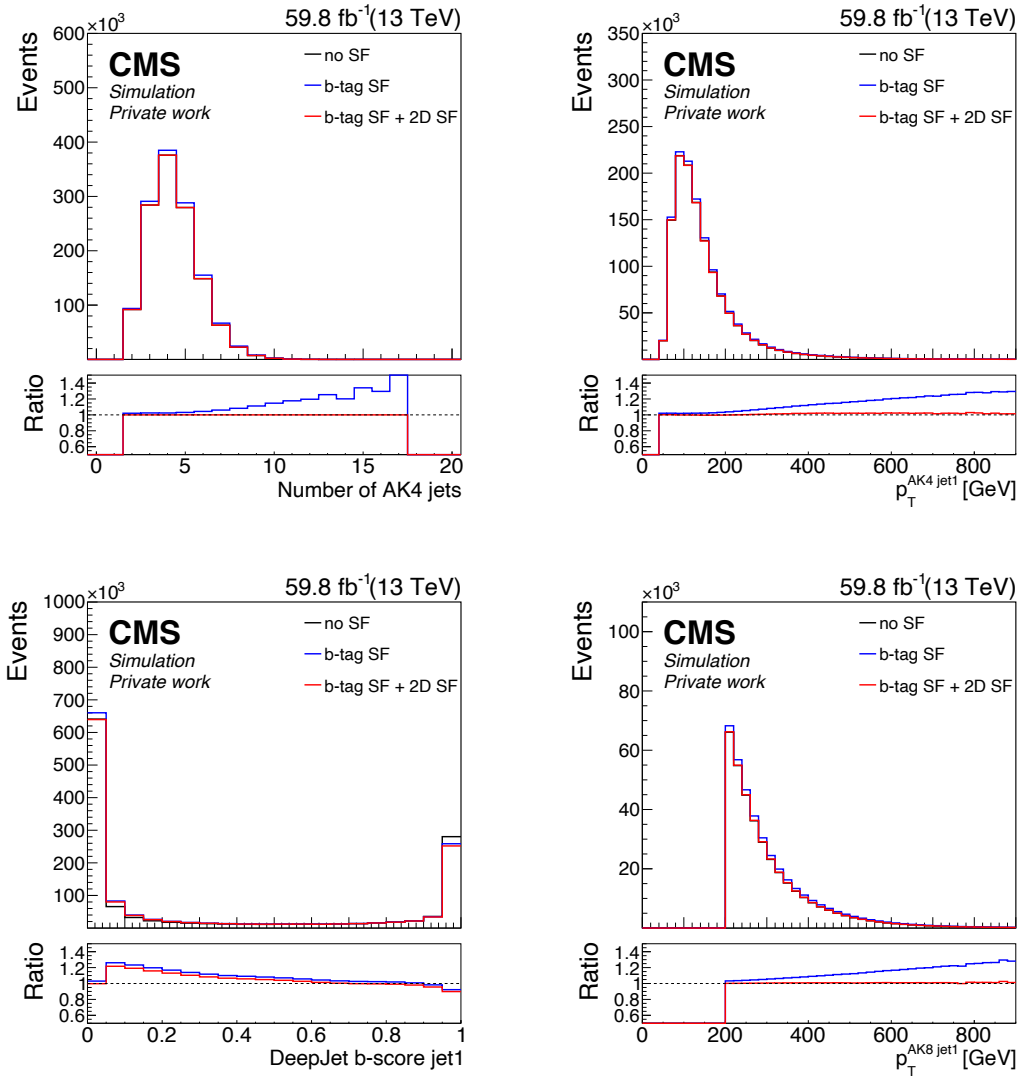

Figure 7.7: The number of AK4 jets (upper left), the p_T (upper right) and the DeepJet b-score (lower left) of the leading AK4 jet, and the p_T of the leading AK8 jet (lower right) for the $t\bar{t}$ sample in the μ+jets channel UL18. Events are shown without b-tagging SFs (black), with b-tagging SFs (blue) and with both b-tagging and 2D SFs (red). The lower panels show the ratio to the case without the b-tagging SFs.

7.3.4 t-tagging and t-mistag rate measurement

The large-radius jets are t-tagged with the following requirements:

- $p_T > 400$ GeV,
- $105 < m_{\mathrm{SD}} < 210$ GeV,
- DeepAK8-MD tagger with 1% mistag rate.

Jets passing these criteria are labelled as t-tagged AK8 jets and they are dominated by $\mathrm{t\bar{t}}$ production.

The choice of the DeepAK8-MD algorithm for the tagging of top jets has been made together with the team working on searches for resonances in the all hadronic $\mathrm{t\bar{t}}$ final state. Different t-tagging algorithms, introduced in Sec. 5.6, have been studied: the traditional cut-based tagger, which uses a cut on the jet substructure variable τ_{32}, the HOTVR tagger, based on jets with variable radius, and the DeepAK8 tagger. In the lepton+jets final state, the sensitivity of the search with the use of the different taggers has been studied for one signal model as benchmark, the Z' boson with 10% relative width, in the μ+jets channel on the UL18 dataset. The sensitivity of the three algorithms is similar, as can be seen in Fig. 7.8. The all hadronic analysis, which requires two t-tagged jets in the final state and is thus more sensitive to the choice of the tagger, found that the best performance is obtained with DeepAK8. The mass-decorrelated version of the tagger is used to avoid mass sculpting (c.f. Subsec. 5.6.2). Data-to-MC SFs are applied to correct for the difference in t-tagging efficiency between data and simulation.

Nevertheless, non-top jets could be wrongly identified as top jets by the algorithm, in particular jets originating from the W+jets process. The rate at which light jets are misidentified as top jets is denoted as t-mistag rate and it is measured in the analysis with a dedicated study.

The t-mistag rate is measured in a control region (CR) dominated by W+jets events, which is obtained with the event classifier described in the following Section. The tagging efficiency is defined as the number of jets after the t-tagging over the number of jets before the t-tagging. Only the leading AK8 jet per event is used for this measurement. For data, the number of top jets from the $\mathrm{t\bar{t}}$ and single t processes is subtracted to account for the contributions of other backgrounds. The top jets are obtained by matching the reconstructed AK8 jets to generator-level top quarks with $\Delta\mathrm{R} \leq 0.4$. The efficiency in data is defined as:

$$\varepsilon_{\mathrm{DATA}} = \frac{N^{tagged}_{\mathrm{DATA}} - N^{tagged}_{\mathrm{t\bar{t}(t)}} - N^{tagged}_{\mathrm{ST(t)}}}{N_{\mathrm{DATA}} - N_{\mathrm{t\bar{t}(t)}} - N_{\mathrm{ST(t)}}} \tag{7.5}$$

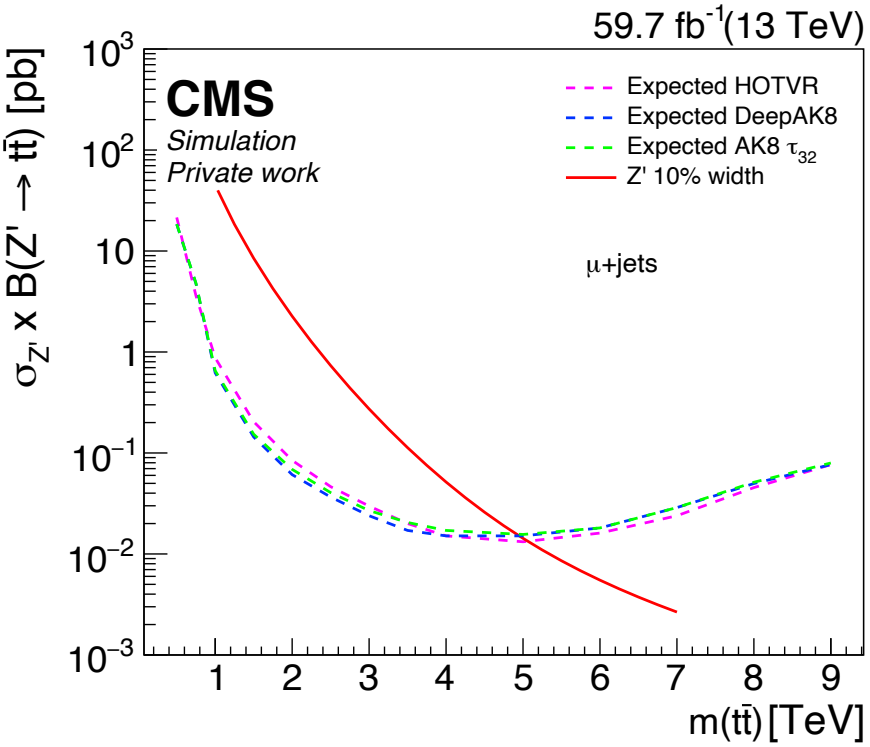

Figure 7.8: The expected exclusion limits at 95% CL on the production cross section for the Z′ boson with 10% relative width as a function of the Z′ boson mass, obtained with different t-tagging algorithms: HOTVR (pink), DeepAK8 (blue) and the cut-based tagger (green). Events from the μ+jets channel UL18 are used.

where N_{DATA} is the number of AK8 jets in data, and $N_{\mathrm{t\bar{t}(t)}}$ and $N_{\mathrm{ST(t)}}$ are the numbers of AK8 jets in the $\mathrm{t\bar{t}}$ and single t samples, respectively, that are matched to generator-level top quarks. $N^{tagged}_{\mathrm{DATA}}$, $N^{tagged}_{\mathrm{t\bar{t}(t)}}$ and $N^{tagged}_{\mathrm{ST(t)}}$ are the numbers of AK8 jets that pass the t-tagging criteria.

For simulation, the efficiency is calculated as:

$$\varepsilon_{\mathrm{MC}} = \frac{N^{tagged}_{\mathrm{W+jets}} + N^{tagged}_{\mathrm{t\bar{t}(l)}} + N^{tagged}_{\mathrm{ST(l)}} + N^{tagged}_{\mathrm{DY}} + N^{tagged}_{\mathrm{QCD}}}{N_{\mathrm{W+jets}} + N_{\mathrm{t\bar{t}(l)}} + N_{\mathrm{ST(l)}} + N_{\mathrm{DY}} + N_{\mathrm{QCD}}} \tag{7.6}$$

where all the MC processes that contribute to the event sample after the t-tagging are added to the W+jets process. Only the light (non-top) jets are considered, in order to measure the mistag rate. The light jets (l) are defined as the ones that are not matched in ΔR to generator-level top quarks. The data-to-MC SF is obtained as: $SF = \varepsilon_{DATA}/\varepsilon_{MC}$. For each data taking period, the mean value of the SFs of the μ+jets and e+jets channel is taken, resulting in:

- UL16preVFP: SF $= 1.14 \pm 0.32$,

- UL16postVFP: SF $= 1.08 \pm 0.36$,
- UL17: SF $= 0.96 \pm 0.21$,
- UL18: SF $= 1.09 \pm 0.21$.

The control distributions after the baseline selection and the application of all SFs and corrections are shown in Figures 7.9 and 7.10 for the μ+jets and e+jets channels, respectively, for the full Run 2 dataset. A good data-to-simulation agreement is observed. The main background is $t\bar{t}$, which makes up around 76% of the total backgrounds, followed by the W+jets ($\sim$ 11%), single t ($\sim$ 7%) and QCD (4% for μ+jets, 1.7% for e+jets) processes. The remaining minor backgrounds are Diboson and DY+jets. After this selection, the $t\bar{t}$ pairs can be reconstructed.

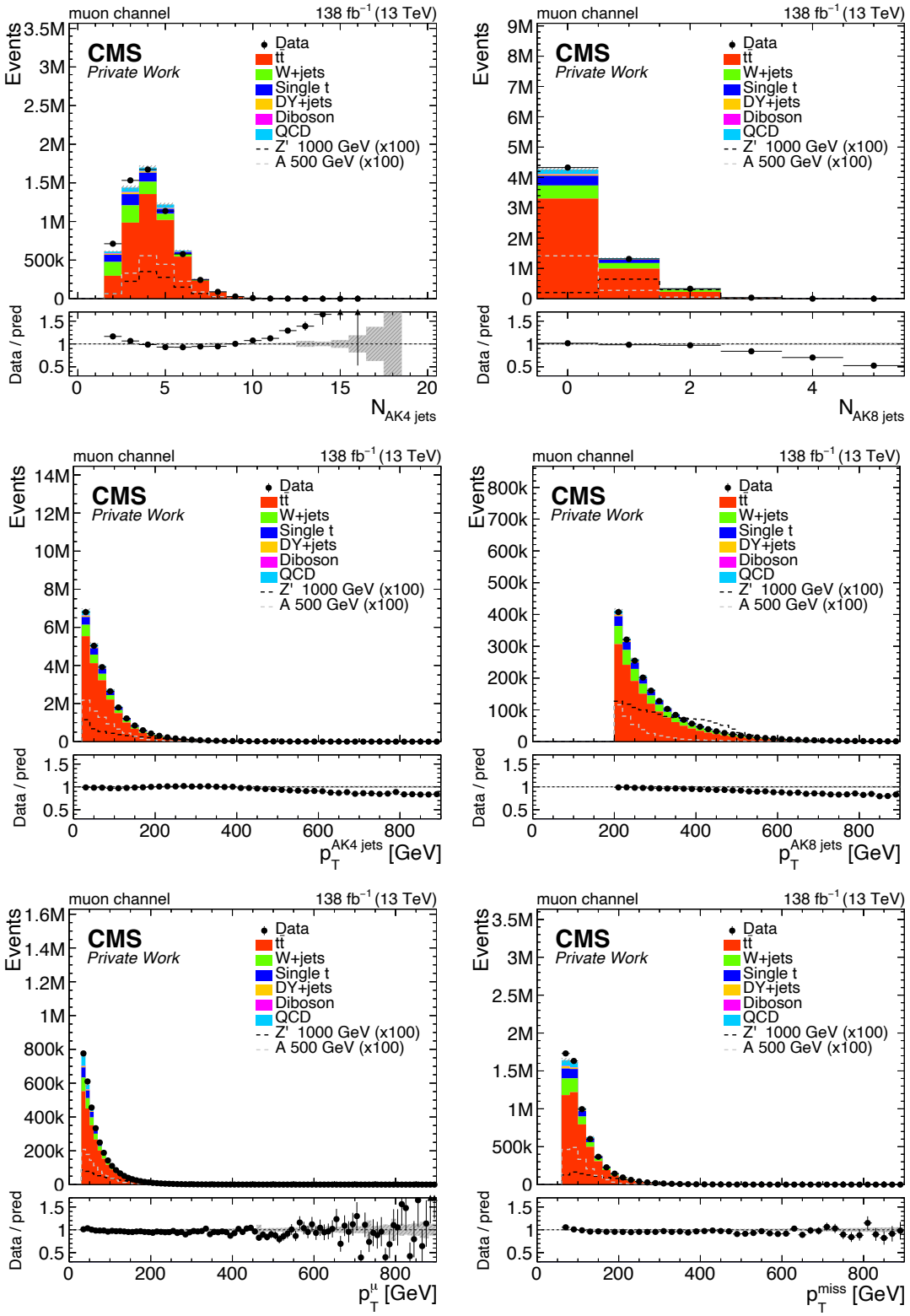

Figure 7.9: The distributions of $N_{AK4\ jets}$ (upper left), $N_{AK8\ jets}$ (upper right), $p_T^{AK4\ jets}$ (middle left), $p_T^{AK8\ jets}$ (middle right), p_T^{μ} (lower left) and p_T^{miss} (lower right) in the μ+jets channel after the baseline selection. The grey band represents the statistical uncertainty.

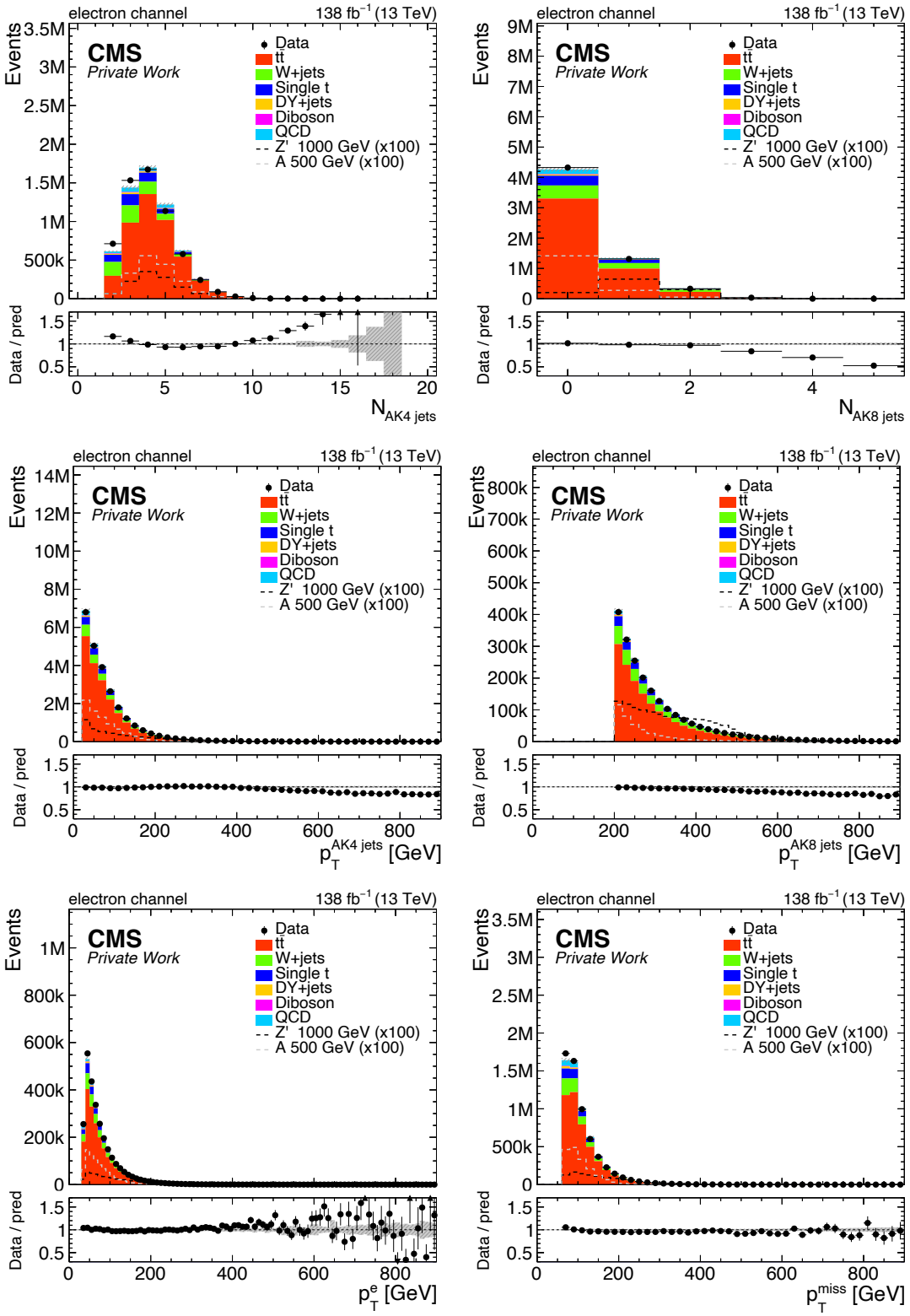

Figure 7.10: The distributions of $N_{AK4\ jets}$ (upper left), $N_{AK8\ jets}$ (upper right), $p_T^{AK4\ jets}$ (middle left), $p_T^{AK8\ jets}$ (middle right), p_T^e (lower left) and p_T^{miss} (lower right) in the e+jets channel after the baseline selection. The grey band represents the statistical uncertainty.

7.4 Reconstruction of the $t\bar{t}$ system

The reconstructed mass of the $t\bar{t}$ system is used to discriminate the presence of a potential signal from the SM backgrounds. While the backgrounds show a falling distribution in the $m_{t\bar{t}}$ spectrum, the signals show a peak or a peak-dip structure at the value corresponding to the mass of the new particle. In order to reconstruct the $t\bar{t}$ pair, the two top quarks are identified starting from the reconstructed final state objects. The leptonically decaying top is reconstructed with the charged lepton, the neutrino and small-radius jets. The hadronically decaying top can be reconstructed either with a t-tagged large-radius jet, or with a combination of small-radius jets. Since different combinations of jets are possible, for each event the candidate with highest probability to originate from two top quarks is selected with a χ^2 approach. In the following, the procedure to reconstruct the $t\bar{t}$ pair is described in detail.

7.4.1 Reconstruction of the neutrino

In the final state considered, exactly one neutrino is present. Neutrinos cannot be directly measured in CMS, but they can be reconstructed starting from p_T^{miss}. A common procedure is to interpret p_T^{miss} as the transverse component of the neutrino's momentum and derive the z-component by assuming that the neutrino and the charged lepton originate from the W boson:

$$P_{\text{W}}^2 = M_{\text{W}}^2 = (P_\nu + P_l)^2, \tag{7.7}$$

where P_{W} is the four-momentum of the W boson, M_{W} its mass, P_ν is the four-momentum of the neutrino and P_l the four-momentum of the charged lepton. It is possible to solve this quadratic equation for the z-component of the neutrino's momentum:

$$p_{z,\nu}^{\pm} = \frac{\mu p_{z,l}}{p_{\text{T},l}^2} \pm \sqrt{\frac{\mu^2 p_{z,l}^2}{p_{\text{T},l}^4} - \frac{E_l^2 p_{\text{T},\nu}^2 - \mu^2}{p_{\text{T},l}^2}}\ , \tag{7.8}$$

where $\mu = M_{\text{W}}^2/2 + p_{\text{T},l}\ p_{\text{T},\nu} \cos(\Delta\phi)$ and $\Delta\phi$ is the azimuthal angle between $\vec{p}_T^{\,\text{miss}}$ and the charged lepton. Equation 7.8 can have zero, one or two real solutions. In the case of no real solution, the real part of the complex solution is used, while in the case of two real solutions, both of them are tested.

7.4.2 Reconstruction of the top quark candidates

The following step is the reconstruction of the two top quark candidates: the top decaying leptonically (t_{lep}) and the one decaying hadronically (t_{had}). The t_{lep} candidate is recon-

structed using the charged lepton (electron or muon), the neutrino and AK4 jets. The t_{had} candidate is reconstructed either with one t-tagged AK8 jet (boosted regime) or with AK4 jets (resolved regime).

Each AK4 jet in the event can be assigned to t_{had}, t_{lep} or neither, and for events with more that ten AK4 jets, only the leading ten are used. All the possible combinations are tested, with the condition that at least one AK4 jet is assigned to t_{lep} and at least one jet, either AK4 or t-tagged AK8, is assigned to t_{had}. Moreover, the following conditions have to be fulfilled: the t-tagged AK8 jet is separated from the charged lepton with $\Delta R(\text{AK8 jet}, l) > 0.8$ and only AK4 jets separated from the t-tagged AK8 jet with $\Delta R(\text{AK8 jet}, \text{AK4 jet}) > 1.2$ are considered for t_{lep}.

The four-momenta of the top quark candidates are given by the sum of the four-momenta of their constituents and the final $t\bar{t}$ candidates result from the sum of the four-momenta of t_{had} and t_{lep}:

$$P_{t_{lep}} = P_\nu + P_l + \sum_i P_{\text{AK4 jet},i} \tag{7.9}$$

$$P_{t_{had}} = \sum_i P_{\text{AK4 jet},i} \quad \text{or} \quad P_{t_{had}} = P_{\text{AK8 jet}} \tag{7.10}$$

$$P_{t\bar{t}} = P_{t_{lep}} + P_{t_{had}}. \tag{7.11}$$

The reconstructed masses of the t_{lep} and t_{had} candidates, separated for the resolved and boosted topology, are shown in Fig. 7.11.

7.4.3 Selection of the $t\bar{t}$ candidate

After constructing all the possible top quark pairs, the correct $t\bar{t}$ candidate for each event must be selected. Among all the possibilities, the $t\bar{t}$ pair whose top quark candidates have mass closest to the true top quark mass are chosen. This selection is made by choosing the pair with smallest χ^2:

$$\chi^2 = \chi^2_{lep} + \chi^2_{had} = \left[\frac{M_{lep} - \bar{M}_{lep}}{\sigma_{\bar{M}_{lep}}}\right]^2 + \left[\frac{M_{had} - \bar{M}_{had}}{\sigma_{\bar{M}_{had}}}\right]^2 \tag{7.12}$$

where M_{lep} and M_{had} are the invariant masses of t_{lep} and t_{had}, respectively. In the boosted case, the value of M_{had} is given by the soft-drop mass of the t-tagged AK8 jet.

