Vrije Universiteit Brussel


Faculteit Wetenschappen en
Bio-ingenieurswetenschappen

Departement Natuurkunde

---

# When charm and beauty adjoin the top

First measurement of the cross section of top quark pair production with additional charm jets with the CMS experiment

---

Seth Moortgat

*Promotor:* Prof. Dr. Jorgen D'Hondt
*Co-promotor:* Prof. Dr. Alberto Mariotti

*Proefschrift ingediend met het oog op het behalen van de academische graad Doctor in de Wetenschappen*

May 20th, 2019

## Doctoral examination commission:

**Chair**: Prof. Dr. Alexander Sevrin
*Theoretische Natuurkunde (TENA) – Vrije Universiteit Brussel*

**Secretary**: Prof. Dr. Petra Van Mulders
*Interuniversity Institute for High Energies (IIHE) – Vrije Universiteit Brussel*

**Promotor**: Prof. Dr. Jorgen D'Hondt
*Interuniversity Institute for High Energies (IIHE) – Vrije Universiteit Brussel*

**Co-promotor**: Prof. Dr. Alberto Mariotti
*Interuniversity Institute for High Energies (IIHE) – Vrije Universiteit Brussel*

Prof. Dr. Steven Lowette
*Interuniversity Institute for High Energies (IIHE) – Vrije Universiteit Brussel*

Prof. Dr. Philippe Claeys
*Analytical, Environmental and Geo-Chemistry (AMGC) – Vrije Universiteit Brussel*

Prof. Dr. Florencia Canelli
*Physik-Institut – University of Zurich*

Prof. Dr. Nuno Castro
*Departamento de Física – University of Minho*
*and Laboratory of Instrumentation and Experimental Particles Physics (LIP)*

Prof. Dr. Fabio Maltoni
*Centre for Cosmology, Particle Physics and Phenomenology (CP3) –Université Catholique de Louvain (UCL)*

Cover Illustration:

The event display [1] shown on the cover of this book is extracted from a proton–proton collision event recorded with the CMS detector on Sunday, the 28$^{\text{th}}$ of August 2016 at 17h37 (CEST). The event shows one isolated muon and four reconstructed jets of which three contain a secondary vertex (white dots) and one contains a non–isolated muon. Two jets are tagged as $b$–jets and one jet is tagged as a $c$–jet. This event is therefore a perfect semileptonic top quark pair candidate.



The research presented in this manuscript was funded by "Fonds voor Wetenschappelijk Onderzoek - Vlaanderen".

"We live in a wonderful world that is full of beauty, charm and adventure. There is no end to the adventures that we can have if only we seek them with our eyes open."

— Jawaharlal Nehru

Image:

View from Little Adam's Peak, Ella, Sri Lanka. Picture taken by the author on December 6th 2018.

# Contents

# Introduction

The most fundamental questions on our existence in the universe can be split into those which are of philosophical nature and those which are of scientific nature. The thin line between these two labels is defined by the concept of *falsifiability*, first discussed in depth by the science philosopher Karl Popper [2]. Theoretical statements that lack any falsifiable claims are usually considered to be of philosophical nature. Instead we prefer to focus on those theories that provide predictions which can be experimentally tested to confirm or refute the theory. But even an intrinsically falsifiable theory may at this very moment not be experimentally testable due to limited technological advances. Such a theory can also be considered of philosophical nature today, but is expected to evolve into a scientific theory once the technology is at hand to test its predictions. Even Einstein, who had himself written down the equations that predicted the existence of gravitational waves, was skeptical on whether they existed and if so, whether they could ever be measured. In 2016 the LIGO and Virgo Collaborations reported the first ever direct observation of a gravitational wave from a binary black hole merger [3]. This anecdote perfectly illustrates why we should never stop trying to better understand the universe we live in. Our ultimate goal should be to transform as many philosophical questions as possible into scientific ones and provide experimentally motivated answers!

In our quest to do so, by trying to describe the elementary building blocks of nature and how they interact with each other, the Standard Model (SM) of Particle Physics came to being. The SM incorporates the laws of quantum mechanics and special relativity to describe the universe on the smallest scales and for a large range of energies. Over the past decades, more powerful particle accelerators were developed which allowed to test the SM to an unprecedented precision. Throughout time, it has passed all these tests and whenever new elementary particles were discovered (predicted or not), the theory was flawlessly extended and further verified. The discovery of the Brout-Englert-Higgs boson by the CMS [4, 5] and ATLAS [6] collaborations at the Large Hadron Collider (LHC) at CERN, Switzerland, was considered the final piece of the puzzle. This particle, whose discovery had been one of the main motivations behind the LHC design and construction, was predicted independently by Robert Brout and Francois Englert [7] as well as Peter Higgs [8] in 1964.

However, there is clear evidence, mainly from cosmological observations and neutrino physics, that some pieces are missing from the puzzle and do not seem to fit in anywhere. Some of the most profound open questions in physics include for example the existence and nature of Dark Matter (DM) and Dark Energy, non-zero neutrino masses and the matter–antimatter asymmetry in the universe which currently remain unanswered by the SM as we know it. Either the SM should be extended with a currently unknown sector of new physics, or it is only valid up to a certain energy frontier

and nature is described by a more general and more complete theory beyond that. The story is far from over, and the coming years will without a doubt be very exciting. The LHC allows us to explore elementary particle physics at energies we have not explored before and provides us with very large datasets that could potentially contain the clues to understanding new physics phenomena beyond the Standard Model (BSM).

In the proton–proton ($pp$) collisions that take place in the LHC, many top quarks ($t$) are produced and the unprecedented size of the collected top–quark dataset often gives the LHC the name of a *top–quark factory*. The top quark, being the heaviest known elementary particle today and therefore coupling the strongest to the Higgs field, may play a special role in this search for new physics phenomena. Therefore precision measurements of top–quark related processes are of crucial importance. This includes measurements of the top–quark mass and width, of the production cross section of a single top quark, a pair of top quarks or even four top quarks together, or associated production of the top quark with other known elementary particles. Recently the first direct observation of the interaction between the Higgs boson and the top quark was announced [9, 10]. An indispensable component in this discovery, but also a very interesting measurement on its own, is to determine how often a top quark pair is produced in association with a pair of either of the two next heaviest quarks, namely bottom[1] ($b$) or charm ($c$) quarks. This question defines the core objective of the work described in this thesis. How often is such a combination of particles produced in the $pp$ collisions at the LHC and is it in line with the predictions from the SM, or is there room for contributions from new physics phenomena? Whereas the production of a top quark pair with additional bottom quark jets ($t\bar{t}b\bar{b}$) has been measured before by the CMS and ATLAS [11–15] Collaborations, this thesis presents the first measurement of the production of a top quark pair with additional charm quark jets ($t\bar{t}c\bar{c}$) using 41.5 fb$^{-1}$ of 13 TeV proton–proton collision data collected with the CMS experiment.

In Chapter 1 the theoretical framework of the SM will be introduced together with some of the most profound unanswered questions. Also a model–independent approach to extend the SM with new interactions will be discussed by making use of an effective field theory. Chapter 2 deals with a description of the Large Hadron Collider and the Compact Muon Solenoid detector which are used in the analysis presented in this thesis. The simulation and reconstruction of proton collisions is detailed in Chapter 3. This is followed by an overview of Machine Learning algorithms and their applications in heavy–flavor jet identification in Chapter 4. The measurement of top quark pair production with additional heavy–flavor jets is described in Chapter 5, followed by an effective field theory interpretation of the obtained results in Chapter 6. Finally, Chapter 7 closes this thesis with a conclusion on the obtained results and an outlook for the future.

[1] Sometimes also called beauty quark.

# Chapter 1 The Standard Model of particle physics

The divisibility of matter into smaller and smaller parts has been a question posed by mankind for a long time. It was already about 400 years B.C. that the Greek philosopher Democritus [16] first introduced the concept of "*atoms*" as being indivisible building blocks of nature, existing in different sorts and shapes. Over the years, the concept of the elementary building blocks of nature evolved, moving from the idea of air, fire, water and earth by Aristotle to the Periodic Table of Elements by Mendelejev [17]. The substructure of atoms was later confirmed by the scattering experiment of Rutherford [18], and J.J. Thomson's cathode ray experiments [19] demonstrated the existence of the electron, being one of the elementary particles we know today. In the late 1960's, the substructure of the proton was probed through deep inelastic scattering experiments [20, 21] at the Stanford Linear Accelerator Center (SLAC). These experiments consisted of shooting high-energy electrons at protons and neutrons in atomic nuclei and confirmed the existence of quarks. Larger and stronger particle accelerators such as the Super Proton–Antiproton Synchrotron (Sp$\overline{\text{p}}$S), the Large Electron-Positron collider (LEP), the Tevatron and the Large Hadron Collider have further revealed the existence and properties of other elementary particles and the forces that act amongst them.

The model describing these elementary particles and their interactions through different forces is known as the Standard Model of particle physics. In this chapter, the theoretical framework of this theory will be introduced. Starting from the currently known particle content of the SM, the mathematical formulation as a quantum field theory will be introduced together with the concept of gauge invariance. The top quark sector will then be discussed in more detail given its importance in the rest of this thesis and finally, the Standard Model Effective Field Theory (SMEFT) will be discussed as a model independent description of beyond the SM interactions that can potentially be probed at the LHC.
Sections 1.1 and 1.5 are largely inspired by Ref. [22], and Sections 1.2 and 1.3 contain information from Refs. [23–25].

## 1.1 Particles and forces

The elementary matter particles of the SM are fermions with a value of the quantum mechanical spin of $\pm 1/2$. They obey the Fermi-Dirac statistics and consequently follow the Pauli exclusion principle, meaning no two fermions can occupy the exact same quantum state at a given time. In the SM, the elementary fermions are subdivided

into the leptons and the quarks, each of which appear in three generations. An overview of the SM fermions is given in Tab. 1.1, together with some elementary properties like their electric charge Q and their mass. They will be discussed in more detail below.

### Leptons

Perhaps the most well known lepton is the electron, being the particle that is responsible for electricity. It is indeed a first generation lepton with an electric charge[1] of $-1 \cdot q_e$. As confirmed by experiments performed in 1956 by Cowan and Reines involving radioactive $\beta$–decay [26], for each charged lepton[2] there exists an electrically neutral (and nearly massless) fermion called a neutrino. Therefore, the electron–neutrino accompanies the electron in the first generation of leptons.
Additionally, from cosmic ray observations [27] it became apparent that another particle existed with the same properties as the electron, but which was much heavier. This marks the discovery of the second generation charged lepton, called the muon. Later also the corresponding muon–neutrino was observed [28].
Finally, through indirect detection in electron-positron collisions at the SLAC accelerator [29], the third generation charged lepton, called the tau lepton, was discovered together with the corresponding tau–neutrino [30]. The tau lepton is special compared to the electron and the muon in the fact that is has enough mass to decay hadronically into quarks. Detailed measurements of the invisible Z decays into neutrinos at LEP have ruled out the existence of a fourth generation of leptons in the SM [31].

### Quarks

Quarks behave quite differently from leptons in nature, given that they are held together by a much stronger force that does not affect leptons, as will be discussed later on. We do not observe individual quarks in nature, but rather see the collection of multiple quarks confined in what are called hadrons. They can be divided into mesons (one quark and one antiquark), for example pions and kaons, or baryons (groups of three quarks) such as the proton and the neutron[3]. As a matter of fact, many of these hadrons were observed before the underlying quark structure was known. Through deep inelastic scattering experiments at SLAC [20, 21] in 1969, where high energy electrons were collided onto protons and neutrons in atomic nuclei, the existence of the up quark and the down quark was confirmed. They form the first generation of quarks and instead of having an integer electric charge, their charges have fractional values of $+2/3$ for the up quark and $-1/3$ for the down quark. Given the clear structure that was observed in the classification of the known hadrons, also known as the *Eightfold Way*, it was clear that a third quark had to exist for example to explain the existence of kaons. Therefore this discovery was also supporting evidence for a second generation quark known as the strange quark (with electric charge $-1/3$). Predicted by the GIM mechanism [32], evidence of a fourth quark was indeed found in 1974 with the discovery of a charmed meson, named the $J/\psi$ meson [33, 34], which is composed of a charm quark and a charm–antiquark. The charm quark thus joins the strange quark in the second generation. The premier of observing a third–generation

[1] The elementary charge $q_e$ is equal to the absolute value of the electric charge of an electron: $q_e = 1.602 \times 10^{-19}$ C.

[2] The original experiment confirmed the existence of the (anti)electron-neutrino as a decay product in radioactive $\beta$–decay.

[3] Recently even groups of four (tetraquarks) or five (pentaquarks) quarks have been observed.

quark happened in 1977 through a measurement of a dimuon resonance at around 9.5 GeV, corresponding to a bottom–quark and bottom–antiquark state called the Υ meson [35]. The final quark in our present picture of the SM, the top quark, was discovered quite a bit later in 1995. The CDF [36] and DØ [37] experiments at the Tevatron collider finally succeeded in finding the last and heaviest of the known quarks. The top quark is much heavier than all the other quarks, with a mass of around 173 GeV. Its short lifetime makes it impossible for the top quark to hadronize and therefore it is never observed in hadrons, but rather as a resonance decaying into other SM particles. This will further be discussed in Sec. 1.4. Searches for extra generations of heavy quarks are putting very stringent constraints on the masses (above the TeV scale) of such quarks (see for example Refs. [38, 39]).

TABLE 1.1: Summary of the fermions (leptons and quarks) of the SM [22] with their electric charge (in units of the elementary charge $q_e$) and an indication of the currently best estimate of their mass [40] (uncertainties are not shown).

| Fermions | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Leptons | | | | Quarks | | | |
| | Particle | | Q [$q_e$] | mass [GeV] | Particle | | Q [$q_e$] | mass [GeV] |
| 1st gen. | electron | ($e$) | -1 | $5 \times 10^{-4}$ | down | ($d$) | $-1/3$ | $5 \times 10^{-3}$ |
| | $e$ neutrino | ($\nu_e$) | 0 | $< 10^{-9}$ | up | ($u$) | $+2/3$ | $2 \times 10^{-3}$ |
| 2nd gen. | muon | ($\mu$) | -1 | 0.106 | strange | ($s$) | $-1/3$ | 0.1 |
| | $\mu$ neutrino | ($\nu_\mu$) | 0 | $< 10^{-9}$ | charm | ($c$) | $+2/3$ | 1.2 |
| 3rd gen. | tau | ($\tau$) | -1 | 1.78 | bottom | ($b$) | $-1/3$ | 4.2 |
| | $\tau$ neutrino | ($\nu_\tau$) | 0 | $< 10^{-9}$ | top | ($t$) | $+2/3$ | 173 |

### Antiparticles

A remarkable enrichment of the SM particle content lies in the existence of antiparticles. Each of the SM fermions has a partner with the exact same properties, but with opposite electric charge. Only for the neutrinos, an open question remains on whether they are Dirac fermions, such as the quarks and charged leptons, or whether they are Majorana fermions, in which case they would behave as their own antiparticle (for a comprehensive overview, see for example Ref [19], Addendum 17.8).
The existence of antiparticles was first inferred from the Dirac equation, which in its Lorentz covariant form can be written as

$$(i\gamma^\mu \partial_\mu - m)\,\psi(x) = 0. \tag{1.1}$$

In this notation, $\psi(x)$ denotes the Dirac fermion field, $\gamma^\mu$ ($\mu \in 0, 1, 2, 3$) are the gamma-matrices and $\partial_\mu$ is the space–time derivative $(\partial/\partial t, \partial/\partial x, \partial/\partial y, \partial/\partial z)$. The reader is referred to Sec. 1.2 for a more detailed quantum field theory description of particle fields. For this equation to be Lorentz covariant, the gamma matrices are required to be $4 \times 4$ matrices and more importantly, $\psi(x)$ is a four component spinor, naturally describing spin $1/2$ particles where two of the components have positive energy solutions, while the other two have negative energy solutions. These correspond to the particle and antiparticle spinors respectively. The negative energy solutions to the Dirac equation were historically thought to be of no physical relevance, but later received an appropriate interpretation as an antiparticle in a quantum field theory description. The first antiparticle to be discovered (in 1932) was the anti–electron or so–called positron [41]. The positron revealed itself as a track in a bubble chamber with an opposite curvature when compared to normal electrons passing through the

magnetic field, indicating an opposite electric charge. By now, antiparticles are produced copiously in particle accelerators and even dedicated antiparticle experiments[4] perform precision measurements on their properties to look for small deviations in the behavior of antimatter compared to normal matter.

### Forces and bosons

The SM not only describes the elementary particles that make up our universe, but also the fundamental forces that govern the interactions among these particles. In the SM, interactions between particles are described by the exchange of force–carrier particles known as bosons. They are integer spin particles that obey Bose–Einstein statistics. Four fundamental forces are currently known:

- **The electromagnetic force**: This force was described originally (in a classical way) by the Maxwell equations. In the quantum mechanical description, this force describes the interaction between electrically charged particles through the exchange of massless photons. The electromagnetic force is a long–range force, making it the dominant force at large distances (much larger than the size of atomic nuclei).

- **The weak nuclear force**: Being responsible for some types of radioactive decay, the weak interaction acts among particles that carry a so–called weak isospin (see Sec. 1.3) through the exchange of charged W bosons (called charged current interactions) or neutral Z bosons (called neutral current interactions). As expressed in its name, this force is relatively weak and acts only on small distances ($\sim 10^{-18}$ m).

- **The strong nuclear force**: The force that holds quarks together inside hadrons (for example protons) is called the strong nuclear force. It acts only between particles that carry so–called color charge (see again Sec. 1.3) and happens through the exchange of massless gluons. In the SM, this force acts only between quarks and gluons. It is by far the strongest force, but it is also short–ranged, meaning it only acts at distances smaller than (approximately) the proton size ($\sim 10^{-15}$ m).

- **The gravitational force**: Gravity is unique in many ways. It is the force we experience most directly in every–day life, though it is by far the weakest of all forces. But more importantly, it is currently not described by the SM and no corresponding boson has ever been observed. This is one of the big open questions in physics. The current description of gravity is based on Einsteins theory of general relativity, in which gravity is described as curvature of space-time (see for example Ref. [43]). Nevertheless, at the smallest scales probed by particle colliders such as the LHC, the gravitational force is so weak it can safely be ignored.

The forces with the corresponding bosons and their properties are also summarized in Tab. 1.2. It is important to keep in mind that each force (currently described by the SM) comes with its corresponding bosons and a specific charge or quantum number (electric charge, weak isospin or color). Whether or not a particle can interact through a specific force depends on the values of these charges, which differ for different types of particles.

[4] Examples of the large variety of antimatter experiments at CERN are AEGIS, ALPHA, ASACUSA, ATRAP and BASE [42].

TABLE 1.2: Summary of the forces in the SM with the corresponding bosons [22] with their electric charge (in units of the elementary charge $q_e$) and an indication of the currently best estimate of their mass [40] (uncertainties are not shown).

| | Bosons | | | |
|---|---|---|---|---|
| | Particle | | Q [$q_e$] | mass [GeV] |
| Electromagnetic force | photon | ($\gamma$) | 0 | 0 |
| Weak force | $W^{\pm}$ bosons | ($W^{\pm}$) | $\pm1$ | 80.4 |
| | Z boson | (Z) | 0 | 91.2 |
| Strong force | 8 gluons | (g) | 0 | 0 |
| Gravitational force$^{(*)}$ | unknown (graviton?) | | ? | ? |

$^{(*)}$ Gravity is not included in the SM, but is rather described by the theory of general relativity.

## 1.2 Quantum Field Theory and Gauge invariance

The mathematical framework in which the SM is expressed, is that of Quantum Field Theory (QFT). In such a theory, particles are presented as fields with specific transformation properties under a set of symmetry groups. The dynamics of those fields, as well as the interactions among them are completely determined by the Lagrangian density, $\mathscr{L}$. As a first example, consider the Lagrangian of a free spinor field $\psi$:

$$\mathscr{L}_{\text{Dirac}} = \overline{\psi}(x)\left(i\gamma^{\mu}\partial_{\mu} - m\right)\psi(x), \tag{1.2}$$

with the same notation as in Eq. (1.1) and where the adjoint spinor $\overline{\psi}$ is defined as $\psi^{\dagger}\gamma^{0}$, with $\psi^{\dagger}$ the hermitian conjugate of $\psi$. Not surprisingly, the equation of motion corresponding to this Lagrangian can be obtained through the Euler–Lagrange equations and yields the Dirac equation in Eq. (1.1). Therefore, $\mathscr{L}_{\text{Dirac}}$ expresses the kinematics of a free Dirac spinor.

According to Noethers theorem [44], symmetries in the theory correspond to conserved currents. Indeed, interacting fields through currents that correspond to the bosons expressed in Tab. 1.2, are introduced in the QFT by the principle of *gauge symmetries* and *gauge invariance*. Referring back to Eq. (1.2) as an example, it is clear that this Lagrangian is invariant under a global U(1) phase transformation

$$\psi(x) \to \psi'(x) = e^{-i\omega}\psi(x), \quad \overline{\psi}(x) \to \overline{\psi}'(x) = e^{i\omega}\overline{\psi}(x), \tag{1.3}$$

where $\omega$ is a constant, independent of $x$. However, a much stronger requirement is for this Lagrangian to be invariant under a *local U(1) transformation*

$$\psi(x) \to \psi'(x) = e^{-i\omega(x)}\psi(x), \quad \overline{\psi}(x) \to \overline{\psi}'(x) = e^{i\omega(x)}\overline{\psi}(x), \tag{1.4}$$

where $\omega(x)$ is now a function of the space–time point $x$. Indeed, due to the fact that the derivative $\partial_{\mu}$ will now act also on $\omega(x)$, $\mathscr{L}_{\text{Dirac}}$ is not anymore invariant under such a local phase transformation. In order to restore the invariance under these local U(1) transformations, a new vector field, $A_{\mu}$, is introduced and an interaction term is added to the Lagrangian (omitting the space–time point dependence $(x)$ from the

notation):

$$\mathscr{L}_{\text{Dirac}} = \overline{\psi}\left(i\gamma^{\mu}\left(\partial_{\mu} + igA_{\mu}\right) - m\right)\psi, \tag{1.5}$$
$$\equiv \overline{\psi}\left(i\gamma^{\mu}\mathcal{D}_{\mu} - m\right)\psi, \tag{1.6}$$
$$= i\overline{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\overline{\psi}\psi - g\overline{\psi}\gamma^{\mu}A_{\mu}\psi. \tag{1.7}$$

Here, $g$ is an interaction strength between the spinor field $\psi$ and the gauge field $A_{\mu}$ and we also introduced the so–called covariant derivative $\mathcal{D}_{\mu} = \partial_{\mu} + igA_{\mu}$. This Lagrangian is again invariant under the local phase transformation, if we require $A_{\mu}$ to transform as

$$A_{\mu} \to A'_{\mu} = A_{\mu} + \frac{1}{g}\left[\partial_{\mu}\omega(x)\right]. \tag{1.8}$$

This example has shown that by requiring the Lagrangian to be gauge invariant under a specific gauge group, one naturally introduces gauge fields in the Lagrangian that describe interactions between the particles. The Lagrangian is finally extended with gauge invariant kinetic terms for the gauge field $A_{\mu}$, using the field strength tensor $F_{\mu\nu} = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu}$, that corresponds to the Lagrangian of the Maxwell equations to describe electromagnetism. The final gauge invariant Dirac Lagrangian can thus be written as

$$\mathscr{L}_{\text{Dirac}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \overline{\psi}\left(i\gamma^{\mu}\left(\partial_{\mu} + igA_{\mu}\right) - m\right)\psi. \tag{1.9}$$

Notice that a bare mass term for the gauge field ($\propto m_A^2 A_{\mu}A^{\mu}$) is not gauge invariant and therefore not allowed. Up to this point, the gauge bosons remain massless and in order to allow them to obtain mass, the idea of spontaneous symmetry breaking needs to be introduced, as will be done in Sec. 1.3.

The example above has shown the principle for a simple U(1) unitary transformation, which can be fully described by one degree of freedom, or one generator of the U(1) gauge group. The more general case of a special unitary group of order $n$, SU($n$), will be important to describe the SM of particle physics in the next section. Such a group is described by a set of $n^2 - 1$ generators, $T^a$ where $a \in \{1, 2, \cdots, n^2 - 1\}$. In order to make the Lagrangian gauge invariant under an SU($n$) transformation, a gauge vector field $A_{\mu}^a$ will have to be introduced for each generator of the group. The covariant derivative will then be defined as $\mathcal{D}_{\mu} = \partial_{\mu} + igT^aA_{\mu}^a$ and will become an $n \times n$ matrix. The interaction or force described by such a SU($n$) gauge group will therefore have $n^2 - 1$ gauge bosons that make up the currents and mediate the interaction. If the gauge group is non-abelian[5], the definition of the field strength tensor and the kinetic term for the gauge fields in the Lagrangian naturally allows for self–interactions of the gauge bosons. This is not the case for an abelian symmetry group.

After imposing gauge invariance in a QFT, one ends up with a Lagrangian that dictates the allowed interactions between the elementary particles. In order to calculate the value of an observable quantity related to a given process, one needs to construct a matrix element $\mathcal{M}$ that describes the kinematics of the particles (i.e. their four–momenta) and the interaction vertices that are involved. Such a matrix element should take into account all possible ways of achieving a certain final–state for a given initial–state of particles, as well as energy and momentum conservation. In a perturbative QFT the matrix element is derived through a set of *Feynman rules*

[5]For non-abelian groups, the generators of the group do not commute.

that are associated to *Feynman diagrams*, which are visual representations of how the process can evolve from its initial–state (left side of the diagram) to its final–state (right side of the diagram). An example is shown in Fig. 1.1, showing a gluon fusion process that results in the production of a top quark – top antiquark ($t\bar{t}$) pair. Each of these top quarks then decays into a bottom quark and a charged W boson, which subsequently decays into a charged lepton and a (anti)neutrino. This process is often referred to as the dileptonic decay channel of the $t\bar{t}$ production in proton collisions.

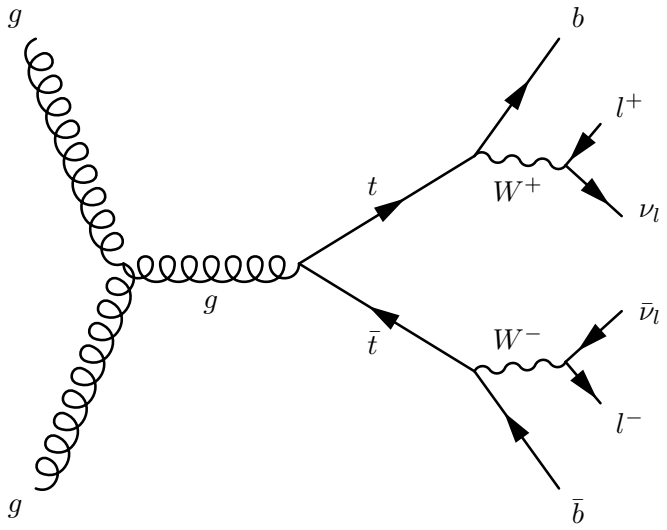


FIGURE 1.1: Feynman diagram describing the production of a top quark pair from gluon fusion. The top quarks decay into a bottom quark and a W boson, which in turn decay into a charged lepton and a (anti)neutrino. This is referred to as the dileptonic decay channel of the top quark pair production.

The calculation of observable quantities then happens through a perturbative expansion in terms of the coupling constants of the interactions that take place. This means that the computation starts from those diagrams which need the minimal amount of interaction vertices to produce the required final–state. These minimal diagrams are called the tree–level diagrams and allow for calculations at first–order or leading–order (LO) accuracy. Adding additional interaction vertices (and consequently more loops in the diagrams) allows for more accurate predictions at next–to–leading order (NLO) or even next–to–next–to–leading order (NNLO). The higher–order corrections to the LO diagrams are often referred to as *radiative corrections*. In theory one should include an infinite amount of orders, though practically the most accurate predictions today are typically limited to NLO or NNLO accuracy. Such predictions are however precise enough to match the experimental precision that can be achieved in most measurements today. The value of an observable quantity, such as a cross section, is proportional to the square of the full matrix element, including in the sum all possible diagrams that describe the process up to a required order in perturbation theory. The amount of contributing diagrams quickly grows with increasing orders in the perturbative expansion, explaining the limitations of the most accurate predictions (even with the rapid growth of computing power over the last decades).

It is finally worth noting that the calculation of radiative corrections in the SM naturally leads to divergent integrals (See Ref. [24] Chapters 9 and 10 for a comprehensive overview). There exist a variety of regularization procedures that modify the

theory such that it remains finite to all orders in perturbation theory. Such artificial changes can then be rectified by a technique known as *renormalization*. This concept starts from the realization that the non–interacting ("bare") SM particles that appear in the Lagrangian are not the same as the real physical states that interact with each other. For example, the observable mass or electric charge of a physical (interacting) electron obtains radiative corrections to its bare mass or charge. These corrections can be interpreted as a cloud of surrounding electrons and photons that are constantly created and annihilated (or as additional loops added to the tree–level diagrams describing these properties). The regularized infinities now appear in these relations between the bare properties and the physically observable ones. One then reverts the regularized theory back to its original form, resulting in finite and well-behaved predictions for the observable quantities, whereas the original infinities are absorbed in the unphysical and therefore unobservable bare quantities.

## 1.3 The Standard Model gauge group

The important question is of course which symmetries manifest themselves in nature. The Standard Model of particle physics is described by a $\mathrm{SU(3)_C} \bigotimes \mathrm{SU(2)_L} \bigotimes \mathrm{U(1)_Y}$ gauge symmetry group. The subscript "C" refers to the color quantum number of the theory of Quantum Chromodynamics (QCD), which describes the strong force. The theory of QCD indeed follows from the requirement of gauge invariance under the SU(3) group. One of the elegant features of the SM is the fact that the electromagnetic and the weak force are collectively described by the Electroweak (EW) theory, which follows from requiring gauge invariance under the $\mathrm{SU(2)_L} \bigotimes \mathrm{U(1)_Y}$ gauge group. The subscript "L" refers to the left–handed chiral structure of the SU(2) group and the subscript "Y" refers to the weak hypercharge as a quantum number of the theory. Each of these two theories will be discussed in more detail below. Afterwards, the concept of electroweak symmetry breaking (EWSB) will be introduced that allows to put mass terms in the theory in a gauge invariant way. Finally, a brief discussion of the flavor structure of the SM will be included.

### The Electroweak theory

The Electroweak theory provides a unified description of the electromagnetic and the weak forces, by requiring gauge invariance under the $\mathrm{SU(2)_L} \bigotimes \mathrm{U(1)_Y}$ gauge group. The abelian U(1) gauge group, described by the hypercharge Y, has been discussed already in Sec. 1.2 and gives rise to a single gauge field $B_\mu$, with gauge coupling $g_1$. The non-abelian $\mathrm{SU(2)_L}$ group characterizes the chirality structure of the electroweak theory. By making use of the chiral projection operators[6]

$$P_L = \frac{1}{2}\left(1-\gamma^5\right), \qquad P_R = \frac{1}{2}\left(1+\gamma^5\right), \tag{1.10}$$

the four-component Dirac spinors can be projected onto the left–handed (LH) chiral states and the right–handed (RH) chiral states. Supported by the clear experimental evidence for the parity–violating nature of the EW theory, the corresponding gauge fields only interact with the LH fermion fields, which are consequently grouped into doublets. For example, the first–generation LH leptons are grouped together in an $\mathrm{SU(2)_L}$ doublet ($e_L$,$\nu_{eL}$), or the first–generation LH quarks form the doublet ($u_L$,$d_L$). The right–handed fermion fields are represented by singlets under the SU(2)

[6] In this notation one defines $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$.

symmetry. The $\mathrm{SU(2)_L}$ gauge group has three generators (most often represented by the $2 \times 2$ Pauli–matrices $\sigma^a$) and correspondingly gives rise to three gauge vector boson fields, which will be denoted $W_\mu^a$ ($a \in \{1, 2, 3\}$). The corresponding gauge coupling will be denoted $g_2$.

However, the physically observable gauge fields of the EW theory are the photon field $A_\mu$, the neutral Z boson field $Z_\mu^0$ and the two charged W boson fields $W_\mu^\pm$, which turn out to be linear superpositions of the four gauge fields of the $\mathrm{SU(2)_L} \bigotimes \mathrm{U(1)_Y}$ gauge group

$$A_\mu = \sin\theta_W \, W_\mu^3 + \cos\theta_W \, B_\mu, \tag{1.11}$$

$$Z_\mu^0 = \cos\theta_W \, W_\mu^3 - \sin\theta_W \, B_\mu, \tag{1.12}$$

$$W_\mu^\pm = \sqrt{1/2} \left( W_\mu^1 \mp i W_\mu^2 \right), \tag{1.13}$$

where $\theta_W$ is the weak mixing angle or the Weinberg angle and is defined as

$$\tan\theta_W = \frac{g_1}{g_2}. \tag{1.14}$$

Putting all of this together, one obtains a theory of massless fermions and EW bosons, including electromagnetic interactions between the photon field and the charged fermions, charged and neutral current interactions between the fermions and the W and Z bosons respectively and the allowed self–interactions between the gauge bosons.

### Quantum Chromodynamics

The theory of Quantum Chromodynamics (QCD) describes how the strong nuclear force is included in the SM. It follows from imposing an $\mathrm{SU(3)_C}$ invariance of the SM Lagrangian. This non–abelian gauge group has eight corresponding generators, typically represented by the $3 \times 3$ Gell-Mann matrices $\lambda^a$ ($a = \{1, \cdots, 8\}$) and therefore also eight massless gauge boson fields $G_\mu^a$ called gluons. The quantum property related to this gauge symmetry is referred to as the color charge and comes in threefold: red, green and blue (together with their anticolors). The strong interaction therefore only affects colored particles, namely quarks and gluons, whereas other fermions and bosons are singlets under this symmetry. The coupling constant that enters the covariant derivative related to this symmetry is called the strong coupling constant $g_s$. One of the striking features of this theory is that this coupling strength decreases as the energy of the probed interactions increases[7] (or as the distance between interacting colored particles decreases). This feature is known as *asymptotic freedom*. This leads to the principle of *color confinement*, meaning that quarks are not observed as individual asymptotic states, but are always grouped together in color–neutral combinations known as hadrons. At particle colliders such as the LHC, the quarks that are produced in the high–energy particle collisions therefore immediately form hadrons, which is usually referred to as *hadronization*. These hadrons undergo a chain of decays and additional radiations, and eventually form a spray of particles that moves through the detector. Such a collimated shower of particles is called a "*jet*" and therefore, instead of observing for example a $b$ quark, we rather observe a $b$ jet in the detector. This is further outlined in Chapter 3.

[7] The value of the coupling constant is not a constant as its name may suggest, but rather varies with the energy of the interaction that is probed. This is known as the running of the coupling constant.

### Electroweak symmetry breaking

The requirement of gauge invariance forbids any explicit mass terms for the bosons or the fermions in the SM Lagrangian. The principle of Electroweak symmetry breaking (EWSB) allows to generate in a gauge invariant way mass terms for the bosons and fermions through the insertion of a complex scalar field in the theory. The EWSB principle was first written down in 1964 by Robert Brout and Francois Englert [7] and around the same time (but independently) by Peter Higgs [8]. A complex scalar field $\phi$, represented as a doublet under the $\mathrm{SU(2)_L}$ gauge group and which is charged under the $\mathrm{U(1)_Y}$ gauge group, is added to the SM Lagrangian:

$$\mathscr{L}_\phi = (\partial^\mu \phi)^\dagger (\partial_\mu \phi) - V(\phi), \tag{1.15}$$

$$\text{with} \quad V(\phi) = -\frac{1}{2}\mu^2 \phi^\dagger \phi + \frac{1}{4}\lambda \left(\phi^\dagger \phi\right)^2. \tag{1.16}$$

$\mathrm{V}(\phi)$ is referred to as the scalar potential, where $\mu$ is a real constant with units of mass, and $\lambda$ a dimensionless parameter that quantifies the self–interaction strength of the scalar field quartic coupling. The potential is explicitly written in this way to emphasize the negative coefficient in front of the quadratic term. This gives the potential a degenerate set of minima, where the scalar field acquires a non–zero vacuum expectation value[8] (VEV) $v = \sqrt{\mu^2/\lambda}$, which is measured to be $\sim 246$ GeV [40]. Expanding the scalar field around one of its minima, it can be written as:

$$\phi = -\frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} + \begin{pmatrix} \phi^1 + i\phi^2 \\ h^0 + i a^0 \end{pmatrix}, \tag{1.17}$$

where its four degrees of freedom, $\phi^1, \phi^2, h^0$ and $a^0$ are explicitly written. The fact that this VEV is not symmetric under the $\mathrm{SU(2)_L} \bigotimes \mathrm{U(1)_Y}$ gauge symmetry, implies that this symmetry is spontaneously broken down to $\mathrm{U(1)_{EM}}$. Indeed, the principle of gauge invariance dictates to transform the derivatives in Eq. (1.15) into covariant derivatives that follow the $\mathrm{SU(2)_L} \bigotimes \mathrm{U(1)_Y}$ gauge symmetry:

$$\partial_\mu \rightarrow \mathcal{D}_\mu = \partial_\mu + \frac{i}{2} g_1 Y B_\mu + \frac{i}{2} g_2 \sigma^a W^a_\mu. \tag{1.18}$$

Through the mixing of these bosons defined in Eqs. (1.11) – (1.13), one naturally obtains mass terms for the massive bosons. Therefore also the VEV is connected to the masses of the $\mathrm{W^\pm}$ and $\mathrm{Z^0}$ bosons and the electroweak couplings $g_1$ and $g_2$ via the following relations:

$$m_W^2 = \frac{v^2 g_2^2}{4}, \qquad m_Z^2 = \frac{v^2 \left(g_1^2 + g_2^2\right)}{4}. \tag{1.19}$$

The three components $\phi^1, \phi^2$ and $a^0$, usually referred to as massless Goldstone bosons, do not appear as physical states in nature. In a specific gauge choice, known as the *unitary gauge*, these massless Goldstone bosons vanish. However the remaining degree of freedom, which was denoted $h^0$, represents a physical scalar boson known as the Brout-Englert-Higgs boson (BEH) or Higgs (H) boson in short. The $\mathrm{SU(2)_L} \bigotimes \mathrm{U(1)_Y}$ gauge symmetry is said to be broken down to a $\mathrm{U(1)_{EM}}$ symmetry with only one remaining unbroken symmetry. This symmetry corresponds to the photon that indeed remains massless. Also the gluons remain massless given

[8] The general set of minima can be expressed as $v = \sqrt{\mu^2/\lambda} \cdot e^{i\theta}$. For simplicity the scalar field is usually expanded around the minimum at $\theta = 0$.

that the scalar is a singlet under the $\mathrm{SU}(3)_\mathrm{C}$ gauge group.

What remains is to generate mass terms for the SM fermions. This can be achieved by introducing Yukawa interactions between the fermions and the scalar field into the SM Lagrangian:

$$\mathscr{L}_{\text{Yukawa}} = -\lambda_d^{ij}\,\bar{q}_L^i \phi d_R^j - \lambda_u^{ij}\,\bar{q}_L^i \tilde{\phi} u_R^j - \lambda_\ell^{ij}\,\bar{\ell}_L^i \phi e_R^j + \text{hermitian conjugate.} \tag{1.20}$$

Here $\tilde{\phi} = i\sigma_2\phi^*$, $q_L$ is a LH quark doublet, $u_R$ and $d_R$ are the RH up and down quark singlets respectively, $\ell_L$ is the LH doublet of leptons and $e_R$ is the RH lepton singlet, $i$ and $j$ are generation indices and perhaps most importantly, $\lambda_u^{ij}$, $\lambda_d^{ij}$ and $\lambda_\ell^{ij}$ are the up–quark, down–quark and lepton Yukawa coupling strengths respectively. They are directly related to the fermion masses $m_f$, which after rotating into the mass–eigenstates (see the next paragraph) can be expressed through

$$m_{f_i} = \frac{\lambda_{f_i} v}{\sqrt{2}}. \tag{1.21}$$

### Flavor in the SM

As discussed in Sec. 1.1, there exist three generations of quarks, each containing an up–type and a down–type quark. Therefore a total of six quark flavors are known in the current SM (up, down, strange, charm, bottom and top). The electroweak sector of the SM has an interesting interplay between the different quark flavors, as it allows for flavor–changing charged–current interactions through the exchange of the charged $W$ bosons. The neutral current interactions are however diagonal in flavor space, and therefore no flavor–changing neutral–current (FCNC) interactions are allowed at tree-level[9] in the SM. As an example, the predicted branching fraction of a top quark decaying into a charm or up quark through the exchange of a neutral gauge boson or a Higgs boson is predicted to be of the order of $\sim 10^{-12}$ to $10^{-17}$ [45] and therefore lies far beyond the experimental sensitivity that can be reached today. It is however interesting to note that in some BSM scenarios (for example a 2–Higgs doublet model) this branching fraction is enhanced up to the order of $\sim 10^{-3}$, which makes these processes an extremely interesting place to search for new physics signatures (see for example Refs. [46–56] for an overview of the most up–to–date CMS and ATLAS analyses on this topic).

Focusing again on the charged current interactions, it is important to realize that the quarks that carry the same quantum numbers can mix. Therefore one can for example express the down–type quarks in two different bases, related by a unitary rotation in three–dimensional space. They can be either expressed in their mass eigenstates $\begin{pmatrix} d & s & b \end{pmatrix}$ to describe their free propagation, or in their weak eigenstates $\begin{pmatrix} d_w & s_w & b_w \end{pmatrix}$ when they take part in a charged–current weak interaction. These two bases are related via

$$\begin{pmatrix} d_w \\ s_w \\ b_w \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \equiv \mathcal{V}_{\text{CKM}} \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \tag{1.22}$$

[9] FCNC interactions can be generated at loop level, which are highly GIM suppressed [32].

where $\mathcal{V}_{\text{CKM}}$ is the unitary Cabibbo–Kobayashi–Maskawa matrix that describes the quark mixing and quantifies the CP violation in the charged–current weak interactions. This matrix shows up in the charged weak current $J_\mu$ as

$$J_\mu = \begin{pmatrix} \overline{u} & \overline{c} & \overline{t} \end{pmatrix} \gamma_\mu \left(1 - \gamma_5\right) \mathcal{V}_{\text{CKM}} \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \tag{1.23}$$

from which it is clear that the square of each entry $|V_{qq'}|^2$ of $\mathcal{V}_{\text{CKM}}$ defines the transition probability from one quark flavor $q$ into another $q'$ via a charged current interaction. The experimentally determined values of the CKM matrix are given by [40]

$$\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} = \begin{pmatrix} 0.97420 \pm 0.00021 & 0.2243 \pm 0.0005 & (3.94 \pm 0.36) \times 10^{-3} \\ 0.218 \pm 0.004 & 0.997 \pm 0.017 & (42.2 \pm 0.8) \times 10^{-3} \\ (8.1 \pm 0.5) \times 10^{-3} & (39.4 \pm 2.3) \times 10^{-3} & 1.019 \pm 0.025 \end{pmatrix}.$$

An important conclusion from these results in the rest of this work is the fact that $|V_{tb}|$ is consistent with unity [57] and much larger than $|V_{td}|$ or $|V_{ts}|$, indicating that the top quark will decay almost exclusively into a bottom quark and a charged W boson (as illustrated in Fig. 1.1).

Finally, it should be noted that in the lepton sector a similar neutrino mixing matrix exists known as the Pontecorvo–Maki–Nakagawa–Sakata (PMNS) matrix [58]. It parametrizes the neutrino mixing, which will briefly be touched upon in Sec. 1.5, and will not be discussed further here.

## 1.4 The top-quark sector

Being the heaviest of all known SM particles, the top quark is of particular interest to study at the LHC where it is abundantly produced. Its width is measured to be $1.36^{+0.14}_{-0.11}$ GeV [59], leading to a lifetime of the order of $10^{-24}$ seconds. This particularly small lifetime is responsible for the rapid decay of the top quark before it can hadronize, making it special from other quarks. Due to its large mass, it has by far the strongest interaction with the H boson due to the large Yukawa coupling strength as defined in Eq. (1.21). It decays almost exclusively to a bottom quark and a W boson, due to the value of $|V_{tb}|$ in the CKM matrix being so close to unity. All of these properties give the top quark a special place in the SM, and have motivated the high–energy physics community to unroll an ambitious program in top quark precision measurements at particle colliders such as the Tevatron and the LHC.

### The top quark mass

The mass of the top quark is a crucial ingredient in global fits of all the SM parameters together. Such global fits submit the SM to a rigorous test and provide information on whether or not the SM is a consistent theory[10]. Fig. 1.2 shows the results of such a fit in the parameter space of the W boson mass and the top quark mass. So far it seems that direct measurements of these masses are indeed consistent with the results of such a global fit from which these measurements are excluded, confirming the predicted relations between the parameters of the SM.

[10]The SM has 19 free parameters that can only be determined by measuring them, and the fermion masses are amongst those. It is therefore extremely important to measure these parameters precisely and to see whether their values obey the predicted relations in the SM.

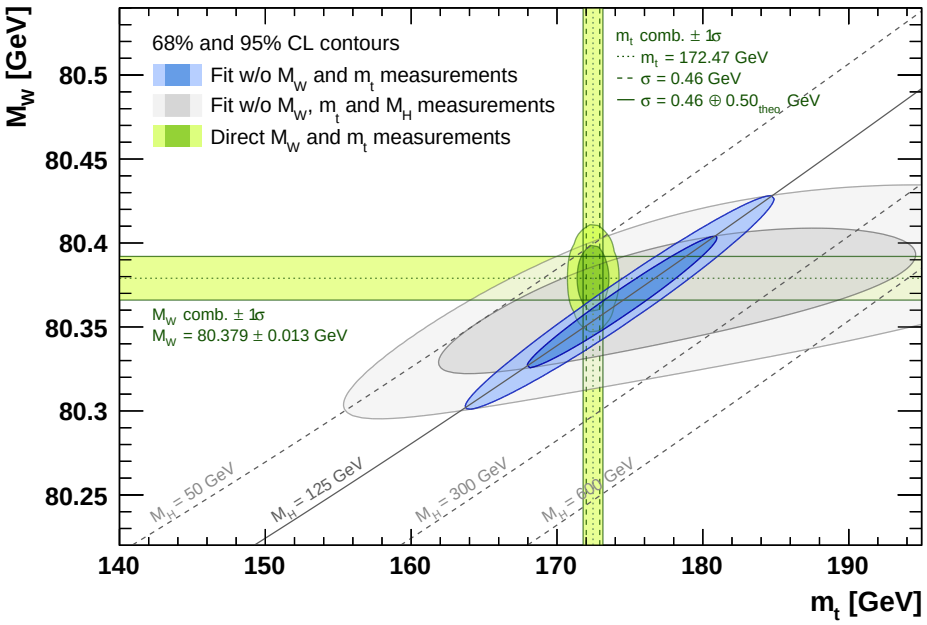


FIGURE 1.2: Results of a fit of the SM parameters as a function of the top–quark mass and the W boson mass. The coloured areas compare the allowed regions from direct measurements of these masses (green) to a global fit excluding these mass measurements (blue and grey). The consistency between these approaches signifies a strong test of the SM and its predicted relations between the parameters. Figure taken from [60].

The value of the top quark mass, $m_t$, has been measured in many different final–states and through a large variety of methods. Some of these analyses try to reconstruct the top quark from its decay products and fit the invariant mass of that combination of particles, others infer the value of $m_t$ through its dependence in the cross section of top quark pair production or single top quark production. There are even dedicated analyses that try to construct more involved observables that are particularly sensitive to $m_t$. An overview of the state–of–the–art measurements is given in Fig. 1.3, resulting in the most accurate values from CMS of $m_t = 172.44 \pm 0.48$ GeV [61] and ATLAS obtaining $m_t = 172.69 \pm 0.48$ GeV [62]. The accuracy of these measurements has reached the level of less than 500 MeV, raising the necessity of posing the question which theoretical parameter is effectively being measured [63]. Despite the ongoing discussions on the interpretation of these numbers, this can safely be labelled as a remarkable precision measurement and improving further the precision will be a non–trivial challenge in the future.

### The production of top quarks at the LHC

In proton–proton collisions at the LHC, top quarks are produced abundantly either alone, in pairs or even four of them at the same time. Each of these cases proceeds through different initial–states, with different probabilities (i.e. with a different cross section) and results in different final–states. Before discussing the production mechanism of these processes at the LHC, it is important to emphasize the particle content of the proton. Even though the proton is classically composed of two up quarks and one down quark, in the quantum mechanical description all quarks and even gluons are present in the proton with a certain probability. The proton Parton

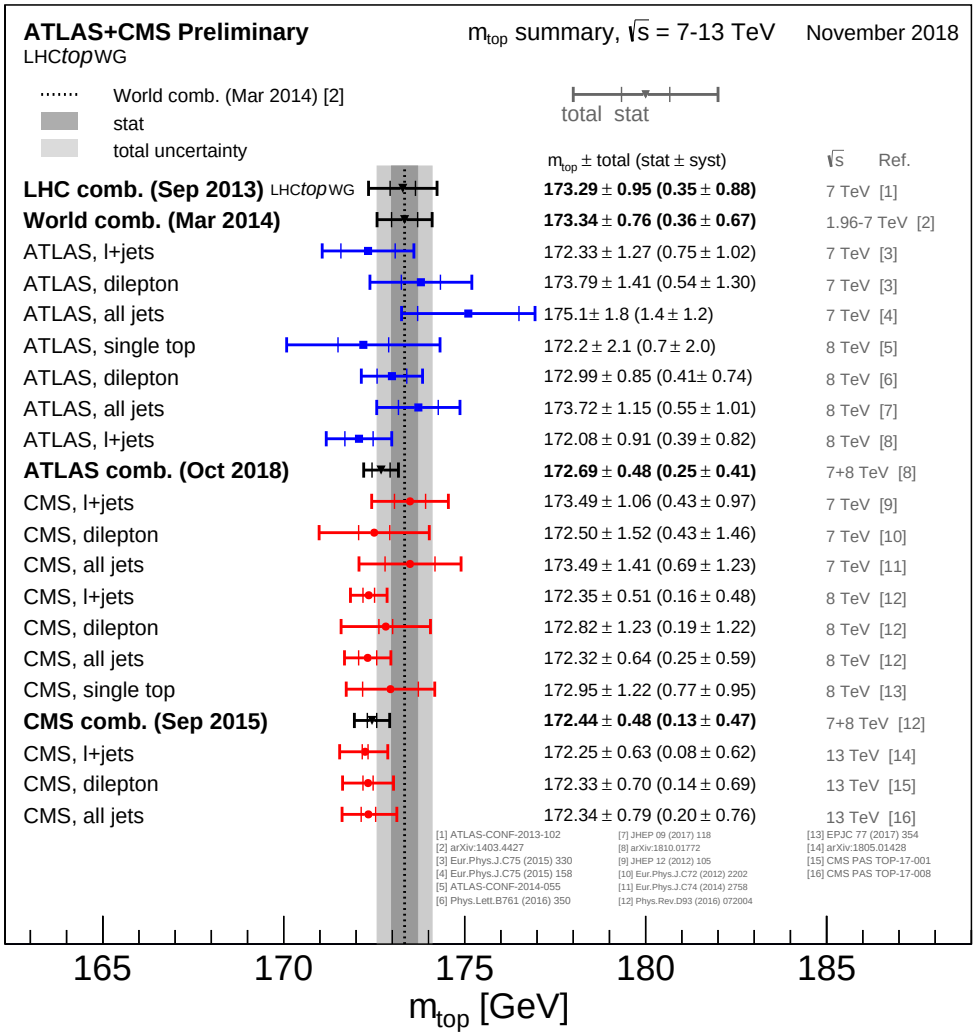


FIGURE 1.3: Summary of the state-of-the-art top–quark mass measurements from CMS and ATLAS. Figure taken from [64].

Density Function, or proton PDF in short, describes how likely it is for a given parton (quark or gluon) to take part in an interaction. This PDF is a function of the fractional momentum $x$ of the proton that the parton carries. Therefore a collision of two protons is effectively a collision of two partons inside the proton, of which the energy is only a fraction of the initial proton energy the collider can deliver. This concept is further outlined in Sec. 3.1.1.

In proton–proton collisions at the LHC, the process with the largest cross section is the production of a top quark pair ($t\bar{t}$). This final–state is dominantly produced through gluon fusion, as illustrated with the middle and right Feynman diagram in Fig. 1.4. Another subdominant production mechanism is through the fusion of a quark and an antiquark[11], illustrated by the left Feynman diagram in Fig. 1.4.

The $t\bar{t}$ production cross section $\sigma_{t\bar{t}}$ has been measured at different particle colliders and at different center–of–mass collision energies ($\sqrt{s}$). Fig. 1.5 summarizes

[11] This quark fusion mechanism was in fact the dominant $t\bar{t}$ production mechanism at the Tevatron, where protons were collided with antiprotons.

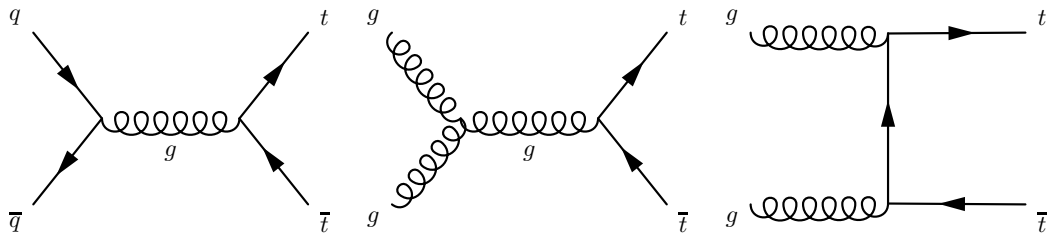


FIGURE 1.4: Feynman diagrams of the different (leading order) top quark pair production mechanisms at the LHC. The left diagram represents the quark fusion process, whereas the middle and right diagrams illustrate the dominant gluon fusion production mechanism.

these measurements and compares them to NNLO theoretical predictions, showing a good agreement with the SM expectations. These theoretical calculations predict a value of $\sigma_{t\bar{t}} \approx 831$ pb [65] at the LHC at a center–of–mass collision energy of $\sqrt{s} = 13$ TeV and for a top quark mass hypothesis of $m_t = 172.5$ GeV. This means that by now the LHC has produced over a 100 million top quark pairs!