The parameters $\bar{M}_{lep}$, $\bar{M}_{had}$, $\sigma_{\bar{M}_{lep}}$ and $\sigma_{\bar{M}_{had}}$ are obtained from simulation by matching the reconstructed top quark candidates to generator-level particles from the l+jets $t\bar{t}$ decay. The matching is performed using the MC $t\bar{t}$ sample and it is defined by:

- for t_{had}:

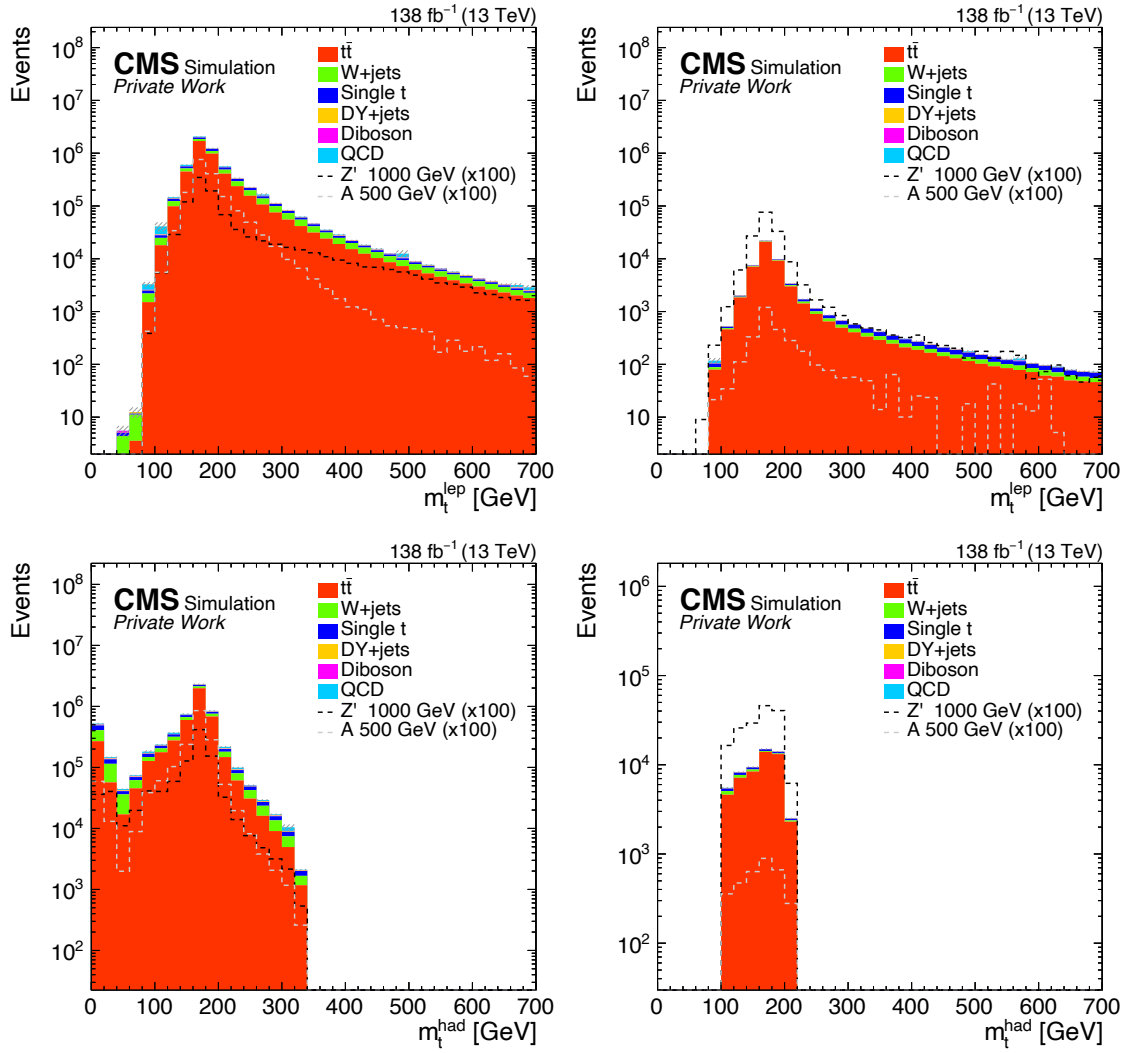

Figure 7.11: The reconstructed masses of the t_{lep} (upper) and t_{had} (lower) candidates for the resolved (left) and boosted (right) regimes, for the SM backgrounds and two signals. Events passing the baseline selection are used.

  - $\Delta R(\text{AK4 jet}, \text{quark}) < 0.4$ for the resolved case,
  - $\Delta R(\text{AK8 jet}, \text{quark}) < 0.8$ for the boosted case,
  - each generator-level quark has to be matched to a jet and one jet can be matched to more than one quark.

- for t_{lep}:
  - $\Delta R(\text{AK4 jet}, \text{b quark}) < 0.4$,
  - $\Delta R(\text{lepton}_{\text{reco}}, \text{lepton}_{\text{gen}}) < 0.1$,
  - $\Delta\phi(\vec{p}_T^{\,\text{miss}}, \text{neutrino}) < 0.3$.

The mass distributions M_{lep} and M_{had} of the matched events are fitted with a gaussian function and the mean and width values are extracted. The average values from the fits in the μ+jets and e+jets channels in the four data-taking periods are used for the χ^2 calculation. The obtained values are summarized in Table 7.4.

category	$\bar{M}_{\text{had}}$ [GeV]	$\bar{M}_{\text{lep}}$ [GeV]	$\sigma_{\bar{M}_{\text{had}}}$ [GeV]	$\sigma_{\bar{M}_{\text{lep}}}$ [GeV]
resolved	173.0	173.6	21.2	24.6
boosted	180.6	171.4	15.6	22.0

Table 7.4: The mean mass and width values used in the χ^2 for the $\mathrm{t\bar{t}}$ pair selection.

The reconstructed invariant mass of the $\mathrm{t\bar{t}}$ pair for three different Z' signals is shown in Fig. 7.12 (left). In particular, it is possible to see the difference between signals at low masses and high masses. For a new resonance of 500 GeV, the distribution shows a clear peak at the value of the generated mass. With increasing mass, the off-shell production becomes more and more important, due to the falling PDFs of the proton and the convolution with the Breit-Wigner of the resonance. The off-shell contribution can be seen as an enhancement in the lower part of the mass spectrum, which has a shape similar to the backgrounds. In Fig. 7.12 (right) the χ^2 distribution is shown for the same signals. The χ^2 has values close to zero for correctly reconstructed t quarks and a second peak at around $\chi^2 = 60$, present especially for low mass signals, which comes from the misreconstruction of one of the two top quarks. Figure 7.13 shows the $m_{\mathrm{t\bar{t}}}$ and χ^2 spectra for the SM backgrounds and two signals. The SM backgrounds show a falling $m_{\mathrm{t\bar{t}}}$ distribution. The χ^2 distribution peaks at zero and has a shoulder at around 60, as seen for the signals. The second peak originates ofter in hypotheses missing a jet: it can happen that t_{lep} is reconstructed, but the hadronic hypothesis fails, resulting in a small t_{had} mass. To reduce the non-$\mathrm{t\bar{t}}$ background contribution and to remove events for which one of the two t quarks was not correctly reconstructed, a cut on $\chi^2 < 30$ is applied to the events in the signal region, which will be defined in the Sec. 7.6.

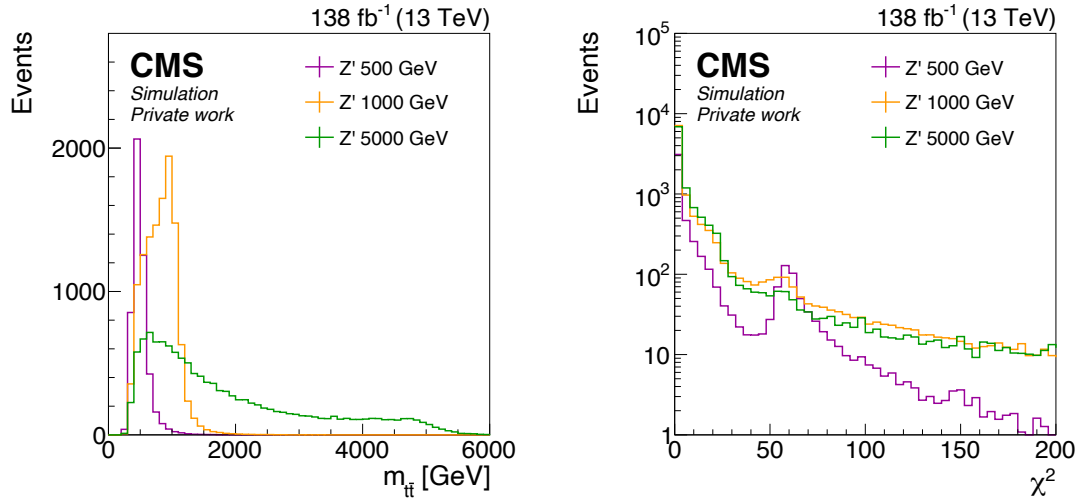

Figure 7.12: The $m_{t\bar{t}}$ (left) and χ^2 (right) distributions for three Z′ signals with different masses. Events passing the baseline selection are used.

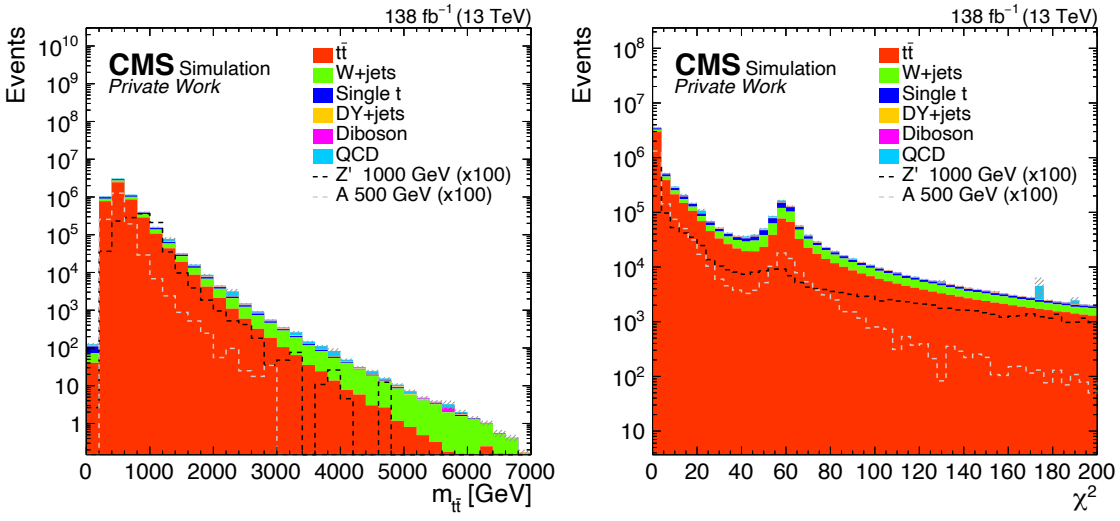

Figure 7.13: The $m_{t\bar{t}}$ (left) and χ^2 (right) distributions for SM backgrounds and two signals. Events passing the baseline selection are used.

7.5 Deep neural network for event classification

In order to maximize the sensitivity of the search, events are categorized into various regions, each enriched with a specific physical process. This is done because, on one hand, it is important to select $t\bar{t}$ events with high purity, and on the other hand, regions enriched with specific backgrounds can be exploited to constrain the systematics uncertainties and to extract the normalization of the processes that are poorly modelled in simulation. Moreover, dividing all the events into classes does not hurt the signal efficiency, as can be the case with the use of hard selections. Neural networks are a powerful tool for event classification, as they can treat a great number of variables and automatically learn information from them. Selections based on neural network outputs can be more efficient compared to traditional cuts on a limited number of specific variables. In the analysis presented in this thesis, a feed-forward, fully-connected deep neutral network (DNN) is used. A sketch of the DNN structure used in the analysis is presented in Fig. 7.14. It consists of an input layer, two hidden layers, and an output layer. The events after the baseline selection are fed into the DNN and classified into three classes, corresponding to the three main processes in the analysis: $t\bar{t}$, V+jets and single t. The neural network approach followed in this search is model-agnostic, which implies that no assumption on a signal model is made and the network can be applied to other signal processes as well. The signals considered here present the same final state as the irreducible $t\bar{t}$ background, and they are naturally categorized into the $t\bar{t}$ class. In the following, the DNN used in this analysis will be presented, together with its performance. Then the application of the DNN in the search will be discussed. The network used in this work has been implemented with the Keras API [125] with the TensorFlow interface [126]. For a complete overview on machine-learning techniques see e.g. Ref. [127].

7.5.1 DNN structure and training

DNNs can be used to extract information from a set of inputs with a varying degree of complexity. The way in which the output is extracted from the input defines the type of network. In this thesis, a *feed-forward* network is used. Feed-forward networks allow the flow of information from the input to the output, through a number of internal layers, in one direction only, meaning that no information can be back-propagated.

The input variables, also referred to as *features*, enter the network in the input layer. Afterwards, a number of *hidden layers* is present and the depth of the network is represented by the number of hidden layers. Each layer is composed of a number of *nodes*, which represent the dimensionality of the layer, and the nodes are connected to one another

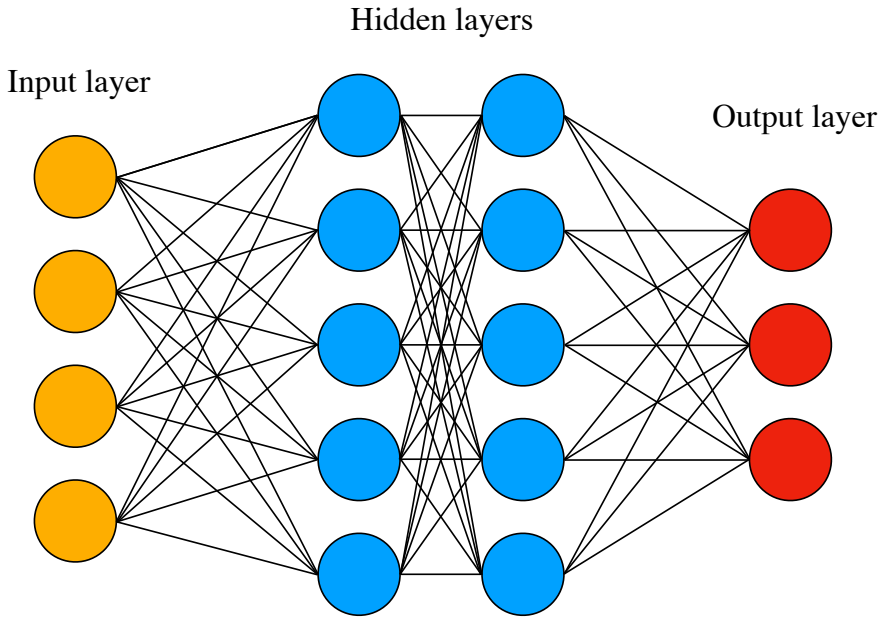

Figure 7.14: A sketch of the DNN structure with two hidden layers used in the analysis.

in a way that resembles the neural connections. All the nodes of a layer are connected to all the nodes of the following layer (*fully-connected* network). Finally, the last layer of the network is the output layer. In a classification task, the number of nodes of the output layer represents the number of output classes in which the events are classified. The number of layers and of nodes are *hyperparameters* of the network and can be optimized.

The activation function used in the DNN for the input and the hidden layers is the rectified linear unit function (ReLU), one of the most commonly used functions in neural networks. The ReLU function is defined as: $h(x) = \max\{0, x\}$. For the output layer the *softmax activation function* is used instead. The softmax function is defined as:

$$\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}, \tag{7.13}$$

where N is the number of output nodes. The softmax function is the most suitable choice for this task, as the values of all the output nodes sum up to 1 and they can be interpreted as a probabilities.

The training of the network consists in the optimization of the *loss function*. For a classifier with multiple classes, the *categorical cross-entropy* is the natural choice as loss function:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i), \tag{7.14}$$

where the sum runs over the number of output nodes N, y_i is the target value of the ith node and $\hat{y}_i$ its predicted value. For an event belonging to the class m, its target values are $y_{j=m} = 1$ and $y_{i\neq m} = 0$. The predicted values are real numbers between 0 and 1 for all the classes m. In the training, the loss function is minimized and the Adam minimizer [128], based on a stochastic gradient descent method, is used with a learning rate of 0.0005.

The events that enter the network are divided into three exclusive sets: *training*, *validation* and *test* set. The training set is used for the actual training of the DNN. The validation set is used to monitor the performance of the network during the training. The test set is used to evaluate the final performance of the DNN on a sample that has not been seen by the network during training. The inputs in this search are split into the three sets in the ratio 60%, 20% and 20% for training, validation and test, respectively. Events in the training set pass through the DNN multiple times, or *epochs*. The weights of the events are updated after a certain number of events, a *batch*, is processed.

The input variables used in the training are low-level quantities of the reconstructed objects - charged leptons, jets, p_T^{miss} - and event quantities. The variables are saved after the baseline selection described in Section 7.3. In total 59 input variables are used and they are summarized in Table 7.5.

Object	variable
lepton	p_T, η, ϕ, E
neutrino	p_T, ϕ
AK4 jets	N, p_T, η, ϕ, E, m, b-tag score
AK8 jets	N, p_T, η, ϕ, E, m_{SD}, τ_{21}, τ_{32}

Table 7.5: The input variables used in the DNN training. The leading 5 AK4 jets and the leading 3 AK8 jets are considered.

If an object is missing, e.g. no AK8 jet is present in the event, the corresponding default value is set to -10. This value has been chosen as it lies outside the allowed ranges of all the variables considered. The distributions of all the input variables are shown in Appendix B for the μ+jets and e+jets channels. In Fig. 7.15 some of the input distributions are shown for the μ+jets channel.

Before entering the DNN, the input features are pre-processed to ensure the stability of the training. The pre-processing algorithm applied scales the variables so that the distribution of each variable has mean of 0 and standard deviation of 1.

To prevent the network from learning the features belonging to the specific training set employed, referred to as *overtraining*, two regularization methods are used. The first is the application of a dropout: after each hidden layer, a dropout layer is placed, in which a

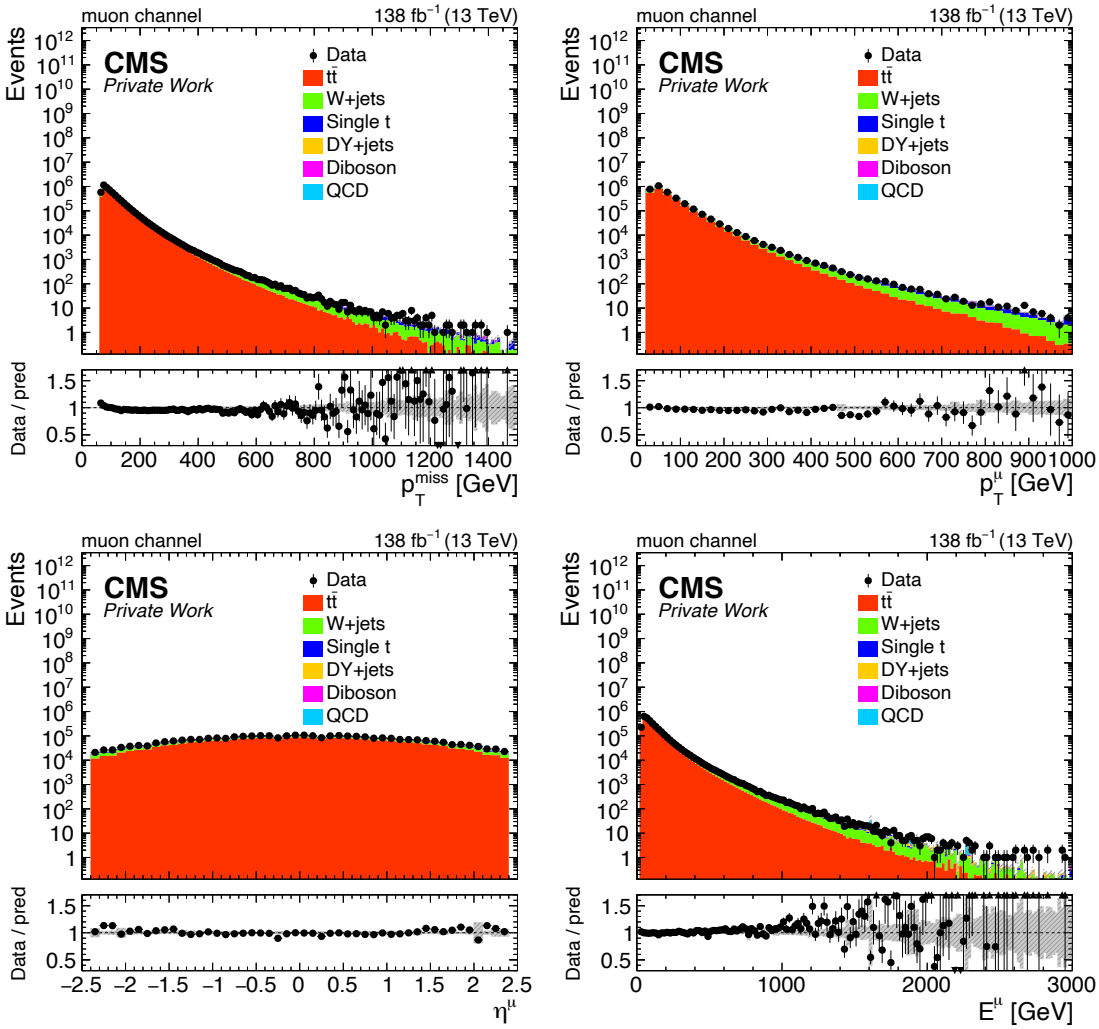

Figure 7.15: The distributions of p_T^{miss} (upper left), p_T^{μ} (upper right), η^{μ} (lower left) and E^{μ} (lower right) for the μ+jets channel used as input variables for the DNN. The grey band represents the statistical uncertainty.

certain fraction (*dropout rate*) of nodes is removed from the network. The dropout rate of 50% is used. The second way to prevent overtraining is to monitor the value of the loss on the validation sample after each epoch. If the loss calculated on unseen events is increasing with respect to the loss calculated on the training sample, it means the network is overtraining. A condition has been implemented, that checks the value of the loss on the validation set and the best model is chosen to be the one for which the validation loss is minimum. The hyperparameters of the DNN developed for the analysis presented in this thesis are summarized in Table 7.6. They have been optimized to achieve the best performance.

The DNN has been trained for each channel, e+jets and μ+jets, using the UL17

Number of hidden layers	2
Number of nodes per hidden layer	512
Activation function hidden layers	ReLU
Activation function output layer	Softmax
Number of epochs	500
Batchsize	2^{15}
Regularization	Dropout (50%)
Optimizer	Adam
Loss	Categorical cross-entropy
Metric	Categorical accuracy

Table 7.6: The hyperparameters of the DNN used in this search.

simulation and it has been applied to all the analyzed periods, as no differences in the DNN performance are expected between different years. The DNN training has been monitored with the loss and accuracy, shown in Fig. 7.16. The loss decreases with the number of epochs both for the training and validation samples, and the one for the validation sample is lower than the one for the training sample, which means that there is no overtraining. The accuracy indicates for how many events the model predicts the correct output. The accuracy of both the training and validation sets reaches a plateau rapidly, with a value around 87%(87.5%) for the training (validation) set.

The performance of the DNN can be represented in terms of the ROC curve, in Fig. 7.17 (left). The area under the ROC curve (AUC) value is used as a measure of the DNN performance in classification tasks. A value of AUC= 0.5 represents random classification, while the value AUC= 1 indicates that all the events have been classified correctly. The values obtained in the DNN employed in this search show a very good performance. In Fig. 7.17 (right) the purity of the sample as a function of the efficiency for each process is presented.

7.5.2 DNN performance

The best models trained on the e+jets and μ+jets channels are applied to data and simulation in the analysis. The output score distributions are shown in Fig. 7.18. A very good classification is obtained with the DNN: in each output node the corresponding SM process has values close to 1, while the other backgrounds have values close to 0. Each event is then categorized exclusively into one of the three classes by taking the highest of the output scores and assigning the event to the corresponding class. In this way, it is

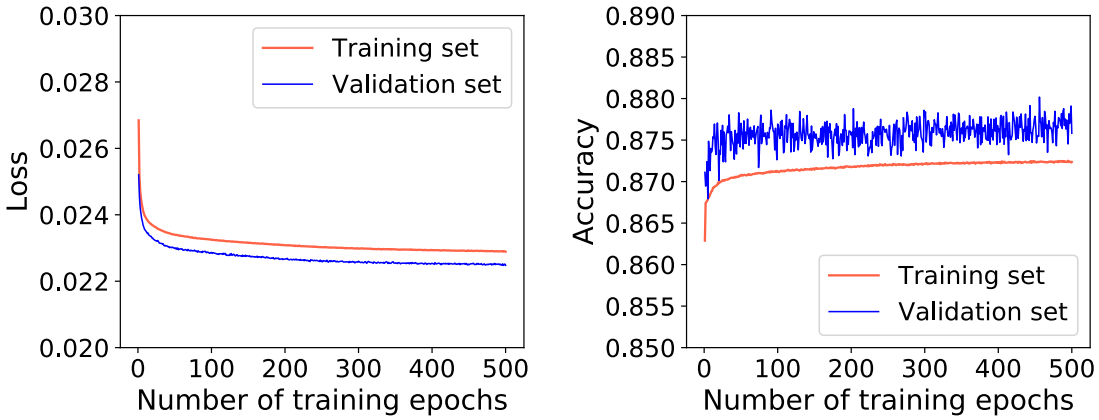

Figure 7.16: The loss (left) and accuracy (right) measured in the DNN training as a function of the number of epochs for the μ+jets channel UL17.

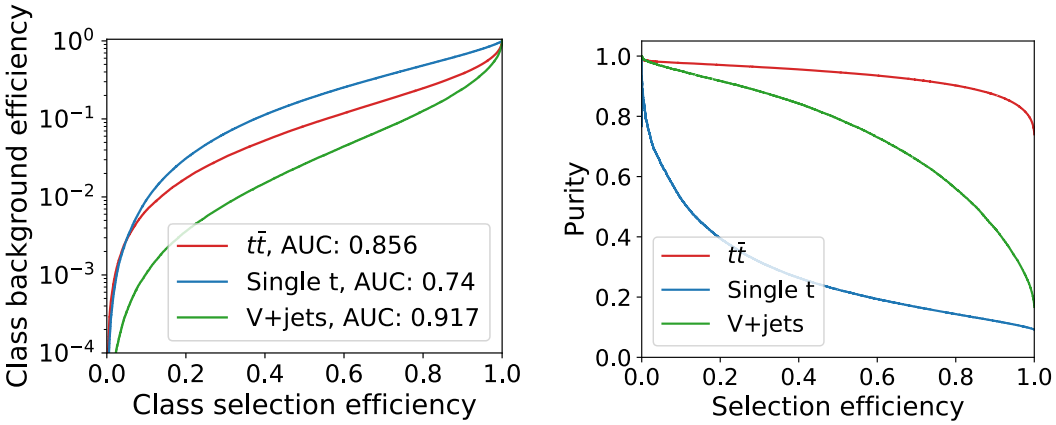

Figure 7.17: Left: The ROC curves of the DNN with the misclassification rate as a function of the efficiency for each process. The values of the AUC are reported. Right: The purity of the samples as a function of the efficiency for each process. DNN trained on the μ+jets channel UL17.

possible to keep the signal selection efficiency high, which is not the case if one-dimensional selections are applied. The three categories of events are used to define the signal and control regions of the search, as will be explained in the following Section.