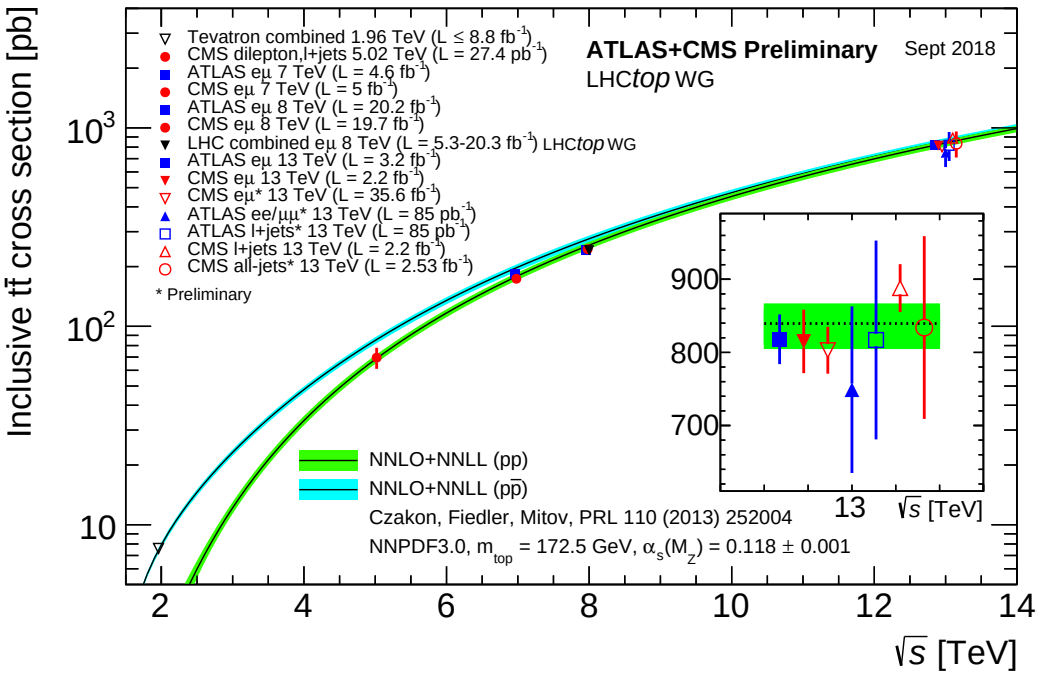


FIGURE 1.5: Summary of the measured values of the $t\bar{t}$ production cross section at different center–of–mass energies, compared to NNLO theory calculations. Figure taken from [64].

At the LHC, also the production of a single top quark (in association with other particles) is studied in detail. This process is usually subdivided in three production mechanisms: the s–channel production, the t–channel production and the associated $tW$ production, for which the LO feynman diagrams are shown in Fig. 1.6. In contrast to the strong couplings involved in the $t\bar{t}$ production, these diagrams involve mostly electroweak processes. On top of that, some LO diagrams of the t–channel and $tW$ production require an initial–state $b$ quark, for which the proton PDF is

particularly small (see Sec. 3.1.1). Additionally, there is a clear asymmetry between the production of a single top quark and a single top antiquark in the t–channel and $s$–channel, resulting from a different proton PDF for initial–state quarks and antiquarks. The predicted cross sections for the summed single top quark and single antitop quark production ($\sigma_{t+\bar{t}}$) at the LHC at center–of–mass collision energy of $\sqrt{s} = 13$ TeV and for a top quark mass hypothesis of $m_t = 172.5$ GeV are given in Tab. 1.3.

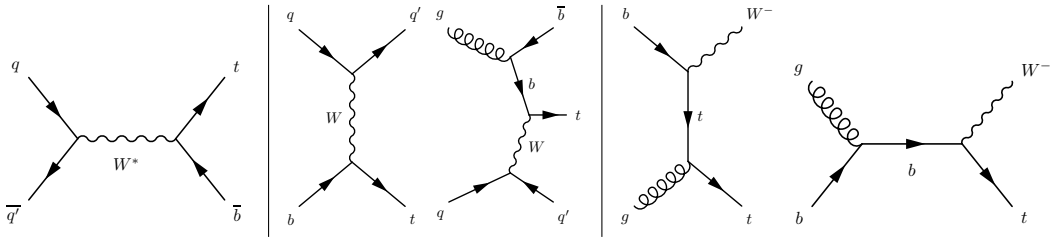


FIGURE 1.6: Feynman diagrams of the different (leading order) single top quark production mechanisms at the LHC. The left diagram represents the s–channel production, the middle two diagrams represent the t–channel production and the right two diagrams illustrate the associated $tW$ production.

TABLE 1.3: Cross section predictions for single top quark production in the different production channels at the LHC for $\sqrt{s} = 13$ TeV, $m_t = 172.5$ GeV. Numbers taken from [66], but dedicated references from therein are quoted in the last column of the table.

| Single top quark cross section $\sigma_{t+\bar{t}}$ @ LHC, $\sqrt{s} = 13$ TeV, $m_t = 172.5$ GeV | | | | |
|---|---|---|---|---|
| | Value [pb] | Uncertainty [pb] | Order | Reference |
| s-channel | 10.32 | $^{+0.40}_{-0.36}$ | NLO | [67, 68] |
| t-channel | 216.99 | $^{+9.04}_{-7.71}$ | NLO | [67, 68] |
| tW-channel | 71.7 | $^{+1.80}_{-1.80}$ | NNLO | [69, 70] |

Finally it is also possible to produce four top quarks at the same time. With a predicted cross section of $\sigma_{4t} = 9.2$ fb [71], this extremely rare process is very challenging to extract from the overwhelming backgrounds at the LHC. Considering the many possible decay modes for the collection of the four top quarks there exist a variety of analysis strategies to search for this rare signal. The CMS and ATLAS Collaborations are taking on this challenge [72–74], so far only putting upper limits on the production cross section. Nevertheless, with the large dataset that the LHC has provided so far, it is expected that this signature will be discovered in the not too far future. This final–state is particularly interesting to study as it is largely sensitive to new physics effects that might enhance the four top quark cross section and change its kinematics.

### The top-Higgs interplay in the SM

Since the discovery of the H boson in 2012 [4, 6], the Higgs physics program has moved from the discovery to the precision frontier. One of the crucial tests of the SM H boson is to measure its couplings to the other massive SM particles. Given

that the Higgs field couples to the massive fermions with a strength proportional to their mass (see Eq. (1.21)), by far the strongest Yukawa coupling is expected to be the one from the top quark, $\lambda_t$. The value of $\lambda_t$ is very close to unity and much larger than that of the second heaviest quark, the bottom quark. In order to verify the Yukawa sector, measuring the interactions between the H boson and the quarks is an indispensable test of the SM. Global combined measurements of the H boson couplings (to both fermions and massive bosons) have been performed by CMS [75, 76] and ATLAS [77] and result in the most sensitive (but often indirect) measurements of the coupling strengths, as shown in Fig. 1.7. This figure shows the agreement between the observation and the SM prediction, and parametrizes also the allowed deviation due to BSM effects (see green and yellow bands).

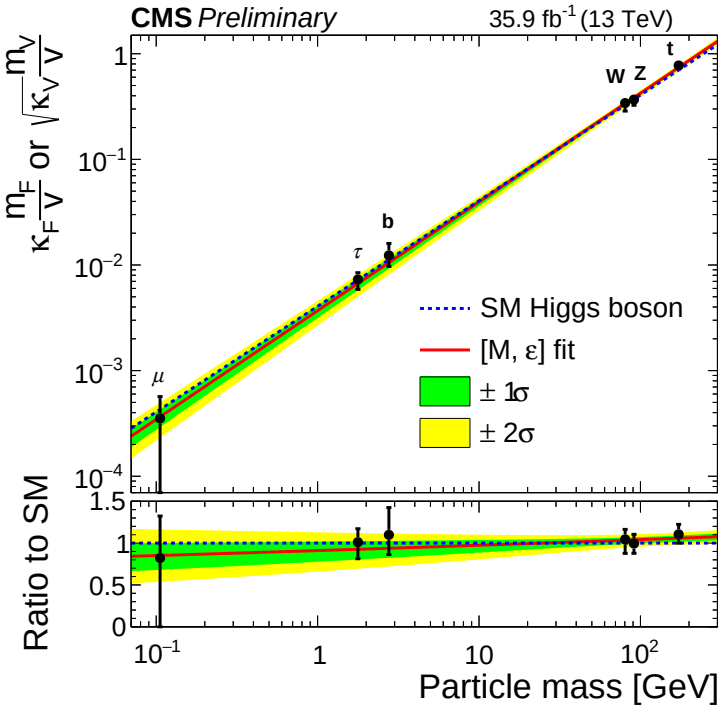


FIGURE 1.7: Summary of the couplings of the H boson to the SM fermions and massive bosons from a combined fit of the measured H boson properties. The green (yellow) band shows the $\pm 1\sigma$ ($\pm 2\sigma$) allowed deviations from the SM predictions due to possible BSM effects. Figure taken from [75].

Recently, in 2018, both the ATLAS and CMS collaborations succeeded in directly measuring the coupling of the H boson to the top quark [9, 10] and to the bottom quark [78, 79]. One very interesting topology to be studied in this context is the associated production of a H boson with a top quark pair, where the H boson decays into a pair of $b$ quarks ($pp \to t\bar{t}H$, $H \to b\bar{b}$), as it is sensitive to both the top quark and the bottom quark Yukawa coupling. This particular topology has been studied by CMS [80] and ATLAS [81] in the last few years. It turns out to be an extremely challenging final–state due to overwhelming and irreducible backgrounds of SM top quark pair production with additional heavy–flavor (HF) jets. The uncertainty on the obtained signal strength[12] of this process is of the order of 50–60%, making

[12]The signal strength $\mu$ of a given process is defined as its measured cross section, divided by its SM expectation, i.e. $\mu = \sigma/\sigma_{\mathrm{SM}}$.

the measurement not yet significant to claim an actual observation of the process (the observed result is still consistent with a signal strength of 0). The dominant uncertainty in these analyses comes from the normalizations of the SM $t\bar{t}$+HF cross section. This background is composed of the SM production of $t\bar{t}b\bar{b}$ and to a lesser extend of $t\bar{t}c\bar{c}$ and $t\bar{t}$+light–flavor ($t\bar{t}$LF) events. Currently, these uncertain background normalizations are taken into account by a conservative 50% uncertainty assigned to each of these $t\bar{t}$+ HF components. To improve this estimate, a large effort has been made in both the CMS [11, 12] and ATLAS [13–15] Collaborations to conduct measurements of the dominant SM $t\bar{t}b\bar{b}$ background at different center–of–mass energies. These analyses focus on measuring the inclusive $t\bar{t}jj$ cross section (where $j$ is a jet of any flavor) and the ratio $R_b = \sigma_{t\bar{t}b\bar{b}}/\sigma_{t\bar{t}jj}$, from which the absolute value of $\sigma_{t\bar{t}b\bar{b}}$ is then extracted. Most of the existing measurements are conducted in the dileptonic decay channel of the top quarks. The results of these measurements are summarized in Tab. 1.4. However, there exist also ongoing measurements in the single lepton or fully hadronic final–states.

Table 1.4: Summary of the $\sigma_{t\bar{t}b\bar{b}}$ measurements of CMS and ATLAS in the dilepton channel at different center–of–mass energies ($\sqrt{s}$). Notice the fact that the phase space definitions in each of the analyses are slightly different, resulting in different values for the measured quantities.

| | $\sqrt{s}$ | Quantity | Value | Uncertainty | Ref. |
|---|---|---|---|---|---|
| **CMS** | 8 TeV | $\sigma_{t\bar{t}b\bar{b}}/\sigma_{t\bar{t}jj}$ [%] | 2.2 | $\pm 0.3(\text{stat.}) \pm 0.5(\text{syst.})$ | [11] |
| | | $\sigma_{t\bar{t}b\bar{b}}$ [fb] | 29 | $\pm 3(\text{stat.}) \pm 8(\text{syst.})$ | [11] |
| | 13 TeV | $\sigma_{t\bar{t}b\bar{b}}/\sigma_{t\bar{t}jj}$ [%] | 2.4 | $\pm 0.3(\text{stat.}) \pm 0.7(\text{syst.})$ | [12] |
| | | $\sigma_{t\bar{t}b\bar{b}}$ [fb] | 88 | $\pm 12(\text{stat.}) \pm 29(\text{syst.})$ | [12] |
| **ATLAS** | 7 TeV | $\sigma_{t\bar{t}+\geq 1\,b/c}/\sigma_{t\bar{t}+\geq 1j}$ [%] | 6.2 | $\pm 1.1(\text{stat.}) \pm 1.8(\text{syst.})$ | [14] |
| | | $\sigma_{t\bar{t}+\geq 1\,b/c}$ [fb] | 160 | $\pm 30(\text{stat.})$ | [14] |
| | 8 TeV | $\sigma_{t\bar{t}b\bar{b}}/\sigma_{t\bar{t}jj}$ [%] | 1.3 | $\pm 0.33(\text{stat.}) \pm 0.28(\text{syst.})$ | [13] |
| | | $\sigma_{t\bar{t}b\bar{b}}$ [fb] | 13.5 | $\pm 3.3(\text{stat.}) \pm 3.6(\text{syst.})$ | [13] |
| | 13 TeV | $\sigma^{e\mu}_{t\bar{t}b\bar{b}}$ [fb] | 25 | $\pm 3(\text{stat.}) \pm 7(\text{syst.})$ | [15] |

These measurements are typically based on the use of HF jet identification algorithms to discriminate the flavors of the additional jets (i.e. those jets that do not originate from the top quark decays). These algorithms are discussed in more detail in Chapter 3. The existing analyses however do not have the power to discriminate between additional charm jets and additional LF (up, down, strange or gluon) jets. This thesis focuses therefore on using dedicated charm jet identification algorithms to simultaneously distinguish between the $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ and $t\bar{t}$LF events and extract their cross sections. Measuring separately the $t\bar{t}c\bar{c}$ cross section is also particularly important, with the eye on a future possible measurement of the $pp \to t\bar{t}H$, $H \to c\bar{c}$ process. On top of that, $t\bar{t}$+HF production can provide sensitivity to new physics models. A generic framework to parametrize these BSM interactions is the Standard Model Effective Field Theory, in which these final–states can obtain contributions from four–quark point–like interaction vertices. This will be discussed in detail in

Sec. 1.6. Before diving into possible extensions of the SM model, it is advisable to first consider the reasons why it is believed that the SM is not a complete and consistent theory that describes the entire universe as we observe it.

## 1.5 The open questions of the Standard Model

The SM has been thoroughly tested over the past decades in a variety of experiments with ever increasing energy and precision. So far it has not shown any conclusive signs of where it might fail. This should make even an optimist feel ill at ease, because we know there are certain phenomena that the SM can so far not explain. These questions point towards the existence of new physics beyond the SM, either by extending the current model with new undiscovered particles and/or symmetries, and/or by providing a more general theory that is valid at much higher energies, and of which the SM can be seen as a low–energy description.
Below, a few of the most profound open questions in physics are very briefly summarized, and references are given where applicable to more comprehensive overviews.

### Dark matter and dark energy

It is often said that the universe itself is the largest and most powerful laboratory, always at our disposal. We however do not get to choose what it will show us and when. Indeed, perhaps one of the biggest mysteries in physics results from indisputable cosmological observations. Namely that only $\sim 5$ % of the energy content of the universe is made up of the baryonic matter as we know it and can be described by the SM. On top of that, about 27 % is composed of another type of matter that does not seem directly visible (in the electromagnetic spectrum) and therefore has been given the name *Dark Matter* (DM). The remaining stunning 68 % is called *Dark Energy* and is even less well understood, but is believed to cause the accelerated expansion of the universe. These numbers are derived from observations of the cosmic microwave background (CMB) by the Planck satelite [82]. The CMB is a relic background of photons that resides from the decoupling era[13], about 300.000 years after the creation of the universe through the Big Bang. Small anisotropies in the temperature of the CMB, together with the corresponding power spectrum, provide a treasure of information on cosmological parameters, including the abundance of DM. Even though this already provides compelling evidence, there are more cosmological observations that support the existence of DM: from rotation curves of galaxies to gravitational lensing observations to the matter distributions in colliding galaxies such as the "*bullet cluster*" or even the structure formation in the universe. None of these can be consistently described without the existence of DM.
All of these cosmological observations are based on the gravitational effects of DM, which is so far the only type of interaction we know that DM has with the known matter. The nature of dark matter, namely which kind of particle it is (if it is a particle at all) and how it interacts with the known matter besides gravitationally, is currently unknown. A large variety of experiments exist that either try to detect DM directly (through collisions of DM with matter in the detectors), indirectly (through the annihilation of DM into known SM particles, for example in the center of our

[13] Before the decoupling of radiation, the energy, temperature and density of the universe were too high for photons to move around freely, causing them to interact constantly with surrounding matter. At a certain point in time, the expansion of the universe allowed the photons to escape and move away freely. Through time, the universe continued expanding, cooling down this relic background of photons to what we now observe as the CMB.

galaxy) or via dedicated collider experiments, for example at the LHC, in which DM could be produced from SM particle collisions.
Even though none of these experiments have successfully identified a DM candidate, many theories have been developed. One of the most popular ones describes DM as a *Weakly Interacting Massive Particle* or WIMP (see for example Ref. [83], Chapter 13.2.3). This general name is used to describe a fermionic DM candidate $\chi$, that has weak interactions with the known matter on top of its gravitational effects. An interesting scenario could be to assume that DM interacts with the SM through FCNC interactions with the top quark [84]. Such an interaction would show up in proton–proton collisions at the LHC and would lead to interesting single top quark + missing energy signatures and flavor–changing top quark decays into a charm or up quark and DM ($t \to c/u + \chi\overline{\chi}$).
In any case, whatever scenario is considered, the constraints on the interaction cross section of DM with the known matter are already very strong. It will be interesting to see what future experiments, with their increasing accuracy, will reveal.

### Matter–antimatter asymmetry and CP violation

The only reason why antimatter has been (and perhaps still is) such an illusive concept is because we do not observe it directly in our surrounding in every–day life. The bed we sleep in, the stars we look at and the air we breathe is almost exclusively made out of ordinary matter. Even though we are used to that idea, this is a puzzling thought since it would mean some imbalance had to exist between the matter and antimatter abundance in the early universe [85]. This can only be explained if the SM allows for a violation of simultaneous charge–conjugation symmetry (turning a particle into its antiparticle) and parity symmetry (inversion of spatial coordinates). This so–called CP violation has indeed been observed in 1964 in dedicated experiments that measure the decay rate of neutral kaons [86]. Later, in 2001, also CP violation in neutral B meson systems was observed by the BaBar and Belle Collaborations [87,88] and very recently, in march 2019, the LHCb Collaboration announced the discovery of CP violation in neutral D meson decays [89].
The unitary $3 \times 3$ CKM matrix introduced in Eq. (1.22) can be parametrized with three mixing angles and a complex phase that quantifies the amount of CP violation in the quark sector. This phase turns out to be non–zero. The allowed CP violation due to this non–zero complex phase is however very small and does not explain the large asymmetry observed in the universe. In the lepton sector, a non–zero complex phase in the PMNS matrix has not yet been conclusively measured, though some combined analyses suggest a non–zero value may be favored [90]. It is anyway safe to state that more work is needed to try and explain the large matter – antimatter asymmetry in the universe.

### Neutrino masses

From the Yukawa Lagrangian in Eq. (1.20), one may wonder why there is no mass term included for the neutrinos. Even though one could in principle write down such a mass term in the Lagrangian, the issue resides in the existence of a right–handed neutrino. So far such a RH neutrino has not been observed, and at first it seems like an impossible mission given that such a particle is a singlet under the entire SM gauge group (it would be an electrically neutral, colorless particle that does not interact with the weak W and Z bosons)[14]. It was therefore long thought that the neutrino

[14]Such a non–interacting type of neutrino is more generally referred to as a *sterile neutrino*.

was a massless particle. However, in 1968 a team of researchers at the Brookhaven National Laboratory observed a deficit of neutrinos coming from the sun [91], which lead to the *solar neutrino problem*, that could at that point not be explained. They had in fact unknowingly discovered the existence of neutrino oscillations, which was only directly and conclusively observed much later (in 2001) through oscillation experiments in the Sudbury Neutrino Observatory [92]. These observations confirmed that neutrinos can change their flavor as they travel, which can only be explained if they indeed have a non–zero mass[15]. Their mass must however be incredibly small and a new further issue arises in explaining the hierarchy of lepton masses[16]. Dedicated neutrino oscillation experiments nevertheless search for the existence of sterile neutrinos through the disappearance of a certain neutrino flavor into an undetected type of sterile neutrino.
Theoretically, the biggest question remains whether the neutrino should be described as a Dirac or a Majorana fermion. In the former case it should be described just like the charged leptons, but in the latter case the neutrino would be its own antiparticle. The Majorana description is particularly interesting given that it allows to write down a bare mass term in the SM Lagrangian without breaking gauge invariance. This hypothesis is tested for example through neutrinoless double beta decay experiments [93].

### Gravity as a quantum field theory

As mentioned before, the gravitational force is not included in the SM, but is rather described by the theory of General Relativity [43], which describes gravity as a bending of space–time itself. Attempts have been made to describe gravity as a quantum field theory, with a corresponding boson that is referred to as the *graviton*. Unfortunately, it turns out that when adding such a theory to the existing SM it breaks the renormalizability of the SM, which means that perturbative calculations to an arbitrary order in the expansion do not any longer necessarily give finite answers. Therefore a QFT description of the SM in curved spacetime [94] does not seem to be straightforward.

### Unification of the couplings

The SM gauge group comes with three couplings: two corresponding to the electroweak $\mathrm{SU(2)_L} \bigotimes \mathrm{U(1)_Y}$ gauge group, $g_1$ and $g_2$, and one corresponding to the $\mathrm{SU(3)_C}$ gauge group, $g_\mathrm{s}$. As mentioned before in Sec. 1.3, these couplings depend on the energy at which they are probed. At some very large energy scale ( $\sim 10^{16}$ GeV), one may hope that these couplings converge to one and the same value, possibly indicating the unification of the three forces. The idea of unification has always been very appealing in light of the persuit of "*one theory of everything*". Unfortunately, in the SM as we know it, the three couplings do not converge[17]. Nevertheless, extensions of the SM have been proposed in which the couplings actually converge (see for example Ref. [95] Chapter 10.2.6 and Ref. [23] Chapter 22.2). Examples are the supersymmetric SM (SUSY) or grand unified theories (GUTs). These theories predict either new particles or new interactions that have so far not been observed.

[15] It suffices that two out of the three known generations are massive, since only a mass difference between the flavors is needed.

[16] In fact, an open question in the SM remains to explain in general the mass hierarchy between all fermions.

[17] It should be realized that this requires an extrapolation of the running of the coupling constants over around 8 orders of magnitude that we have not been able to probe so far!

### The hierarchy "*problem*"

The hierarchy problem refers to the perturbative corrections to the mass of the H boson, which are particularly sensitive to the scale at which new physics is expected to appear beyond the SM. The quotes around *problem* in the title of this paragraph are intensionally included to point out the fact that is not so much of a problem, but rather an inconvenient and unappealing fine–tuning of the H boson mass. Indeed, the H boson mass needs to be corrected by loop contributions to its self–energy, with fermions running in the loops as depicted in Fig. 1.8. The H boson mass $m_h$ can be expressed as its bare mass $m_h^0$ corrected with additional terms describing the loop contributions (in principle up to infinite orders in perturbation theory). This is expressed in Eq. (1.24), from which it can be seen that the corrections scale with the fermion Yukawa coupling $\lambda_f$ squared multiplied by the square of some high–energy cut–off scale[18] $\Lambda_{\text{UV}}$. This scale reflects some high–energy frontier at which new physics is expected to appear and is typically taken to be around the Planck scale ( $\sim 10^{19}$ GeV), where quantum gravitational effects become relevant.

$$m_h^2 = \left(m_h^0\right)^2 + \Delta m_h^2 \ , \tag{1.24}$$
$$\text{with} \quad \Delta m_h^2 = -\frac{|\lambda_f|^2}{8\pi^2}\Lambda_{\text{UV}}^2 + \cdots$$

From this formula it is clear that in order to obtain a measured value of the H boson mass of $m_h \approx 125$ GeV [96], a miraculous cancellation needs to happen between the bare mass term and its loop corrections. This seems unnatural and may point towards some new physics that regulates this fine–tuning. For example in SUSY extensions of the SM, each fermion gets a scalar counterpart that would also run in the loop in Fig. 1.8 and introduce a term in Eq. (1.24) with opposite sign but similar magnitude, such that the radiative corrections are (partially) cancelled and the fine–tuning is reduced. Unfortunately, these scalar counterparts have not yet been observed.

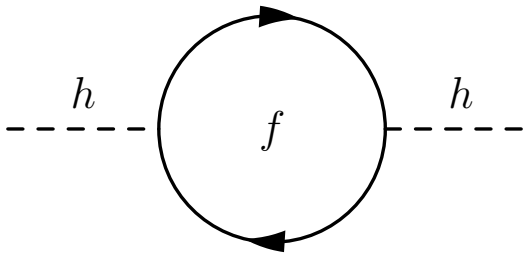


FIGURE 1.8: Feynman diagram describing the fermion loop correction to the H boson mass.

## 1.6 The Standard Model Effective Field Theory

Having summed up some of the most profound open questions that the SM still has not answered in Sec. 1.5, the question arises on how we can extend the SM with new

[18] The cut–off scale is introduced to regularize divergences in the calculations of the integrals that appear in the loop calculations, see for example Ref. [23], Appendix A.4.

particles and interactions. A wide landscape of model–specific BSM theories has been developed in the last decades, making it almost impossible to test the predictions of each and every one of them at particle colliders such as the LHC. Therefore a more generic and model–independent approach has been developed to parametrize a wide range of new interactions through effective interactions, known as the Standard Model Effective Field Theory or SMEFT in short. In this section, a theoretical overview of the SMEFT framework will be provided, including discussions on the assumptions and limitations. We will work towards a set of new operators that could potentially affect the production of top quarks with additional heavy–flavor jets. This information will be used in Chapter 6, where novel methods will be described to optimally constrain (or potentially discover) new physics interactions in the SMEFT.

### Prologue: Fermi's theory of beta decay

In 1933, Enrico Fermi was the first to describe the process of radioactive beta decay through a contact interaction between four particles: a proton, a neutron, an electron and an electron–antineutrino [97, 98]. The Feynman diagram that describes such a point–like interaction is shown in Fig. 1.9 on the left. It was in fact the up quark in the proton that was converted into a down quark of the resulting neutron, though this was not yet known at that time. The coupling constant that describes the strength of this interaction was later denoted the Fermi constant $G_F$. This theoretical description turned out to be very accurate and for a low energy process such as radioactive beta decay, it was an appropriate description. It is now known that this process actually proceeds through a weak interaction, mediated by a W boson as shown on the right in Fig. 1.9. Nevertheless, as long as the energy of the process stays well below the mass of the W boson, the effective point–like interaction is a suitable approximation with accurate predictive power. The correspondance between the Fermi constant and the weak coupling constant $g_w$ is given in Eq (1.25). From this equation it is clear that the Fermi constant in fact has the dimension of an inverse mass squared[19].

$$G_F = \frac{\sqrt{2}}{8}\frac{g_w^2}{m_W^2} \tag{1.25}$$

This was one of the first effective field theories (EFT) to describe a low–energy approximation through a point–like interaction that effectively proceeds through a mediator with a mass far above the typical energy of the process.

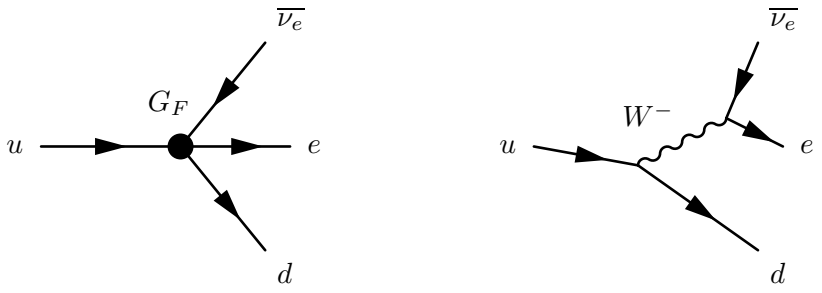


FIGURE 1.9: Feynman diagrams describing beta decay through the effective Fermi interaction (left) and through the weak interaction via the exchange of a W boson (right).

[19]Throughout this work, natural units are used expressing all dimensions in terms of mass. Often "mass" will be omitted and dimensions will be expressed as bare numbers.

### SMEFT in a nutshell

Despite the many searches for new physics signatures at the LHC, no significant deviations from the SM predictions have so far revealed themselves. If a new particle would indeed exist with a mass below the energy reach of the LHC, it would most likely appear as a clear peaked excess in an invariant mass spectrum of its decay products, assuming it decays into SM particles that we can detect. This is schematically depicted in Fig. 1.10 with the red excess on top of the blue SM expectation. Many of these searches, sometimes referred to as *bump hunts*, have been conducted and no significant and unexpected excesses have been observed. It is however possible that the mass of a new particle lies well beyond the current reach of the LHC, resulting in a resonance that can not yet be observed with the current collision energy. Such a resonance would however show up in the high–energy tails of energy–dependent distributions, as illustrated again in Fig. 1.10 in dark green. Such a scenario, in which we would only be able to observe the low–energy regime of the BSM contributions, can be described by an EFT.

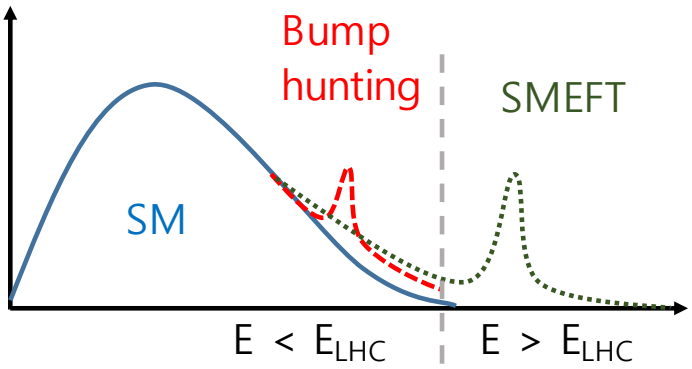


FIGURE 1.10: Schematic illustration of new physics searches within the energy range of the LHC (red) and beyond that reach (green).

Theoretically, if a new particle $\Omega$ with mass $\Lambda$ mediates the interaction between SM particles with a strength given by a new physics coupling $g_*$, its propagator is expressed on the left side of the arrow in Eq. (1.26). If the mass $\Lambda$ is much larger than the typical momentum $p$ of the interaction, this can be expanded as shown on the right side of Eq. (1.26) and the higher order terms, suppressed by higher orders of mass scale $\Lambda$, can be neglected as a good first order approximation.

$$\frac{g_*^2}{p^2-\Lambda^2} \quad \xrightarrow[\Lambda^2 \gg p^2]{} \quad -\frac{g_*^2}{\Lambda^2}\left[1+\cancel{\frac{p^2}{\Lambda^2}}+\cancel{\frac{p^4}{\Lambda^4}}+\cdots\right] \tag{1.26}$$

Graphically, this corresponds to the Feynman diagrams depicted in Fig. 1.11, in which the interaction through a massive mediator is replaced by an effective point–like vertex.

The SM Lagrangian is of dimension four. All of the operators in the SM are therefore of dimension four and are scaled with a dimensionless coupling constant[20]. To construct the Standard Model Effective Field Theory, one extends the SM Lagrangian

[20]The only exception lies in the Higgs sector, where the scalar potential has a dimension two operator with a dimensionful parameter $\mu^2$ in front, as shown in Eq. (1.16).

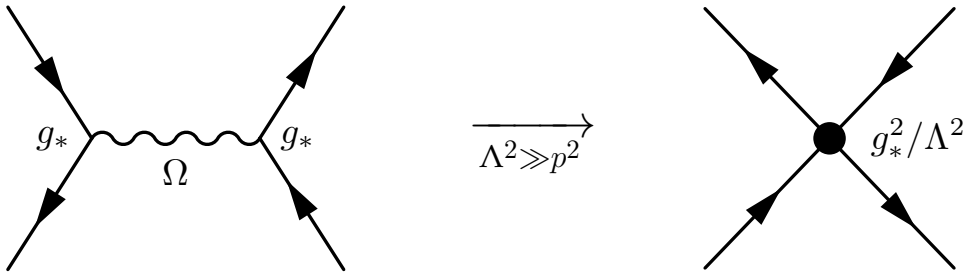


FIGURE 1.11: Feynman diagrams describing a new mediator $\Omega$ with mass $\Lambda$ that couples to the SM particles with a new physics coupling $g_*$ (left) and the corresponding EFT vertex describing the point–like interaction (right).

with new operators of dimension larger than four [99–101], which are suppressed by powers of an energy scale $\Lambda$ that represents the typical energy scale of the new physics resonance that the EFT describes. This is expressed by Eq. (1.27), where the index $d$ denotes the dimensionality of the operator $O$ and the index $i$ runs over all allowed operators of a given dimension. The coefficient $C_i$ is a dimensionless coupling constant that is called the *Wilson coefficient*.

$$\mathscr{L}_{\text{SMEFT}} = \mathscr{L}_{SM} + \sum_{d>4} \sum_{i} \frac{C_i}{\Lambda^{d-4}} O_i^{(d)} \tag{1.27}$$

The allowed operators need to obey the gauge invariance of the SM gauge group. There exists only one such operator of dimension five, which is a lepton–number-violating operator that could provide a mass term for the neutrinos [102]. At dimension six however, a whole new world of operators opens up which are suppressed by the new physics scale squared ($\sim \Lambda^{-2}$). Depending on the flavor assumptions, the number of dimension six operators can go as high as a few thousands if one assumes full flavor–non–universality. However, the minimal set of operators needed in a fully flavor–universal scenario is 59. There exists a freedom in choosing a particular basis of operators to fully describe the SMEFT at dimension six. A popular choice is the so–called Warsaw basis [101]. Higher order operators are suppressed by even higher orders of $\Lambda$, and can often be neglected to first order. Nevertheless for some processes the higher order operators are relevant, or even the first relevant contributions to consider. Therefore one should always validate whether or not higher orders can safely be neglected. It is useful to note that all operators with odd-numbered dimensions can only generate baryon or lepton number violating processes.

Any observable, such as a cross section, or a number of observed events in a given phase space region, can be expressed in terms of its SM value and its additional contributions due to the SMEFT effects. For contributions of dimension six operators, the functional form of an observable $\sigma$ can be expressed as in Eq. (1.28), where the indices $i$ and $j$ run over the number of operators that are considered for the given process and $\tilde{\sigma}_i$ and $\tilde{\delta}_{i,j}$ are coefficients to be determined. In this notation, $\tilde{\sigma}_i$ signifies the strength of the interference of the SMEFT operators with the SM, whereas $\tilde{\delta}_{i,j}$ represents the pure EFT contribution, including quadratic terms for a single operator and the interference effects amongst SMEFT operators.

$$\sigma = \sigma_{SM} + \sum_{i} \frac{C_i}{\Lambda^2} \tilde{\sigma}_i + \sum_{i,j} \frac{C_i\, C_j}{\Lambda^4} \tilde{\delta}_{i,j} \tag{1.28}$$

The interference term is only suppressed by two orders of $\Lambda$, making it usually the dominant one with respect to the squared term, which is suppressed by $\Lambda^{-4}$. Nonetheless this is not always the case, as the quadratic term can become significant or even dominant for larger values of the Wilson coefficients, or in the absence of interference between the SM and the SMEFT operators.

From the discussion above, it is clear that an EFT description is only valid under some assumptions, which should always be validated:

- SMEFT operators should always comply to the gauge invariance of the SM gauge group.
- The EFT description is only valid as long as the energy that is involved in the effective vertex of the considered process lies below the mass of the new physics resonance that it describes. Given that this mass scale is a priori unknown, an equivalent statement could be that one can only interpret the results of an EFT in terms of a UV completion[21] with a mediator mass well above the energy of the process under consideration. In order to ensure this requirement in a measurement, one solution is to put an upper limit on the energy scales ($M_{\text{cut}}$) that are relevant in a given process under consideration and to ignore all information from measurements that exceed this energy.
- As with any quantum field theory, the *perturbativity* of the theory needs to be ensured in order to be able to perform perturbative calculations and make predictions. This means that the EFT coupling should not exceed the value of $(4\pi)^2$ [103]. This is expressed in Eq. (1.29) for dimension six operators.

$$\underline{\texttt{dimension six}}: \qquad \frac{|C_i|\, M_{\text{cut}}^2}{\Lambda^2} < (4\pi)^2 \tag{1.29}$$

### The top quark in the SMEFT

In the last years, a large effort has been made by the top quark community at the LHC to agree on a common guideline in interpreting top quark measurements in the SMEFT. Such common guidelines have recently been written down in Ref. [103], together with a model that allows to generate matrix elements including the SMEFT operators that can be used to generate event simulations, as will be discussed in Chapter 3. In this paragraph the relevant set of dimension six operators in which the top quark is involved are outlined. It should be noted that in principle one could add a set of FCNC effective operators as well. Due to the extreme rarity of FCNC processes in the SM, dedicated measurements of such processes can put very stringent constraints on the allowed parameter space of the FCNC SMEFT [46–51]. In this work, such operators will however not be considered any further.

As mentioned before, the amount of operators depends on the flavor assumptions that are made. Already in the SM, the Yukawa sector breaks flavor universality, given that the top quark couples much stronger to the scalar fields than the other quarks. Therefore any new physics sector is not necessarily expected to be fully flavor–universal. A more likely scenario is that the flavor–structure indeed follows the hierarchy in the Yukawa sector. With this in mind, the third generation quarks

[21] The term UV completion is used to refer to a full theory at high energies for which the EFT is a low–energy approximation. The electroweak theory could be considered a UV completion of the Fermi theory, as described in the prologue of this section.

are given independent couplings with respect to the first and second generation. This is known as the principle of Minimal Flavor Violation [104] (MFV) and is used throughout this work. The light–quark sector (first and second generation) is treated symmetrically, following a $\mathrm{SU(2)_q} \bigotimes \mathrm{SU(2)_u} \bigotimes \mathrm{SU(2)_d}$ flavor symmetry. We adopt the following notation: $q$ denotes the LH light–quark electroweak $\mathrm{SU(2)_L}$ doublet, whereas $u$ and $d$ represent the RH light–quark singlets. Similarly $Q$ is used to refer to the LH third–generation quark doublet and $t$ ($b$) is used to indicate the RH top (bottom) quark singlets.

A complete overview of all dimension six operators in the Warsaw basis can be found in Ref. [101]. Extracting those that involve a top quark results in a specific categorization: operators involving the top quark and the scalar H boson, operators involving other gauge bosons, four–fermion operators involving top quarks and leptons and four–quark operators. A comprehensive overview can be found in Refs. [103,105]. In this work, we will restrict ourselves to those operators possibly contributing to top quark pair production with additional HF jets. This restricts the set of operators to the four–quark operators and operators involving top or bottom quarks coupling to gluons[22]. The latter are referred to as the chromomagnetic dipole operators and can be written as:

$$O_{tG} = (\bar{Q}\,\sigma^{\mu\nu}\,T_A\,t_R)\,\tilde{\phi}\,G^A_{\mu\nu} + \text{h.c.}\,, \tag{1.30}$$

$$O_{bG} = (\bar{Q}\,\sigma^{\mu\nu}\,T_A\,b_R)\,\phi\,G^A_{\mu\nu} + \text{h.c.}\,. \tag{1.31}$$

These operators are strongly constrained by multijet and direct top quark pair production [106] and $t\bar{t}$+HF production is not expected to yield a comparable sensitivity to these operators. Therefore we do not consider them, but rather focus on the extensive set of four–quark operators that have not yet been strongly constrained in the past. Under the MFV assumption, these operators can be subdivided into contributions including four heavy quarks (4H) and those involving two heavy and two light quarks (2H2L). Following the recommended basis choice by the LHC Top Working Group [103], these operators are summarized in Tab. 1.5. The last three columns indicate whether or not these operators contribute to $t\bar{t}t\bar{t}$, $t\bar{t}b\bar{b}$ and $t\bar{t}c\bar{c}$ production at the LHC. Besides the separation into the 4H and 2H2L categories, it is clear that the Lorentz structures of the operators follow a pattern: they can either be color singlet (1) or color octet (8) operators, indicated by the presence or absence of the corresponding SU(3) generators $T^A$; they either involve LH or RH quark currents (or a combination of both) and they can either be vector currents (including $\gamma_\mu$) or scalar currents. Furthermore in Tab. 1.5, $\varepsilon$ is the totally antisymmetric Levi–Civita tensor in SU(2)–space and $\tau^I$ are the Pauli–matrices. Currently, only a subset of these operators are constrained from measurements of $t\bar{t}$ and $t\bar{t}t\bar{t}$ production [73,105,107,108]. In Chapter 6 it will be shown how a measurement of the $t\bar{t}c\bar{c}$ and $t\bar{t}b\bar{b}$ processes can provide additional sensitivity as well as the first constraints on currently unconstrained operators [109].

[22] Another operator that does not involve directly a top quark, but that can contibute significantly to top quark pair production is the triple gluon operator: $O_G = f_{\mathrm{ABC}} G^{A\nu}_{\mu} G^{B\rho}_{\nu} G^{C\mu}_{\rho}$.

TABLE 1.5: Summary of the four–quark SMEFT operators of dimension six, subdivided into the four–heavy operators (4H) and two–heavy–two–light operators (2H2L). The index $i = 1, 2$ represents the generation of the light quarks. The last three columns indicate whether or not these operators contribute to $t\bar{t}t\bar{t}$, $t\bar{t}b\bar{b}$ and $t\bar{t}c\bar{c}$ production at the LHC.

| Four-quark operators | | | | |
|---|---|---|---|---|
| | Operator | $t\bar{t}t\bar{t}$ | $t\bar{t}b\bar{b}$ | $t\bar{t}c\bar{c}$ |
| **4H** | $O^1_{QQ} = \frac{1}{2}\left(\overline{Q}\ \gamma_\mu\ Q\right)\left(\overline{Q}\ \gamma^\mu\ Q\right)$ | ✓ | ✓ | |
| | $O^8_{QQ} = \frac{1}{2}\left(\overline{Q}\ \gamma_\mu\ T^A\ Q\right)\left(\overline{Q}\ \gamma^\mu\ T^A\ Q\right)$ | ✓ | ✓ | |
| | $O^1_{tb} = \left(\bar{t}\ \gamma_\mu\ t\right)\left(\bar{b}\ \gamma_\mu\ b\right)$ | | ✓ | |
| | $O^8_{tb} = \left(\bar{t}\ \gamma_\mu T^A\ \ t\right)\left(\bar{b}\ \gamma_\mu\ T^A\ b\right)$ | | ✓ | |
| | $O^1_{tt} = \left(\bar{t}\ \gamma_\mu\ t\right)\left(\bar{t}\ \gamma_\mu\ t\right)$ | ✓ | | |
| | $O^1_{bb} = \left(\bar{b}\ \gamma_\mu\ b\right)\left(\bar{b}\ \gamma_\mu\ b\right)$ | | | |
| | $O^1_{Qt} = \left(\overline{Q}\ \gamma_\mu\ Q\right)\left(\bar{t}\ \gamma^\mu\ t\right)$ | ✓ | ✓ | |
| | $O^8_{Qt} = \left(\overline{Q}\ \gamma_\mu\ T^A\ Q\right)\left(\bar{t}\ \gamma^\mu\ T^A\ t\right)$ | ✓ | ✓ | |
| | $O^1_{Qb} = \left(\overline{Q}\ \gamma_\mu\ Q\right)\left(\bar{b}\ \gamma^\mu\ b\right)$ | | ✓ | |
| | $O^8_{Qb} = \left(\overline{Q}\ \gamma_\mu\ T^A\ Q\right)\left(\bar{b}\ \gamma^\mu\ T^A\ b\right)$ | | ✓ | |
| | $O^1_{QtQb} = \left(\overline{Q}\ t\right)\varepsilon\left(\overline{Q}\ b\right)$ | | ✓ | |
| | $O^8_{QtQb} = \left(\overline{Q}\ T^A\ t\right)\varepsilon\left(\overline{Q}\ T^A\ b\right)$ | | ✓ | |
| **2H2L** | $O^{(8,3)}_{Qq} = \left(\overline{Q}\gamma_\mu T^A\tau^I Q\right)\left(\overline{q^i}\gamma_\mu T^A\tau^I q^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(8,1)}_{Qq} = \left(\overline{Q}\gamma_\mu T^A Q\right)\left(\overline{q^i}\gamma_\mu T^A q^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(1,3)}_{Qq} = \left(\overline{Q}\gamma_\mu \tau^I Q\right)\left(\overline{q^i}\gamma_\mu \tau^I q^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(1,1)}_{Qq} = \left(\overline{Q}\gamma_\mu Q\right)\left(\overline{q^i}\gamma_\mu q^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(8)}_{tu} = \left(\bar{t}\gamma_\mu T^A t\right)\left(\overline{u^i}\gamma_\mu T^A u^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(8)}_{td} = \left(\bar{t}\gamma_\mu T^A t\right)\left(\overline{d^i}\gamma_\mu T^A d^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(1)}_{tu} = \left(\bar{t}\gamma_\mu t\right)\left(\overline{u^i}\gamma_\mu u^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(1)}_{td} = \left(\bar{t}\gamma_\mu t\right)\left(\overline{d^i}\gamma_\mu d^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(8)}_{tq} = \left(\bar{t}\gamma_\mu T^A t\right)\left(\overline{q^i}\gamma_\mu T^A q^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(8)}_{Qu} = \left(\overline{Q}\gamma_\mu T^A Q\right)\left(\overline{u^i}\gamma_\mu T^A u^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(8)}_{Qd} = \left(\overline{Q}\gamma_\mu T^A Q\right)\left(\overline{d^i}\gamma_\mu T^A d^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(1)}_{tq} = \left(\bar{t}\gamma_\mu t\right)\left(\overline{q^i}\gamma_\mu q^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(1)}_{Qu} = \left(\overline{Q}\gamma_\mu Q\right)\left(\overline{u^i}\gamma_\mu u^i\right)$ | ✓ | ✓ | ✓ |
| | $O^{(1)}_{Qd} = \left(\overline{Q}\gamma_\mu Q\right)\left(\overline{d^i}\gamma_\mu d^i\right)$ | ✓ | ✓ | ✓ |

# Chapter 2 The CMS experiment at the LHC

In the past decades, building up the SM as we know it today has required the construction of more powerful particle accelerators. The design of these accelerators depends greatly on the objectives posed by the most challenging and urgent questions at a certain moment in time. Two main considerations are in order: which particles will be collided and will they be accelerated in a circular or a linear collider? For example, in the 1990s precision measurements of the electroweak sector were needed, resulting in the construction of the Large Electron–Positron collider [110, 111] at CERN. This was a circular collider in which point–like particles, electrons and positrons, were accelerated and collided at energies chosen specifically to probe the resonant production of Z and W bosons. The large amount of synchrotron radiation of these particles has resulted in the construction of the largest particle accelerator tunnel currently in existence ($\sim$ 26.7 km in circumference). Near the end of its scheduled program, hints of a scalar boson were showing up, but the results were not conclusive [112]. The choice was then made not to extend the LEP program, and build a new machine aimed at the discovery of the H boson: the Large Hadron Collider (LHC). During the construction of the LHC, experiments were conducted at the Tevatron collider [113] at Fermilab, which collided protons with antiprotons at energies up to 2 TeV. But also here the searches for the H boson were inconclusive [114]. Eventually in 2012, the proton–proton collisions in the LHC revealed the existence of the H boson [4,6], confirming the theory that was proposed around 50 years earlier by R. Brout, F. Englert and P. Higgs. The LHC is now the most powerful particle accelerator on earth. It is therefore the best place to study the SM processes in more detail, and above all to look for hints of new unexplored physics!

In this chapter the experimental setup used to accumulate the collision data for the analysis in Chapter 5 will be discussed. The design and performance of the LHC at CERN will be detailed in Sec. 2.1. Then the Compact Muon Solenoid experiment (CMS) and the corresponding detector will be discussed in full detail in Sec. 2.2.

## 2.1 The Large Hadron Collider

### 2.1.1 The CERN accelerator complex

CERN[1], the European Organization for Nuclear Research, has played a world–leading role in state–of–the–art fundamental research in particle physics. It was founded in 1954 on the French–Swiss border near Geneva. The Laboratory has grown over the last decades, hosting a large number and variety of particle physics experiments. Among these experiments is the worlds most powerful particle accelerator, the LHC, located underground at a depth between 50 and 175 m. Built in the existing 26.7 km long tunnel from its predecessor, LEP, this circular accelerator is today capable of colliding protons at a center–of–mass energy of up to 13 TeV. Additionally, it hosts a rich program of heavy–ion physics in which collisions of heavy–nuclei, such as Lead and Xenon, can be studied with dedicated experiments. Many smaller accelerators have served in the past and are today still operational as a pre–acceleration chain for the LHC. Protons consecutively make their way through the LINAC2 (a linear accelerator), the Booster, the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS), after which the beam of protons is divided in two beams that enter the LHC in opposite directions. This is illustrated in Fig. 2.1, where other experiments at CERN are also indicated. These other experiments are for example aimed at studying antimatter (such as the Antiproton Decelerator), radioactive nuclei (ISOLDE), neutron–nucleus interactions (n-ToF) and future plans exist to host a rich program in neutrino physics (CENF).

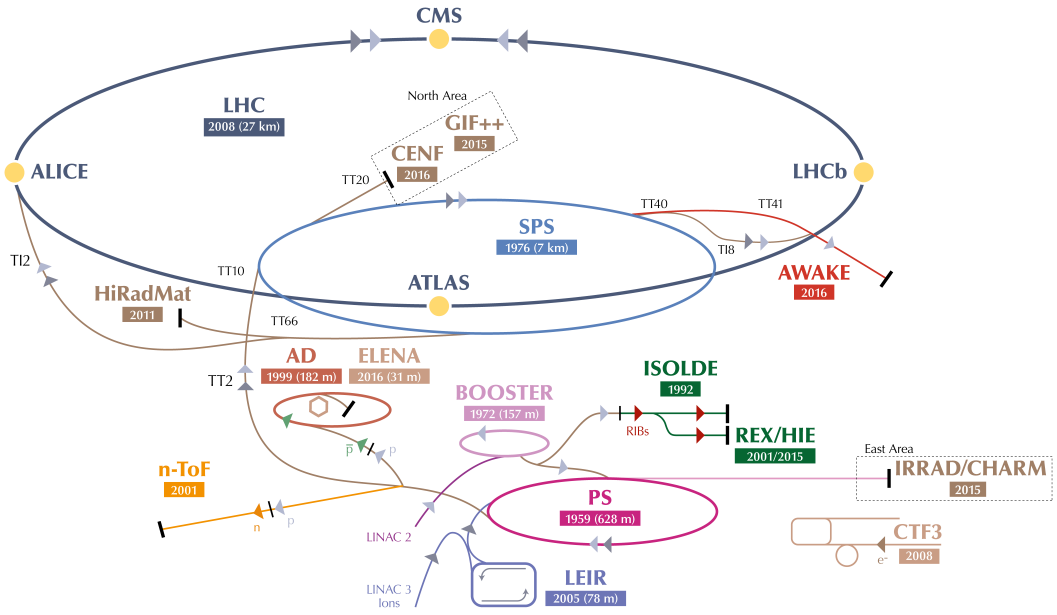


FIGURE 2.1: Schematic overview of the accelerator complex at CERN. Figure taken from Ref. [115].

[1]The acronym CERN resides from the original french name of the organization: *Conseil Européen pour la Recherche Nucléaire.*

### 2.1.2 The LHC experiments

The protons running in opposite directions in the accelerator are made to collide at specific collision points. Around these points, particle detectors are built that are capable of capturing the resulting debris that emerges from the collision. More accurately, their goal is to reconstruct the interesting physics phenomena from a debris of subsidiary particles that are produced. The LHC hosts four main experiments at its collision points (see also Fig. 2.1):

1. **LHCb (The Large Hadron Collider Beauty experiment):** LHCb [116] has as a main objective to study $b$ hadron physics, trying to understand the nature of CP violation in (and beyond) the SM. These studies could provide insight in the matter–antimatter asymmetry in the universe by revealing the mechanism that are responsible for the known CP violation and potentially unfolding new sources of CP violation in rare decays of $b$ and $c$ hadrons. The detector covers mostly the forward direction to exploit the enhanced $b$ hadron production in that region.

2. **ALICE (A Large Ion Collider Experiment):** ALICE [117] is a dedicated experiment to investigate the collisions of heavy–ions in the LHC during the dedicated heavy–ion runs. It is believed that in the collisions of these heavy–nuclei, for a very short time a special state of matter known as the quark–gluon plasma (QGP) is created. This is the state of matter that could also have existed in the moments right after the creation of the universe through the Big Bang. By studying the behavior of the outgoing particles, the properties of this QGP could be revealed and better understood. The LHC has even injected a run of proton–lead collisions to study the behavior in this "intermediate case".

3. **ATLAS (A Toroidal LHC Apparatus):** ATLAS [118] is one of the two general–purpose detectors, initially built to discover the H boson, but serving a large variety of other physics studies as well. It has a typical cylindrical layered layout, hermetically sealing the collision point to detect as many of the outgoing particles as possible. It is the largest of the LHC experiments (though not the most dense one) and features a large toroidal magnet in which many of the sub–detector layers are housed.

4. **CMS (Compact Muon Solenoid):** CMS [119] is the other general–purpose detector, built with the same objectives but a slightly different layout compared to the ATLAS detector. This experiment is used throughout the rest of this work and its architecture and performance will be discussed in much more detail in Sec. 2.2.

Finally, it should be mentioned that three other, smaller experiment exist at the LHC. Two of those focus on very forward physics phenomena. These experiments are called TOTEM [120] and LHCf [121] and are housed in the very forward parts of the CMS and ATLAS detectors respectively. The third one, MoEDAL [122], is located near the LHCb detector and searches for magnetic monopoles.