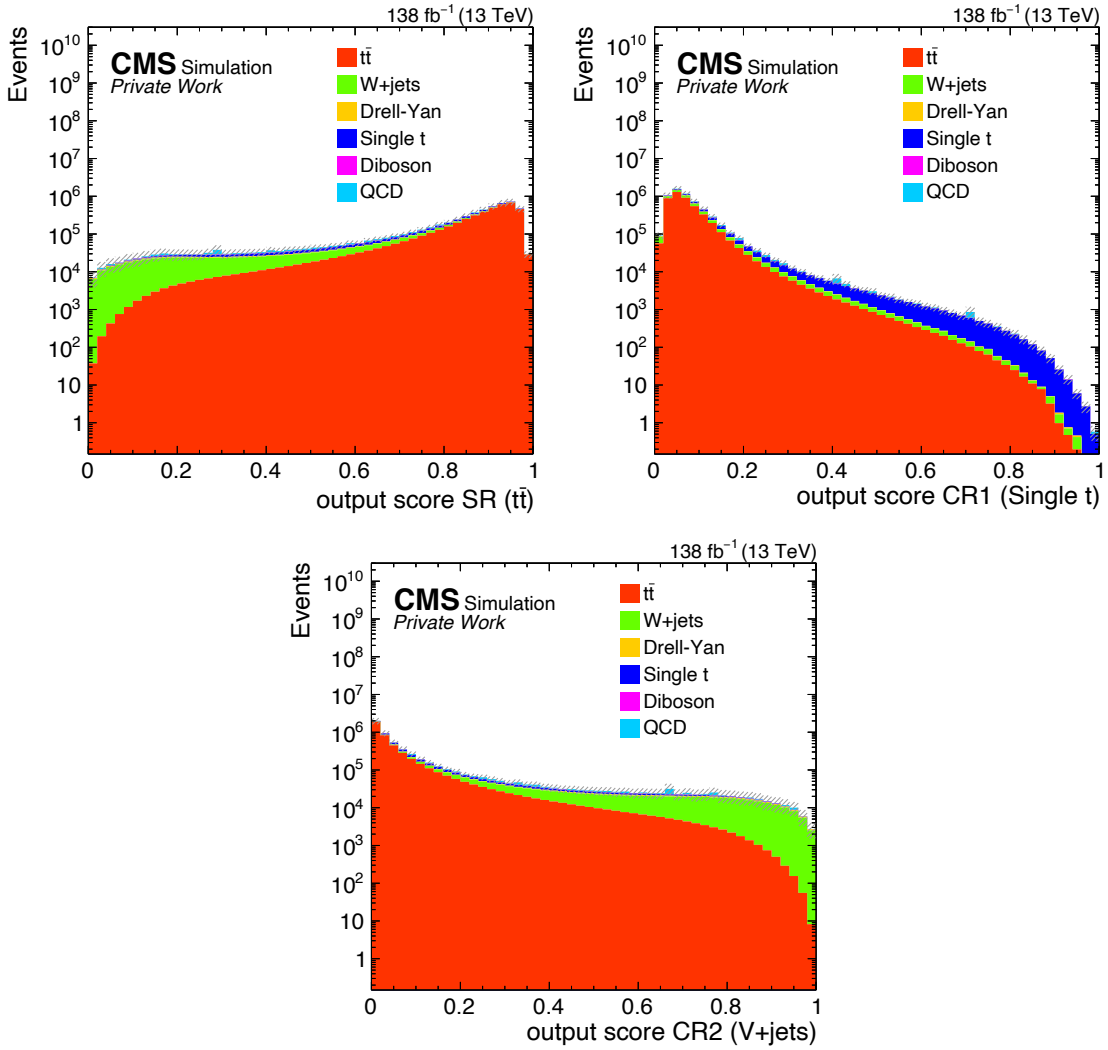

Figure 7.18: The three DNN output scores: the $t\bar{t}$ node (upper left), the single t node (upper right) and the V+jets node (lower).

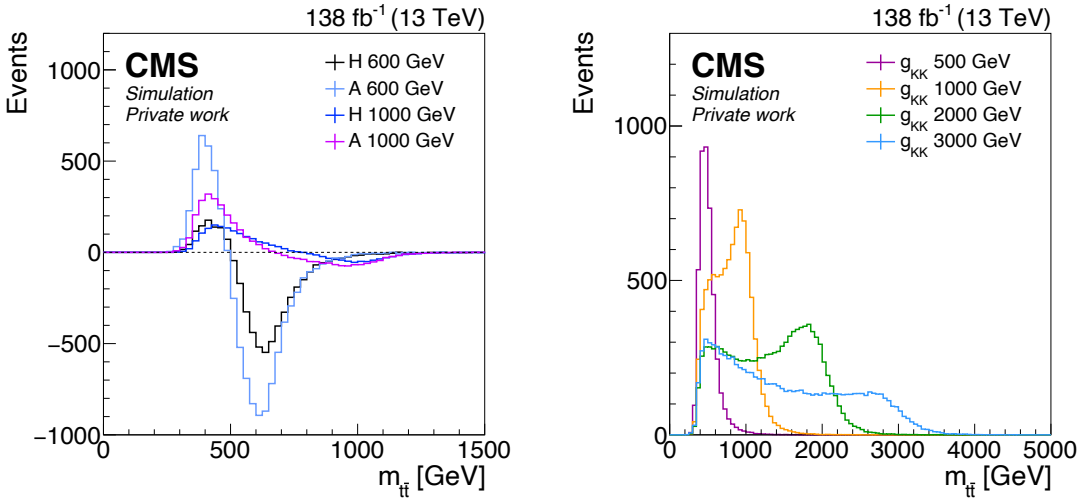

Figure 7.19: The reconstructed $m_{t\bar{t}}$ distribution for scalar H and pseudoscalar A Higgs bosons (left) and for g_{KK} gluons (right) for different values of the new particle mass. Events passing the baseline selection are used.

7.6 Search variables and event categorization

There are two variables that are most sensitive to the presence of a signal decaying to top quarks and are exploited in the search. The first one is the invariant mass of the reconstructed $t\bar{t}$ pair. The SM backgrounds exhibit a falling distribution in $m_{t\bar{t}}$, whereas the signals show a peak in correspondence of the value of the simulated mass. A particular case is given by the heavy Higgs bosons, that present a peak-dip structure, caused by interference effects with the $t\bar{t}$ background. The exact shape follows the interference pattern, which can vary with the mass and width of the heavy Higgs bosons. The reconstructed $m_{t\bar{t}}$ distribution for different signals is presented in Fig. 7.19. Moreover, two categories are considered, that depend on the way the hadronic top quarks are reconstructed: the 1 t-tag category, where t_{had} is reconstructed with a t-tagged AK8 jet, and the 0 t-tag category, where t_{had} is reconstructed with AK4 jets. Figure 7.20 shows the $m_{t\bar{t}}$ distributions for the resolved and boosted categories for the SM backgrounds and two signal processes.

The second variable is $\cos(\theta^*)$, where θ^* is the angle between the momentum of t_{lep}, boosted in the $t\bar{t}$ rest frame, and the momentum of $t\bar{t}$, calculated in the laboratory frame. The distribution of $\cos(\theta^*)$ is shown in Fig. 7.21 for different signal processes. In Fig. 7.22 the same distribution is shown for the SM backgrounds and two signal processes, divided in resolved and boosted categories. A clear shape difference is visible between backgrounds and signals. In the resolved category, the SM $t\bar{t}$ has an asymmetric distribution that peaks at 1, that is partly given by the the s-channel gluon exchange contribution [57]. The signal

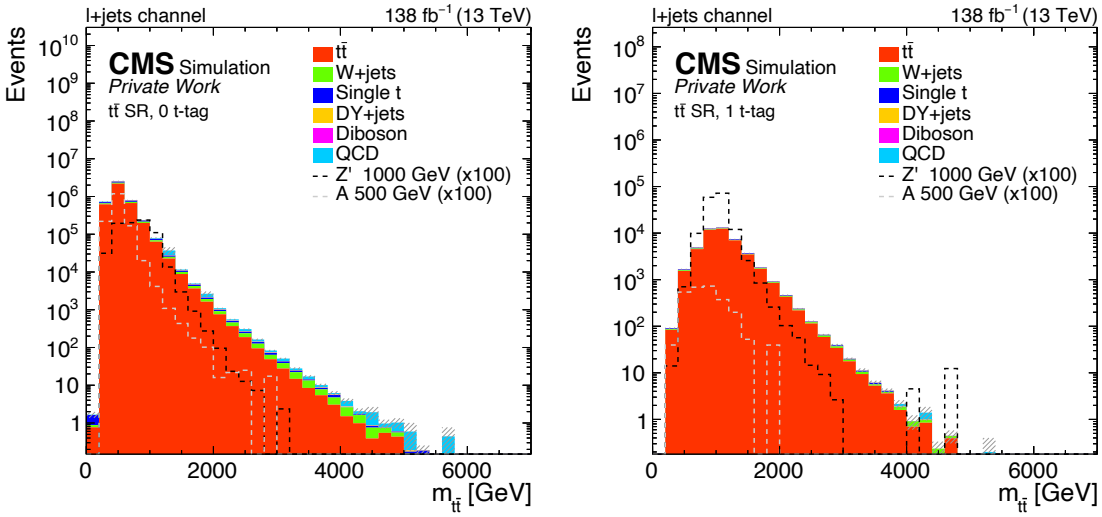

Figure 7.20: The $m_{t\bar{t}}$ distribution for the SM backgrounds, a pseudoscalar A boson with mass of 500 GeV and the Z′ boson with mass of 1000 GeV, for the resolved (left) and boosted (right) regime. Events in the $t\bar{t}$ category are used.

distributions present different structures, and the specific shape depends on the spin and mass of the new particle. In the boosted regime, the bulk of the $\cos(\theta^*)$ is at negative values, with a deficit of events close to 1. Both the $m_{t\bar{t}}$ and $\cos(\theta^*)$ variables are used to fully exploit the differences of the signals with respect to the backgrounds. The final distributions that are used in the statistical analysis are the $m_{t\bar{t}}$ spectra in bins of $\cos(\theta^*)$. Six bins are defined, which have been chosen to maximise the sensitivity of the search, and their edges are $[-1, -0.7, -0.5, 0, 0.5, 0.7, 1]$.

The final event categorization is described in the following. Events are divided into a signal region (SR) and two control regions (CRs) based on the output nodes of the DNN, described in Sec. 7.5. In the SR, dominated by $t\bar{t}$ events, a further cut of $\chi^2 < 30$ is applied (see. Eq. 7.12) to remove misreconstructed top quark pairs. The two CRs are dominated by single t events and V+jets events, respectively. They are used in the fit to data to better constrain the non-$t\bar{t}$ background normalizations. The events in the SR are further divided into categories based on the presence of a t-tagged jet and on the binning in $\cos(\theta^*)$. The distinction in 0 t-tag and 1 t-tag categories is made only for the first four bins in $\cos(\theta^*)$, while it is not the case for the last two bins of $\cos(\theta^*)$ given the lack of events in the 1 t-tag region with high $\cos(\theta^*)$ values (see Fig. 7.22). The final categories used in the statistical interpretation of the results are summarized in Table 7.7.

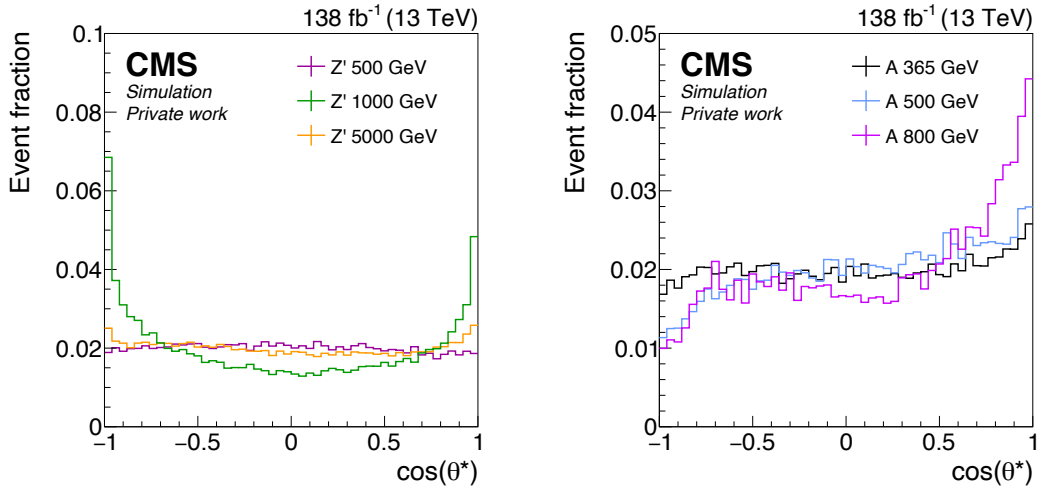

Figure 7.21: The $\cos(\theta^*)$ distribution for the Z′ (left) and pseudoscalar Higgs (right) signals. The relative width of the A bosons is set to 25%. Events passing the baseline selection are used.

7.7 Systematic uncertainties

Different sources of systematic uncertainties can affect the $m_{t\bar{t}}$ distribution in the search. They can have an effect on the normalization of the $t\bar{t}$ mass spectra, on the shapes, or both. For each of the uncertainties, two additional $m_{t\bar{t}}$ distributions are derived, that correspond to the up and down variation by 1 standard deviation (σ) of the given uncertainty. The sources of systematic uncertainties are summarized in Table 7.8 and they are discussed in detail in the following.

- **Integrated luminosity** The integrated luminosity of 137.62 fb^{-1} recorded by the CMS experiment during the 2016-2018 period is assigned a normalization uncertainty of 1.6% [129–131]. The luminosity has a normalization effect only on the $m_{t\bar{t}}$ distribution.

- **SM production cross sections** The following values are used for the uncertainties on the cross sections of SM processes: 20% for $t\bar{t}$ production, 50% for single t production and 50% for W+jets production, to which the subdominant backgrounds (DY+jets, QCD and VV) are added. The values are based on Ref. [75]. They affect only the normalization of the $m_{t\bar{t}}$ distribution and account also for the normalization of the factorization and renormalization scales and PDFs uncertainties. The large values used are a consequence of the poor theoretical modelling of the backgrounds in the highly boosted regime.

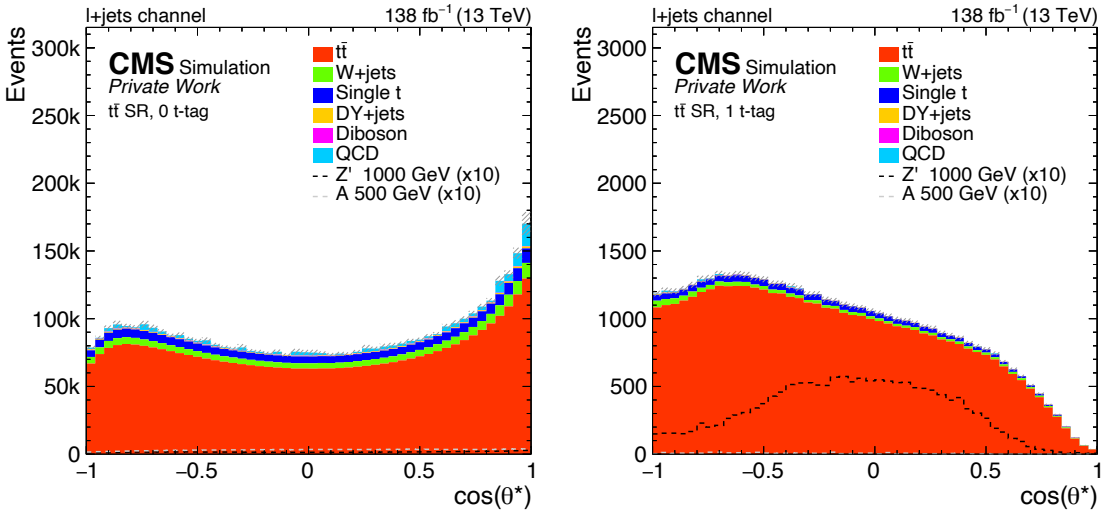

Figure 7.22: The $\cos(\theta^*)$ distribution for the SM backgrounds, a pseudoscalar A boson with mass of 500 GeV and the Z′ boson with mass of 1000 GeV, for the resolved (left) and boosted (right) regime. Events in the $t\bar{t}$ class are used.

- **Pileup reweighting** As described in Sec. 7.2.2, simulated samples are reweighted to match the number of pileup interactions in data. The minimum-bias cross section of 69.2 mb is used [120], and the reweighting is repeated by varying this value by ±4.6%. The uncertainty affects both the shape and normalization.

- **Trigger prefiring** The L1 prefiring weights [121] are applied to simulated events in 2016 and 2017 and the related uncertainty affects the shape and normalization of the $m_{t\bar{t}}$ distributions.

- **Muon and electron efficiencies** The lepton identification, isolation, reconstruction and trigger efficiencies are corrected with the application of dedicated data-to-simulation SFs. The corresponding uncertainties are estimated by varying each scale factor independently by $\pm 1\sigma$. The uncertainties depend on the lepton p_T and η, except for the muon reconstruction uncertainty which depends on the muon p and η. The shape and normalization of the final distributions are affected by these uncertainties.

- **b-tagging** The difference in the b-tagging efficiency in data and simulation is accounted for with the application of SFs on MC samples. The corrections are provided for b-, c- and light-flavour jets. The uncertainty is estimated by varying the SFs within their uncertainties, which are provided as a function of the jet b-tag score, flavour, p_T and η. The shape and normalization of $m_{t\bar{t}}$ are affected.

Region name	DNN node	χ^2 selection	$\cos(\theta^*)$ bin	t-tag category
SRbin1_0T	$t\bar{t}$	$\chi^2 < 30$	$[-1, -0.7]$	0 t-tag
SRbin1_1T	$t\bar{t}$	$\chi^2 < 30$	$[-1, -0.7]$	1 t-tag
SRbin2_0T	$t\bar{t}$	$\chi^2 < 30$	$[-0.7. - 0.5]$	0 t-tag
SRbin2_1T	$t\bar{t}$	$\chi^2 < 30$	$[-0.7, -0.5]$	1 t-tag
SRbin3_0T	$t\bar{t}$	$\chi^2 < 30$	$[-0.5, < 0]$	0 t-tag
SRbin3_1T	$t\bar{t}$	$\chi^2 < 30$	$[-0.5, < 0]$	1 t-tag
SRbin4_0T	$t\bar{t}$	$\chi^2 < 30$	$[0, 0.5]$	0 t-tag
SRbin4_1T	$t\bar{t}$	$\chi^2 < 30$	$[0, 0.5]$	1 t-tag
SRbin5	$t\bar{t}$	$\chi^2 < 30$	$[0.5, 0.7]$	-
SRbin6	$t\bar{t}$	$\chi^2 < 30$	$[0.7, 1]$	-
CR1	single t	-	-	-
CR2	V+jets	-	-	-

Table 7.7: The definition of the categories used in the analysis.

- **t-tagging** The DeepAK8 top tagging efficiency is corrected in simulation with data-to-simulation correction factors as a function of the jet p_T. The related uncertainty in each of the three p_T bins is treated as an unconstrained parameter. The reason is that the dataset in which the SFs are calculated overlaps with the signal region of the analysis. The uncertainty affects both the shape and the normalization of the final distributions.

- **t-mistag rate** The t-mistag rate is measured in a CR in the analysis and applied as a SF. Its value is varied within the uncertainty to obtain the corresponding systematic error. This uncertainty has a shape and normalization effect.

- **Jet energy corrections** The uncertainties on the jet energy scale (JES) and resolution (JER) are obtained by varying the corrections within their uncertainties, simultaneously for AK4 and AK8 jets, and the analysis is repeated using the modified jet energies. The JES uncertainties depend on jet p_T and η, while the JER on the jet η. The variations of AK4 jets are propagated to the Type-I correction of p_T^{miss}. The shape and normalization of $m_{t\bar{t}}$ are affected by the uncertainties.

- **PDFs** The simulated samples are generated using PDFs from the NNPDF 3.1 set [119]. The systematic uncertainty on the choice of the PDF set is estimated by using 100 replicas of the PDFs and constructing 100 corresponding $m_{t\bar{t}}$ distributions. The shape variation is obtained by taking the root-mean-square (RMS) of the replicas in each bin of the distribution with respect to the nominal distribution. To take into

source	uncertainty	type
luminosity	$\pm 1.6\%$	norm
$t\bar{t}$ cross section	$\pm 20\%$	norm
single t cross section	$\pm 50\%$	norm
W+jets + others cross section	$\pm 50\%$	norm
pileup reweighting	$\pm 1\sigma$	norm & shape
trigger prefiring	$\pm 1\sigma$	norm & shape
muon identification	$\pm 1\sigma(p_T, \eta)$	norm & shape
muon reconstruction	$\pm 1\sigma(p, \eta)$	norm & shape
muon isolation	$\pm 1\sigma(p_T, \eta)$	norm & shape
muon trigger	$\pm 1\sigma(p_T, \eta)$	norm & shape
electron identification+isolation	$\pm 1\sigma(p_T, \eta)$	norm & shape
electron reconstruction	$\pm 1\sigma(p_T, \eta)$	norm & shape
electron trigger	$\pm 1\sigma(p_T, \eta)$	norm & shape
b-tagging	$\pm 1\sigma$(b-score, flavour, p_T, η)	norm & shape
t-tagging	unconstrained	norm & shape
t-mistag rate	$\pm 1\sigma$	norm & shape
PDF (signal)	NNPDF 3.1	shape
PDF (backgrounds)	NNPDF 3.1	shape
μ_R (signal)	$\pm 1\sigma$	shape
μ_R (backgrounds)	$\pm 1\sigma$	shape
μ_F (signal)	$\pm 1\sigma$	shape
μ_F (backgrounds)	$\pm 1\sigma$	shape
JES	$\pm 1\sigma(p_T, \eta)$	norm & shape
JER	$\pm 1\sigma(\eta)$	norm & shape

Table 7.8: The list of systematic uncertainties considered in the analysis.

account acceptance effects, the distributions are normalized to the cross sections and the overall normalization variation is included in the SM production cross section uncertainties. The PDF uncertainties are treated as uncorrelated among signal and background processes and among all signal and control regions.

- μ_R **and** μ_F To the choice of the μ_R and μ_F scales used in the sample generation is associated a shape uncertainty. Each scale is varied independently by a factor of 1/2 and 2. As in the case of the PDF uncertainty, the distributions are normalized to take into account the acceptance and only the $m_{\mathrm{t\bar{t}}}$ shape is affected. The normalization effect is included in the SM cross section uncertainties. The μ_R and μ_F uncertainties are treated as uncorrelated among signal and background processes and among all signal and control regions.

7.8 Statistical interpretation

A statistical interpretation of the results is performed to probe the presence of a possible signal. The statistical method used is based on the CL_S method [132]. The likelihood $\mathcal{L}$ is defined as:

$$\mathcal{L}(\mu, \boldsymbol{\nu}) = \prod_i \frac{\lambda_i^{n_i}(\mu, \boldsymbol{\nu})}{n_i!} e^{-\lambda_i(\mu,\boldsymbol{\nu})} \tag{7.15}$$

where n_i is the number of observed events in each bin i of the $m_{\mathrm{t\bar{t}}}$ distribution and λ_i is the expected number of events, μ is the parameter of interest (POI) and $\boldsymbol{\nu}$ is the vector of nuisance parameters. The expected number of events in each bin can be expressed as:

$$\lambda_i(\mu, \boldsymbol{\nu}) = \mu \cdot S_i(\boldsymbol{\nu}) + B_i(\boldsymbol{\nu}) \tag{7.16}$$

with S_i the number of signal events and B_i the number of background events per bin.

For the heavy Higgs boson models, the interference with SM $\mathrm{t\bar{t}}$ has to be explicitly taken into account in the statistical analysis. This leads to the modification of the expected number of events as:

$$\lambda_i(\mu, \boldsymbol{\nu}) = \mu \cdot S_{\mathrm{RES},i}(\boldsymbol{\nu}) + \sqrt{\mu} \cdot S_{\mathrm{INT},i}(\boldsymbol{\nu}) + B_i(\boldsymbol{\nu}) \tag{7.17}$$

where S_{RES} and S_{INT} are the resonant and the interference part of the signal, respectively. In this case, the POI is identified with the 4th power of coupling strength modifier, $\mu = g^4_{\Phi \mathrm{t\bar{t}}}$, with $\Phi = \mathrm{H/A}$.

The Combine tool [133] is used for the statistical analysis. A simultaneous binned maximum-likelihood (ML) fit is performed of the $m_{\mathrm{t\bar{t}}}$ distribution in the categories defined

in Sec. 7.6. The systematic uncertainties described in the previous section are included as nuisance parameters. The uncertainties that affect the normalization only are assigned a log-normal prior distribution, while for the shape uncertainties a Gaussian prior distribution is used. The statistical uncertainty is treated with a simplified version of the Barlow-Beeston approach [134].

Optimizations for statistical analysis

While performing the statistical analysis, a set of checks has been carried out to ensure the robustness of the fitting procedure. Some issues have been encountered during this step, which were due to the complexity of the fit, outlined in the following.

Many signal and control regions are used simultaneously in the fit of $m_{t\bar{t}}$, which cover very different phase space regions, defined by the angular variable $\cos(\theta^*)$ and by the possible presence of a large-radius t-tagged jet. Moreover, the extremely high $m_{t\bar{t}}$ regime analyzed, up to several TeV, is known not to be well described by MC simulation. The main difference among the considered regions was found in the data/MC agreement in the SR1 0 t-tag, which covers the lowest bin in $\cos(\theta^*)$, from -1 to -0.7, and the resolved regime (cf. Appendix C).

In order to allow more flexibility to the fit, the μ_R and μ_F scale uncertainties, the PDF uncertainty and the minor backgrounds cross section uncertainties have been treated as uncorrelated among the signal and control regions and among signal and background processes, as described in Sec. 7.7.

Furthermore, in order to reduce the impact of large statistical fluctuations that can affect the systematics templates or the background contributions, a smoothing procedure has been applied. First, the $m_{t\bar{t}}$ distribution in the single t CR has been smoothed for the QCD background, which shows the largest fluctuations. Secondly, the templates of the jet energy scale, jet energy resolution and pileup variations have been smoothed using a constant function.

Finally, the binning of the $m_{t\bar{t}}$ distributions in the SRs and CRs has been optimized. Different bin values are chosen for the 0 t-tag SRs, the 1 t-tag SR, the CR1 and the CR2, while maintaining the same binning for all signal interpretations. The main difference can be found in the 1 t-tag SRs, where the $m_{t\bar{t}}$ distribution starts at a higher value (600 GeV) with respect to the other regions, starting at 350 GeV, close to the $t\bar{t}$ production threshold. The choice is driven by the lack of statistics at low $m_{t\bar{t}}$ in events with a t-tagged jet.

To validate the soundness of the model after the optimizations applied, a goodness-of-fit (GOF) test has been performed; it evaluates how well the data agree with the predictions from the simulation with a test statistics based on the saturated model [135]. In Fig. 7.23

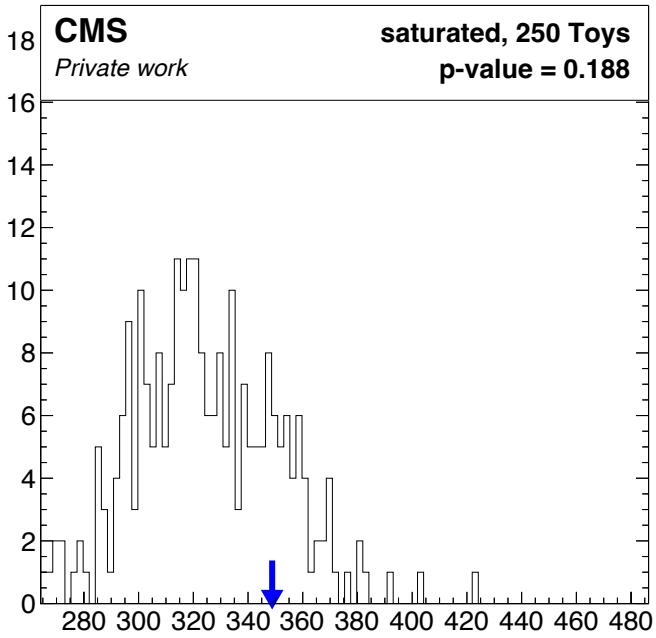

Figure 7.23: The GOF test for a 2 TeV Z′ boson with 10% relative width calculated for 250 toy data. The blue arrow shows the value of the test statistic in data.

the GOF test for a Z′ signal with mass of 2 TeV is presented: it shows the distribution of the test statistic $f(t)$ for 250 generated toy data sets, while the arrow indicates the value of the test statistic in data t_0. The p-value is calculated as $p = \int_{t=t_0}^{+inf} f(t)dt$, and in this case a value of around 19% is obtained, indicating good compatibility.

7.9 Results

The $m_{t\bar{t}}$ distributions in the CRs are shown in Fig. 7.24 after the fit to data. They are used in the statistical analysis to constrain the normalization and shape of the background processes. The $m_{t\bar{t}}$ distributions in the SRs after the fit to data are presented in Figures 7.25 and 7.26. No significant deviation is observed from the SM expectation.

Upper limits at 95% CL are set on the product of cross section times branching fraction $\sigma(pp \to X) \times \mathrm{BR}(X \to t\bar{t})$, for the Z′ and g_{KK} particles. The observed and expected limits are shown in Figures 7.27-7.28 as a function of the new particle's mass. The Z′ bosons are excluded with masses up to 4.3 TeV (3.8 TeV expected), 5.3 TeV (5.4 TeV expected) and 6.7 TeV (6.7 TeV expected) for 1%, 10% and 30% relative widths, respectively. The g_{KK} are excluded with masses up to 4.7 TeV (4.4 TeV expected). These are the most stringent limits to date and improve the previous CMS result [75] (see Sec. 3.3.3), which

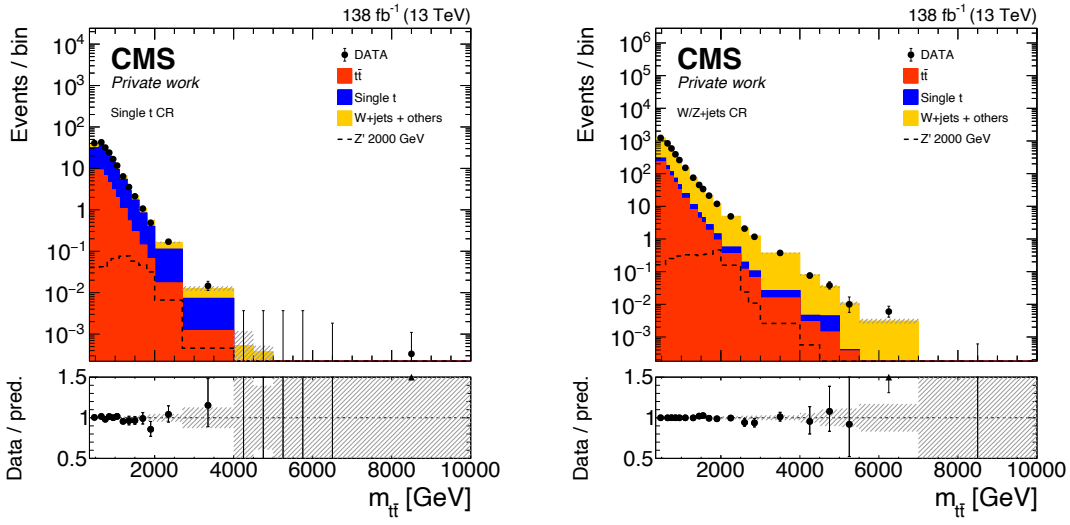

Figure 7.24: The $m_{t\bar{t}}$ distributions in the single t (left) and W/Z+jets (right) CRs after the background-only fit to data. The grey area in the bottom panel corresponds to the total uncertainty on the background prediction.

includes the combination of all three $t\bar{t}$ decay channels, of up to 500 GeV. An example of the comparison of the expected limits with the 2016 $t\bar{t}$ combination is presented in Fig. 7.29 for the Z′ signals with 1% relative width. In general, a better sensitivity is seen especially at low and at high masses, while for masses in the range around $1.5 - 4.5$ TeV, the sensitivity is similar. At lower masses, the great improvement is reached thanks to the improved selections for the resolved regime, due mainly to the inclusion of low-p_T, isolated leptons. The foreseen combination with the other two $t\bar{t}$ decay channels with the Run 2 dataset is expected to improve even more the exclusion limits.

Furthermore, constrains are set on the coupling strength modifiers $g_{\Phi t\bar{t}}$ for the heavy Higgs bosons with 2.5% relative width, shown in Fig. 7.30. The limits are presented for the narrow signals only, while the results for broader signals (10 and 25% relative width) are still under investigation. The reason is the non-monotonic behaviour of CL_S as a function of the parameter of interest, caused by the quartic and quadratic dependence of the likelihood function on $g_{\Phi t\bar{t}}$, which leads to multiple crossings of the $CL_S = 0.05$ threshold used to obtain the upper exclusion limits. Comparing the results with the ones from the resolved CMS analysis [80] (cf. Sec. 3.3.3), which includes the combination of lepton+jets and dilepton finals states, a similar sensitivity is obtained at the highest probed $m_{t\bar{t}}$ masses, while the sensitivity at low values of $m_{t\bar{t}}$ is worse. The resolved CMS analysis [80] presents most stringent limits in the mass range $365 - 1000$ GeV, thanks to the optimization for the resolved regime and the improved theoretical description of the SM $t\bar{t}$ background. The ATLAS result [78] extends the constrains for masses from 1000 GeV up to 1400 GeV.

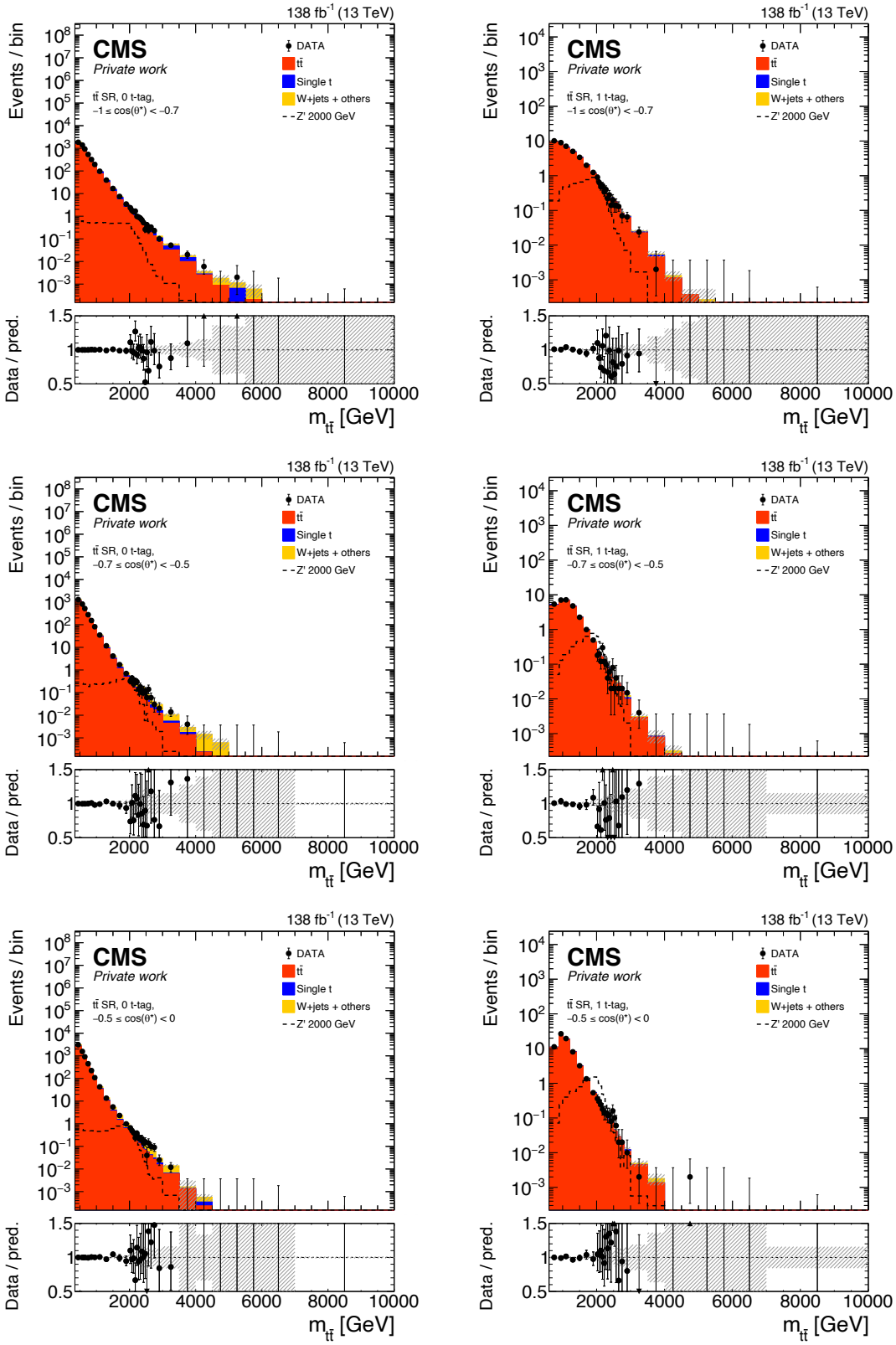

Figure 7.25: The $m_{t\bar{t}}$ distributions in the first three bins of $\cos(\theta^*)$ in the $t\bar{t}$ SR, for events in the resolved (0 t-tag) and boosted (1 t-tag) categories, after the background-only fit to data. The grey area in the bottom panel corresponds to the total uncertainty on the background prediction.

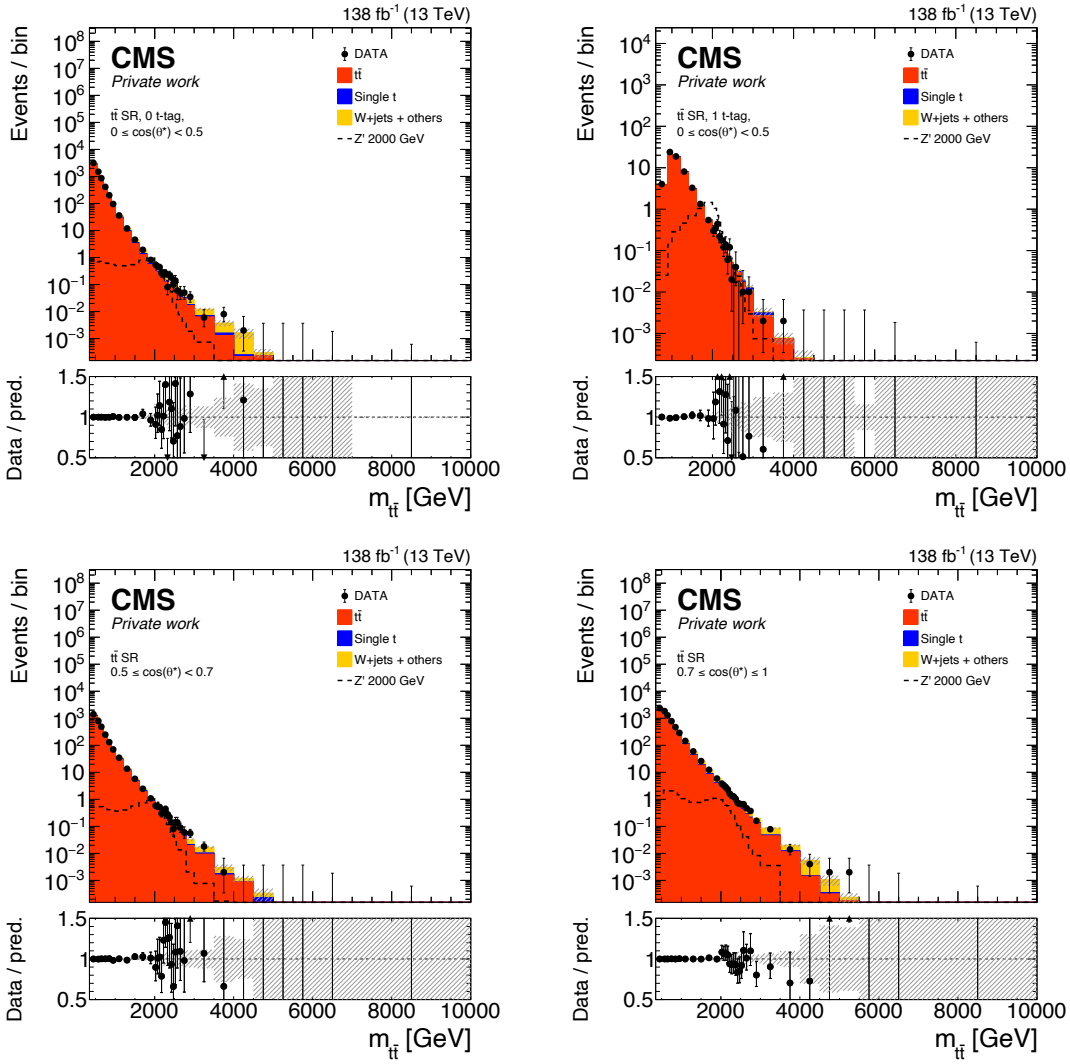

Figure 7.26: The $m_{\mathrm{t\bar{t}}}$ distributions in the last three bins of $\cos(\theta^*)$ in the $\mathrm{t\bar{t}}$ SR after the background-only fit to data. In the region $0 \leq \cos(\theta^*) < 0.5$ (top) events are divided in the resolved (0 t-tag) and boosted (1 t-tag) categories. The grey area in the bottom panel corresponds to the total uncertainty on the background prediction.

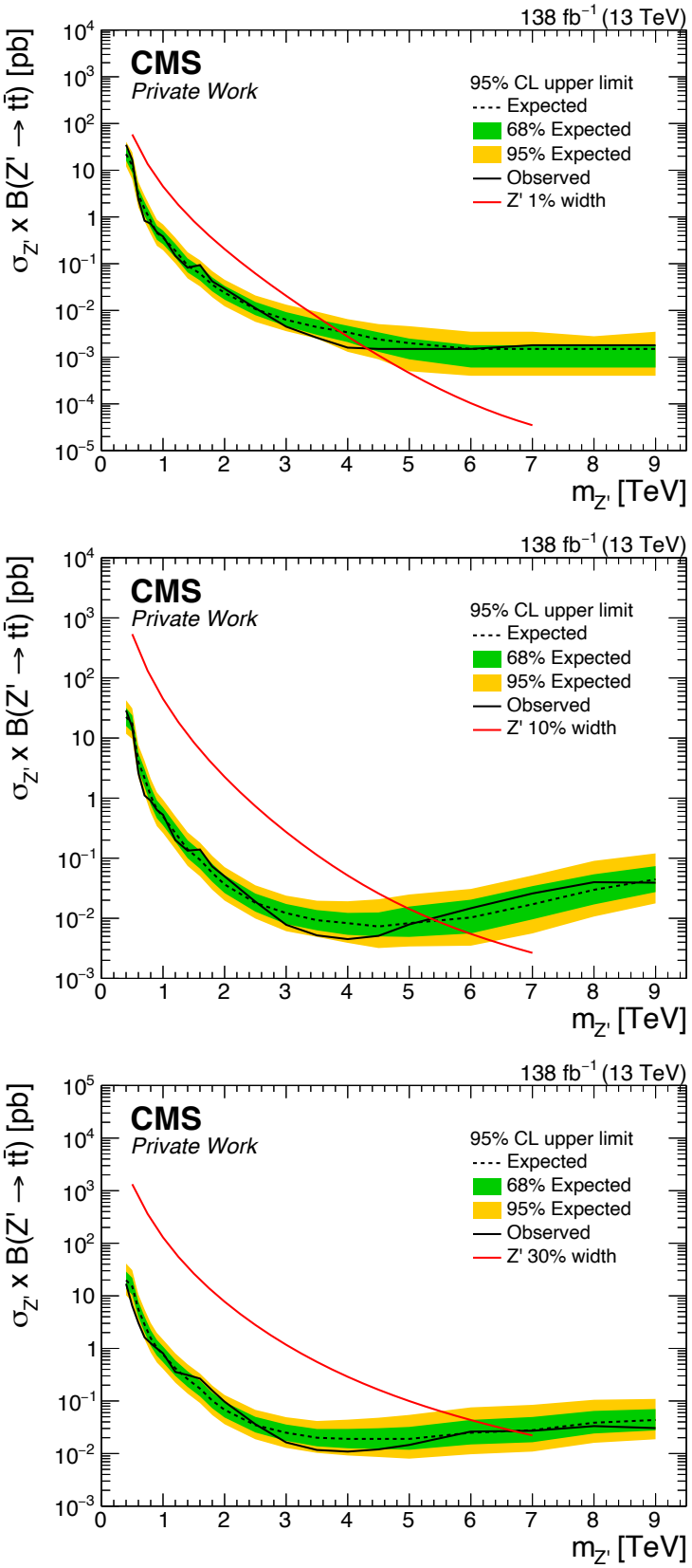

Figure 7.27: Expected and observed upper limits at 95% CL on the production cross section times branching fraction for Z′ bosons with 1% (upper), 10% (middle) and 30% (lower) relative widths, as a function of the Z′ mass.

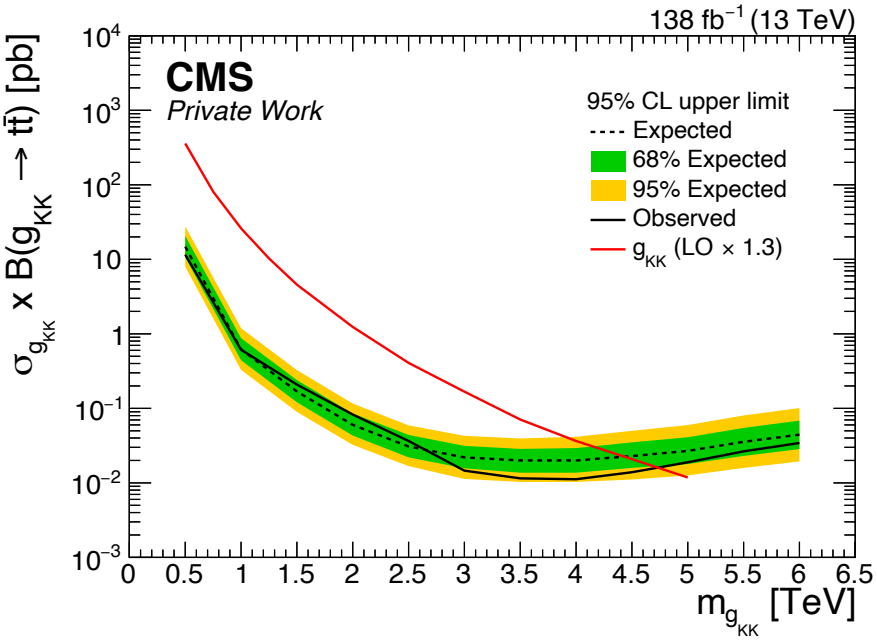

Figure 7.28: Expected and observed upper limits at 95% CL on the production cross section times branching fraction for g_{KK} gluons, as a function of the g_{KK} mass.

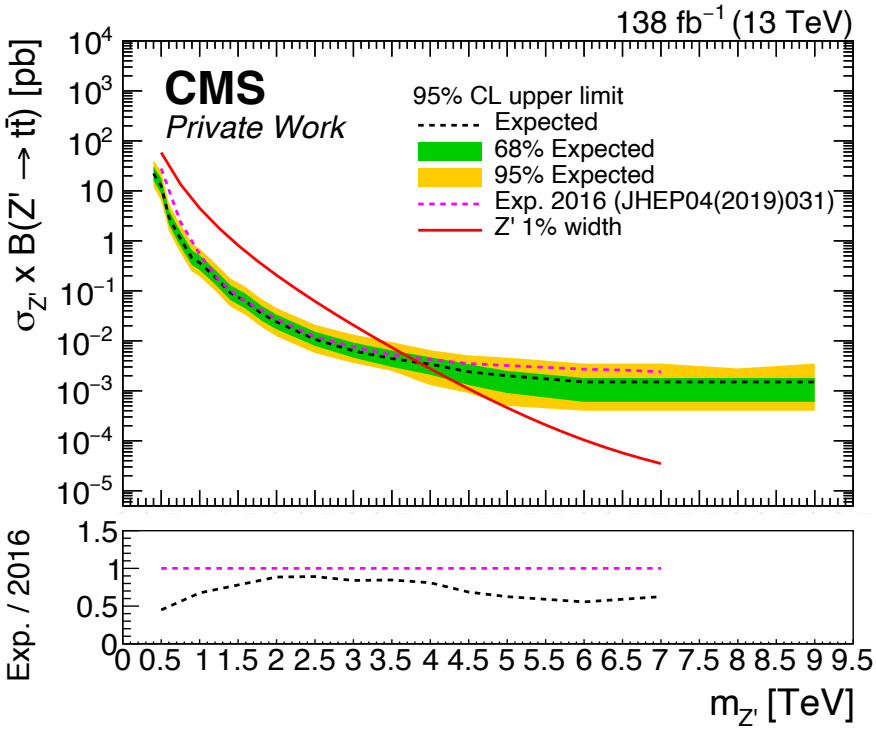

Figure 7.29: Expected upper limits at 95% CL on the production cross section times branching fraction for Z′ bosons with 1% relative width compared to the previous result from Ref. [75], which includes the combination of the three $\mathrm{t\bar{t}}$ decay channels.