### 2.1.3 Design and performance of the LHC

The conceptual design of the LHC [123] was written in 1995, resulting in the construction of the strongest particle accelerator on earth between 1998 and 2008. The machine has been built to supply proton–proton collisions at center–of–mass

energies[2] of up to 14 TeV, and a design luminosity of $\sim 10^{34}$ cm$^{-2}$s$^{-1}$ [124]. Its first physics run took place in 2010, colliding protons at $\sqrt{s} = 7$ TeV, followed by a run at $\sqrt{s} = 8$ TeV in 2011 and 2012. The provided dataset after this period turned out to be large enough to discover a new scalar boson with properties consistent with the SM Higgs boson. After a long shutdown period of two years, the LHC resumed its operation at the end of 2015 at a stunning $\sqrt{s} = 13$ TeV, which continued until the end of 2018. Currently another long shutdown is ongoing, preparing the LHC to operate at $\sqrt{s} = 14$ TeV in 2021. In the future, a big luminosity upgrade is planned (between 2023 and 2025) in order for the LHC to be able to operate at higher luminosities (HL–LHC) from 2026 onwards.

Accelerating charged particles requires electric fields, whereas magnetic fields bend the particles on their circular trajectories. The acceleration happens in 16 radio frequency (RF) cavities in which an electrical field oscillates at a frequency of 400 MHz [123, 125], resulting in an increased proton energy of 0.5 MeV per revolution. On the other hand, 1232 dipole magnets which are each approximately 15 m long are responsible for the bending of the particle trajectories. A schematic vertical slice of such a dipole magnet is shown in Fig. 2.2. Two beam pipes run through the dipole magnets, each surrounded by coils that generate the required magnetic field. The coils are cooled down by liquid helium to a temperature of 1.9 K, such that the superconductive properties of the cooled coils allows to send a current of up to 11.7 kA through them. Throughout the graduate acceleration procedure, the magnetic field in the dipoles can be increased from 0.535 T at injection up to 8.4 T at collision energy (in case of 14 TeV collisions). Additionally, quadrupole and higher order multipole magnets are placed at the injection points of the LHC and at specific places along the ring to squeeze the proton beams and confine them within the beam pipe, as well as to deflect the beams towards a frontal collision at the collision points.

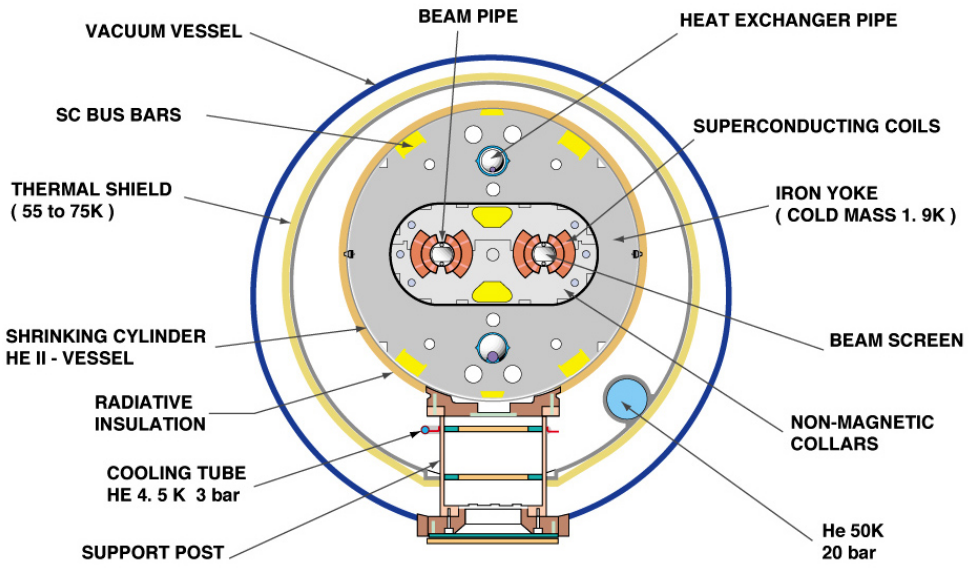


FIGURE 2.2: Schematic overview of a dipole magnet of the LHC. Figure taken from Ref. [126].

[2]The centre–of–mass energy is twice the individual energy of each proton beam.

Key towards discovering new physics phenomena, is to record as much high–quality collision data as possible. The rate at which protons collide at the interaction points is expressed in terms of the (instantaneous) luminosity $\mathcal{L}$. A beam of protons is composed of separate bunches, each containing $\sim 10^{11}$ protons. The luminosity for a Gaussian beam distribution [125] is defined as

$$\mathcal{L} = \frac{f N_b^2 n_b}{4\pi} F, \tag{2.1}$$

where $N_b$ represents the number of protons per bunch and $n_b$ the number of bunches per beam. The luminosity scales with the revolution frequency of the beam $f$. A geometrical factor $F$ (in units of [$\text{cm}^{-2}$]) is taken into account to accommodate for the transversal dimensions of the beam, as well as the crossing angle at the collision point. The rate for any given process of interest ($N_\alpha$) can then easily be calculated from the luminosity and the process–specific cross section $\sigma_\alpha$ through $N_\alpha = \mathcal{L}\ \sigma_\alpha$. Often the absolute number of events over a given period of data–taking is expressed in terms of the integrated luminosity $\mathcal{L}_{\text{int}} = \int \mathcal{L}\ dt$ and is usually expressed in units of *inverse femtobarn* ($\text{fb}^{-1}$).

In the collision of two bunches, multiple proton–proton collisions can occur, which is known as (in–time) pileup (PU). The average number of proton–proton interactions in 2017 is estimated to be around ∼32 (see Sec. 2.2.8). This number scales with the luminosity and has increased over the years. These multiple interactions often make it harder to extract the interesting physics signals from the large background of low–energy particles flying around. On top of that, the integration of the signals in the electronics may take longer than the time between consecutive collisions (25 ns), resulting in so-called out–of–time pileup. This problem is expected to become problematic during the high–luminosity upgrade of the LHC (HL–LHC) [127], where the luminosity is expected to increase by a factor of at least 5, leading to an average pileup of around ∼200. Updates to the CMS detector are indeed foreseen to cope with this high–PU environment [128].

The data used in this thesis were collected during the 2017 run of the LHC. During this period, proton–proton collisions at 13 TeV were delivered with a peak instantaneous luminosity of $\sim 2 \times 10^{34}\ \text{cm}^{-2}\text{s}^{-1}$ [129,130], twice the design luminosity of the LHC! This remarkable luminosity results from a stunning bunch crossing rate of 40 MHz, or equivalently a bunch spacing of 25 ns. Each beam can contain up to 2808 bunches, though the specific filling scheme of the LHC varies over time. An overview of the integrated luminosity provided by the LHC can be found in Fig. 2.3, showing the per–day cumulative integrated luminosity for each of the data–taking periods over the past years. A total integrated luminosity of almost 50 $\text{fb}^{-1}$ was delivered by the LHC in 2017, which is a remarkable achievement given that the run started relatively late in the year (around the end of May).

## 2.2 The CMS experiment

The Compact Muon Solenoid detector [119, 131] is one of the two general–purpose detectors surrounding the collision points at the LHC. Though not as large as the ATLAS detector, it is more dense hence the “Compact” in its name. The detector has a total length of 21.6 m, a diameter of 14.6 m and weights 12,500 tons. Its cylindrical architecture is composed of consecutive detector layers that each provide a different

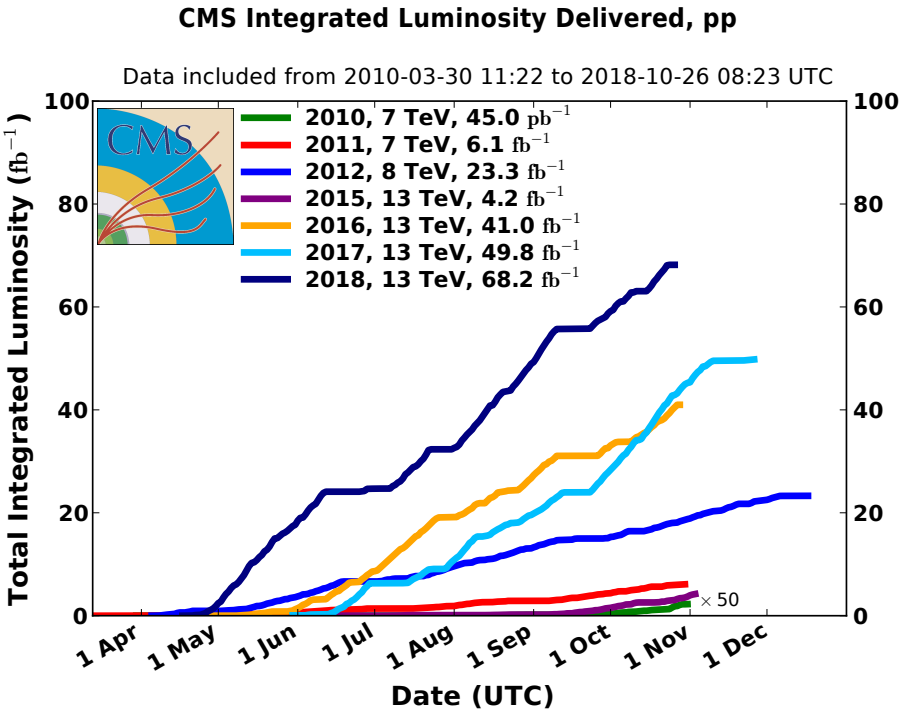


FIGURE 2.3: Cumulative integrated luminosity delivered by the LHC to the CMS experiment during the different run periods over the past years. Figure taken from Ref. [129].

kind of information to be used for the particle identification. These sub–detectors are typically composed of a cylindrical barrel part, and two endcap compartments to ensure full geometric coverage. A schematic view of the CMS detector is shown in Fig. 2.4. At the heart of the detector lies a superconducting solenoid magnet that bends the charged particle trajectories and allows for a precise determination of their momentum. Inside of this solenoid, CMS houses a tracking detector closest to the beam pipe, followed by an electromagnetic calorimeter (ECAL) and a hadronic calorimeter (HCAL). On the outside of the magnet, the muon detection systems were installed. Each of these detector components will be discussed in more detail in the following sections.

From the electronic signals that are generated by the detector compartments, the particles emerging from the collisions need to be reconstructed, and their four–momenta need to be accurately determined. Due to the very large collision rate of 40 MHz, a triggering system is needed to filter out the interesting collision events and keep the data acquisition and storage manageable. This will be discussed in more detail in Sec. 2.2.7.

In light of the $t\bar{t}$+HF analysis that is reported in Chapter 5, it is of crucial importance to identify jets from HF quarks, known as $b$– and $c$–tagging. These algorithms rely heavily on the tracking detector for charged particle reconstruction and the reconstruction of secondary decay vertices very close to the interaction point. The full jet reconstruction needs the information from the tracker as well as the calorimeter systems. Furthermore, the identification of the dileptonic top quark pair decay makes the identification of electrons (tracker + ECAL) and muons (tracker +

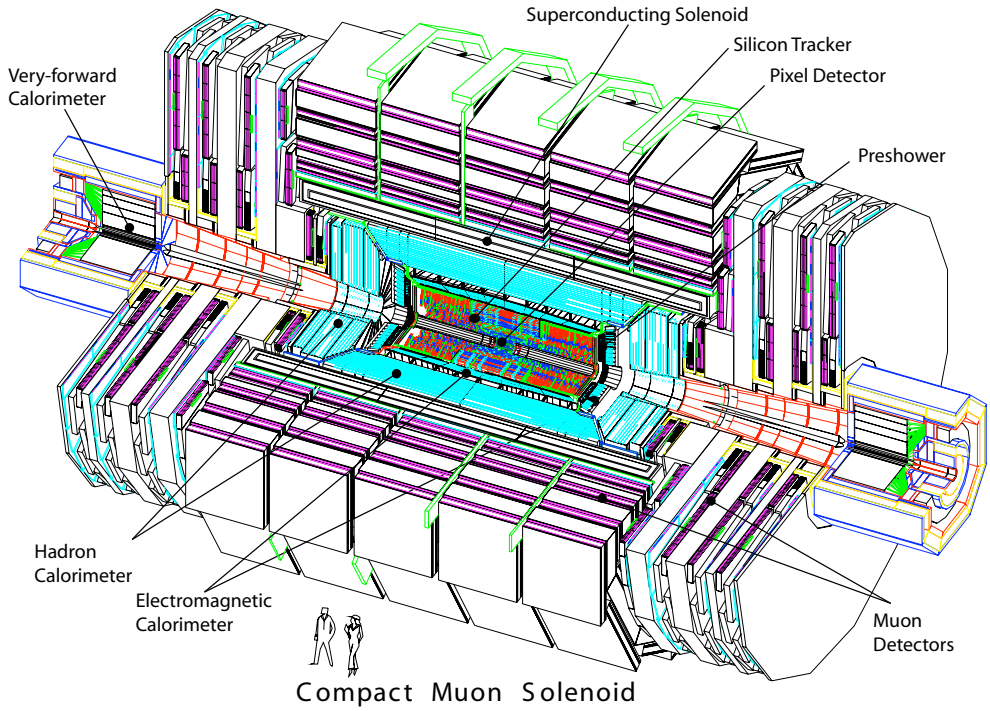


FIGURE 2.4: Schematic view of the CMS detector indicating its different sub–detector components. Figure taken from Ref. [131].

muon systems) indispensable. Therefore it is clear that all these sub–detectors are crucial to successfully perform the $t\bar{t}$+HF analysis.

### 2.2.1 The CMS coordinate system

The coordinate system [119] used in the CMS detector has its origin at the collision point in the middle of the experiment, the $y-$axis pointing upwards toward the surface, and the $x-$axis aimed towards the center of the LHC accelerator ring. The coordinate system is right–handed, resulting in the $z-$axis pointing along the beam axis (toward the Jura mountains from LHC Point 5 near the town of Cessy where the CMS detector is located). Positions inside the detector are measured in cylindrical coordinates: the azimuthal angle $\phi$ measures the inclination from the $x-$axis in the perpendicular plane with respect to the beam axis (i.e. in the $x-y$ plane), the radial distance $r$ is also measured in the $x-y$ plane and the polar angle $\theta$ measures the inclination from the $z-$axis. More often, the pseudorapidity

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right], \tag{2.2}$$

is used instead of the polar angle. For energies much larger than the mass of the outgoing particles, the pseudorapidity coincides well with the rapidity

$$y = \frac{1}{2}\ \ln\left(\frac{E+p_z}{E-p_z}\right). \tag{2.3}$$

Pseudorapidity is often preferred since differences in $\eta$ are Lorentz invariant.

### 2.2.2 The solenoid magnet

For a precise measurement of the charged particle momenta (as well as their charge), the curvature of the resulting tracks in a strong magnetic field needs to be determined. Therefore, a superconducting solenoid magnet was installed in the heart of the CMS detector [131], at a radius of approximately three meter around the beam pipe. With a length of 12.9 m it covers the barrel region of the detector. The solenoid is composed of 2168 turns of a high–purity aluminium–stabilized conductor through which runs a current of 19.5 kA (after cooling it down to around 4.6 K). In order to confine the magnetic field lines outside of the solenoid volume, the outer muon systems are interleaved with an iron return yoke. This way, an approximately homogeneous magnetic field of 3.8 T is obtained inside the solenoid in the barrel region, whereas in the endcap regions as well as in the outer barrel regions (outside the solenoid) the magnetic field is inhomogeneous.

### 2.2.3 The charged–particle tracker

The innermost part of the CMS detector is occupied by the tracking detector (tracker) [132, 133]. As charged particles fly through the magnetic field inside the detector, their trajectories describe a helix, of which the parameters are determined by the position, momentum, mass and charge of the particle. In order to identify the most interesting physics interactions it is crucial to correctly identify the primary interaction vertex of the hard collision between two protons, and to disentangle its position from those of vertices from pileup collisions that mostly give rise to low–energy tracks. Equally important in the production of top quark pair events with HF jets is the identification and reconstruction of secondary decay vertices, typically displaced by at most several millimeters with respect to the primary interaction position. The challenge boils down to dismantle the vast amount of tracker information as shown for example in Fig. 2.5, and identify the interesting tracks and vertices, which in this figure (presumably) reside from $b$ and $c$ hadron decays in semileptonic top quark pair events.

In order to accurately reconstruct primary and secondary vertices and the tracks which reside from those, a tracking system is placed very close to the beam pipe and which is composed of several layers of silicon–based semiconductor technology. The considerations on the architecture and material usage are driven by the requirement for the tracker to be radiation resistant. Given its close proximity to the collision point, it has to deal with a large flux of particles passing through each of its modules. A general schematic overview of the original tracking geometry and its different modules is shown in Fig. 2.6.
The region closest to the beam pipe, up to a distance of about 20 cm, has to deal with the largest flux of passing particles ($\sim$ 1 MHz at a radius of 15 cm), but is at the same time one of the most crucial parts for an optimal resolution of the track parameters. In this region silicon pixel modules have been installed, each with a surface area of $\sim 100 \times 150$ µm$^2$. The original design consisted of three barrel layers and two endcap disks placed on each side of the detector. However, during the technical stop at the end of 2016 and the start of 2017, this pixel detector has been upgraded with additional layers and lighter technology. A detailed discussion of the upgraded geometry will be discussed below.
After the pixel layer, the silicon strip detector starts which is composed of several submodules. In the barrel region, the inner barrel tracker (TIB) is composed of four layers of longer and thicker silicon strips (surface area of 10 cm $\times$ 80 µm, and thickness

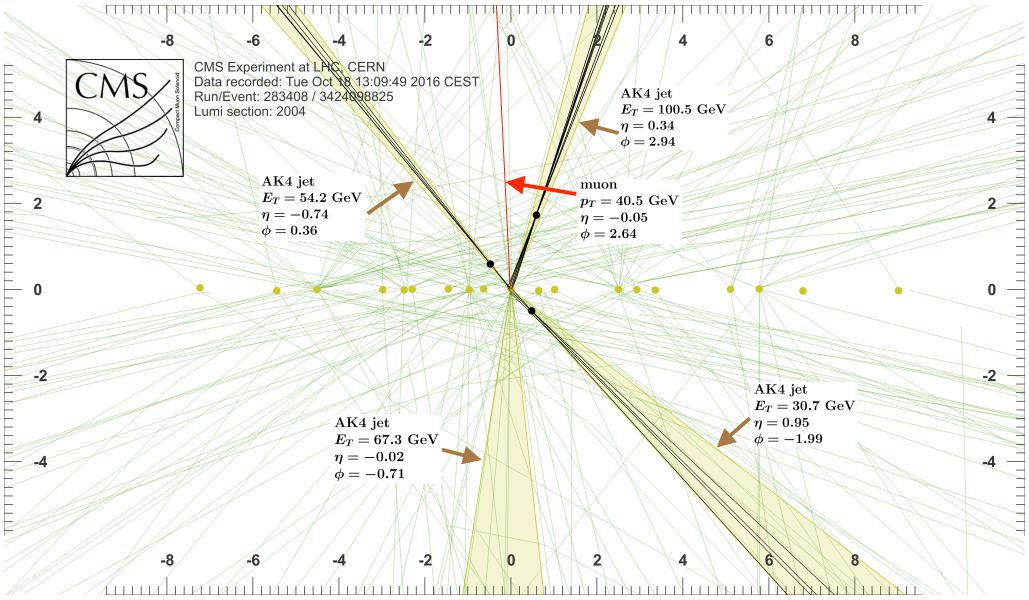


FIGURE 2.5: CMS recorded event in 2016 data collected at 13 TeV. Three jets have a displaced reconstructed vertex. Two of the jets are $b$–tagged, whereas another one is $c$–tagged using dedicated $b$–/$c$–tagging algorithms developed in the CMS collaboration (see Sec. 3.3 and 4.4). Tracks with $p_T > 0.5$ GeV are shown. The number of reconstructed primary vertices is 19. Reconstructed primary vertices are shown in yellow along the beam direction, whereas reconstructed displaced vertices and associated tracks are presented in black. Dimensions on the x and y axes are given in cm. Figure taken from Ref. [1].

of $\sim 320\,\mu$m). The outer barrel tracker (TOB) deals with an even more reduced flux, resulting in six layers of even thicker ($\sim 500\,\mu$m) silicon strips. The endcap tracker also uses silicon strip technology and is composed of an inner disc tracker (TID) consisting of thee small discs at each side of the TIB, and a larger endcap tracker (TEC) which is comprised of nine more layers on each side of the entire barrel tracker. All together the strip tracker is composed of over 15 thousand modules, mounted on a carbon–fibre structure and cooled to a temperature of around $-15$°C.

### Phase I upgrade of the pixel tracker

During the end–of–the–year technical stop of the LHC at the end of 2016, the CMS pixel detector has received a full upgrade [133]. This upgrade was performed in order to deal with the increased luminosity, resulting in higher pileup multiplicity and more severe radiation damage. The main goal was to obtain as good or better performance with the upgraded pixel tracker in these high luminosity conditions, compared to the original tracker geometry performance in the lower luminosity conditions.

The original three–layer barrel, two–disc endcap architecture has been elevated to a four–layer barrel, three–disc endcap geometry, as illustrated in Fig. 2.7. The upgraded pixel detector consists of 1440 silicon pixel modules. Whereas the original three–layer barrel had its layers positioned a radius of 4.4, 7.3 and 10.2 cm from the interaction point, the new four–layer geometry has a larger radial coverage, with layers positioned at radii of 3.0, 6.8, 10.2 and 16.0 cm. It covers a pseudorapidity of up to $|\eta| < 2.5$. Not only is there a clear advantage of four–hit track reconstruction

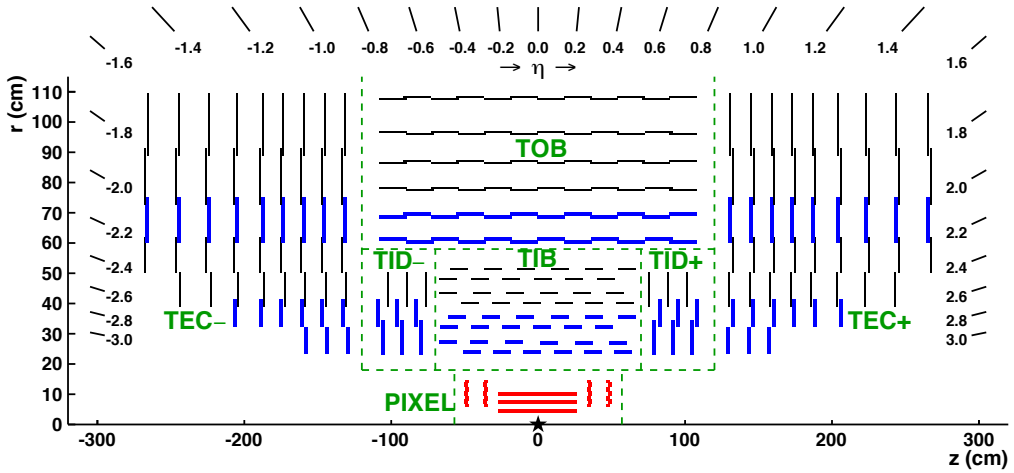


FIGURE 2.6: Schematic overview of the original tracker system. Figure taken from Ref. [132].

opposed to using only three hits, but the fact that the innermost layer is even closer to the interaction point increases a lot the accuracy of the track reconstruction algorithms.

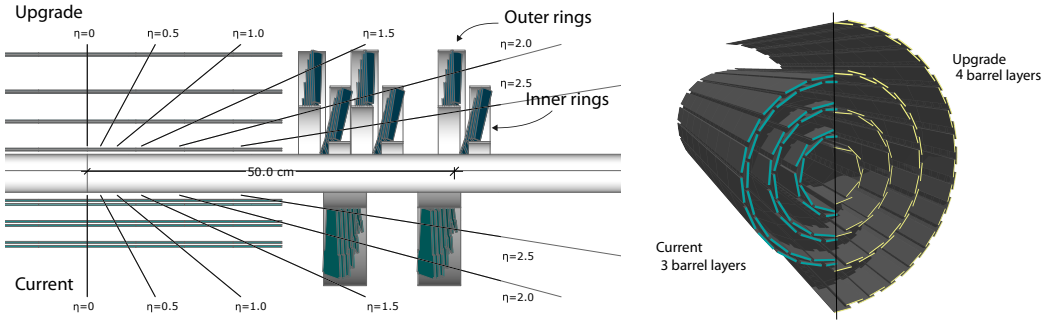


FIGURE 2.7: Upgrade of the pixel detector geometry: the old 3–layer barrel (BPIX), 2–disk endcap (FPIX) design has been replaced in 2017 with a 4–layer barrel, 3–disk endcap layout. (left) Schematic overview of the old pixel design (lower half) and the new pixel design (upper half). (right) Three–dimensional visualization of the original and the upgraded pixel barrel layout. Figures taken from Ref. [133].

In order to demonstrate the improved performance of the upgraded pixel tracker, Fig. 2.8 compares the performance between the original and the upgraded pixel tracker in terms of a reduced impact parameter[3] resolution of muon tracks, in a high pileup multiplicity environment. An improved resolution of more than 50% is observed over a large range of track momenta. Later in Sec. 3.3.3, a clear improvement in the performance of HF tagging algorithms will be demonstrated with the upgraded pixel detector, which will be of great value in the $t\bar{t}$+HF cross section measurement discussed in this thesis.

[3]The track impact parameter (IP) defines the distance of closest approach between a reconstructed track and the reconstructed primary interaction vertex it was assigned to.

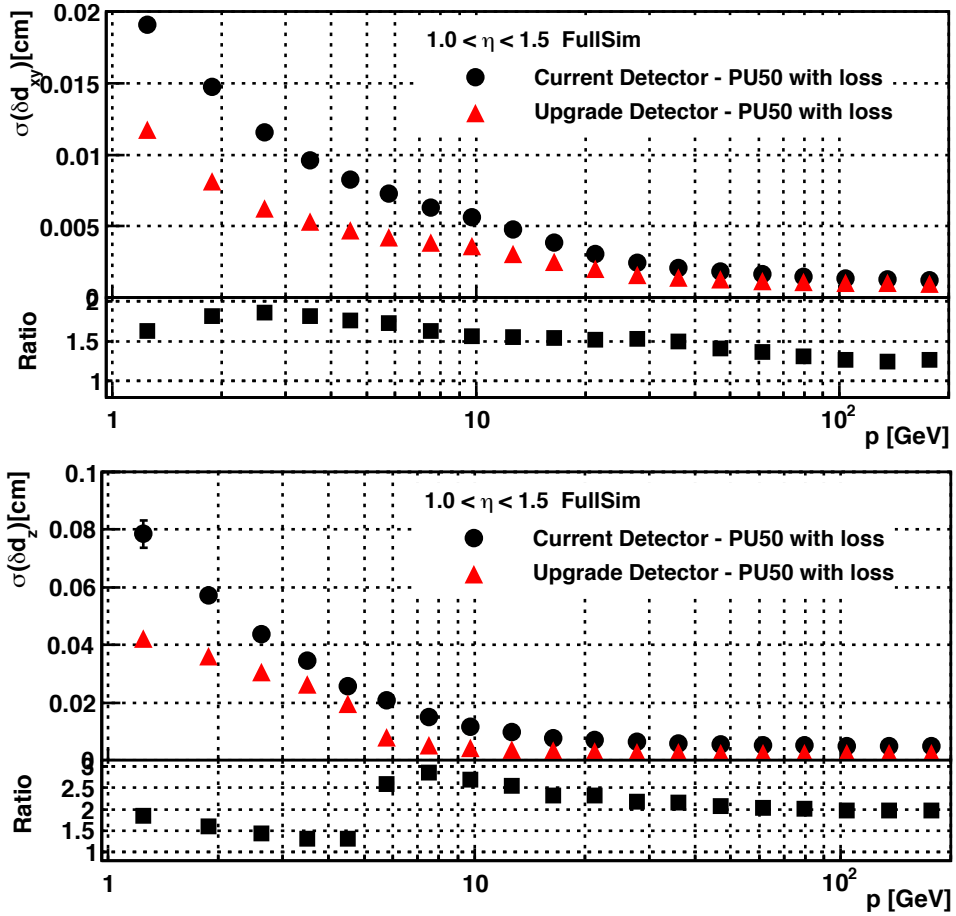


FIGURE 2.8: (top) Transverse impact parameter resolution, $\sigma(\delta d_{xy})$ (in simulation) for muon tracks as a function of the track momentum for $1.0 < |\eta| < 1.5$ for the original pixel detector (black circles) and the Phase I upgrade detector (red triangles) for an average pileup multiplicity of 50.
(bottom) Same as the upper figure, but for the longitudinal track impact parameter resolution $\sigma(\delta d_z)$. Figures taken from Ref. [133].