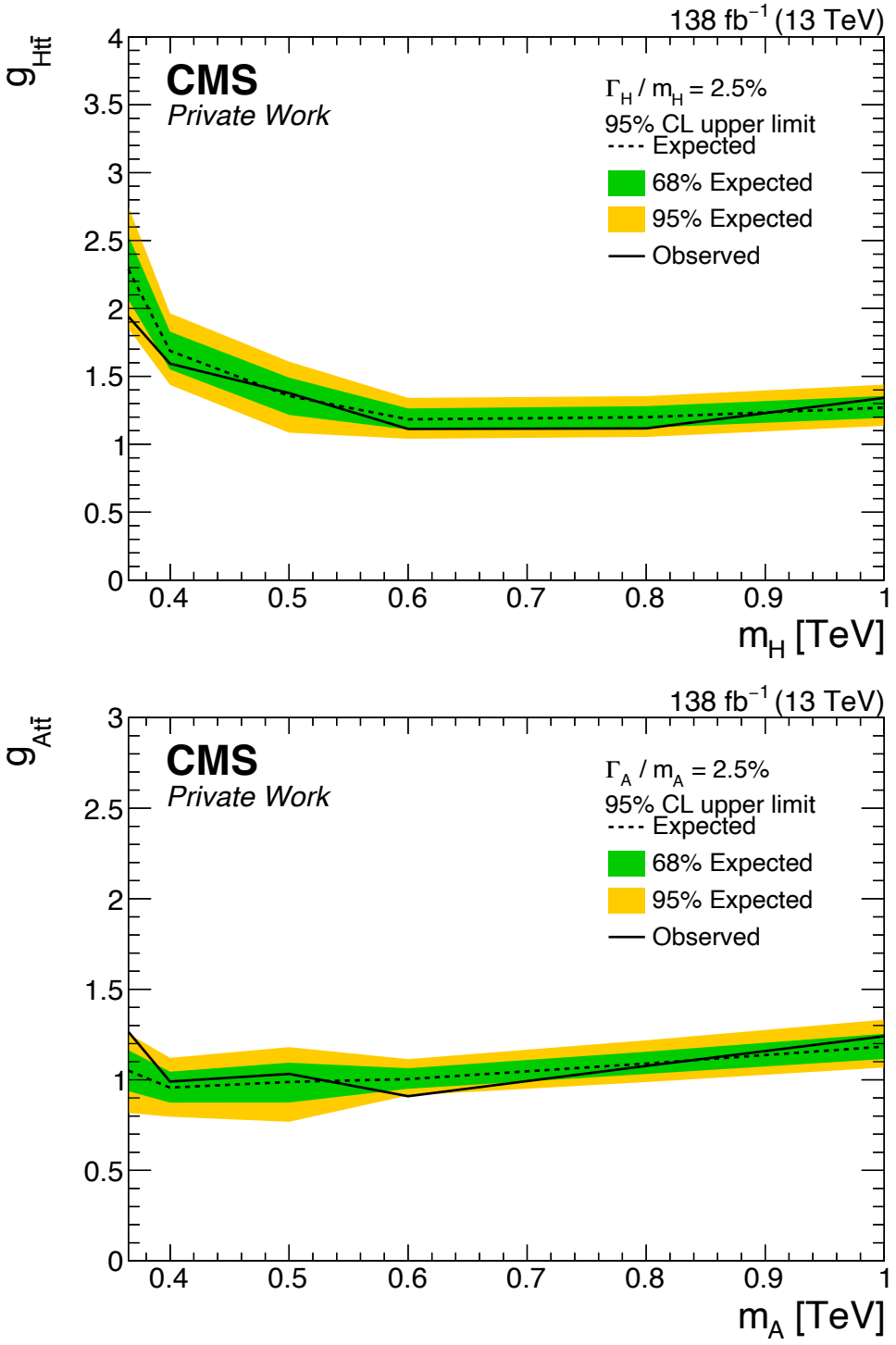

Figure 7.30: Expected and observed upper limits at 95% CL on the coupling strength modifier for scalar H (upper) and pseudoscalar A (lower) Higgs bosons with 2.5% relative width, as a function of the boson mass.

7.10 Summary and outlook

A search has been performed for new particles decaying to top quark pairs in the lepton+jets final state. Data collected at the center-of-mass energy of 13 TeV with the CMS experiment at the LHC during 2016-2018 have been used, which correspond to an integrated luminosity of 138 fb^{-1}. The search is performed in a model-agnostic manner, to be reinterpretable in different BSM models which could manifest themselves with different modification of the $m_{\text{t}\bar{\text{t}}}$ spectrum. The signal models considered include Z′ bosons, Kaluza-Klein gluons and scalar H and pseudoscalar A heavy Higgs bosons. Spin-1 particles appear as resonances in the $m_{\text{t}\bar{\text{t}}}$ spectrum, and at high masses at the multi-TeV scale the non-resonant component becomes more and more important. A different case is presented by spin-0 heavy Higgs bosons, which interfere with the SM $\text{t}\bar{\text{t}}$ background resulting in a peak-dip structure in the $\text{t}\bar{\text{t}}$ mass distribution. The search targets both the resolved and the boosted regimes, in order to be sensitive to particles with low mass, below 1 TeV where scalar and pseudoscalar bosons are predicted, up to several TeV, the highest masses probed in the $\text{t}\bar{\text{t}}$ final state. For the boosted regime, non-isolated leptons are considered and the hadronically decaying top quarks are reconstructed with large-radius jets, which are t-tagged with a novel ML-based technique. To improve the signal efficiency with respect to the previous published results, a DNN for event classification has been trained and applied to the analysis. The aim is to divide all the events into a SR and two CRs, each dominated by a SM process: the SR dominated by $\text{t}\bar{\text{t}}$, the first CR dominated by single t and the second CR dominated by V+jets processes. The CRs are used to constrain the normalization and shape of the backgrounds processes, which are not well modelled by simulation in a highly boosted scenario.

In the SR, the $\text{t}\bar{\text{t}}$ mass distribution is binned in $\cos(\theta^*)$ to exploit the difference in the spin of the new particles and the templates have been used to perform the statistical analysis. No deviation from the SM expectation was observed and upper exclusion limits have been placed. For spin-1 resonances observed and expected exclusion limits at 95% CL are derived on the cross section times branching fraction for Z′ and g_{KK}, and the excluded mass ranges are: $0.4 - 4.3$ TeV ($0.4 - 3.8$ TeV expected), $0.4 - 5.3$ TeV ($0.4 - 5.4$ TeV expected) and $0.4 - 6.7$ TeV ($0.4 - 6.7$ TeV expected) for Z′ bosons with 1%, 10% and 30% relative widths, respectively. The excluded mass range for g_{KK} gluons is $0.5 - 4.7$ TeV ($0.5 - 4.4$ TeV expected). Compared to the previous CMS analysis based on 2016 data [75], which includes the combination of the three $\text{t}\bar{\text{t}}$ decay channels, a similar sensitivity is reached for extra-wide resonances, while it improves for narrow and wide resonances of up to 500 GeV. These results are the most stringent limits to date. Exclusion limits at 95% CL are places on the coupling strength modifier $g_{\Phi\text{t}\bar{\text{t}}}$ for scalar and pseudoscalar Higgs

bosons, for masses in the range $365 - 1000$ TeV and 2.5% relative width.

With the Run 3 at $\sqrt{s} = 13.6$ TeV, CMS is expected to double the recorded data with respect to Run 2. The search with Run 3 data could benefit from the larger dataset and from the improvements presented in this thesis, namely the extension al lower masses, sensitive to a scalar or pseudoscalar particle, the novel ML-based techniques for the top tagging of jets and the event classification with the DNN approach. However, with the higher instantaneous luminosity a higher PU scenario in Run 3 could impair the search, due to a higher jet multiplicity, which would impact especially the combinatorics of the $\mathrm{t\bar{t}}$ reconstruction. The use of the PUPPI algorithm in all CMS analysis in Run 3 is the first step towards the reduction of PU jets, also thanks to the new tune developed for PUPPI for Run 2 UL which is the basis of the tune used in Run 3. An alternative to the χ^2 approach used for the identification of the $\mathrm{t\bar{t}}$ candidates is a fundamental development for future iteration of the analysis. Machine-learning techniques have proven to be extremely powerful tools and could be exploited in the $\mathrm{t\bar{t}}$ reconstruction.

Moreover, looking at the results of the spin-1 resonances, the sensitivity improves with the mass of the probed particle only up to around 5 TeV, in particular for g_{KK} and broad Z' signal (10% and 30% widths). This is due to the non-resonant component of the extremely high mass resonances. Improvements in the reconstruction of the particles in this regime are crucial for future analysis, as well as the study of new possible sensitive variables.

Chapter 8

Conclusions

The Standard Model is the most successful particle physics theory, corroborated over decades by extremely precise measurements of its parameters and by the discovery of all the predicted particles. Nonetheless, many open questions remain, given by shortcomings of the theory and experimental observations in clear contradiction with the predictions. Theories of new physics are postulated to extend the Standard Model and answer to one or more of the open questions. Many theories Beyond the Standard Model are linked to the top quark: being the most massive elementary particle, it is the perfect portal to new physics at high energies. Such theories predict the existence of new heavy particles that modify the $t\bar{t}$ mass spectrum.

In this thesis a search for new heavy particles decaying to $t\bar{t}$ in the lepton+jets final state has been presented. The search is based on pp data collected during 2016-2018 by the CMS experiment at the LHC and correspond to an integrated luminosity of 138 fb^{-1}. A variety of new physics models has been tested, including spin-1 resonances at multi-TeV scale, as Z' bosons and g_{KK} gluons, which present a peak in the $t\bar{t}$ mass spectrum, and spin-0 heavy Higgs bosons, scalar (H) or pseudoscalar (A). The heavy Higgs boson signals interfere with the $t\bar{t}$ production, which can be seen as the peak-dip structure in the $t\bar{t}$ mass spectrum.

Both the resolved and the boosted regimes were considered. The search in the resolved regime has been optimized for H/A bosons at masses below 1 TeV. This regime is characterized by isolated leptons and well separated small-radius jets. On the other hand, in the boosted regime the decay products of the top quarks are collimated, resulting in leptons which are close to jets, thus non-isolated, and in large-radius jets that are used to reconstruct the hadronic decay of the top quarks. The PUPPI algorithm has been used for both small- and large-radius jets: this is one of the few analyses in CMS to use PUPPI for small-radius jets in Run 2. The version of the PUPPI algorithm used, PUPPI v15, has

been developed to optimize the jet energy and p_T^{miss} resolution of PUPPI jets in Run 2 and it is the starting version of the algorithm for Run 3. Given the excellent performance of PUPPI, it is now the official pileup mitigation technique in CMS for Run 3 and beyond.

The sensitivity of the search has been improved with the use of novel, machine-learning based techniques. First, to identify the jets originating from the hadronic decay of top quarks, the DeepAK8 tagger has been used. Moreover, to increase the signal efficiency, a deep neural network for event classification has been developed and applied to the analysis.

Finally, two search variables were used to probe the presence of new particles: the $\mathrm{t\bar{t}}$ mass distribution $m_{\mathrm{t\bar{t}}}$ and the angular variable $\cos(\theta^*)$, which is sensitive to the spin of the decaying particle. No excess over the SM expectation was found and stringent exclusion limits have been placed for the various signal models. Upper exclusion limits have been derived on the cross section times branching fraction for Z' and g_{KK} signals. The excluded mass ranges are: $0.4-4.3$ TeV, $0.4-5.3$ TeV and $0.4-6.7$ TeV for Z' bosons with 1%, 10% and 30% relative widths, respectively, and $0.5-4.7$ TeV for g_{KK} gluon. They correspond to the most stringent limits to date for $\mathrm{t\bar{t}}$ resonances. Furthermore, exclusion limits have been placed on the coupling strength modifiers for scalar and pseudoscalar Higgs bosons, for masses in the range $365-1000$ TeV and 2.5% relative width.

The results presented in the thesis will be combined with the two complementary analyses which target the all hadronic and dileptonic decays of the $\mathrm{t\bar{t}}$ pair using the Run 2 dataset, extending even more the exclusion regions. Given the model-independent approach used in this search, the results can be re-interpreted in many other models which predict a deviation in the $\mathrm{t\bar{t}}$ mass distribution.

Appendix A

Tables of simulated samples

Process	$\sigma \times$ BR	Number of weighted events [10^6]			
	[pb]	UL16preVFP	UL16postVFP	UL17	UL18
$t\bar{t}$ semileptonic	$3.64 \cdot 10^2$	39046.23	42833.98	104665.74	140602.84
$t\bar{t}$ all hadronic	$3.80 \cdot 10^2$	30289.25	33636.22	72581.78	105390.93
$t\bar{t}$ dileptonic	$8.73 \cdot 10^1$	2680.29	3091.96	7545.28	10319.11
W($\to l\nu$)+jets, $70 < H_T < 100$ GeV	$1.27 \cdot 10^3$	16.82	19.17	43.97	65.41
W($\to l\nu$)+jets, $100 < H_T < 200$ GeV	$1.25 \cdot 10^3$	21.27	19.48	46.73	50.90
W($\to l\nu$)+jets, $200 < H_T < 400$ GeV	$3.36 \cdot 10^2$	17.59	14.68	41.67	57.39
W($\to l\nu$)+jets, $400 < H_T < 600$ GeV	$4.52 \cdot 10^1$	2.47	2.09	5.42	7.24
W($\to l\nu$)+jets, $600 < H_T < 800$ GeV	$1.10 \cdot 10^1$	2.29	2.19	5.37	7.53
W($\to l\nu$)+jets, $800 < H_T < 1200$ GeV	$4.94 \cdot 10^0$	2.49	2.06	5.06	7.14
W($\to l\nu$)+jets, $1200 < H_T < 2500$ GeV	$1.16 \cdot 10^0$	2.07	2.06	4.86	6.43
W($\to l\nu$)+jets, $H_T > 2500$ GeV	$2.62 \cdot 10^{-2}$	0.81	0.71	1.19	2.08
DY($\to ll$)+jets, $70 < H_T < 100$ GeV	$1.40 \cdot 10^2$	6.57	5.85	11.97	16.65
DY($\to ll$)+jets, $100 < H_T < 200$ GeV	$1.40 \cdot 10^2$	9.45	8.25	18.46	25.63
DY($\to ll$)+jets, $200 < H_T < 400$ GeV	$3.84 \cdot 10^1$	5.75	5.58	12.23	17.92
DY($\to ll$)+jets, $400 < H_T < 600$ GeV	$5.21 \cdot 10^0$	2.65	2.49	5.38	8.69
DY($\to ll$)+jets, $600 < H_T < 800$ GeV	$1.27 \cdot 10^0$	2.63	2.25	5.18	6.92
DY($\to ll$)+jets, $800 < H_T < 1200$ GeV	$5.68 \cdot 10^{-1}$	2.39	2.32	4.41	6.49
DY($\to ll$)+jets, $1200 < H_T < 2500$ GeV	$1.33 \cdot 10^{-1}$	2.12	1.97	4.68	5.95
DY($\to ll$)+jets, $H_T > 2500$ GeV	$2.98 \cdot 10^{-3}$	0.72	0.70	1.36	1.90
WW	$7.59 \cdot 10^1$	15.74	15.80	15.49	15.46
WZ	$2.76 \cdot 10^1$	7.91	7.54	7.79	7.87
ZZ	$1.21 \cdot 10^1$	1.28	1.15	2.71	3.50
single t/$\bar{t}$ s-channel	$3.36 \cdot 10^0$	19.34	19.26	48.68	67.08
single t t-channel	$1.36 \cdot 10^2$	5839.19	6550.10	13637.10	18666.37
single $\bar{t}$ t-channel	$8.10 \cdot 10^1$	1951.75	1917.63	4382.09	6065.21
single t tW-channel	$1.95 \cdot 10^1$	106.90	106.20	269.09	353.99
single $\bar{t}$ tW-channel	$1.95 \cdot 10^1$	100.96	115.72	268.79	347.37
QCD, $50 < H_T < 100$ GeV	$1.86 \cdot 10^8$	35.73	11.08	26.03	38.23
QCD, $100 < H_T < 200$ GeV	$2.36 \cdot 10^7$	65.50	72.64	53.30	82.21
QCD, $200 < H_T < 300$ GeV	$1.55 \cdot 10^6$	17.97	42.72	42.32	56.30
QCD, $300 < H_T < 500$ GeV	$3.24 \cdot 10^5$	13.59	45.50	42.91	60.99
QCD, $500 < H_T < 700$ GeV	$3.03 \cdot 10^4$	55.50	15.07	35.75	48.64
QCD, $700 < H_T < 1000$ GeV	$6.44 \cdot 10^3$	15.24	13.72	33.65	47.93
QCD, $1000 < H_T < 1500$ GeV	$1.12 \cdot 10^3$	13.56	12.42	10.14	14.24
QCD, $1500 < H_T < 2000$ GeV	$1.08 \cdot 10^2$	9.66	9.24	7.53	10.75
QCD, $H_T > 2000$ GeV	$2.20 \cdot 101$	4.83	4.84	4.09	5.28

Table A.1: List of SM simulated samples used in the analysis. The cross section of each process in given in pb. The weighted number of generated events is given for each data-taking period.

$M_{Z'}$ [GeV] ($\Gamma/m = 1\%$)	σ [pb]	Number of weighted events			
		UL16preVFP	UL16postVFP	UL17	UL18
400	$2.54 \cdot 10^{1}$	545137	570127	195849	200177
500	$2.56 \cdot 10^{1}$	537434	526301	188083	191959
600	$1.78 \cdot 10^{1}$	509112	510991	198225	190965
700	$1.17 \cdot 10^{1}$	444498	444141	197226	199814
800	$7.73 \cdot 10^{0}$	425095	474239	193169	191956
900	$5.16 \cdot 10^{0}$	463051	461916	199844	197077
1000	$3.53 \cdot 10^{0}$	428524	449628	167704	201398
1200	$1.74 \cdot 10^{0}$	464061	497321	195613	198217
1400	$9.03 \cdot 10^{-1}$	481700	548913	208708	217763
1600	$5.0 \cdot 10^{-1}$	534876	548012	219802	207930
1800	$2.83 \cdot 10^{-1}$	495243	533048	208252	203473
2000	$1.66 \cdot 10^{-1}$	457884	527279	211425	196868
2500	$4.70 \cdot 10^{-2}$	456081	508484	203518	201970
3000	$1.49 \cdot 10^{-2}$	482947	490008	192535	171190
3500	$5.09 \cdot 10^{-3}$	466444	426819	189565	187591
4000	$1.90 \cdot 10^{-3}$	422942	429056	184083	184021
4500	$7.64 \cdot 10^{-4}$	413509	417527	174039	179341
5000	$3.22 \cdot 10^{-4}$	109600	90300	191207	193733
6000	$6.06 \cdot 10^{-5}$	107673	91498	202610	192170
7000	$1.15 \cdot 10^{-5}$	107295	89723	192854	197279
8000	$1.81 \cdot 10^{-6}$	105321	89233	191551	170678
9000	$1.93 \cdot 10^{-7}$	106005	91144	182122	201986

Table A.2: List of Z′ signal samples with $\Gamma/m = 1\%$ used in the analysis. The cross section σ in given in pb. The number of generated events is given for each data-taking period.

$M_{Z'}$ [GeV] ($\Gamma/m = 10\%$)	σ [pb]	Number of weighted events			
		UL16preVFP	UL16postVFP	UL17	UL18
400	$2.45 \cdot 10^{0}$	201986	206000	473000	482000
500	$2.44 \cdot 10^{0}$	246000	205000	491000	500000
600	$1.74 \cdot 10^{0}$	244000	230000	494000	467000
700	$1.16 \cdot 10^{0}$	270000	230000	500000	464000
800	$7.75 \cdot 10^{-1}$	266000	230000	434000	482000
900	$5.25 \cdot 10^{-1}$	246000	230000	500000	428000
1000	$3.62 \cdot 10^{-1}$	270000	206000	452000	488000
1200	$1.82 \cdot 10^{-1}$	270000	230000	464000	455000
1400	$9.68 \cdot 10^{-2}$	262000	197000	488000	467000
1600	$5.40 \cdot 10^{-2}$	246000	230000	470000	456000
1800	$3.15 \cdot 10^{-2}$	264000	230000	464000	458000
2000	$1.90 \cdot 10^{-2}$	270000	230000	473000	500000
2500	$5.91 \cdot 10^{-3}$	267000	230000	462000	476000
3000	$2.11 \cdot 10^{-3}$	270000	230000	467000	470000
3500	$8.53 \cdot 10^{-4}$	246000	182000	500000	452000
4000	$3.89 \cdot 10^{-4}$	267000	202000	500000	476000
4500	$1.97 \cdot 10^{-4}$	270000	230000	497000	497000
5000	$1.08 \cdot 10^{-4}$	108000	72000	188000	194000
6000	$4.16 \cdot 10^{-5}$	88000	92000	192000	200000
7000	$1.93 \cdot 10^{-5}$	98000	92000	200000	200000
8000	$1.05 \cdot 10^{-5}$	84000	90000	166000	200000
9000	$6.30 \cdot 10^{-6}$	108000	72000	200000	194000

Table A.3: List of Z′ signal samples with $\Gamma/m = 10\%$ used in the analysis. The cross section σ in given in pb. The number of generated events is given for each data-taking period.

$M_{Z'}$ [GeV]	σ [pb]	Number of weighted events			
($\Gamma/m = 30\%$)		UL16preVFP	UL16postVFP	UL17	UL18
400	$7.35 \cdot 10^{-1}$	270000	230000	476000	482000
500	$6.74 \cdot 10^{-1}$	270000	182000	446000	485000
600	$4.84 \cdot 10^{-1}$	270000	228000	452000	497000
700	$3.29 \cdot 10^{-1}$	270000	230000	500000	494000
800	$2.24 \cdot 10^{-1}$	246000	230000	452000	500000
900	$1.54 \cdot 10^{-1}$	267000	225000	437000	407000
1000	$1.08 \cdot 10^{-1}$	244000	229000	500000	363000
1200	$5.61 \cdot 10^{-2}$	246000	228000	452000	465000
1400	$3.08 \cdot 10^{-2}$	270000	230000	481000	452000
1600	$1.78 \cdot 10^{-2}$	228000	230000	497000	476000
1800	$1.07 \cdot 10^{-2}$	256000	228000	486000	476000
2000	$6.69 \cdot 10^{-3}$	249000	230000	476000	500000
2500	$2.34 \cdot 10^{-3}$	247000	206000	455000	465000
3000	$9.52 \cdot 10^{-4}$	270000	230000	464000	408000
3500	$4.40 \cdot 10^{-4}$	270000	221000	500000	476000
4000	$2.27 \cdot 10^{-4}$	246000	228000	500000	500000
4500	$1.28 \cdot 10^{-4}$	268000	230000	473000	500000
5000	$7.69 \cdot 10^{-5}$	108000	88000	200000	200000
6000	$3.30 \cdot 10^{-5}$	108000	90000	200000	200000
7000	$1.68 \cdot 10^{-5}$	108000	86000	200000	200000
8000	$9.52 \cdot 10^{-6}$	108000	92000	200000	197000
9000	$5.78 \cdot 10^{-6}$	108000	92000	176000	194000

Table A.4: List of Z′ signal samples with $\Gamma/m = 30\%$ used in the analysis. The cross section σ in given in pb. The number of generated events is given for each data-taking period.

$M_{g_{KK}}$ [GeV]	σ [pb]	Number of weighted events			
($\Gamma/m \sim m/6$)		UL16preVFP	UL16postVFP	UL17	UL18
500	$2.93 \cdot 10^{2}$	235000	200000	500000	500000
1000	$2.10 \cdot 10^{1}$	244000	250000	485000	479000
1500	$3.69 \cdot 10^{0}$	240000	228000	500000	500000
2000	$9.37 \cdot 10^{-1}$	250000	208000	494000	494000
2500	$3.05 \cdot 10^{-1}$	213000	196000	497000	488000
3000	$1.17 \cdot 10^{-1}$	241000	248000	500000	500000
3500	$5.18 \cdot 10^{-2}$	226000	250000	494000	500000
4000	$2.55 \cdot 10^{-2}$	250000	239000	473000	484000
4500	$1.42 \cdot 10^{-2}$	250000	249000	440000	479000
5000	$8.47 \cdot 10^{-3}$	250000	234000	500000	476000
5500	$5.55 \cdot 10^{-3}$	214000	248000	494000	479000
6000	$3.82 \cdot 10^{-3}$	250000	242000	476000	431000

Table A.5: List of g_{KK} signal samples used in the analysis. The cross section σ in given in pb. The number of generated events is given for each data-taking period.

M_H [GeV] ($\Gamma/m = 2.5\%$)	type	σ [pb]	Number of weighted events			
			UL16preVFP	UL16postVFP	UL17	UL18
365	res	$2.02 \cdot 10^{-1}$	44001	45491	86586	80113
	int	$-1.09 \cdot 10^{0}$	-243892	-244752	-455312	-464474
400	res	$5.48 \cdot 10^{-1}$	110433	119834	241956	221418
	int	$-1.08 \cdot 10^{0}$	-217545	-244155	-407172	-381995
500	res	$7.71 \cdot 10^{-1}$	174249	174337	348465	345675
	int	$-4.74 \cdot 10^{-1}$	-98109	-110955	-198807	-199995
600	res	$5.06 \cdot 10^{-1}$	112831	114780	207171	202687
	int	$-1.33 \cdot 10^{-1}$	-28958	-33768	-57695	-54322
800	res	$1.58 \cdot 10^{-1}$	34413	35825	56638	68839
	int	$4.79 \cdot 10^{-2}$	10449	10966	20316	18754
1000	res	$5.01 \cdot 10^{-2}$	9174	11358	22709	22164
	int	$6.00 \cdot 10^{-2}$	13776	13125	24490	23711

Table A.6: List of scalar Higgs boson (H) signal samples with $\Gamma/m = 2.5\%$ used in the analysis. The signals are split into resonant and interference parts. The cross section of each process in given, together with the weighted number of generated events for each data-taking period.

M_H [GeV] ($\Gamma/m = 10\%$)	type	σ [pb]	Number of weighted events			
			UL16preVFP	UL16postVFP	UL17	UL18
365	res	$7.34 \cdot 10^{-2}$	16270	16550	32683	33073
	int	$-9.42 \cdot 10^{-1}$	-190509	-210240	-421756	-381134
400	res	$1.35 \cdot 10^{-1}$	30319	30309	54844	57476
	int	$-9.36 \cdot 10^{-1}$	-211992	-212560	-404689	-416738
500	res	$1.73 \cdot 10^{-1}$	39022	38889	77994	70544
	int	$-4.67 \cdot 10^{-1}$	-107012	-109145	-212459	-205383
600	res	$1.16 \cdot 10^{-1}$	26205	26194	52100	51118
	int	$-1.60 \cdot 10^{-1}$	-37050	-38853	-70831	-68068
800	res	$3.82 \cdot 10^{-2}$	8642	8644	17187	16460
	int	$2.35 \cdot 10^{-2}$	4014	3860	8654	9065
1000	res	$1.28 \cdot 10^{-2}$	2894	2894	5509	4920
	int	$4.53 \cdot 10^{-2}$	8761	9822	19996	19102

Table A.7: List of scalar Higgs boson (H) signal samples with $\Gamma/m = 10\%$ used in the analysis. The signals are split into resonant and interference parts. The cross section of each process in given, together with the weighted number of generated events for each data-taking period.