### 2.2.4 The electromagnetic calorimeter

Behind the tracking layers, an electromagnetic calorimeter (ECAL) [131, 134] system has been installed. The main objective of the ECAL is to measure the energy deposits of electromagnetic showers induced by photons and electrons. The reconstruction and identification of electrons and photons is not a trivial task. It is important to disentangle prompt photon production from for example possible bremsstrahlung photons radiated by electrons or from collinear diphoton signatures from neutral pion decays. This means that a good granularity of the ECAL is a crucial feature. Additionally, the ECAL should be radiation hard, should have a rapid response time and should be compact enough to fit within the solenoid volume. All of these prerequisites have led to the choice of lead tungstate ($PbWO_4$) scintillating crystals to build the ECAL detector. Due to the scintillating properties of this material, light is produced as the electromagnetic shower passes through the crystals. This light is then captured in photodiodes with intrinsic gain to amplify the signal. The $PbWO_4$ crystals are radiation hard and have a short radiation length of $X_0 = 0.89$ cm, allowing for a compact ECAL system. The readout is also fast, as about 80% of

the light is emitted within 25 ns (so within the bunch crossing time of the LHC).

A schematic overview of the ECAL detector is shown in Fig. 2.9. Three components of the ECAL are installed: the barrel ECAL (EB), the endcap ECAL (EE) and the ECAL preshower (ES). The EB has a pseudorapidity coverage of up to $|\eta| < 1.479$. Each of the crystals in the EB covers $1°\times1°$ in $\Delta\eta \times \Delta\phi$, resulting in a front face cross section of $22\times22$ mm$^2$ and a length of 230 mm, which corresponds to $\sim 26$ $X_0$. A total of 61200 crystals is placed in the EB. Additionally, the EE is comprised of 7324 crystals in each of the two endcaps, covering $1.479 < |\eta| < 3.0$. These have a front face cross section of $28.6\times28.6$ mm$^2$ and a length of 220 mm ($\sim 25$ $X_0$). Finally the dedicated preshower system is composed of two layers of lead absorbers, with silicon strip detectors behind them. Through their superior granularity, they help not only to identify neutral pions but additionally improve position determination of electrons and photons in the region $1.653 < |\eta| < 2.6$.

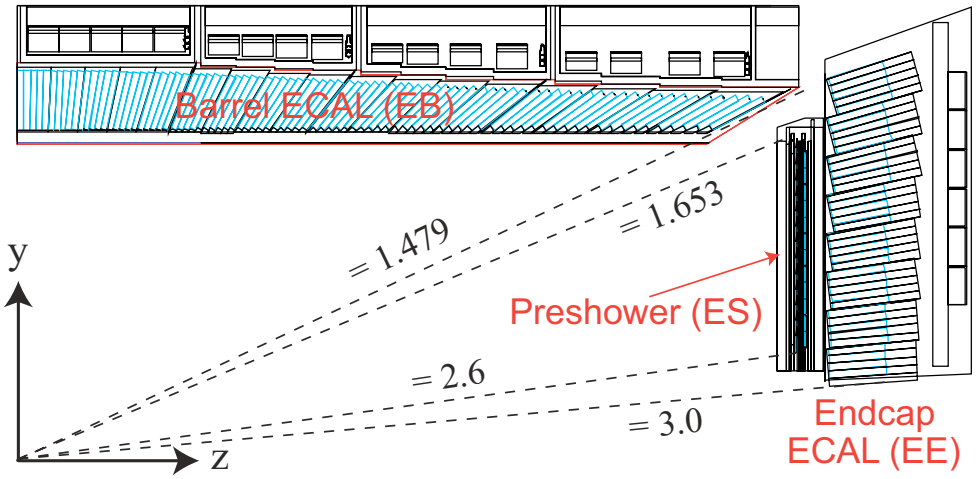


FIGURE 2.9: Schematic visualization of a quarter of the CMS ECAL system, showing the position of the barrel ECAL (EB) and the endcap ECAL (EE). Figure taken from Ref. [131].

The typical momentum resolution for electrons with $p_T \approx 45$ GeV from $Z \to e^+e^-$ decays ranges from 1.7% to 4.5% [135]. Photons are reconstructed with an energy resolution of the order of 1–2% in the barrel, going up to 3–4% in the endcaps [136].

### 2.2.5 The hadronic calorimeter

Charged and neutral hadrons will lose only a very small portion of their energy in the ECAL layers. Therefore a hadronic calorimeter (HCAL) [131, 137] is placed after the ECAL layers to measure the hadronic activity in the detector. Together with the ECAL it makes up an indispensable calorimetry system which is crucial for the reconstruction of jets and missing transverse energy. In the barrel region, this detector comprises the available space between the ECAL and the solenoid magnet (HB), supplemented with an outer barrel HCAL (HO) which is located just outside the solenoid (contained in the barrel muon system). In the endcap regions the barrel system is hermetically closed through the endcap HCAL (HE) and the very forward region is covered by the forward HCAL (HF). These different modules are visualized

in Fig. 2.10.

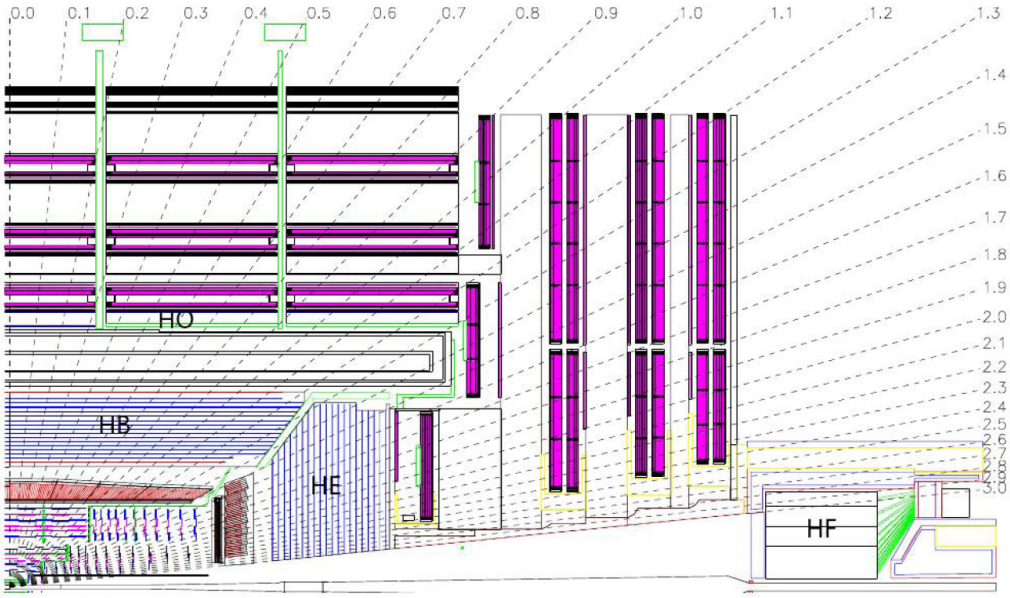


FIGURE 2.10: Schematic visualization of a quarter of the CMS HCAL system, showing the position of the barrel HCAL (HB), the outer HCAL (HO), the endcap HCAL (HE) and the forward HCAL (HF). The magnet is displayed in between the HB and the HO. Figure taken from Ref. [131].

The HB system was constructed using Brass absorber layers to induce the hadronic shower, interleaved with plastic scintillator tiles to measure the energy of the passing particles. The absorber layer has a reasonably short interaction length, such that the majority of the shower is contained in the 15 layers that are placed inside the solenoid magnet volume. This systems covers a pseudorapidity range of $|\eta| < 1.4$ and is composed of 2304 segments of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$, meaning that each HCAL tower overlaps with a 5×5 tile of ECAL crystals. The HO system includes additional scintillator layers outside the magnet volume, increasing the effective thickness of the HCAL system by around ten radiation lengths. This system (covering a pseudorapidity range of $|\eta| < 1.26$) allows to capture the remaining part of the hadronic shower that passed through the solenoid and therefore highly increases the resolution on the jet energy and missing transverse energy measurements.
The HE system has a pseudorapidity coverage of $1.3 < |\eta| < 3.0$ (thus overlapping slightly with the HB) and its 2304 modules have a much broader $\phi$ segmentation (from 5° to 10°), but with the same $\eta$ segmentation as the HB. The HF system covers $3.0 < |\eta| < 5.2$ and is located closest to the beampipe. Therefore it needs to be much more radiation hard, resulting in a design of steel absorbers with quartz scintillating fibres. This technology is based on Cerenkov light produced in the quartz fibres, which is collected by photomultipliers.

The energy resolution of the hadronic calorimeter is typically of the order of 10–20% (depending on the pseudorapidity) for hadrons with an energy of around 50–100 GeV [138].

### 2.2.6 The muon detector

As can be inferred from the name, the CMS detector has a state–of–the–art muon detection system installed [131,139]. This system comprises the outermost layers of the detector interleaved by the return yoke and complements the muon track measurement in the inner tracking detector. The technology and geometry which is employed in the barrel ($|\eta| < 1.2$) and endcap ($0.9 < |\eta| < 2.4$) regions differs due to the different radiation environment and magnetic field configuration. This results in three different detector subsystems based on gaseous detector technologies. The barrel region is equipped with Drift Tubes (DT), whereas Cathode Strip Chambers (CSC) are used in the endcap muon system. Each of these is also supplemented with Resistive Plate Chambers (RPC). The different components of the muon system are visualized in Fig. 2.11 and their geometry is detailed below. For a more comprehensive overview on how these detectors work, see for example Chapter 7 of Ref. [140].

**DT :** The barrel muon system is composed of four layers at radii of around 4.0, 4.9, 5.9 and 7.0 m from the interaction point. The three innermost layers are composed of twelve planes each of which eight measure the position in the $r - \phi$ plane and four measure the $z$–coordinate. The outermost layer has only eight planes to measure the $r - \phi$ position. DTs allow for an excellent spacial resolution, and are supplemented with RPCs to improve also the timing resolution.

**CSC :** The CSC modules in each of the endcaps are multi–wired proportional chambers. These are composed of anode wires oriented parallel to cathode strips in a gaseous volume. The signals from the anode wires are fast and can be efficiently used in the trigger, though the position resolution is more coarse. This is complemented with a more precise (though slower) position measurement on the cathode strips. Again fast and precise timing information is provided by the RPCs.

**RPC :** Supplementing the DT and CSC modules, the RPCs allow for a very precise timing resolution of the order of 1 ns. This makes it possible to unambiguously assign muons to their corresponding bunch crossing. Their spatial resolution is much worse compared to that of DTs and CSCs.

Combining the different muon system technologies with the existing information from the inner tracking system results in an optimal muon momentum resolution of ∼1–2% in the barrel for muons with a $20 < p_T < 100$ GeV and still better than 10% for muons with a $p_T$ up to ∼1 TeV [141]. The fast timing and position information supplied by the muon system is also used in the trigger.

### 2.2.7 Triggering and data acquisition

Knowing that the event size of a single bunch crossing is of the order of a MB, it becomes immediately clear that at a rate of 40 MHz, not every collision event can be saved to storage. Taking into account the bandwidth of the readout electronics and storage capacity, a manageable rate of 100 Hz dictates the need for a rejection factor of the order of $10^5$. The trigger system [131, 142, 143] is therefore responsible for making decisions on which collisions potentially contain interesting physics. This is an absolute vital step in the data acquisition and reconstruction chain, given that an event which is rejected by the trigger is lost forever. On top of that, the available time to make such a crucial decision is limited by the time between adjacent bunch

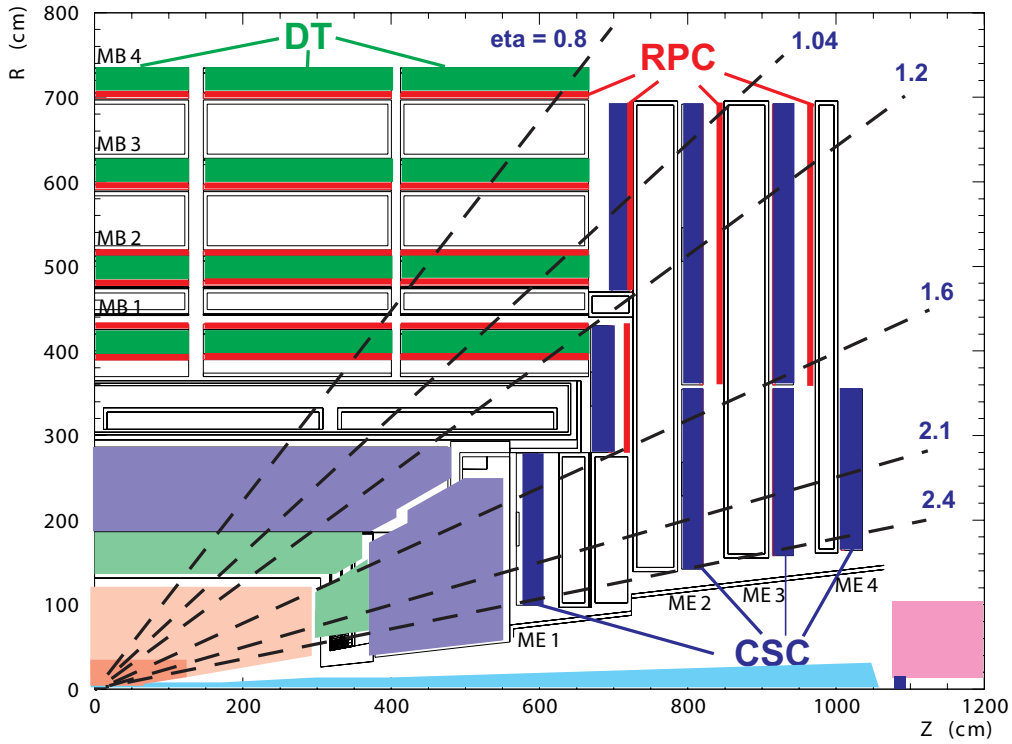


FIGURE 2.11: Schematic visualization of a quarter of the CMS muon detector system, showing the position of the drift tubes (DT), resistive plate chambers (RPC) and the cathode strip chambers (CSC). Figure taken from Ref. [131].

crossings.

The CMS data acquisition chain can be decomposed in three parts: the level–1 trigger (L1), the processor farm (DAQ) and the high–level trigger (HLT).

**L1 :** The L1 trigger system is based on custom hardware processors that are aimed to reduce the recorded event rate to ∼100 kHz. The total time allocated for deciding whether or not to keep an event is 3.2 µs. This includes the transport of the signals to and from the front–end electronics of the L1 trigger hardware. This leaves approximately 1 µs for the actual L1 decision while the raw detector data is stored in buffers. The L1 decision can only be made using primitive objects built from information that is easily available on this short timescale. This includes energy measurements from the calorimeter systems (using reduced granularity) and hit information from the fast muon system. Despite this limited amount of information, the trigger system is capable of doing a first primitive reconstruction of electrons, muons, photons and jets and to use global information on the summed transverse energy of the collision event[4]. All of the information is finally gathered in a global trigger used to decide whether or not the buffered event data can continue their journey along the data acquisition chain. This is summarized in Fig. 2.12.

[4]Tracker information is currently not used at the L1 trigger level. Plans exist on including track–based information in the L1 trigger for the Phase II upgrade of CMS to be used with the HL–LHC [144].

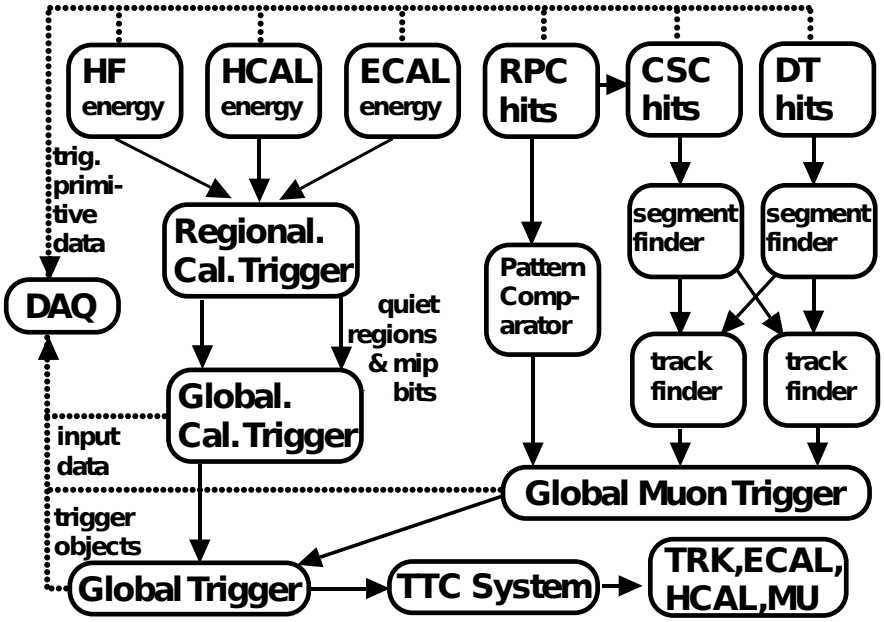


FIGURE 2.12: Schematic overview of the sub–detector information used in the L1 hardware trigger. Figure taken from Ref. [143].

**DAQ :** In case of a favorable L1 trigger decision, the data are passed to the front–end drivers (FEDs) and sent through a switching system. The data of a single event are at this stage contained in several hundred front–end buffers and through a switching system they are recombined into readout processor units that gather all the data of one specific collision. This readout processor farm contains software that describes the second filtering stage, namely of the HLT. A schematic overview of the DAQ system is shown in Fig. 2.13.

**HLT :** In order to make a second trigger decision which lies as close as possible to the offline analysis needs, the HLT is a software–based trigger. Within the available timescale of a few seconds it has the possibility to perform an event reconstruction which is typically a simplified form of the actual offline reconstruction algorithms. A list of dedicated HLT paths are designed to accommodate the needs of many analyses performed by the CMS collaboration. These HLT trigger tables can be constantly updated to accommodate for more advanced analysis needs and the changing environment (increasing luminosity, pileup, collision energy, etc.). The final filtering is based on the decision of the logical "OR" of all these trigger paths and should result in a data throughput of the order of 100 Hz, to be saved to storage.

The events that pass the HLT undergo a full event reconstruction chain after which they are ready to be analyzed (see Chapter 3). This reconstruction and data storage happens through computing resources made available by institutes all over the world that are connected to CERN. This inter–connected network is often referred to as the "Grid" [145] and allows for mass storage and data access from all over the world.

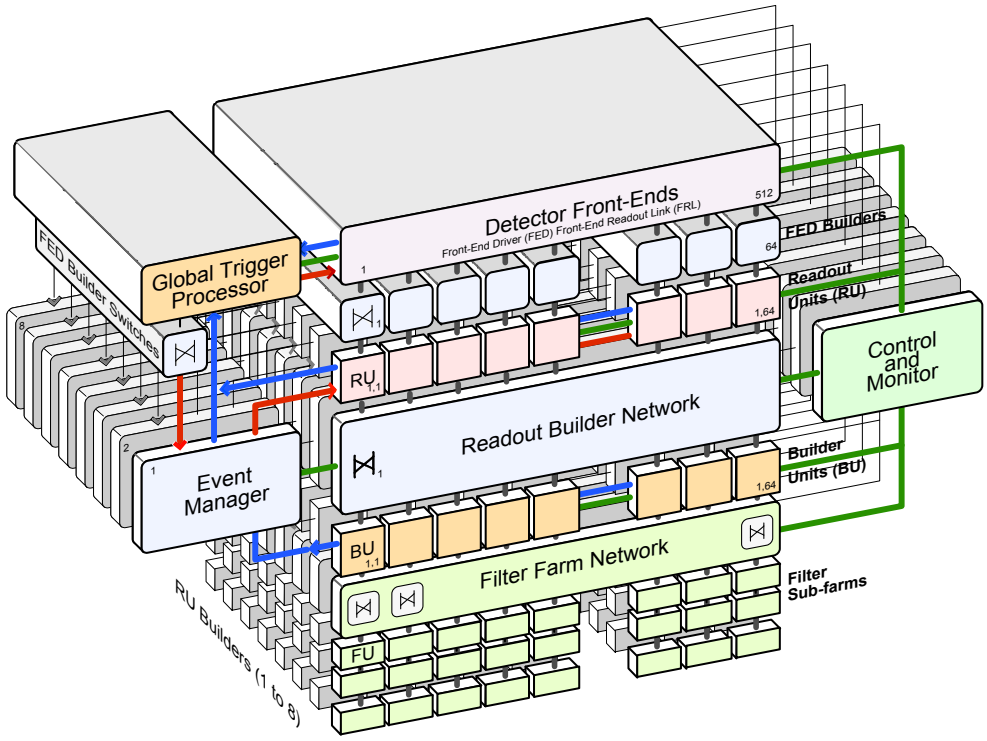


FIGURE 2.13: Structure of the CMS data acquisition farm that takes the information of the L1 trigger and builds the events from the electronic signals of the sub–detectors. Figure taken from Ref. [142].

### 2.2.8 CMS data-taking performance in 2017

After the somewhat extended end–of–the–year technical stop in 2016–2017, the LHC started circulating protons again by the end of May 2017. The first runs were used mainly for commissioning studies to test and validate the newly installed pixel tracker (as well as the rest of the detector). Starting mid–June, the LHC has delivered almost 50 $\text{fb}^{-1}$ of data before the end of 2017, of which a stunning 45 $\text{fb}^{-1}$ has been collected by CMS as shown in Fig. 2.14. After validating the quality of the data, CMS released a so–called "*Golden JSON*" which is a summary of all run numbers which are good for data analysis. This resulted in a high–quality dataset of 41.5 $\text{fb}^{-1}$ of integrated luminosity. To get an idea, this dataset contains around 35 million top quark pair events and roughly 2.1 million H bosons. This is the dataset that will be used in the analysis presented in this thesis.

A peak instantaneous luminosity of $\sim 2 \times 10^{34}$ $\text{cm}^{-2}\text{s}^{-1}$ (twice the design luminosity) was an absolute record in 2017. The downside of this amazing achievement is the increased amount of pileup interactions per bunch crossing. Assuming an inelastic proton–proton scattering cross section of 69.2 mb [146], an average of 32 pileup vertices per bunch crossing has been observed in 2017. The entire pileup distribution is shown in Fig. 2.15. It can be seen that a low–pileup run was performed which is typically used for analyses focusing on forward physics (using CASTOR[5]). The tails of this distribution reach up to around 70 pileup vertices at the highest luminosities.

[5]The CASTOR (Centauro And Strange Object Research) detector is a very forward calorimeter system inside the CMS detector (near TOTEM), used especially to investigate heavy–ion collisions [147].

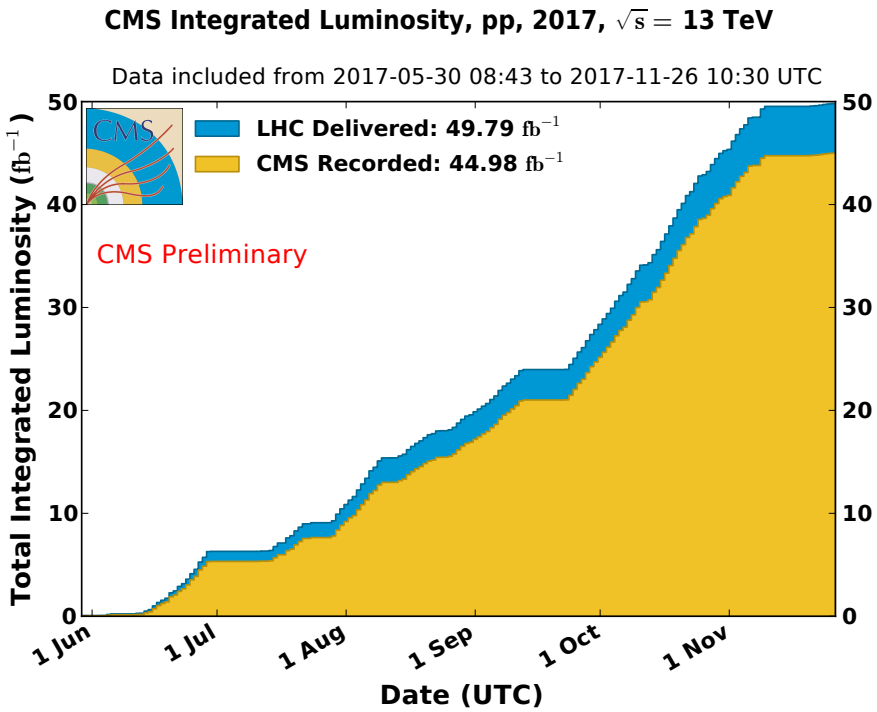


FIGURE 2.14: Cumulative integrated luminosity of the CMS experiment (yellow), compared to the delivered luminosity by the LHC (blue) in 2017 proton–proton collisions at $\sqrt{s} = 13$ TeV. Figure taken from Ref. [129].