M_H [GeV] ($\Gamma/m = 25\%$)	type	σ [pb]	Number of weighted events			
			UL16preVFP	UL16postVFP	UL17	UL18
365	res	$3.99 \cdot 10^{-2}$	8436	9012	17016	17629
	int	$-7.62 \cdot 10^{-1}$	-170319	-171810	-327978	-339091
400	res	$5.31 \cdot 10^{-2}$	11974	4893	23944	21670
	int	$-7.43 \cdot 10^{-1}$	-158972	-168692	-270072	-296150
500	res	$5.71 \cdot 10^{-2}$	12708	12912	25484	24569
	int	$-4.27 \cdot 10^{-1}$	-88237	-97902	-141874	-165434
600	res	$3.89 \cdot 10^{-2}$	8736	8800	14220	16750
	int	$-1.88 \cdot 10^{-1}$	-43744	-44497	-73627	-68137
800	res	$1.39 \cdot 10^{-2}$	2788	3127	5951	5928
	int	$-1.47 \cdot 10^{-2}$	-3450	-4075	-7511	-6125
1000	res	$5.04 \cdot 10^{-3}$	1114	1115	2174	2284
	int	$1.95 \cdot 10^{-2}$	3947	3867	8503	7327

Table A.8: List of scalar Higgs boson (H) signal samples with $\Gamma/m = 25\%$ used in the analysis. The signals are split into resonant and interference parts. The cross section of each process in given, together with the weighted number of generated events for each data-taking period.

M_A [GeV] ($\Gamma/m = 2.5\%$)	type	σ [pb]	Number of weighted events			
			UL16preVFP	UL16postVFP	UL17	UL18
365	res	$5.87 \cdot 10^{0}$	1188521	1303406	2357012	2607046
	int	$-6.09 \cdot 10^{0}$	-1382652	-1390763	-1779358	- 2770825
400	res	$6.04 \cdot 10^{0}$	1274403	1345760	2646386	2380828
	int	$-3.72 \cdot 10^{0}$	-772319	-839949	-1483967	-1576405
500	res	$2.86 \cdot 10^{0}$	647811	648334	1256146	1295688
	int	$-7.87 \cdot 10^{-1}$	-183156	-195531	-374396	-358622
600	res	$1.24 \cdot 10^{0}$	278598	280878	450471	534780
	int	$-2.81 \cdot 10^{-2}$	-5320	-10244	-13291	-22172
800	res	$2.74 \cdot 10^{-1}$	62120	29574	119062	124293
	int	$2.08 \cdot 10^{-1}$	47686	46310	78433	83393
1000	res	$7.54 \cdot 10^{-2}$	16832	16626	34006	31350
	int	$1.76 \cdot 10^{-1}$	39880	39098	74684	77868

Table A.9: List of pseudoscalar Higgs boson (A) signal samples with $\Gamma/m = 2.5\%$ used in the analysis. The signals are split into resonant and interference parts. The cross section of each process in given, together with the weighted number of generated events for each data-taking period.

M_A [GeV] ($\Gamma/m = 10\%$)	type	σ [pb]	Number of weighted events			
			UL16preVFP	UL16postVFP	UL17	UL18
365	res	$1.19 \cdot 10^{0}$	256295	145074	523391	529368
	int	$-4.93 \cdot 10^{0}$	-1112973	-1117514	-2090114	-2171094
400	res	$1.28 \cdot 10^{0}$	288659	286415	574028	546129
	int	$-3.44 \cdot 10^{0}$	-786867	-789101	-1428088	1579689
500	res	$6.61 \cdot 10^{-1}$	149616	149630	284804	296687
	int	$-9.37 \cdot 10^{-1}$	-219111	-227202	-421949	394355
600	res	$2.99 \cdot 10^{-1}$	60298	67583	135700	134971
	int	$-1.67 \cdot 10^{-1}$	-37340	-44194	-79016	-81901
800	res	$7.10 \cdot 10^{-2}$	16088	16094	32174	31988
	int	$1.35 \cdot 10^{-1}$	29431	28100	58928	60225
1000	res	$2.09 \cdot 10^{-2}$	4745	4748	9039	9207
	int	$1.37 \cdot 10^{-1}$	30484	30112	60562	60547

Table A.10: List of pseudoscalar Higgs boson (A) signal samples with $\Gamma/m = 10\%$ used in the analysis. The signals are split into resonant and interference parts. The cross section of each process in given, together with the weighted number of generated events for each data-taking period.

M_A [GeV] ($\Gamma/m = 25\%$)	type	σ [pb]	Number of weighted events			
			UL16preVFP	UL16postVFP	UL17	UL18
365	res	$3.84 \cdot 10^{-1}$	64681	82251	157752	163270
	int	$3.51 \cdot 10^{0}$	-717781	-795880	-1357976	-1358396
400	res	$3.93 \cdot 10^{-1}$	73648	88492	168613	151476
	int	$2.81 \cdot 10^{0}$	-639506	-643623	-1002534	-1154813
500	res	$2.23 \cdot 10^{-1}$	50395	50203	91091	91719
	int	$-1.06 \cdot 10^{0}$	-195455	-246767	-417111	-394124
600	res	$1.08 \cdot 10^{-1}$	21998	23560	46447	44068
	int	$-3.55 \cdot 10^{-1}$	-67339	-77251	-141218	-141738
800	res	$2.87 \cdot 10^{-2}$	6505	6501	12385	12387
	int	$1.32 \cdot 10^{-2}$	2077	1220	2981	3543
1000	res	$9.36 \cdot 10^{-3}$	1916	2069	3817	4039
	int	$6.31 \cdot 10^{-2}$	11379	13722	25293	23694

Table A.11: List of pseudoscalar Higgs boson (A) signal samples with $\Gamma/m = 25\%$ used in the analysis. The signals are split into resonant and interference parts. The cross section of each process in given, together with the weighted number of generated events for each data-taking period.

Appendix B

Deep neural network input variables

B.1 Distributions of the DNN input variables for the μ+jets channel

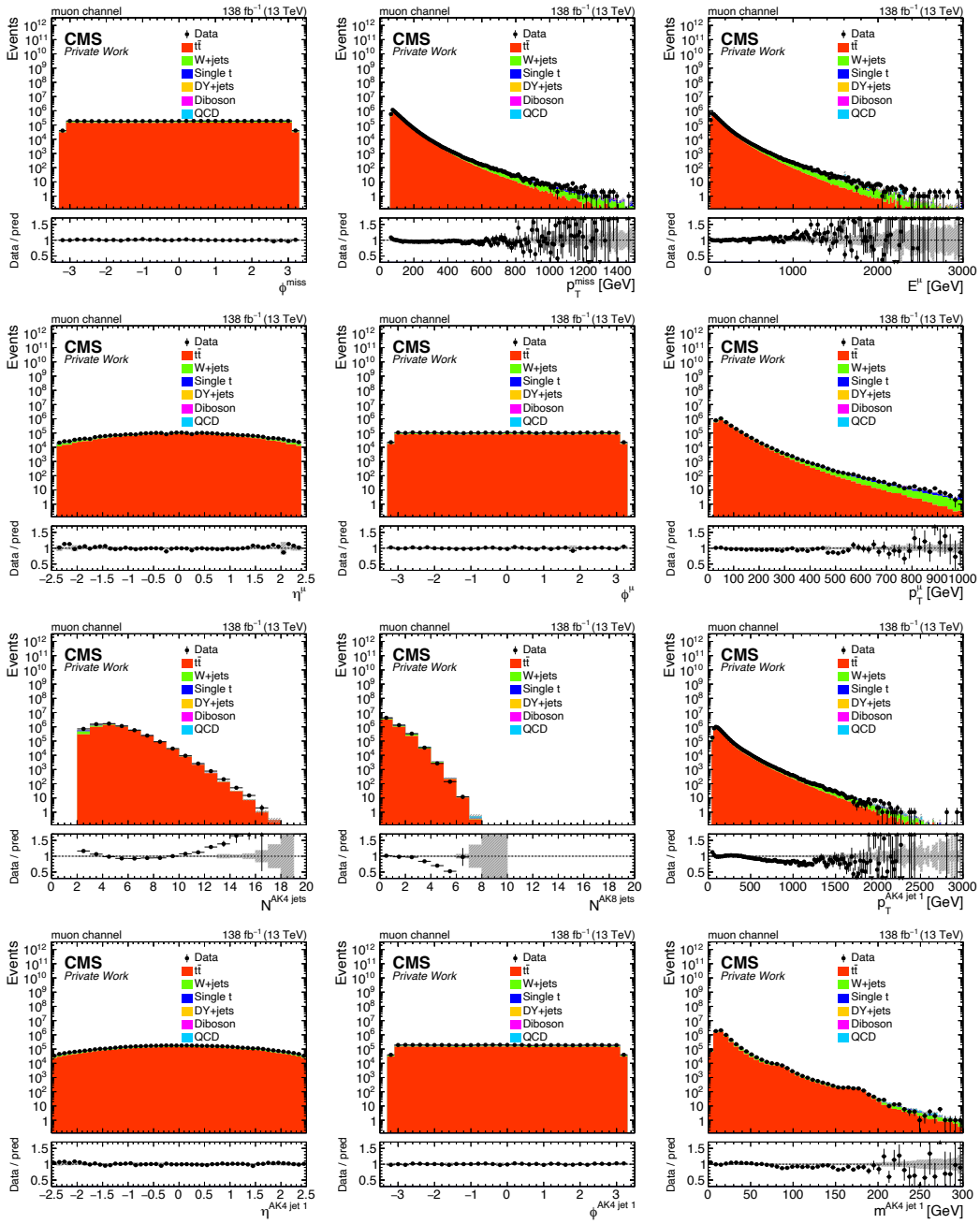

Figure B.1: The distributions of p_T^{miss}, μ, number of jets and first AK4 jet for the μ+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

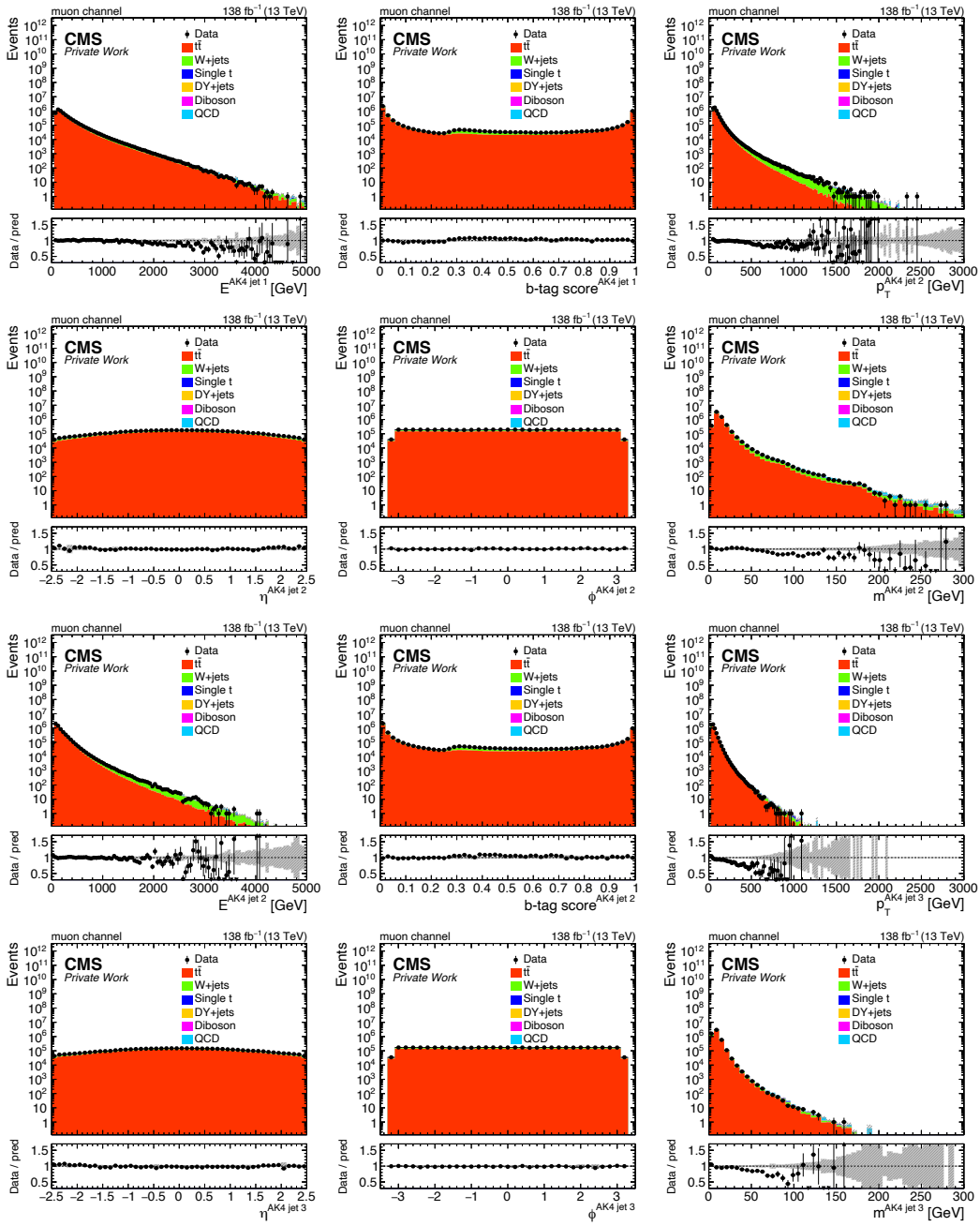

Figure B.2: The distributions of the first, second and third AK4 jet for the μ+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

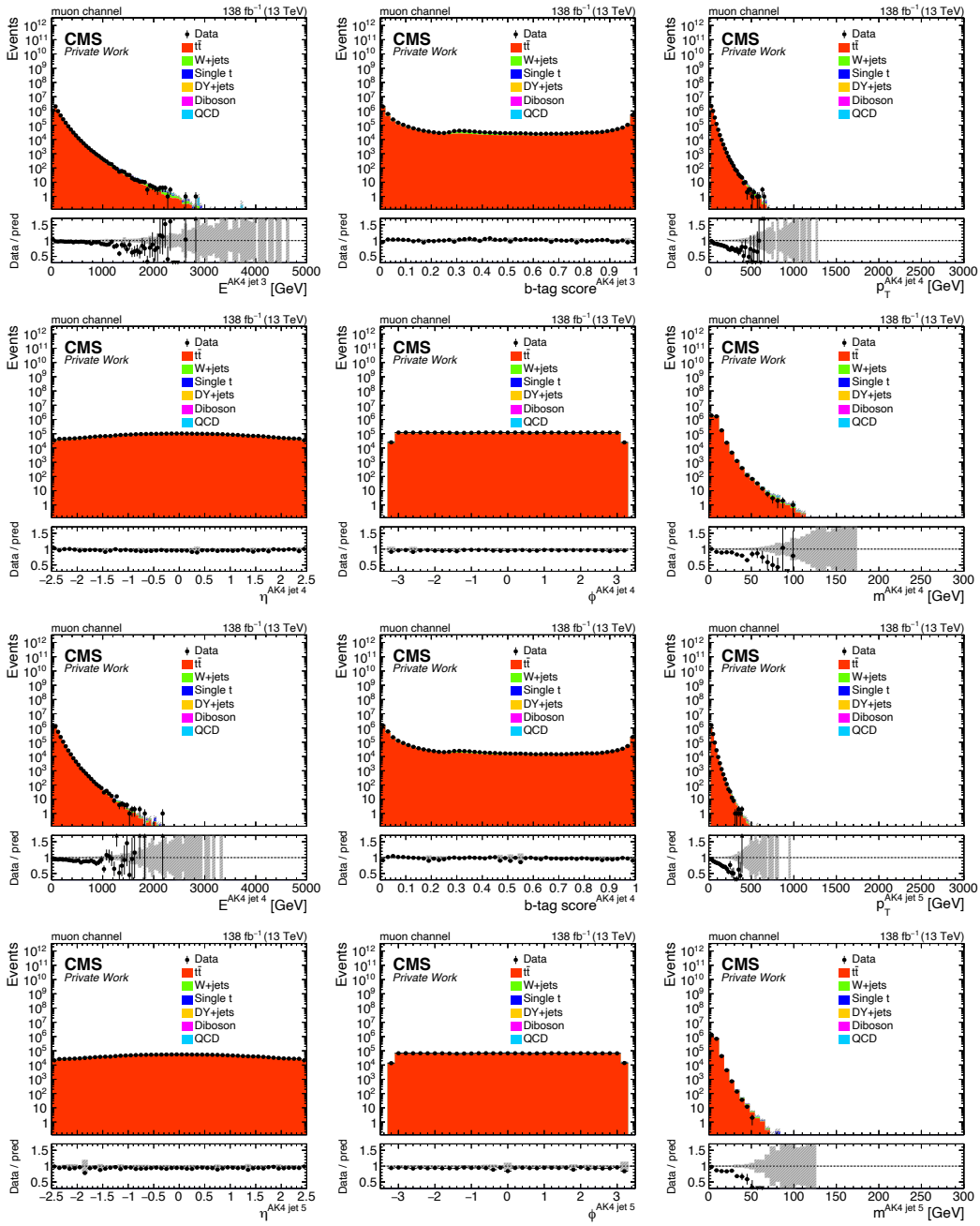

Figure B.3: The distributions of the third, fourth and fifth AK4 jet for the μ+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

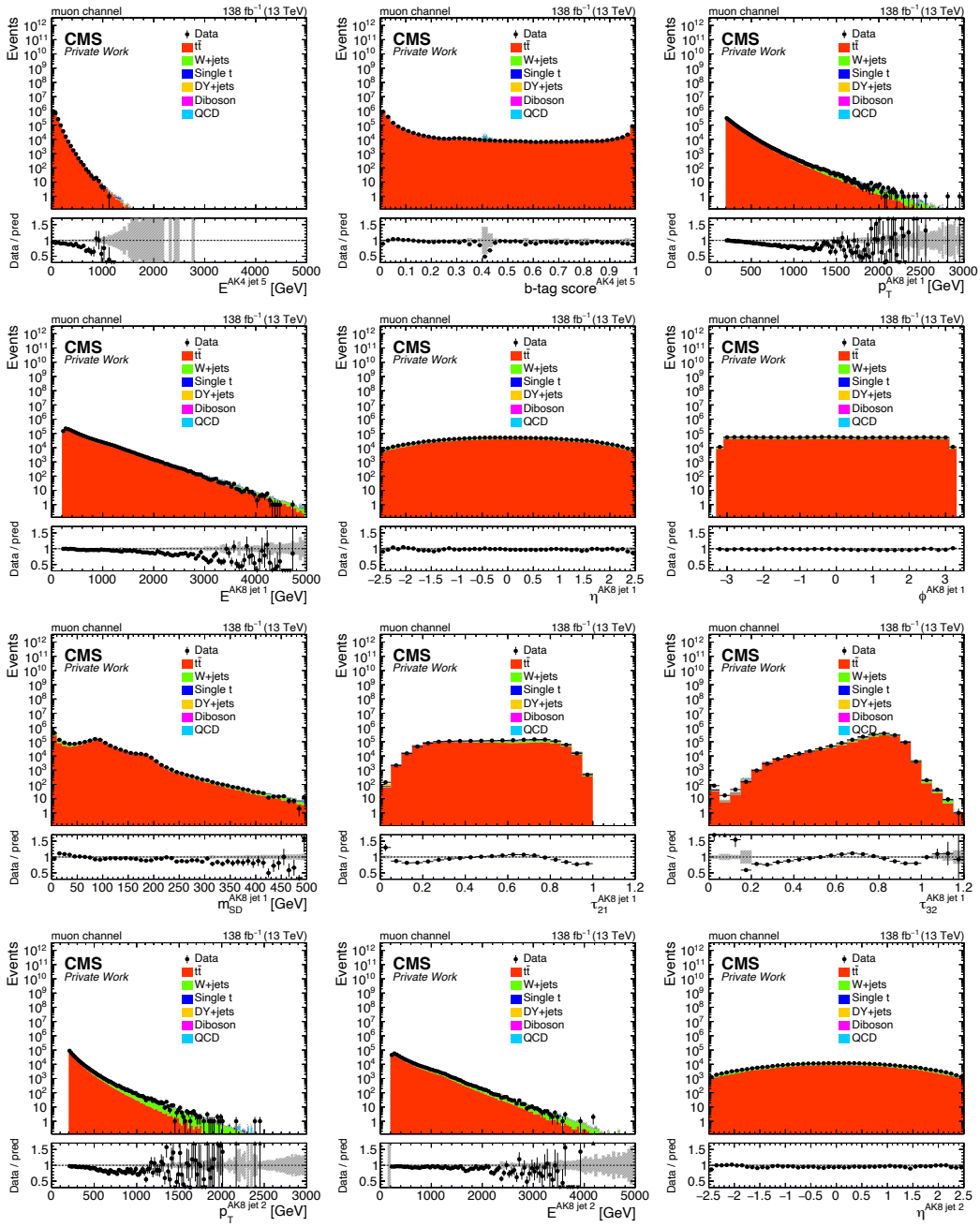

Figure B.4: The distributions of the fifth AK4 jet and the first AK8 jet for the μ+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

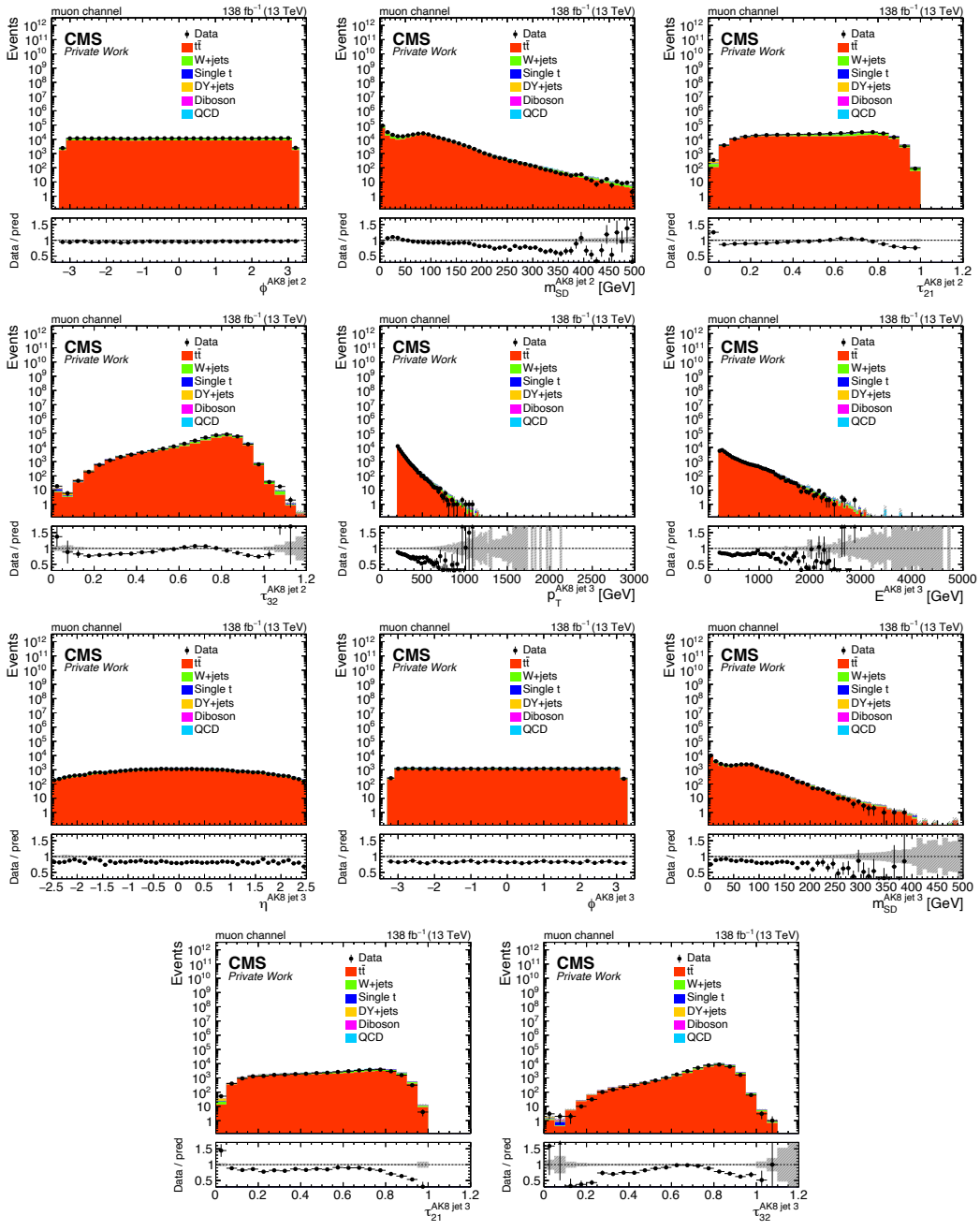

Figure B.5: The distributions of the second and third AK8 jet for the μ+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

B.2 Distributions of the DNN input variables for the e+jets channel

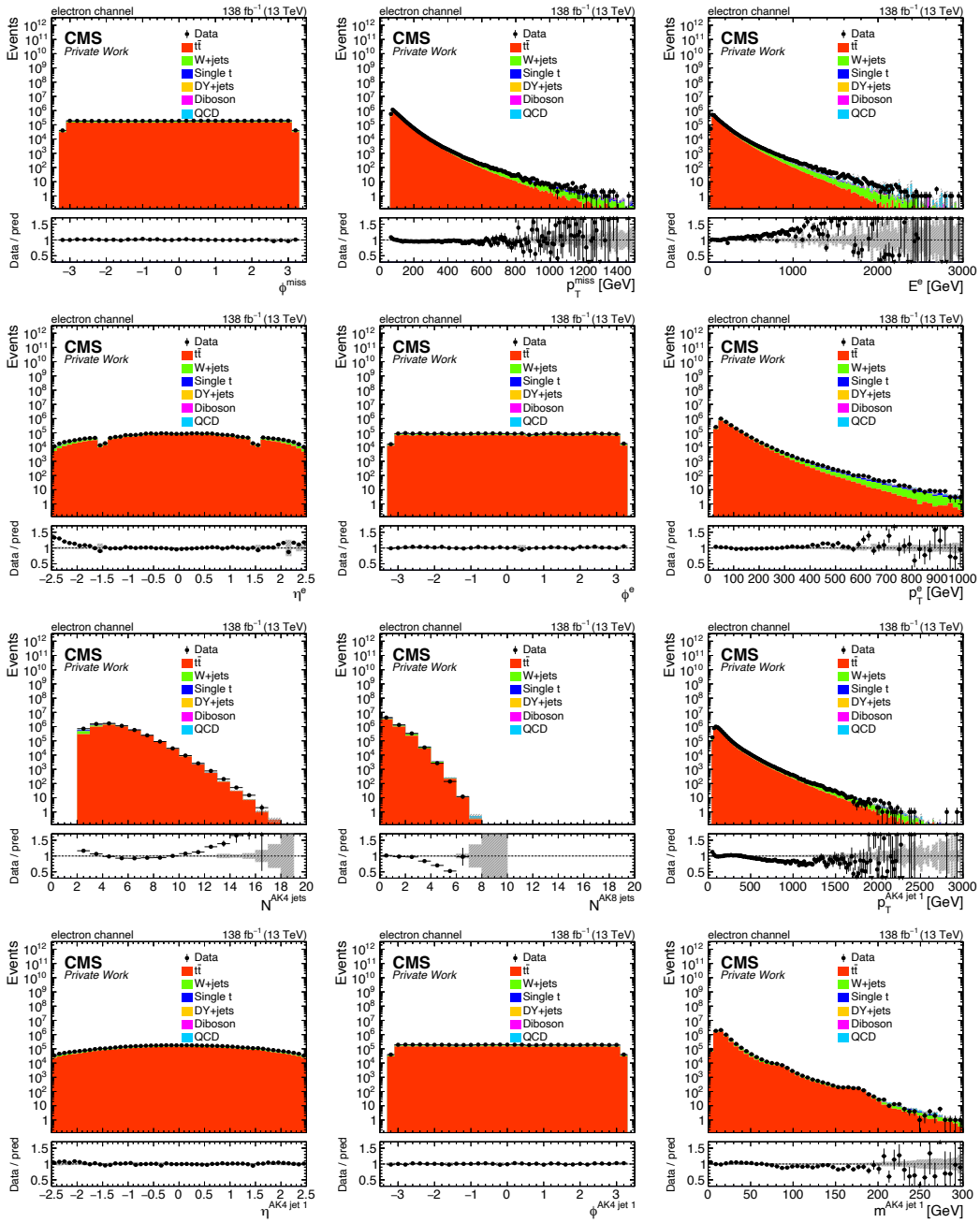

Figure B.6: The distributions of p_T^{miss}, e, number of jets and first AK4 jet for the e+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

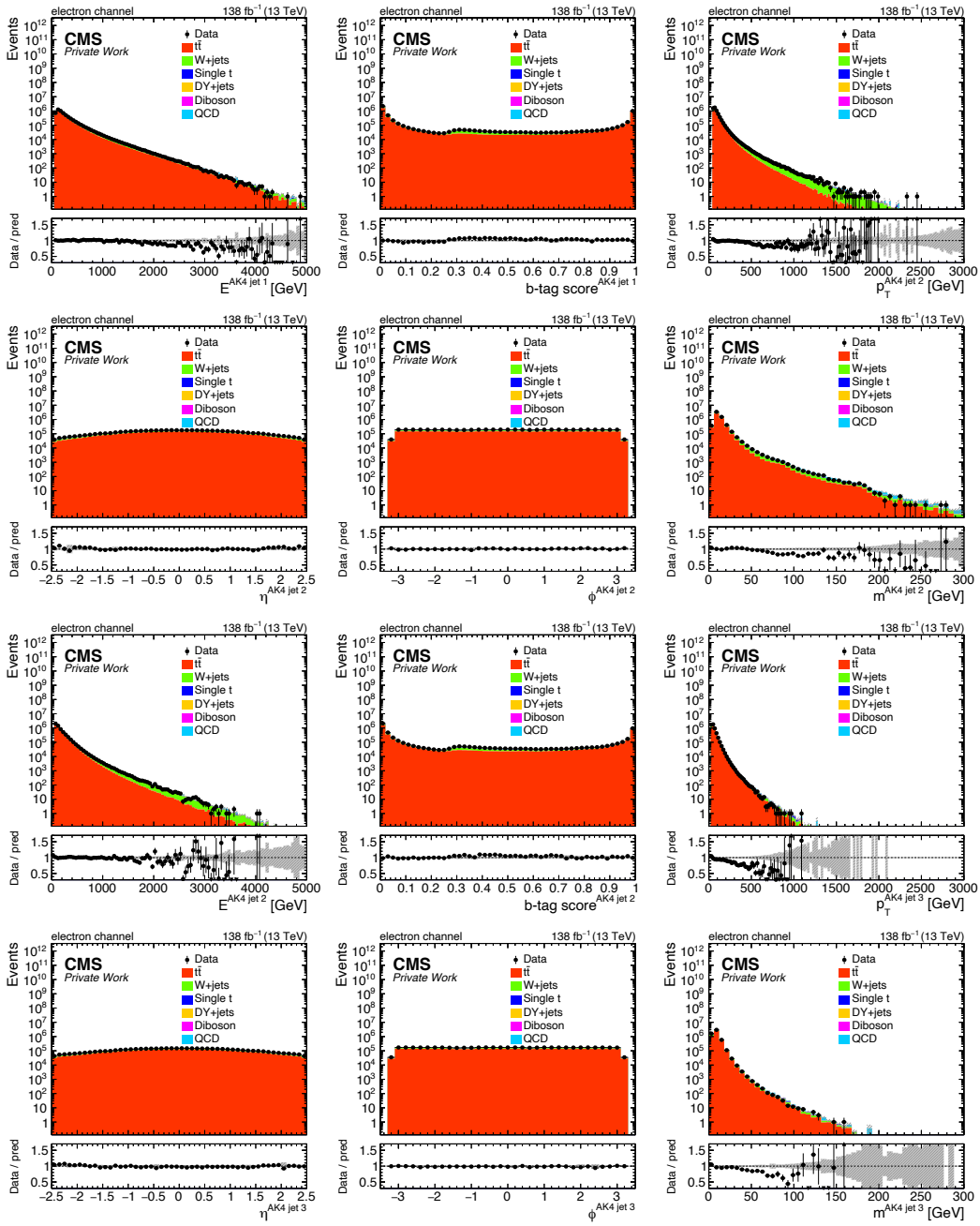

Figure B.7: The distributions of the first, second and third AK4 jet for the e+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

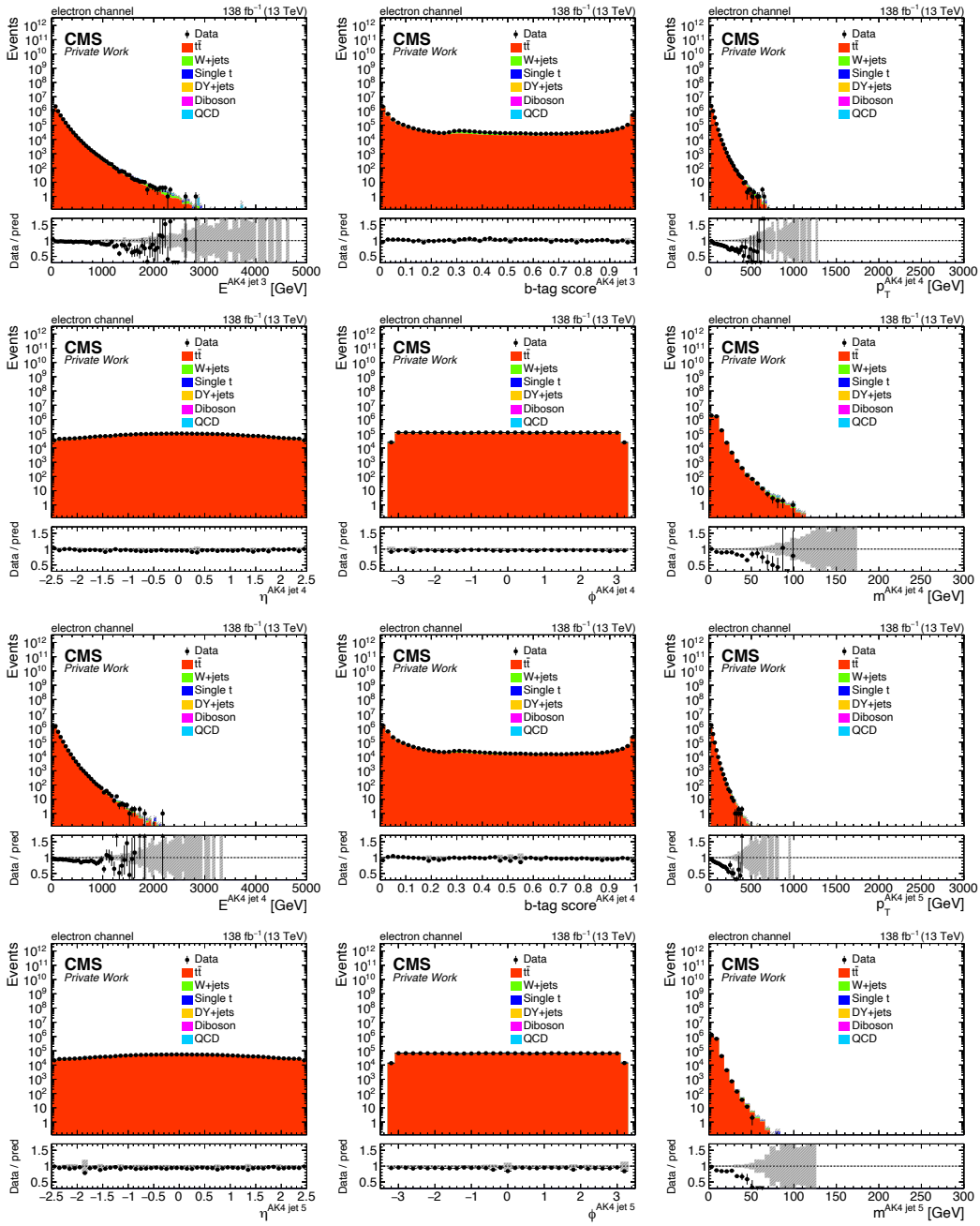

Figure B.8: The distributions of the third, fourth and fifth AK4 jet for the e+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

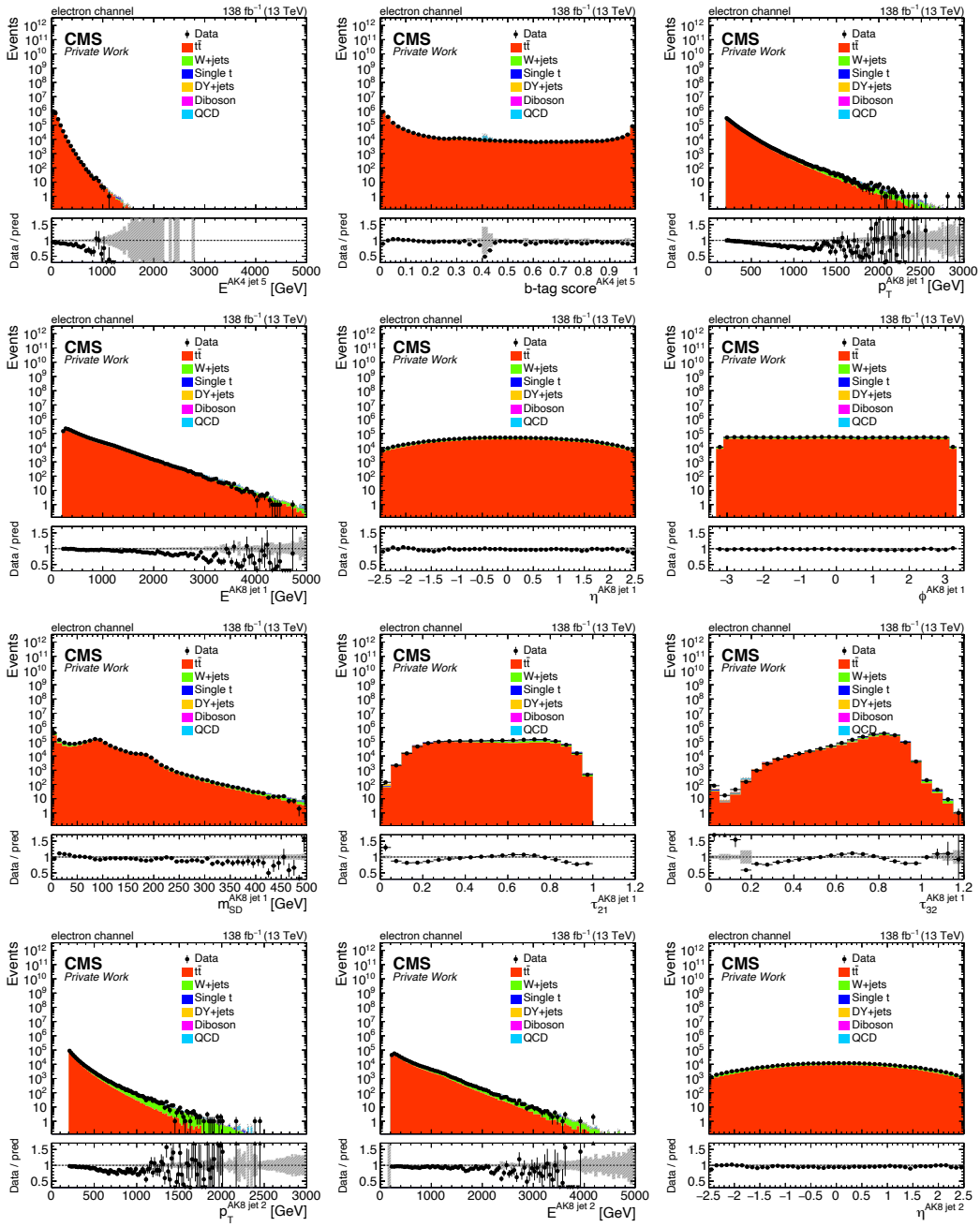

Figure B.9: The distributions of the fifth AK4 jet and the first AK8 jet for the e+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

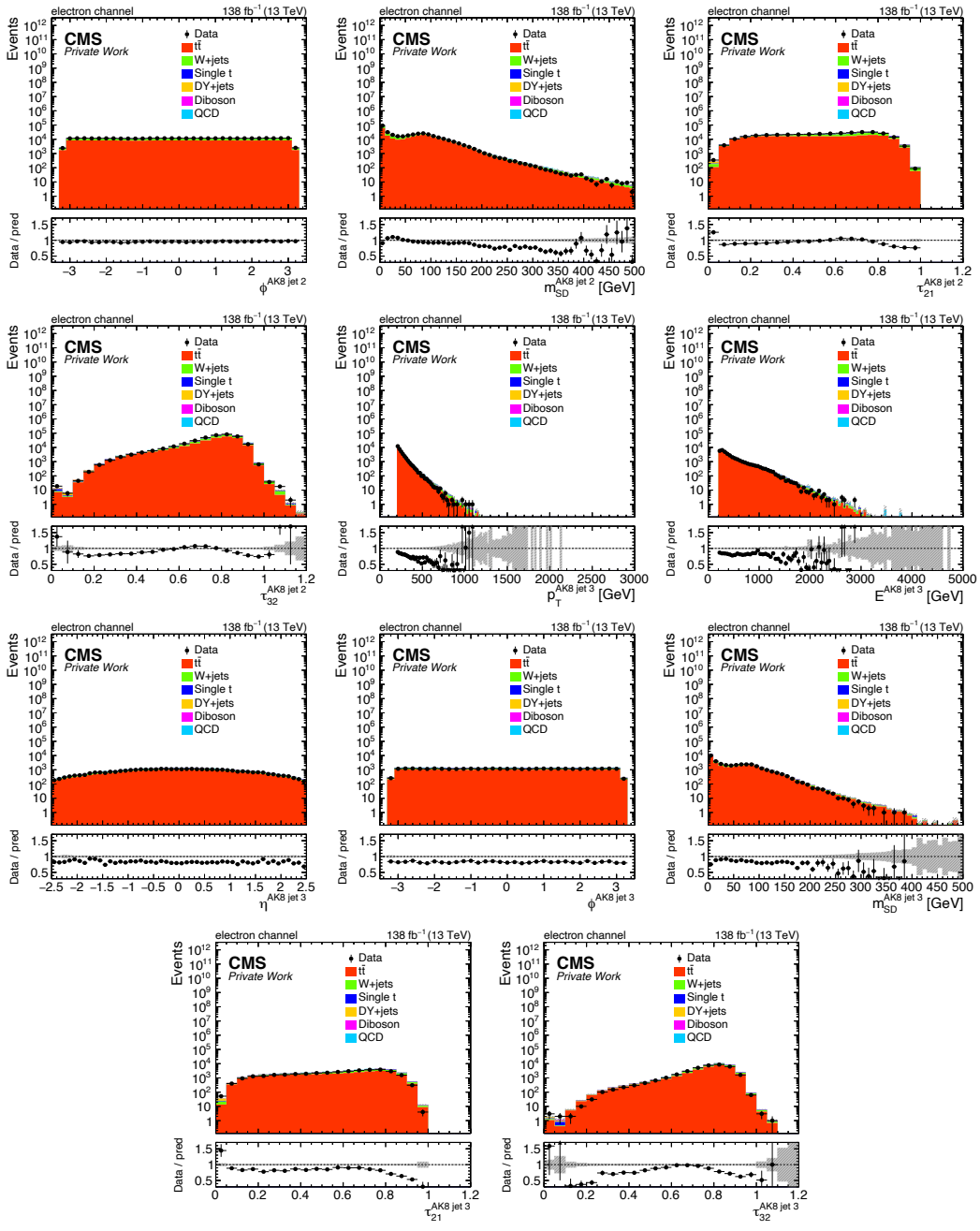

Figure B.10: The distributions of the second and third AK8 jet for the e+jets channel used as inputs to the DNN described in Sec.7.5. The grey band represents the statistical uncertainty.

Appendix C

$m_{t\bar{t}}$ pre-fit distributions

The $m_{t\bar{t}}$ distributions are shown in Fig. C.1 for the control regions and in Figures C.2-C.3 for the signal regions before the fit to data.

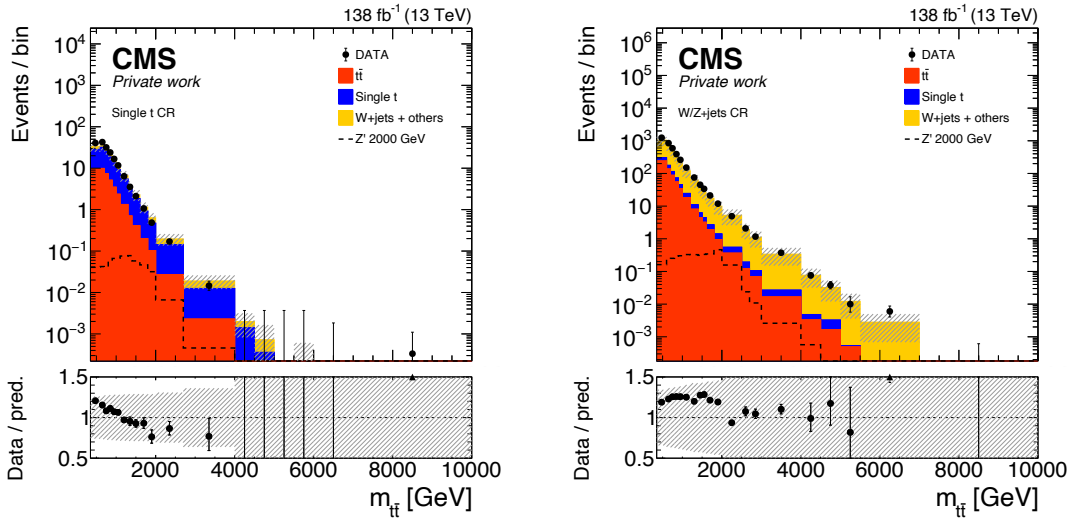

Figure C.1: The $m_{t\bar{t}}$ distributions in the single t (left) and W/Z+jets (right) CR before the fit to data. The grey area in the bottom panel corresponds to the total uncertainty on the background prediction.

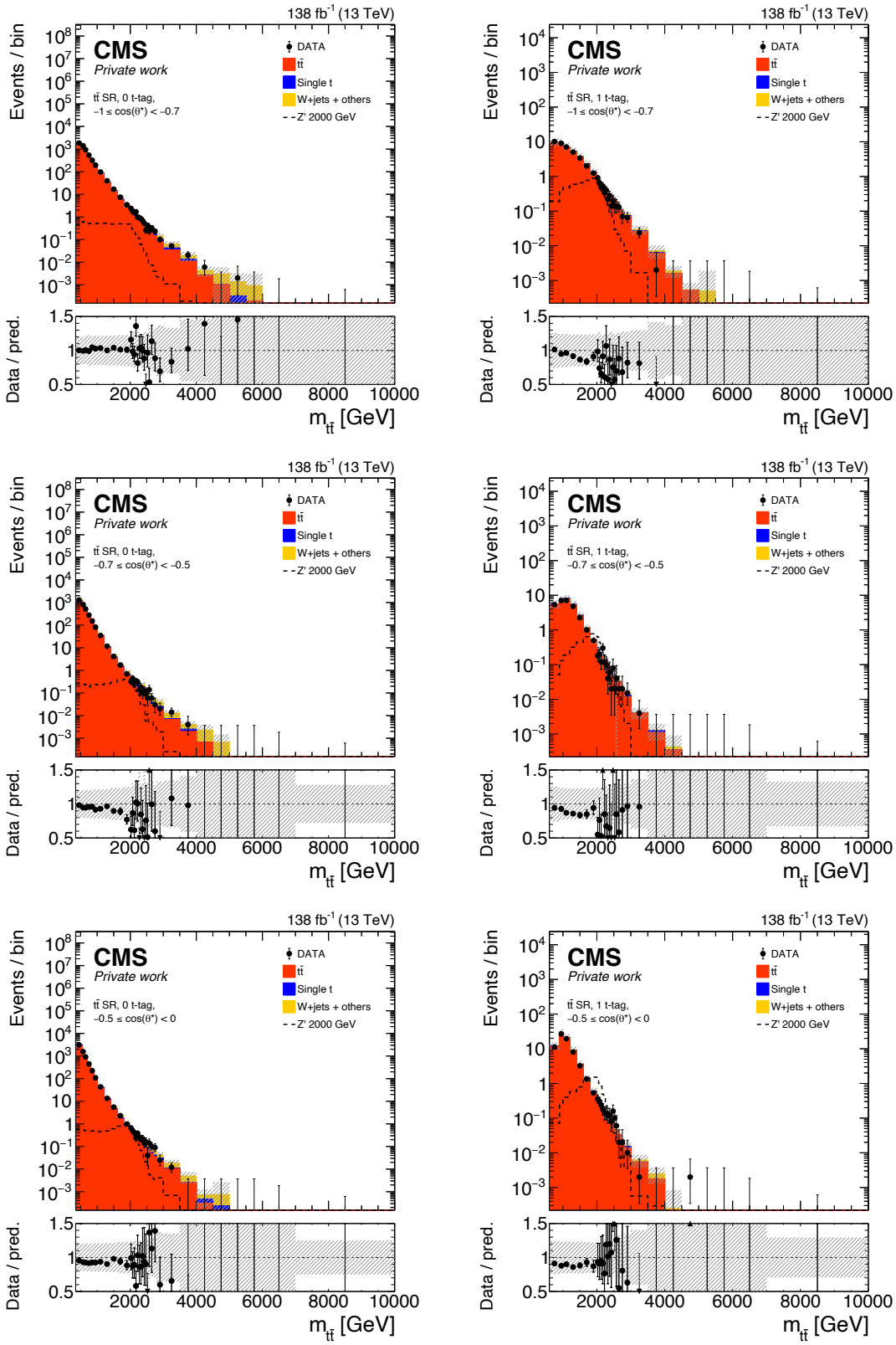

Figure C.2: The $m_{t\bar{t}}$ distributions in the first three bins of $\cos(\theta^*)$ in the $t\bar{t}$ SR, for events in the resolved (0 t-tag) and boosted (1 t-tag) categories, before the fit to data. The grey area in the bottom panel corresponds to the total uncertainty on the background prediction.