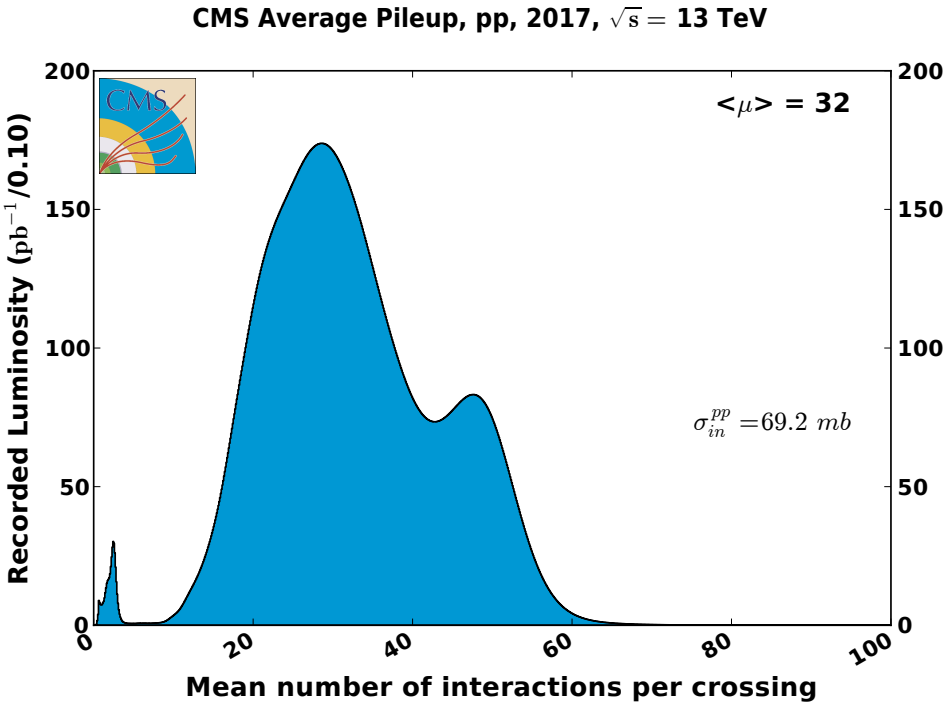


FIGURE 2.15: Mean number of interactions per bunch crossing for the 2017 proton–proton collisions at $\sqrt{s} = 13$ TeV, using a minimum bias cross section value of 69.2 mb, which is determined by finding the best agreement with data and is recommended for CMS analyses. Figure taken from Ref. [129].

# Chapter 3 Simulation and reconstruction of proton–proton collisions

If new physics phenomena are indeed hiding within the vast datasets of $pp$ collisions, the only way to detect it is to look for deviations from the SM expectations. Consequently, one can not expect to actually discover anything new without a proper and accurate prediction of what is already known. The need for accurate theoretical predictions is inevitable and keeps the synergy between the theoretical and experimental community essential. These predictions however go far beyond the calculation of matrix elements and partonic cross sections. What we physically observe are interactions of final stable particles, many of which are composite hadrons, with the material of the detectors. A proper calculation of the matrix element up to the highest possible order in perturbation theory is just the start of this chain. After that the effects of multiple parton interactions, additional radiation, hadronization, decays and detector response need to be predicted as well. This will be discussed in detail in Sec. 3.1.
As we can not directly detect new particles from a voltage extracted from a given detector element, the next challenge presents itself in trying to properly reconstruct the particles that interact with the detector components. Dedicated reconstruction algorithms have been developed to translate the electronic signals into a final set of physics objects: e.g. muons, electrons, photons, jets and missing energy. CMS has made the choice to use a so–called particle flow algorithm to combine information from all sub–detectors to achieve an optimal particle identification and reconstruction. Sec. 3.2 deals with these reconstruction techniques in detail.
Finally, given its central role in the measurement of top quark pair production with additional HF jets, Sec. 3.3 is fully devoted to the description of HF jet identification algorithms, including a discussion on the state–of–the–art flavor–tagging algorithms used by the CMS collaboration.

## 3.1 Simulating proton–proton collisions at the LHC

Providing accurate theoretical predictions on the observable quantities which are extracted from our particle detectors is a highly non–trivial task. Our ability to deal with the complexity of these high–energy proton collisions relies for the largest part on simulation software. These simulations are often not based on exact analytical calculations, but rather rely on Monte Carlo sampling techniques [148,149].

The LHC is a discovery machine, resulting in the choice to collide composite hadronic particles (protons) which suffer less from synchrotron radiation and allow for higher maximum collision energies. The theory of event simulation at hadron

colliders has made big steps forward in the last decade [150–152]. The composite nature of the colliding particles in combination with the relatively poor understanding of the strong force (QCD) that acts among them makes this field of research extremely challenging. The principle of asymptotic freedom dictates that the strong force becomes weaker, and therefore more perturbative, at higher energies. Therefore the predictions from matrix element calculations and partonic cross sections of the hard scattering process in high–energy collisions can be predicted up to a relatively good precision. Typically the most accurate predictions are already at NNLO accuracy in QCD. Unfortunately, the story does not end there as the hadronic particles resulting from these collisions still have a way to go before interacting with the detector material. A proper description of extra radiation (parton shower), hadronization, decays and detector response lies ahead. Additionally the partons that do not participate in the hard scattering can still undergo soft scattering processes (underlying event). All of these additional steps in the chain typically take place at much lower energies and in regimes in which the strong force stops being perturbative and our predictions become less and less accurate. Exact solutions up to a fixed order in perturbation theory are not anymore appropriate to describe these phenomena and other (phenomenological) models need to be used to predict these steps.

Before expanding in more detail on each of the steps in this simulation chain, we summarize once more how a typical *pp* collision proceeds. A schematic representation of a typical *pp* collision is shown in Fig. 3.1, indicating each of the steps in the chain.

- **Hard scattering:** When two protons collide, only one of the partons (quarks or gluons) inside each proton participates in the main interaction of interest. An important ingredient in this process is the underlying structure of the proton that dictates the probability for each parton to participate in this hard scatter. The corresponding matrix elements need to be calculated, from which a (differential) cross section can be determined. This is outlined in more detail in Sec. 3.1.1.

- **Parton shower:** Additional particles may be produced through radiation of either gluons (QCD radiation) or via electromagnetic radiation. This is described by the parton shower (PS), that deals with non–trivial descriptions of very soft or very collinear splittings. As the PS evolution proceeds, it takes into account conservation of energy and momentum at each step until a certain energy scale has been reached. This is discussed further in Sec. 3.1.2.

- **Hadronization and fragmentation:** Due to the principle of color confinement, only color neutral hadrons are observed. Consequently the colored partons after the PS undergo a hadronization process. These hadrons can then further decay into stable particles that make their way through the detector. The non-perturbative nature of QCD at these energy scales makes these calculations highly non–trivial, and solutions to this problem are summarized in Sec. 3.1.3.

- **Underlying event:** The part of the proton that did not participate in the hard interaction can still undergo soft scattering processes. These give rise to multi–parton interactions resulting in the so–called underlying event (UE) which needs to be modeled appropriately, as discussed in Sec. 3.1.4.

- **Detector simulation:** Finally, the stable particles will interact with the detector material to leave their final trace for physicists to observe. A proper simulation of the detector repsonse, including for example energy deposits and showering in the calorimeters, constitutes the final step before the event reconstruction can start. This subject is touched upon in Sec. 3.1.5.

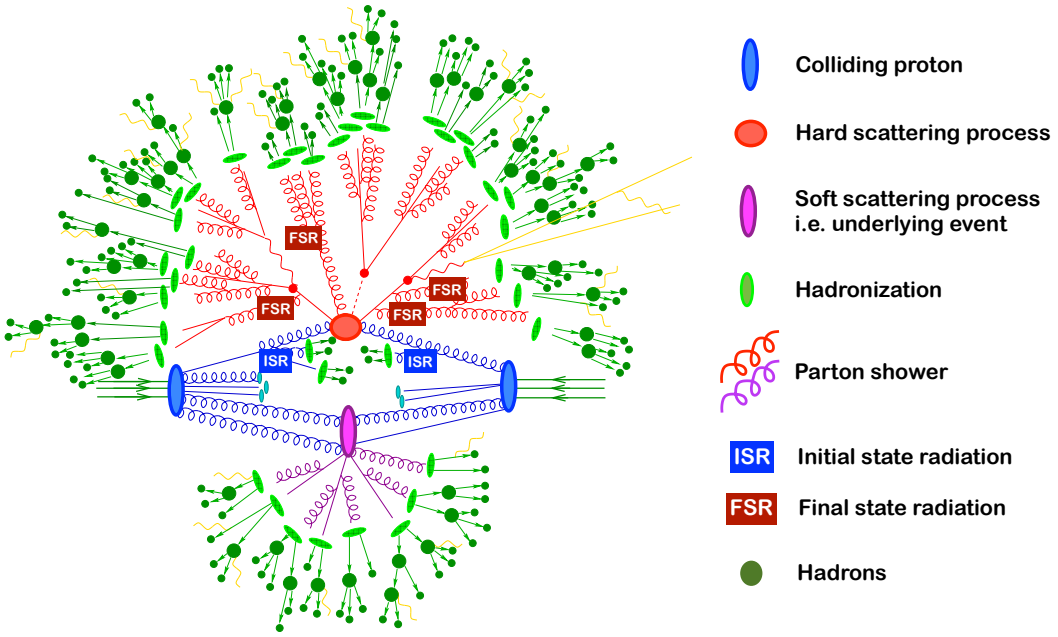


FIGURE 3.1: A schematic representation of a typical $pp$ collision, indicating the different steps of the simulation chain. Figure adapted from Ref. [153].

### 3.1.1 The hard scattering process

Probably the most fundamental step in the simulation chain is to calculate the matrix element of the process of interest. The first challenge is to provide a proper description of the proton, being itself a composite particle from which only a single parton will take part in the hard scattering process. The proton is composed of two up quarks and one down quark. These are called the valence quarks of the proton. However, a quantum–mechanical description of the proton is needed to properly describe its behavior. Rather than being a fixed composition of three quarks, the proton is in fact a dynamical system in which gluons and quarks of all flavors can constantly be created and annihilated. Gluons inside the proton may for example split into quark–antiquark pairs, which after a short period of time annihilate again. Which partons are present and how likely it is to find them inside the proton depends on the energy at which the proton is probed. The proton Parton Density Functions $f_i(x, Q^2)$ (PDFs) describe the probability density to find a certain parton $i$ inside the proton with a fraction $x$ of the proton momentum[1], depending on the probed energy scale $Q^2$. The determination of these PDFs is therefore a very important and active field of study. Collaborations like NNPDF [154], MSTW [155] or CTEQ [156] are focused on fitting these functions to data observed in deep inelastic scattering, Drell-Yan and multijet processes. The advantage of these measurements lies in the fact that the resulting PDFs are not process–specific and can thus be applied in any

[1]This quantity of sometimes referred to as "Björken $x$".

other simulation or calculation. An example of one of the PDF sets used by CMS throughout the simulations for the 2017 data–taking period of the LHC is shown in Fig. 3.2. This figure illustrates that at higher momentum fraction $x$ the valence quarks are indeed most abundant and a clear asymmetry is observed between the valence quarks and their corresponding antiquarks. This asymmetry is clearly not present for strange, charm and bottom (anti)quarks as they are equally likely to be found inside the proton. At lower energy fractions however, the gluon PDF (which in the figure is scaled by a factor $1/10$) is by far the largest, leading on average to a much larger interaction rate for gluon induced processes in proton collisions. This illustrates why the gluon fusion production mechanism for $t\bar{t}$ production (see Fig. 1.4) is the dominant one at the LHC.
The dependence of the PDFs on the energy scale $Q^2$ is described by the so-called DGLAP[2] equations [157–159]. This allows for an interpolation of the fitted PDFs between measurements at different values of $Q^2$.

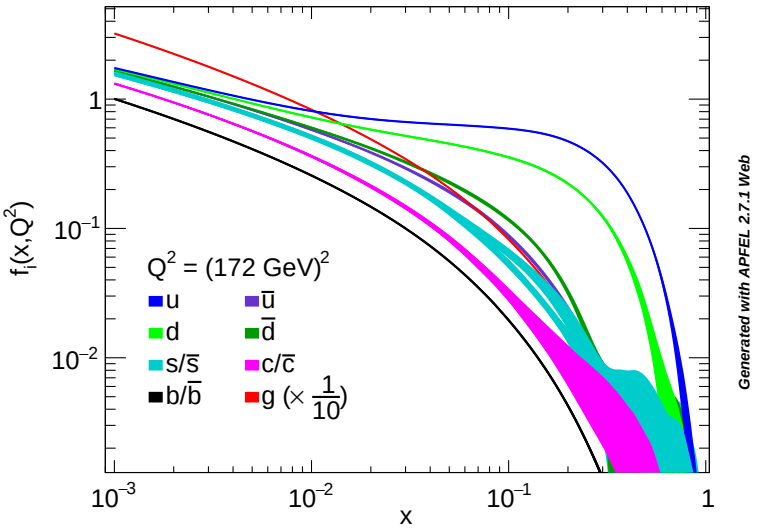


FIGURE 3.2: Parton distribution functions $f_i(x, Q^2)$ as a function of the fractional proton momentum $x$, for the quarks and gluons inside the proton, evaluated with the `NNPDF31_nnlo_as_0118` PDF set [154] at a value of $Q^2 = (172)^2$ GeV$^2$. The gluon PDF is scaled with a factor $1/10$. Uncertainty bands show the $\pm 1\ \sigma$ band on the measured PDFs. Figure generated from Ref. [160, 161].

With a proper description of the proton structure at hand, the ingredients are in place to determine the matrix element of the process under consideration ($pp \to X$) and to derive a cross section $\sigma_{pp\to X}$. The realization of this tedious task relies on the basic theorem of the factorization of hard processes [162]. This theorem states that the total cross section can be factorized through the convolution of the *partonic cross section* ($\hat{\sigma}_{ij\to X}$) of partons $i$ and $j$ and their corresponding PDFs $f_i\left(x_i, Q^2\right)$ and $f_j\left(x_j, Q^2\right)$. In order to accommodate for all possible initial–state configurations, this convolution needs to be integrated over the fractional momenta $x_i$ and $x_j$ and summed over all possible initial–state partons that may result in the final–state $X$ of interest, as shown in Eq. (3.1).

$$\sigma_{pp\to X} = \sum_{i,j} \int dx_i \int dx_j\, f_i\left(x_i, Q^2\right) f_j\left(x_j, Q^2\right) \hat{\sigma}_{ij\to X} \tag{3.1}$$

[2]DGLAP: Dokshitzer–Gribov–Lipatov–Altarelli–Parisi.

The partonic cross section, $\hat{\sigma}_{ij\to X}$, is then determined by the matrix element (ME) calculation and the phase–space integration of the final–state $X$, and is completely factorized from the proton PDFs. This partonic cross section can now be determined up to a fixed order in perturbation theory. This is done with matrix element generator software that provides automatic calculations of the matrix elements up to a given order in perturbation theory (usually restricted to NLO or NNLO) and which uses Monte Carlo sampling techniques to generate a set of simulated events in the desired phase space. Examples of event generators that are used in the CMS simulations are POWHEG (LO+NLO) [163–165], MADGRAPH (LO) or MG_aMC@NLO (NLO) [166] and PYTHIA (LO) [167]. These ME generators have in common that they need as input a model that describes the particle content, couplings, interactions and other constants to be used in the matrix element calculation. Such a model can either describe the SM, or a BMS model that for example implements the SMEFT operators, as will be used in Chapter 6. FEYNRULES [168] is a commonly used software package that provides a common interface to generate such models and encapsulates it in the UNIVERSAL FEYNRULES OUTPUT (UFO) format [169] that can be used as input to each of the above mentioned ME generators.

Finally, it is important to note that the ME calculations have to deal with ultraviolet, infrared and collinear divergences that appear in fixed order calculations in perturbation theory. These divergences are merely artifacts of the perturbative expansion and should always vanish when an infinite number of perturbative orders is taken into account. The limited computing resources however force us to deal with these divergences and to regulate them. Ultraviolet divergences are regulated by the so-called *renormalization scale* $\mu_R$. This energy scale shows up in the evaluation of the strong coupling constant $\alpha_s$. On the other hand, infrared and collinear divergences show up with the emission of low–energy (infrared) gluons or emissions which are almost in the same direction as the parton (collinear). These divergences are regulated through a parameter called the *factorization scale* $\mu_F$ and shows up both in the PDFs (through the DGLAP equations) and in the partonic cross section. These parameters are in principle chosen arbitrarily, though at physically meaningful values for the typical energy scales of the considered process. This turns out to be extremely non–trivial in the case of $t\bar{t}$ + HF production, given the very wide range of energy scales (from the top quark mass to the relatively soft emission of additional HF jets) involved in such a process [170–173]. In any case, the somewhat arbitrary choice of these parameters may seem troublesome at first. Typically the dependence of a physical observable on the value of $\mu_F$ and $\mu_R$ diminishes with higher orders in the perturbative expansion, and the remaining dependence is taken into account as an uncertainty[3].

### 3.1.2 Parton showering

The colored particles that participate in the hard scattering process (and any additional scattering) can undergo a chain of soft radiations or branchings into other particles [153]. This soft radiation and collinear splitting results in additional final–state particles which are initially not included in the matrix element calculation. Adding this extra radiation directly to the ME may seem like the most physically motivated solution, but drastically increases the complexity of the

[3]The uncertainties related to the value of the renormalization and factorization scales are evaluated by varying these scales up or down by a factor of two in the simulation chain. Outside of this range the predictions do not agree anymore with the observations in data.

calculations. Alternatively, it is typically favored (or even necessary) to describe the additional radiation as a *parton shower* (PS). This treatment corresponds to an approximate higher–order correction to the hard scattering in the limits of either very soft radiation of gluons ($\overset{(-)}{q} \to \overset{(-)}{q}g$) or very collinear splitting of a gluon into a quark–antiquark pair ($g \to q\overline{q}$) or into another pair of gluons ($g \to gg$). If this additional radiation happens through the initial–state partons it is referred to as initial–state radiation (ISR), as opposed to final–state radiation (FSR) describing the parton shower for the final–state particles. The parton shower proceeds by considering for each of the partons in the event the probability that it undergoes a branching into two daughter particles. These probabilities are again dictated through the DGLAP formalism and depend on the energy scale of the parton under consideration. The branching takes into account conservation of four–momentum, resulting in a shower of particles with decreasing energy. As long as the energy is large enough, the perturbative description of QCD is valid and possible infrared and collinear divergences are taken care of using so–called Sudakov form factors. The parton shower continues until a fixed energy scale $\Lambda_{\mathrm{QCD}}$ is reached, at which point it is believed that the perturbative description of QCD is not valid anymore. Instead hadronization effects need to be taken into account, which will be discussed in Sec. 3.1.3. This scale is usually taken around 300 MeV [174].
Examples of frequently used parton shower simulators are Pythia [167] and Herwig++ [175]. Even though these simulators also provide ME calculations, their PS functionality is often interfaced with other, more precise ME generators at higher orders.

Finally it is worth mentioning that for final–states with multiple hadronic partons there often exists an ambiguity in the combination of the ME generation and the PS evolution [176]. The final jet multiplicity of each event is determined by the number of partons that result from the combined ME+PS treatment. A given N–jet final–state can be achieved either via a proper N parton description at ME level, with corresponding soft and collinear branching through the PS, or via a N$-$1 parton configuration at ME level, with a wide–angle, hard emission in the PS evolution that results in an extra reconstructed jet. These ambiguities are resolved using matching or merging schemes that need to properly decide which of the above scenarios provides the most accurate description of the kinematics. Most importantly in the effort of predicting differential cross sections, these schemes should avoid double counting of events. An example of such a merging scheme that will be used further along in Chapter 5 is the so–called FxFx procedure, as described in Ref. [177].

### 3.1.3 Hadronization and fragmentation

After the parton shower has reached a scale $\Lambda_{\mathrm{QCD}}$, the value of the strong coupling constant grows to values that do not any longer allow for a perturbative expansion with reliable predictions at fixed order. Instead, due to color confinement, the colored partons that result from the PS have to be combined into color–neutral states. This process is known as hadronization and the non–perturbative nature of QCD at these scales forces us to resort to phenomenological descriptions of these processes. Unfortunately, at this moment in time there exists no description of the hadronization based on first principles.
The first ideas concerning this topic date back from the 1970s (see Ref. [178]). Today, there exist two main types of phenomenological models that are frequently used, namely the Lund string model [179, 180] and the cluster models [181]. The basic

ideas behind each of these models is discussed briefly below, but a more comprehensive overview can be found in Ref. [150].

### Lund string model

The idea behind the string model for hadronization is based on the presence of a colored flux–tube spanned between the colored partons. This can be seen as a collection of field–lines (in analogy with electromagnetic field lines) that is present between the two partons, as illustrated in Fig. 3.3 on the left. As quarks move apart, the string tension gives rise to a linearly rising potential of the form $V(r) = \kappa\, r$, where $\kappa$ is the so–called string constant whose value has been inferred from hadron mass spectroscopy measurements to be around 1 GeV/fm. As the quarks move apart far enough, the potential reaches a value at which it is possible for the string to break and create a new quark–antiquark pair, resulting in two new colour singlet systems with a given invariant mass. This is illustrated in Fig. 3.3 on the right. As long as the invariant mass of the new pairs is large enough, additional new breaks may occur. This continues until a set of color–neutral hadrons with on–shell masses is produced.

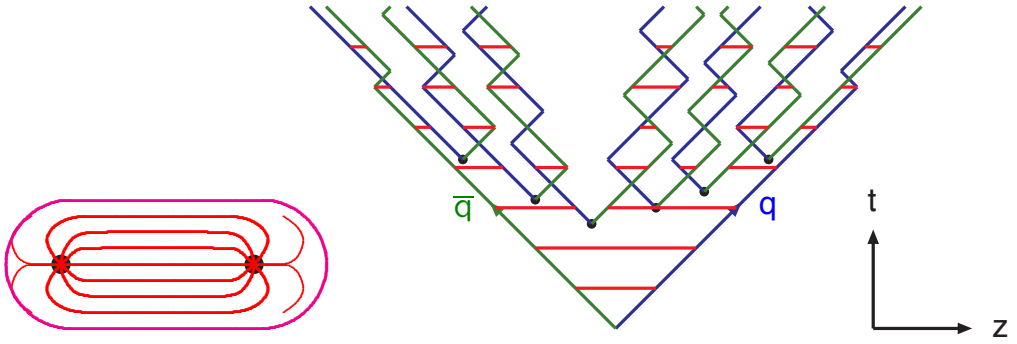


FIGURE 3.3: (left) Pictorial representation of a flux tube of strings between a quark–antiquark pair.
(right) Example of the evolution in time and longitudinal direction of the Lund string system. The breakup of the strings results in the formation of color–neutral hadrons. Figures taken from Ref. [150].

### Cluster model

At the basis of the cluster model lies the concept of preconfinement of QCD [182]. When tracing the color–flow in the parton shower, as illustrated in Fig. 3.4, one ends up with adjacent color–neutral combinations. One then forces the gluons to split into quark–antiquark pairs and recombines adjacent color–anticolor states to form color–neutral clusters. These clusters typically have intermediate mass scales (up to a few GeV) and are interpreted as "excited hadrons" that can further decay into the final set of on–shell hadron states. On average the cluster models are known to give a slightly less accurate description of the observed hadron spectra, though with less free parameters in the model to fit the data compared to the string model.

In both these approaches, the production of HF hadrons requires a slightly adopted treatment due to their special properties (*i.e.* they can not be treated as approximately massless states). Especially relevant in the case of HF production is to model their decays into stable particles. These decays are parametrized in software packages such as EvtGen [183], which predict hadronic branching fractions

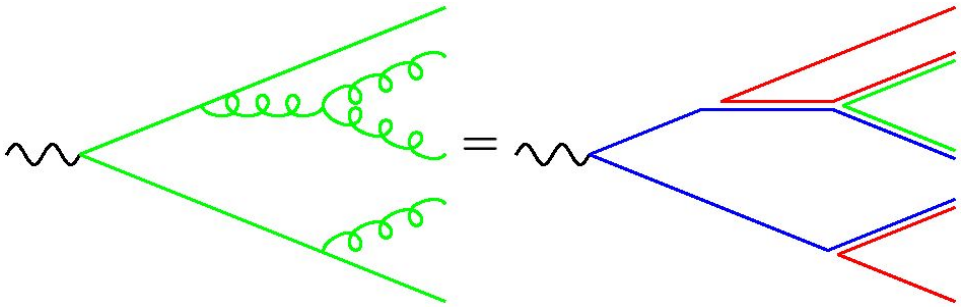

FIGURE 3.4: Schematic view of the color–flow in the parton shower as used in the cluster hadronization models. After tracing the color–flow as shown in the figure, the final gluons are forced to split into quark–antiquark pairs and adjacent states are combined into color–neutral clusters. Figure taken from Ref. [150].

and tune these predictions to observations in data (largely from LHCb results on $b$ hadron measurements).

#### Fragmentation functions

Finally also the predicted energy distributions of the final–state particles need to match the observations in data. Fragmentation functions describe how the momentum of an initial parton is distributed among the final–state particles that result from the partons after the PS and hadronization [184]. These functions are also (partially) described by non–perturbative models of which the parameters are tuned to data. Again a special treatment is needed to describe the momentum distribution of bottom and charmed hadrons. The momentum transfer from the initial heavy quarks to the resulting HF hadrons is expected to be larger compared to LF hadrons [185]. A variety of phenomenological models exist to parametrize the fragmentation of HF hadrons [186–191]. These models exhibit a set of parameters that are fitted to measurements of D and B meson production (see Ref. [40], Chapter 19.9 for an overview).

### 3.1.4 The underlying event

The concept of pileup has already been introduced before as the appearance of multiple proton interactions in the same (or adjacent) bunch–crossings. This typically results in low–energy activity which is spatially separated from the $pp$ interaction that is responsible for the hard scattering process. Similarly, additional event activity is expected from within the hard $pp$ interaction, caused by the remaining proton remnants consisting of partons that did not take part in the hard scattering process directly. This results in additional low–energy activity in the collision events and is known as the underlying event (UE). It is crucial to properly model the UE as well, given that it largely affects for example the observed charged hadron multiplicities measured by CMS and ATLAS [192, 193].
The origin of the UE is found in one of the following phenomena:

- **Proton remnants:** As mentioned before, the part of the proton that did not take part in the hard scattering is itself a colored state that will undergo a parton shower and hadronization process.
- **Multi–parton interactions (MPI):** There exists also the possibility that multiple partons inside the colliding protons undergo a scattering process. Even though the chance of having two or more hard scattering vertices within the

same *pp* collision is small, it happens frequently that the hard scattering is accompanied by at least one soft scattering process (see also Fig. 3.1). Since the $2 \to 2$ scattering cross section diverges at low transverse momenta, a perturbative description of these low–energy processes is not accurate and regularization procedures are needed to control the predicted cross section of the MPI.

- **Color reconnections:** The partons in the proton remnant are not independent from those participating in the hard scattering as there exist also color connections between them. In the hadronization models it turns out that these color connections can cause interference effects between the hard scattering and the UE, which need to be taken into account by phenomenological color–reconnection models.

It is clear that the additional event activity is not simply described by another hard scattering process. Instead simulations of the UE depend on a large number of parameters that are tuned to data (see for example Ref. [194]) as well as on the choice of the PDF set. In the simulations of 2017 *pp* collisions used by the CMS collaboration, the TuneCP5 underlying event tune is used, as described in Ref. [195].

### 3.1.5 Detector simulation

At the end of the simulation chain, a detailed description of the particle detector is needed that is able to describe how the stable final–state particles will interact with the different layers of the detector material. A full simulation of the CMS detector has been integrated in the GEANT4 [196–198] toolkit. This includes a description of each single detector module in each of the detector layers which have been described in Sec. 2.2. The detector simulation does not only include the interactions of the outgoing particles with the active detector layers, but also models the dead zones for example occupied by cables and wiring or support structures. Additionally, the precise configuration of the magnetic field inside different parts of the detectors needs to be correctly modeled. Finally, also pileup interactions are modeled and added at this stage as they may interfere with the signals from the hard scattering processes as the particles pass through the detector.