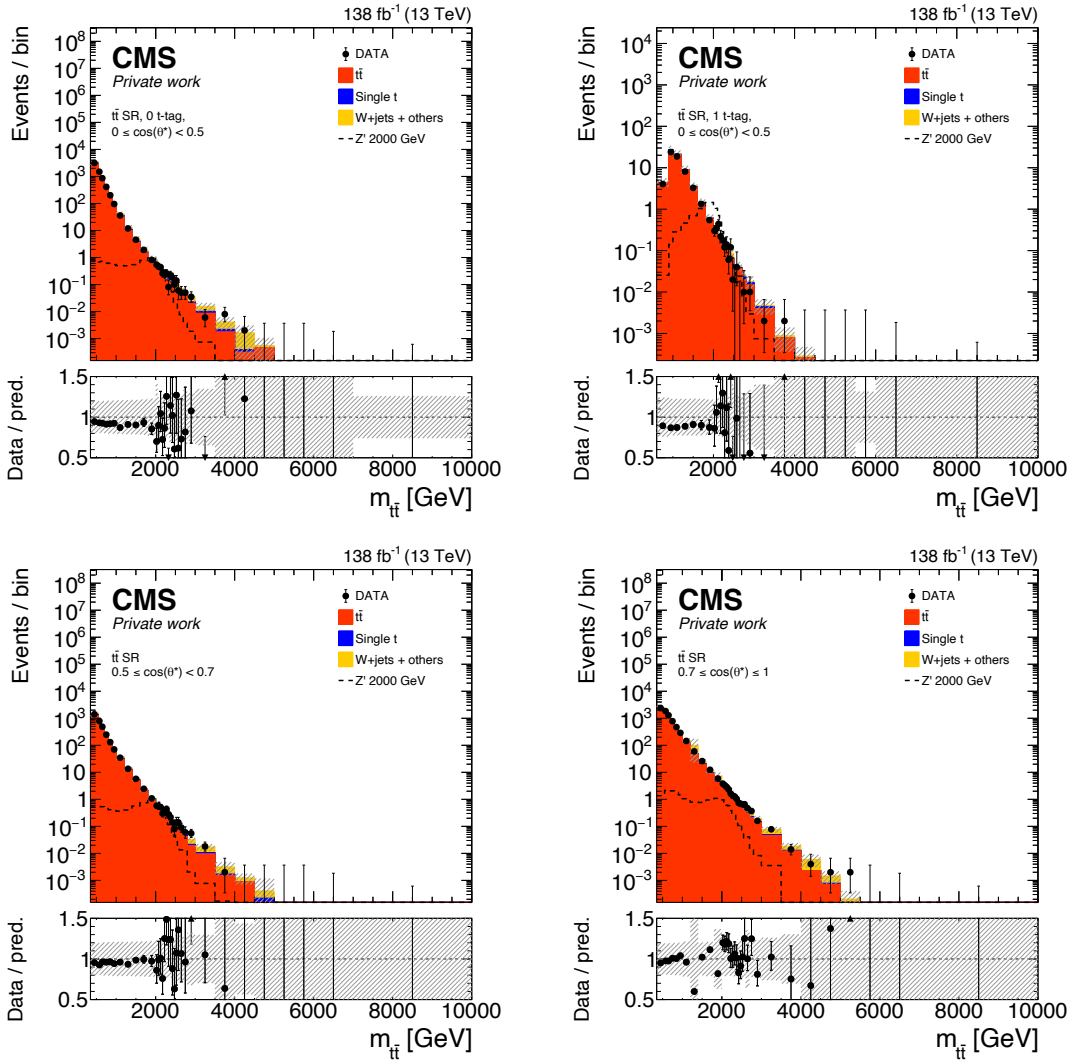

Figure C.3: The $m_{t\bar{t}}$ distributions in the last three bins of $\cos(\theta^*)$ in the $t\bar{t}$ SR before the fit to data. In the region $0 \leq \cos(\theta^*) < 0.5$ (top) events are divided in the resolved (0 t-tag) and boosted (1 t-tag) categories. The grey area in the bottom panel corresponds to the total uncertainty on the background prediction.

Bibliography

[1] H. Jabusch, "Search for Axion-like Particles Decaying to Top Quark-Antiquark Pairs with the CMS Experiment". PhD Dissertation, in preparation, Universität Hamburg, 2025.

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

[3] Super-Kamiokande Collaboration, "Evidence for Oscillation of Atmospheric Neutrinos", *Phys. Rev. Lett.* **81** (1998), no. 8, 1562–1567, doi:10.1103/physrevlett.81.1562.

[4] SNO Collaboration, "Measurement of the Rate of $\nu_e + d \to p + p + e^-$ Interactions Produced by ^8B Solar Neutrinos at the Sudbury Neutrino Observatory", *Phys. Rev. Lett.* **87** (2001) 071301, doi:10.1103/PhysRevLett.87.071301.

[5] B. Pontecorvo, "Mesonium and anti-mesonium", *Sov. Phys. JETP* **6** (1957) 429.

[6] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons", *Phys. Rev. Lett.* **13** (1964) 508–509, doi:10.1103/PhysRevLett.13.508.

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

[8] CMS Collaboration, "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV", *JHEP* **06** (2013) 081, doi:10.1007/JHEP06(2013)081, arXiv:1303.4571.

[9] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B* **716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[10] M. Thomson, "Modern Particle Physics". Cambridge University Press, 2013.

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

[12] M. Kobayashi and T. Maskawa, "CP-Violation in the Renormalizable Theory of Weak Interaction", *Prog. Theor. Phys.* **49** (1973) 652–657, doi:10.1143/PTP.49.652.

[13] N. Cabibbo, "Unitary Symmetry and Leptonic Decays", *Phys. Rev. Lett.* **10** (1963) 531–533, doi:10.1103/PhysRevLett.10.531.

[14] S. L. Glashow, "Partial-symmetries of weak interactions", *Nucl. Phys.* **22** (1961) 579–588, doi:10.1016/0029-5582(61)90469-2.

[15] S. Weinberg, "A Model of Leptons", *Phys. Rev. Lett.* **19** (1967) 1264–1266, doi:10.1103/PhysRevLett.19.1264.

[16] A. Salam, "Weak and electromagnetic interactions", pp. 244–254. doi:10.1142/9789812795915_0034.

[17] D0 Collaboration, "Observation of the Top Quark", *Phys. Rev. Lett.* **74** (1995) 2632–2637, doi:10.1103/PhysRevLett.74.2632.

[18] CDF Collaboration, "Observation of Top Quark Production in $\overline{p}p$ Collisions with the Collider Detector at Fermilab", *Phys. Rev. Lett.* **74** (1995) 2626–2631, doi:10.1103/PhysRevLett.74.2626.

[19] CMS Collaboration, "CMS Twiki - NNLO+NNLL top-quark-pair cross sections". https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TtbarNNLO.

[20] M. Czakon and A. Mitov, "Top++: A program for the calculation of the top-pair cross-section at hadron colliders", *Comput. Phys. Commun.* **185** (2014), no. 11, 2930–2938, doi:https://doi.org/10.1016/j.cpc.2014.06.021.

[21] D0 Collaboration, "Top Pair Branching Fractions". https://www-d0.fnal.gov/Run2Physics/top/top_public_web_pages/top_feynman_diagrams.html.

[22] S. Bailey, T. Cridge, and e. a. Harland-Lang, L. A., "Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs", *Eur. Phys. J. C* **81** (2021), no. 4, 341, doi:10.1140/epjc/s10052-021-09057-0.

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

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

[25] G. Curci, W. Furmanski, and R. Petronzio, "Evolution of parton densities beyond leading order: The non-singlet case", *Nucl. Phys. B* **175** (1980) 27–92, doi:10.1016/0550-3213(80)90003-6.

[26] B. Andersson, "The Lund model". Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 1998.

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

[28] T. Sjöstrand, S. Ask, J. R. Christiansen et al., "An introduction to PYTHIA 8.2", *Comput. Phys. Commun.* **191** (2015) 159–177, doi:https://doi.org/10.1016/j.cpc.2015.01.024.

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

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

[31] M. Bahr et al., "Herwig++ physics and manual", *Eur. Phys. J. C* **58** (2008) 639–707, doi:10.1140/epjc/s10052-008-0798-9, arXiv:0803.0883.

[32] J. Alwall, R. Frederix, S. Frixione et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP* **2014** (2014), no. 7, 79, doi:10.1007/JHEP07(2014)079.

[33] J. Alwall, S. Höche, F. Krauss et al., "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions", *Eur. Phys. J. C* **53** (2008), no. 3, 473–500, doi:10.1140/epjc/s10052-007-0490-5.

[34] R. Frederix and S. Frixione, "Merging meets matching in MC@NLO", *JHEP* **2012** (2012), no. 12, 61, doi:10.1007/JHEP12(2012)061.

[35] GEANT4 Collaboration, "GEANT4– a simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003), no. 3, 250–303, doi:https://doi.org/10.1016/S0168-9002(03)01368-8.

[36] BaBar Collaboration, "Measurement of an Excess of $\bar{B} \to D^{(*)}\tau^-\bar{\nu}_\tau$ Decays and Implications for Charged Higgs Bosons", *Phys. Rev. D* **88** (2013), no. 7, doi:10.1103/physrevd.88.072012.

[37] Belle Collaboration, "Measurement of $R(D)$ and $R(D^*)$ with a semileptonic tagging method", *Phys. Rev. Lett.* **124** (2020), no. 16, doi:10.1103/physrevlett.124.161803.

[38] LHCb Collaboration, "Addendum: Test of lepton universality in beauty-quark decays", *Nature Physics* **19** (2023), no. 10, 1517–1517, doi:10.1038/s41567-023-02095-3.

[39] The Muon $g-2$ Collaboration, "Measurement of the Positive Muon Anomalous Magnetic Moment to 0.20 ppm", *Phys. Rev. Lett.* **131** (2023), no. 16, doi:10.1103/PhysRevLett.131.161802.

[40] E. Corbelli and P. Salucci, "The extended rotation curve and the dark matter halo of M33", *Mon. Not. Roy. Astron. Soc.* **311** (2000), no. 2, 441–447, doi:10.1046/j.1365-8711.2000.03075.x.

[41] V. Trimble, "Existence and Nature of Dark Matter in the Universe", *Ann. Rev. Astron. Astrophys.* **25** (1987), no. 1, 425–472, doi:10.1146/annurev.aa.25.090187.002233.

[42] E. Komatsu, K. M. Smith, J. Dunkley et al., "Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation", *Astrophys. J. Suppl.* **192** (2011), no. 2, 18, doi:10.1088/0067-0049/192/2/18.

[43] Plack Collaboration, "Planck 2018 results. VI. Cosmological parameters", *Astron. Astrophys.* **641** (2020) A6, doi:10.1051/0004-6361/201833910.

[44] H. Georgi and S. L. Glashow, "Unity of All Elementary-Particle Forces", *Phys. Rev. Lett.* **32** (1974) 438–441, doi:10.1103/PhysRevLett.32.438.

[45] S. P. Martin, "A Supersymmetry Primer", pp. 1–98. World Scientific, 1998. doi:10.1142/9789812839657_0001.

[46] C. Csaki, "The Minimal Supersymmetric Standard Model (MSSM)", *Modern Physics Letters A* **11** (1996), no. 08, 599–613, doi:10.1142/s021773239600062x.

[47] L. Randall and R. Sundrum, "Large mass hierarchy from a small extra dimension", *Phys. Rev. Lett.* **83** (1999) 3370–3373, doi:10.1103/PhysRevLett.83.3370.

[48] A. Pérez-Lorenzana, "An introduction to extra dimensions", *J. Phys. Conf. Ser.* **18** (2005) 224–269, doi:10.1088/1742-6596/18/1/006.

[49] K. Agashe, A. Belyaev, T. Krupovnickas et al., "CERN LHC signals from warped extra dimensions", *Phys. Rev. D* **77** (2008) 015003, doi:10.1103/PhysRevD.77.015003.

[50] E. M. V. Barger, W.Y. Keung, "Sequential W and Z bosons", *Phys. Lett. B* **94** (1980), no. 3, 377–380, doi:10.1016/0370-2693(80)90900-4.

[51] C. T.Hill, "Topcolor assisted technicolor", *Phys. Lett. B* **345** (1995), no. 4, 483–489, doi:10.1016/0370-2693(94)01660-5.

[52] K. R. Lynch, S. Mrenna, M. Narain et al., "Finding Z′ bosons coupled preferentially to the third family at CERN LEP and the Fermilab Tevatron", *Phys. Rev. D* **63** (2001), no. 3, doi:10.1103/physrevd.63.035006.

[53] R. M. Harris, C. T. Hill, and S. J. Parke, "Cross Section for Topcolor Z′ decaying to top-antitop", doi:10.48550/arxiv.hep-ph/9911288.

[54] R. M. Harris and S. Jain, "Cross sections for leptophobic topcolor Z′ decaying to top–antitop", *Eur. Phys. J. C* **72** (2012), no. 7, doi:10.1140/epjc/s10052-012-2072-4.

[55] G. Branco, P. Ferreira, L. Lavoura et al., "Theory and phenomenology of two-Higgs-doublet models", *Phys. Rept.* **516** (2012), no. 1-2, 1–102, doi:10.1016/j.physrep.2012.02.002.

[56] O. Eberhardt, U. Nierste, and M. Wiebusch, "Status of the two-Higgs-doublet model of type II", *JHEP* **2013** (2013), no. 7, doi:10.1007/jhep07(2013)118.

[57] D. Dicus, A. Stange, and S. Willenbrock, "Higgs decay to top quarks at hadron colliders", *Phys. Lett. B* **333** (1994), no. 1-2, 126–131, doi:10.1016/0370-2693(94)91017-0.

[58] CMS Collaboration, "CMS Beyond SM particles decaying 2(to) Higgs, top and Gauge bosons (B2G) Public Physics Results". https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsB2G.

[59] CDF Collaboration, "Limits on the production of narrow $t\bar{t}$ resonances in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV", *Phys. Rev. D* **77** (2008), no. 5, doi:10.1103/physrevd.77.051102.

[60] CDF Collaboration, "Search for Resonant $t\bar{t}$ Production in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV", *Phys. Rev. Lett.* **100** (2008), no. 23, doi:10.1103/physrevlett.100.231801.

[61] CDF Collaboration, "Search for resonant production of $t\bar{t}$ pairs in 4.8 fb^{-1} of integrated luminosity of $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV", *Phys. Rev. D* **84** (2011), no. 7, doi:10.1103/physrevd.84.072004.

[62] CDF Collaboration, "Search for resonant production of $t\bar{t}$ decaying to jets in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV", *Phys. Rev. D* **84** (2011), no. 7, doi:10.1103/physrevd.84.072003.

[63] D0 Collaboration, "Search for a narrow $t\bar{t}$ resonance in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV", *Phys. Rev. D* **85** (2012), no. 5, doi:10.1103/physrevd.85.051101.

[64] D0 Collaboration, "Search for $t\bar{t}$ resonances in the lepton plus jets final state in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV", *Phys. Lett. B* **668** (2008), no. 2, 98–104, doi:10.1016/j.physletb.2008.08.027.

[65] CMS Collaboration, "Search for anomalous $t\bar{t}$ production in the highly-boosted all-hadronic final state", *JHEP* **2012** (2012), no. 9, doi:10.1007/jhep09(2012)029.

[66] ATLAS Collaboration, "A search for $t\bar{t}$ resonances in lepton+jets events with highly boosted top quarks collected in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector", *JHEP* **2012** (2012), no. 9, doi:10.1007/jhep09(2012)041.

[67] ATLAS Collaboration, "Search for $t\bar{t}$ resonances in the lepton plus jets final state with ATLAS using 4.7 fb^{-1} of pp collisions at $\sqrt{s} = 7$ TeV", *Phys. Rev. D* **88** (2013), no. 1, doi:10.1103/physrevd.88.012004.

[68] CMS Collaboration, "Search for resonant $t\bar{t}$ production in lepton+jets events in pp collisions at $\sqrt{s} = 7$ TeV", *JHEP* **2012** (2012), no. 12, doi:10.1007/jhep12(2012)015.

[69] CMS Collaboration, "Search for Z$'$ resonances decaying to $t\bar{t}$ in dilepton+jets final states in pp collisions at $\sqrt{s} = 7$ TeV", *Phys. Rev. D* **87** (2013), no. 7, doi:10.1103/physrevd.87.072002.

[70] CMS Collaboration, "Searches for new physics using the $t\bar{t}$ invariant mass distribution in pp collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. Lett.* **111** (2013), no. 21, doi:10.1103/physrevlett.111.211804.

[71] CMS Collaboration, "Search for resonant $t\bar{t}$ production in proton-proton collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. D* **93** (2016), no. 1, doi:10.1103/physrevd.93.012001.

[72] ATLAS Collaboration, "Search for heavy particles decaying into a top-quark pair in the fully hadronic final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", *Phys. Rev. D* **99** (2019), no. 9, doi:10.1103/physrevd.99.092004.

[73] CMS Collaboration, "Search for $t\bar{t}$ resonances in highly boosted lepton+jets and fully hadronic final states in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **2017** (2017), no. 7, doi:10.1007/jhep07(2017)001.

[74] ATLAS Collaboration, "Search for heavy particles decaying into top-quark pairs using lepton-plus-jets events in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", *Eur. Phys. J. C* **78** (2018), no. 7, doi:10.1140/epjc/s10052-018-5995-6.

[75] CMS Collaboration, "Search for resonant $t\bar{t}$ production in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **2019** (2019), no. 4, 31, doi:10.1007/jhep04(2019)031.

[76] ATLAS Collaboration, "Search for $t\bar{t}$ resonances in fully hadronic final states in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", *JHEP* **2020** (2020), no. 10, 61, doi:10.1007/JHEP10(2020)061.

[77] ATLAS Collaboration, "Search for Heavy Higgs Bosons A/H Decaying to a Top Quark Pair in pp Collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector", *Phys. Rev. Lett.* **119** (2017), no. 19, doi:10.1103/physrevlett.119.191803.

[78] ATLAS Collaboration, "Search for heavy neutral Higgs bosons decaying into a top quark pair in 140 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV with the ATLAS detectorr", *JHEP* **2024** (2024), no. 8, doi:10.1007/JHEP08(2024)013.

[79] CMS Collaboration, "Search for heavy Higgs bosons decaying to a top quark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **2020** (2020), no. 4, doi:10.1007/jhep04(2020)171.

[80] CMS Collaboration, "Search for heavy pseudoscalar and scalar bosons decaying to top quark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV", Technical Report CMS-PAS-HIG-22-013, (2024).

[81] L. Evans and P. Bryant, "LHC Machine", *JINST* **3** (2008), no. 08, S08001, doi:10.1088/1748-0221/3/08/S08001.

[82] E. Mobs, "The CERN accelerator complex - August 2018. Complexe des accélérateurs du CERN - Août 2018". https://cds.cern.ch/record/2636343, 2018.

[83] CMS Collaboration, "CMS Luminosity - Public Results". https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[84] T. Sakuma and T. McCauley, "Detector and Event Visualization with SketchUp at the CMS Experiment", *J. Phys. Conf. Ser.* **513** (2014), no. 2, 022032, doi:10.1088/1742-6596/513/2/022032.

[85] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008), no. 08, S08004, doi:10.1088/1748-0221/3/08/S08004.

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

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

[88] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* **12** (2017), no. 10, P10003–P10003, doi:10.1088/1748-0221/12/10/p10003.

[89] CMS HCAL/ECAL Collaborations, "The CMS barrel calorimeter response to particle beams from 2 to 350 GeV/c", *Eur. Phys. J. C* **61** (2009), no. 2, 353–356, doi:10.1140/epjc/s10052-009-1024-0.

[90] W. Adam, B. Mangano, T. Speer et al., "Track Reconstruction in the CMS tracker", Technical Report CMS-NOTE-2006-041, (2006).

[91] CMS Collaboration, "Pileup mitigation at CMS in 13 TeV data", *JINST* **15** (2020), no. 09, P09018–P09018, doi:10.1088/1748-0221/15/09/p09018.

[92] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014), no. 10, P10009–P10009, doi:10.1088/1748-0221/9/10/p10009.

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

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

[95] D. Bertolini, P. Harris, M. Low et al., "Pileup per particle identification", *JHEP* **2014** (2014), no. 10, 59, doi:10.1007/JHEP10(2014)059.

[96] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **13** (2018), no. 06, P06015–P06015, doi:10.1088/1748-0221/13/06/p06015.

[97] CMS Collaboration, "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC", *JINST* **16** (2021), no. 05, P05014, doi:10.1088/1748-0221/16/05/p05014.

[98] S. D. Ellis and D. E. Soper, "Successive combination jet algorithm for hadron collisions", *Phys. Rev. D* **48** (1993) 3160–3166, doi:10.1103/PhysRevD.48.3160.

[99] S. Catani, Y. Dokshitzer, M. Seymour et al., "Longitudinally-invariant $k_\perp$-clustering algorithms for hadron-hadron collisions", *Nucl. Phys. B* **406** (1993), no. 1, 187–224, doi:https://doi.org/10.1016/0550-3213(93)90166-M.

[100] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm", *JHEP* **2008** (2008), no. 04, 063–063, doi:10.1088/1126-6708/2008/04/063.

[101] Y. Dokshitzer, G. Leder, S. Moretti et al., "Better jet clustering algorithms", *JHEP* **1997** (1997), no. 08, 001–001, doi:10.1088/1126-6708/1997/08/001.

[102] M. Wobisch and T. Wengler, "Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering", *Monte Carlo generators for HERA physics Proc.* (1999) doi:10.48550/arxiv.hep-ph/9907280.

[103] R. Kogler, "Advances in Jet Substructure at the LHC". Springer Tracts Mod. Phys. 284, 2021.

[104] D. Krohn, J. Thaler, and L.-T. Wang, "Jets with variable R", *JHEP* **2009** (2009), no. 06, 059–059, doi:10.1088/1126-6708/2009/06/059.

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

[106] M. Stoll, "Vetoed jet clustering: the mass-jump algorithm", *JHEP* **2015** (2015), no. 4, doi:10.1007/jhep04(2015)111.

[107] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual", *Eur. Phys. J. C* **72** (2012), no. 3, doi:10.1140/epjc/s10052-012-1896-2.

[108] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV", *JINST* **12** (2017), no. 02, P02014–P02014, doi:10.1088/1748-0221/12/02/p02014.

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

[110] CMS Collaboration, "Jet energy scale and resolution measurement with Run 2 Legacy Data Collected by CMS at 13 TeV", Technical Report CMS-DP-2021-033, (2021).

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

[112] E. Bols, J. Kieseler, M. Verzetti et al., "Jet flavour classification using DeepJet", *JINST* **15** (2020), no. 12, P12012–P12012, doi:10.1088/1748-0221/15/12/p12012.

[113] A. J. Larkoski, S. Marzani, G. Soyez et al., "Soft drop", *JHEP* **2014** (2014), no. 5, doi:10.1007/jhep05(2014)146.

[114] J. Thaler and K. V. Tilburg, "Identifying boosted objects with N-subjettiness", *JHEP* **2011** (2011), no. 3, doi:10.1007/jhep03(2011)015.

[115] CMS Collaboration, "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques", *JINST* **15** (2020), no. 06, P06005–P06005, doi:10.1088/1748-0221/15/06/p06005.

[116] CMS Collaboration, "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector", *JINST* **14** (2019), no. 07, P07004–P07004, doi:10.1088/1748-0221/14/07/p07004.

[117] CMS Collaboration, "Performance of the pile up jet identification in CMS for Run 2", Technical Report CMS-DP-2020-020, (2020).

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

[119] R. D. Ball, V. Bertone, S. Carrazza et al., "Parton distributions from high-precision collider data", *Eur. Phys. J. C* **77** (2017), no. 10, 663, doi:10.1140/epjc/s10052-017-5199-5.

[120] CMS Collaboration, "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV", *JHEP* **2018** (2018), no. 7, 161, doi:10.1007/JHEP07(2018)161.

[121] CMS Collaboration, "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **15** (2020), no. 10, P10017, doi:10.1088/1748-0221/15/10/P10017.

[122] J. Gao, C. S. Li, B. H. Li et al., "Next-to-leading order QCD corrections to the heavy resonance production and decay into top quark pair at the LHC", *Phys. Rev. D* **82** (2010) 014020, doi:10.1103/PhysRevD.82.014020, arXiv:1004.0876.

[123] B. Hespel, F. Maltoni, and E. Vryonidou, "Signal background interference effects in heavy scalar production and decay to a top-anti-top pair", *JHEP* (2016), no. 10, 16, doi:10.1007/JHEP10(2016)016, arXiv:1606.04149.

[124] J. M. Lindert, S. Pozzorini, R. Boughezal et al., "Precise predictions for V+jets dark matter backgrounds", *Eur. Phys. J. C* **77** (2017), no. 12, doi:10.1140/epjc/s10052-017-5389-1.

[125] F. Chollet et al., "Keras". https://keras.io, 2015.

[126] M. Abadi, A. Agarwal, P. Barham et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems", 2015. Software available from tensorflow.org.

[127] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning". MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[128] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", *3rd International Conference for Learning Representations Proc.* (2017) arXiv:1412.6980.

[129] CMS Collaboration, "Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$TeV in 2015 and 2016 at CMS", *Eur. Phys. J. C* **81** (2021) 800, doi:10.1140/epjc/s10052-021-09538-2.

[130] CMS Collaboration, "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$TeV", CMS Physics Analysis Summary CMS-PAS-LUM-17-004, (2018).

[131] CMS Collaboration, "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$TeV", CMS Physics Analysis Summary CMS-PAS-LUM-18-002, (2019).

[132] A. L. Read, "Presentation of search results: the CLs technique", *J. Phys. G* **28** (2002), no. 10, 2693, doi:10.1088/0954-3899/28/10/313.

[133] CMS Collaboration, "The CMS Statistical Analysis and Combination Tool: Combine", *Computing and Software for Big Science* **8** (2024), no. 1, 19, doi:10.1007/s41781-024-00121-4.

[134] R. J. Barlow and C. Beeston, "Fitting using finite Monte Carlo samples", *Comput. Phys. Commun.* **77** (1993) 219–228, doi:10.1016/0010-4655(93)90005-W.

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

Declaration on oath /

Eidesstattliche Versicherung

I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated.

If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

I declare that this bound copy of the dissertation and the dissertation submitted in electronic form (via the Docata upload) and the printed bound copy of the dissertation submitted to the faculty (responsible Academic Office or the Doctoral Office Physics) for archiving are identical.

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

Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Ich versichere, dass dieses gebundene Exemplar der Dissertation und das in elektronischer Form eingereichte Dissertationsexemplar (über den Docata-Upload) und das bei der Fakultät (zuständiges Studienbüro bzw. Promotionsbüro Physik) zur Archivierung eingereichte gedruckte gebundene Exemplar der Dissertationsschrift identisch sind.

Date | Datum	Signature of doctoral candidate | Unterschrift der Doktorandin