It is not surprising that the simulation of such a complex detector system requires large computational time and power, since at many points the software needs to simulate the stochastic way a given particle will respond to the environment of the detector it passes through. Given that this is often the most time consuming step in the simulation chain, there exists the possibility to use a faster, though less accurate simulation of the detector, labelled as the *fastsim* simulation. The *fastsim* implementation, as opposed to the *fullsim* simulation, uses simplified parameterizations of the reconstruction efficiencies for several physics objects to avoid a full simulation of all the interactions of the particles with the detector layers.

## 3.2 Reconstruction of physics objects

For any collision event, whether it is simulated or from actual data, the signals which are collected from the detector need to be processed and translated into a set of final physics objects which can be used for analyses. A first step in this process consists of reconstructing tracks and clusters from the hits that are recorded in the tracker modules and calorimeter cells respectively (see Sec. 3.2.1). Also the hits in the outer muon systems are combined to form tracks of muon candidates. These basic objects

then serve as input to the so–called Particle Flow (PF) algorithm which is adopted by the CMS collaboration [199]. As will be discussed in Sec. 3.2.2, this algorithm combines all available information from the different sub–detectors to output a set of physics objects suitable for analyses. The remainder of this section is devoted to a detailed overview of the reconstruction of electrons, muons, jets and missing energy, which are crucial ingredients in the analysis presented in Chapter 5.

### 3.2.1 Reconstruction of tracks and calorimeter clusters

As a first step, a set of basic detector–level objects needs to be defined as an input to the PF reconstruction. This consists of charged particle tracks, calorimeter clusters and standalone muon tracks. Each of these will be discussed below in more detail.

#### Tracking

The silicon detectors employed in the tracker of the CMS detector are composed of many small detector cells to achieve a good granularity. As charged particles fly through these innermost detection layers, they will interact with the silicon pixels and strips and generate *hits* in the detector modules. The main objective of an efficient tracking algorithm is to correctly combine these hits to form tracks that describe a helix as they move through the magnetic field. Not only should such an algorithm correctly reconstruct the path of the particles that fly through, it should do so while trying to avoid the accidental reconstruction of fake tracks that do not belong to any real particle. To this end, CMS has adopted an iterative fitting procedure [200, 201] based on four steps:

**Step 1. Seed generation:** At first the innermost pixel layers are used to generate seeds for the tracking algorithm. The newly installed pixel detector allows to employ a four–hit seed finding algorithm. With an excellent position resolution of ∼10–20 µm, the pixel detector is perfectly suitable for this task. A first estimation of the helix parameters can be made from a combination of three or four pixel hits and can then be passed on to the next tracking stage.

**Step 2. Trajectory building:** The next stage, also referred to as the "*pattern finding*", takes as input the seeds from the pixel detector and tries to connect these to hits in the silicon strip layers. The initial helix parameters are used to extrapolate the track onto the next layer. If another hit is found within the extrapolated uncertainty, it is added to the track and through a Kalman Filtering procedure [202], the helix parameters are updated and the track is extrapolated to the next layer. In case no hit is found, the algorithm is still allowed to continue and a *fake* or *invalid* hit is assigned to it. It is also possible that multiple hits match the extrapolated track from the previous layer, in which case a new trajectory is created for each of them. Possible duplicates are removed in the next step.

**Step 3. Ambiguity resolution :** One seed may give rise to multiple tracks, or the same track may be reconstructed from multiple seeds. In order to avoid these ambiguities, a simple arbitration algorithm is deployed based on the fraction of shared hits among two tracks with respect to the minimum number of hits in any of the two tracks. If this shared fraction exceeds a threshold, only the track with the most hits is kept. If they both have an equal amount of hits, the one with the best fitted $\chi^2$ value is kept.

**Step 4. Final track fitting :** The final set of trajectories is then refitted to exploit all of the available hits assigned to it. This refitting procedure is repeated twice, once starting from the innermost hit outwards, and another time from the outermost hits inwards.

After a first iteration, a set of high–quality tracks has been reconstructed and the corresponding hits are removed. The remaining hits that have not been assigned to any track are then used in a new iteration with looser reconstruction criteria to recover some efficiency with a modest fake rate.

The performance of this tracking approach is illustrated in Fig. 3.5. These figures are based on simulated data and show the expected tracking efficiency (left) and fake rate (right) as a function of the track $p_T$. In general a tracking efficiency of 90% can be achieved while keeping the fake rate at the level of ∼5%. Very low–$p_T$ tracks show a much worse tracking performance due to the fact that they will spiral and stay confined within the tracking system, creating many hits and often not even reaching the outer tracking layers. Also at very large transverse momenta the performance degrades. This can be explained by the fact that such tracks are often created by very boosted objects which cause collimated streams of particles in narrow cones. This results in a high density environment that makes the tracking very challenging. Additionally such high–energy tracks have a small curvature, making it harder to measure precisely their momentum. The figure also shows a comparison between the performance with the old pixel detector geometry (2016, blue) and the new four–pixel layer geometry (2017, red). A higher efficiency and lower fake rate is observed with the new pixel detector over the entire range of track transverse momenta.

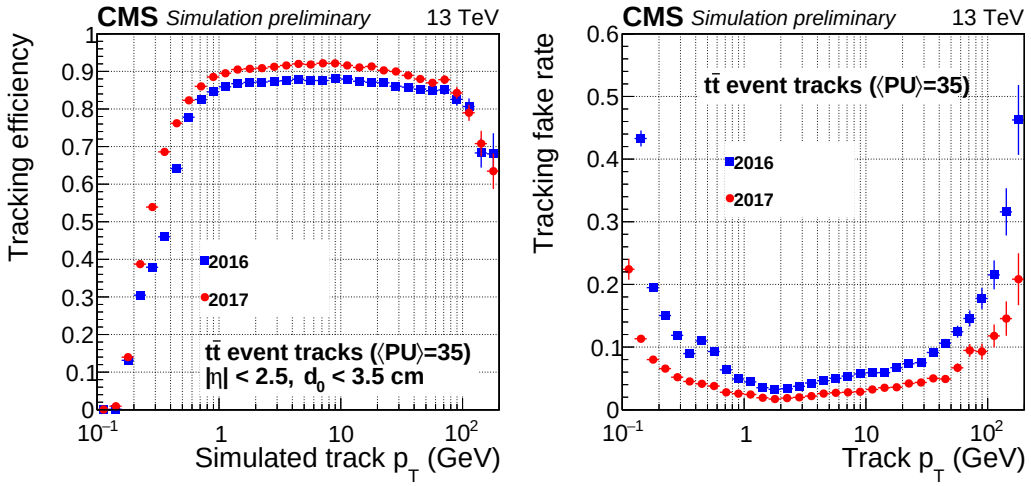


FIGURE 3.5: Simulated tracking efficiency (left) and fake rate (right) as a function of the track $p_T$, assuming an average pileup multiplicity of 35. A comparison is shown between the performance with the old pixel detector geometry (2016, blue) and the new four–layer pixel geometry (2017, red). Figures taken from Ref. [203].

## Primary vertex reconstruction

The reconstruction of primary interaction vertices, both from the hard scatter and from pileup vertices, proceeds through a fitting algorithm known as the *Adaptive Vertex Filter* (AVF) [204,205]. Starting from seed tracks, a clustering algorithms groups

together tracks based on their distance, or more precisely their impact parameter with respect to the fitted vertex position. Tracks which show a large displacement are weighed down by this procedure. The main primary interaction vertex from the hard scatter process is taken to be the one with the largest value of summed $p_T^2$ of the physics objects connected to the vertex. The physics objects in this definition are the clustered jets (see Sec. 3.2.5) with the tracks assigned to the vertex, as well as the missing transverse energy assigned to it (see Sec. 3.2.6).

### Calorimetry

As outlined in Sec. 2.2.4 and 2.2.5, both the ECAL and the HCAL can be seen as a pixelated grid of detector modules or *cells*. A dedicated clustering algorithm has been developed to cluster together calorimeter cells in the calorimeter systems to be used later on in the PF reconstruction (see Ref. [199], Sec. 3.4 for a comprehensive explanation). Due to the different granularity, the clustering is performed separately in the ECAL and the HCAL. First, cluster seeds are identified as calorimeter cells with energy deposits above a certain threshold. Progressively, neighboring cells (sharing a side or corner with the seeds) are added to the cluster if their energy exceeds another (typically lower) threshold. At any stage, the thresholds are chosen above twice the level of electronic noise and specific values can be found in Ref. [199]. The collection of selected neighboring cells is called a *topological cluster*. Given that multiple seeding cells may end up in the same topological cluster, an algorithm is responsible for sharing the energy deposits among the seeds to end up with a collection of exclusive calorimeter clusters. The collection of final HCAL and ECAL clusters serves as an input to the PF algorithm.

### Standalone muon reconstruction

Muons can be reconstructed through their signatures in the outer muon chambers of the CMS detector. However, they will of course also leave their trace in the inner tracker detectors. CMS therefore defines three types of muons [206]:

- **Standalone muons:** By using solely the information from the muon detector systems, muon tracks can be reconstructed with similar tracking techniques as explained in the beginning of this section. Hits from the DT and CSC systems are combined into seeds and are fitter with a Kalman Filter approach including also hits from the RPC detectors.
- **Tracker muons:** An "inside–out" approach is used by extrapolating reconstructed tracks in the tracking system all the way up to the muon systems. If the extrapolated tracks are matched with a DT or CSC segment, it is labelled a tracker–muon–track.
- **Global muons:** As opposed to the tracker muons, the global muons constitute an "outside–in" approach by matching extrapolated standalone muons to tracks in the tracker.

Many muons are reconstructed as both tracker and global muons. If two muons share the same track in the inner tracker, they are merged into one object. This way one ends up with a set of high–quality global/tracker muons and a set of standalone muons that typically have a worse momentum resolution. These collections are served as input to the final PF algorithm.

### 3.2.2 The particle flow algorithm

The Particle Flow algorithm [199] combines the information obtained by the different detector sub–systems in order to reconstruct a collection of physics objects to be used for physics analyses. These physics objects can either be muons, electrons, photons, charged hadrons or neutral hadrons. Each of these objects has a distinct signature when considering the entire detector. This is illustrated in Fig. 3.6. Muons leave clear tracks in the tracker and the muon system, but have only very small energy deposits in the calorimeters. Electrons are identified by a track in the inner tracker which can be matched to an energy cluster in the ECAL, but almost no energy deposits in the HCAL. Photons also deposit almost all of their energy in the ECAL systems, but are electrically neutral and therefore do not leave a track in the tracker. Charged hadrons create tracks in the tracker and deposit the majority of their energy in the HCAL systems. Finally neutral hadrons can only be identified through their energy deposits (mainly) in the HCAL.

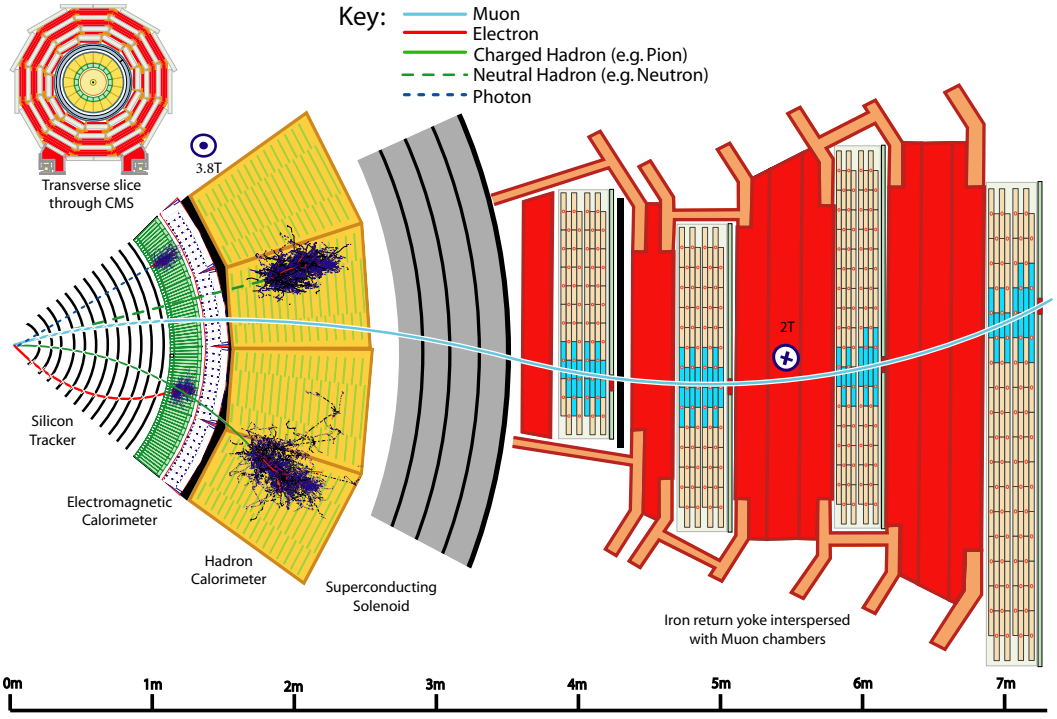


FIGURE 3.6: Schematic representation of a vertical slice through the CMS detector, including the signatures of different physics objects as they make their way through the detector. Figure taken from Ref. [199].

The PF algorithm exploits these distinct signatures throughout the entire detector. First a linking algorithm is deployed to find matching tracks, clusters and muon tracks and combines them into so–called *blocks*. Tracks are extrapolated from the tracker to the calorimeter (ECAL or HCAL) and associated to a cluster if the extrapolated position lies within the cluster boundaries. Clusters in the electromagnetic calorimeter can be associated to overlapping clusters in the hadronic calorimeter. Finally, global muons are constructed from extrapolations of the tracks in the tracker and muon systems as explained in the previous section. In the final step the PF algorithm tries to further refine and interpret these associations in order to identify the physics objects by their distinct detector signatures. The

assignment of physics objects starts from the most straightforward identification, namely that of muons. The global muon definition outlined in the previous section already constitutes a high–quality link between the tracker and the muon systems and allows for an efficient identification of muons. The muons are added to the collection of PF objects and their blocks are removed. Next, matching blocks of tracks and ECAL clusters are identified as electrons and also removed. The PF algorithm is now left to identify photons, neutral and charged hadrons. The energy in the calorimeter clusters is compared to the energy of the associated tracks to that cluster. If the calorimeter energy matches the energy of the tracks, only charged hadrons are identified. If the calorimeter energy is larger than that of the tracks, the charged hadrons are identified by the tracks and their energy is removed. The remaining energy is assigned either to photons (especially when the majority of the energy deposit happened in the ECAL) or eventually to neutral hadrons. It may also happen that the energy of the tracks exceeds the associated calorimeter energy, in which case tracks are removed by looking further for extra muons (with loosened reconstruction criteria) or by eliminating tracks with large uncertainties.

The next sections discuss in more detail the identification requirements of high–quality electron (Sec. 3.2.3) and muon (Sec. 3.2.4) candidates and the jet–clustering algorithms that combine PF objects into collimated particle showers known as *jets* (Sec. 3.2.5). Finally any remaining momentum imbalance in the transverse plane is interpreted as so–called missing energy or momentum, as will be discussed in Sec. 3.2.6.

### 3.2.3 Electron reconstruction

The PF algorithm identifies electrons by looking for blocks that link tracks to large deposits in the ECAL. Therefore it seems a priori straightforward to determine the energy and momentum of electrons from their track parameters and ECAL cluster deposits. However, the difficulty arises from the fact that electrons can radiate significant amounts of their energy in the form of bremsstrahlung photons. This leads to distorted tracks and raises the need to properly recombine the ECAL clusters from the radiated photons with the electrons in order to get a correct estimate of the original electron energy. Two methods are combined within the PF algorithm to deal with electrons in different environments.

1. **ECAL-based approach:** For electrons which are well isolated from other activity in the event, the original strategy was based mostly on the use of the ECAL clusters. Within a certain window around the electron cluster, the energy deposits originating from radiated bremsstrahlung photons are combined with the electron cluster to form a *supercluster*. From the position of this supercluster, the associated inner pixel tracker hits could then be inferred.

2. **Tracker-based approach:** The above strategy clearly fails for electrons which are produced in jets, given the increased ambiguity in correctly identifying the energy deposits and inner tracker hits corresponding solely to the electron. Similarly, electrons with low momenta are highly bent within the tracker volume and consequently radiate their bremsstrahlung photons over a very large area, making it harder to combine them in a supercluster. To deal with these specific cases, a tracker–based approach has been developed. As long as the bremsstrahlung radiation is limited, the iterative tracking procedure outlined in Sec. 3.2.1 is well suited to reconstruct electron tracks with a high efficiency. However, when a significant amount of energy is radiated, this approach will

result in tracks with a worse $\chi^2$ value[4]. Therefore a selection of tracks with worse $\chi^2$ and on average less pixel hits are refitted with a Gaussian Sum Filter (GSF) [207] instead of the usual Kalman Filter. The GSF approach is more adapted to the energy losses along the electron trajectory and therefore results in a large efficiency recovery for electrons when combining the ECAL–based and the tracking–based approach.

The benefit from combining these two complementary methods is illustrated in Fig 3.7, where the left figure illustrates the clear gain in reconstruction efficiency when going from the ECAL–based approach (red) to a combination of the ECAL– and tracker–based approach (blue), whereas the right figure shows that an electron identification efficiency of more than 95% can be achieved in the 2017 data–taking period for electrons with $p_T > 20$ GeV. These efficiencies are measured using $Z \to e^+e^-$ events.

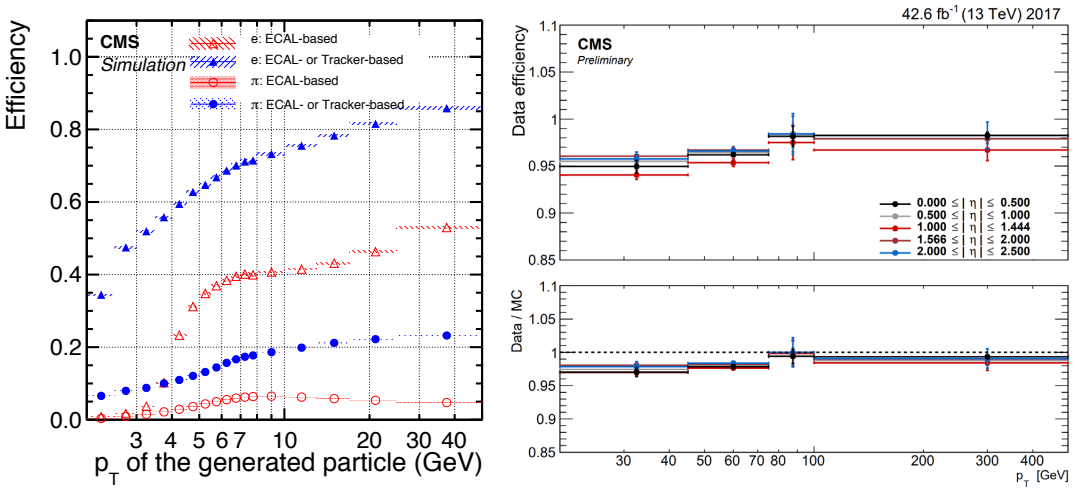


FIGURE 3.7: (left) Simulated seeding efficiency as a function of generator–level $p_T$ for electrons (triangles) and pions (circles), showing the clear gain going from the ECAL–based approach (red) to a combination of the ECAL– and tracker–based approach (blue). Figure taken from Ref. [199].
(right) Electron identification efficiency as a function of electron $p_T$ measured in 2017 data on the top, and the data–to–simulation correction factors on the bottom. Figure taken from Ref. [208].

As the electrons are reconstructed, a set of additional identification requirements (IDs) are typically applied at analysis level to extract a set of high–quality electrons suitable for the analysis. CMS supports a set of predefined electron IDs which are either based on a simple cut–based approach or on a more involved multivariate analysis (MVA) output. In the analysis which will be discussed in Chapter 5, a medium cut–based electron ID is used. The details of the specific quality requirements can be found in Refs. [209, 210] and depend on whether the electron supercluster is located in the barrel (EB, $|\eta| \leq 1.479$) or the endcap (EE, $1.479 < |\eta| < 2.5$). However, the analysis will employ a custom requirement on the electron isolation. The relative isolation of an electron ($\mathcal{I}^e_{\text{rel}}$) is defined as the sum of the $p_T$ of all PF objects that lie within a cone of $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} < 0.3$ around the electron,

[4]The radiated energy will distort the helix shape on which the track parametrization is based.

relative to the electron transverse momentum $p_T^e$. This is defined in Eq. (3.2), where the sum runs over all charged hadrons (CH), neutral hadrons (NH) and photons ($\gamma$) respectively. The neutral component is corrected for possible pileup contributions[5]. This pileup subtraction is based on the mean $p_T$ density due to pileup ($\rho$) and the effective area ($A$) of the isolation cone which denotes the expected amount of neutral pileup per $\rho$ from simulations as a function of the electron pseudorapidity [211, 212]. The values for the effective area $A$ (as a function of the pseudorapidity) can be found in Ref. [213]. The specific value of the upper thresholds on $\mathcal{I}^e_{\text{rel}}$ applied in the analysis in Chapter 5 depends on the absolute value of the pseudorapidity of the electron, as expressed in Eq. (3.2). The distinction between the different $\eta$ regions is based on the transition from the barrel to the endcap ECAL.

$$\begin{aligned}\mathcal{I}^e_{\text{rel}} &\equiv \frac{\sum\limits_{\text{CH}} p_T + \max\left(0, \sum\limits_{\text{NH}} p_T + \sum\limits_{\gamma} p_T - [A \times \rho]\right)}{p_T^e} \\ &< \begin{cases} 0.077 \text{ if } |\eta^e| < 1.48, \\ 0.068 \text{ if } |\eta^e| \geq 1.48, \end{cases}\end{aligned} \tag{3.2}$$

Finally, dedicated correction factors are derived on the identification and isolation efficiencies to correct for the discrepancies observed in simulations with respect to the data. These correction factors are typically derived as a function of $p_T$ and $\eta$. An example is shown on the bottom panel of the right plot in Fig. 3.7.

### 3.2.4 Muon reconstruction

The muon reconstruction has been described already in Sec. 3.2.1. In the analysis which will be discussed in Chapter 5, a tight muon ID [141, 214] will be used that is aimed at reconstructing prompt and well isolated muons from W and Z boson decays. Muons can however also be created in (semi)leptonic decays of hadrons in which case they are surrounded by additional hadronic activity due to jet formation of the hadrons. These muons are typically lower in $p_T$ and are therefore referred to as *soft muons*. The reconstruction and identification of these soft muons (and similarly for soft electrons) will play an important role in the identification of heavy–flavor jets as will be discussed later on in Sec. 3.3. For this purpose, the CMS collaboration also supports a *soft muon ID* which is efficient in identifying soft muons, but suffers from a larger misidentification rate due to hadrons which are misidentified as soft muons.

In the selection of prompt muons, an isolation criterium ($\mathcal{I}^\mu_{\text{rel}}$) is also applied as expressed in Eq. (3.3), where this time the sums run over all PF objects within a cone of $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} < 0.4$ around the muon. The neutral pileup correction is treated differently with respect to the electron isolation. The energy deposits within the isolation cone from charged hadrons originating from pileup vertices ($\sum_{\text{PU}} p_T$) are summed together and scaled by a factor 0.5[6]. This quantity is then subtracted

[5] The charged pileup is already subtracted from the considered PF collection through the impact parameter information of the corresponding tracks with respect to the primary interaction vertex. This procedure is known as charged hadron subtraction (CHS).

[6] The factor 0.5 is estimated from simulations to be the average ratio of neutral to charged particle production in inelastic proton–proton collisions.

from the neutral hadron and photon sums.

$$\mathcal{I}^{\mu}_{\text{rel}} \equiv \frac{\sum\limits_{\text{CH}} p_T + \max\left(0, \sum\limits_{\text{NH}} p_T + \sum\limits_{\gamma} p_T - 0.5 \times \sum\limits_{\text{PU}} p_T\right)}{p^{\mu}_T} < 0.15, \tag{3.3}$$

The efficiency of the tight muon identification in 2017 is shown in Fig. 3.8 as a function of muon $p_T$ and $\eta$. The drops in efficiency around $|\eta| \sim 0.2$ are due to gaps between the wheels in the muon detectors. These efficiencies are measured using a tag–and–probe method in a sample of $Z \to \mu^+\mu^-$ events [206]. Also for muons, dedicated correction factors are derived on the identification and isolation efficiencies to correct for the discrepancies observed in simulations with respect to the data. The bottom panels in Fig. 3.8 show an example of the scale factors for the Tight muon ID as a function of $p_T$ and $\eta$.

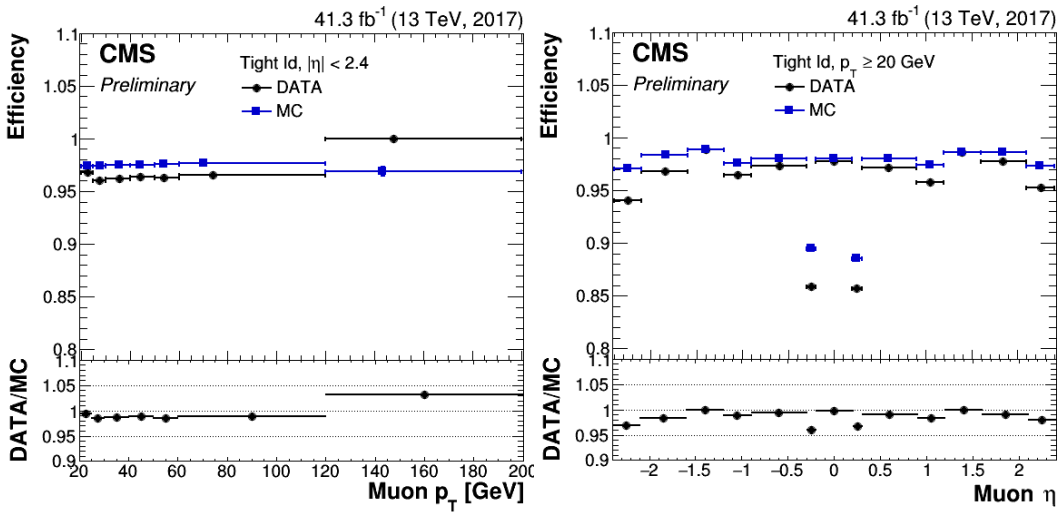


FIGURE 3.8: Muon identification efficiency in data (black) and simulation (blue) according to the tight muon ID as a function muon $p_T$ (left) and $\eta$ (right). The bottom panel shows the data–to–simulation correction factors. The drops in efficiency around $|\eta| \sim 0.2$ are due to gaps between the wheels in the muon detectors. Figures taken from Ref. [215].

### 3.2.5 Jet reconstruction

As already introduced in Sec. 3.1.3, the quarks that are produced in the $pp$ collisions will undergo a process of hadronization during which they will form color–neutral states. These hadrons will move through the detector and may radiate or decay along their trajectory. A single quark produced in the hard scattering process will therefore give rise to a collimated spray of particles, referred to as a *jet*. It is clear from this picture that dedicated jet clustering algorithms are needed to combine again all outgoing particles that resulted from the original quark (or gluon). Such algorithms need to be infrared and collinear safe, meaning that additional soft radiations or collinear splittings of the original partons should ideally not affect the outcome of the clustering algorithm. Below, the anti–k$_T$ algorithm will be discussed which is used in the CMS reconstruction and does indeed comply with the above mentioned criteria. The energy estimate of a clustered jet is prone to several sources of uncertainties which need to be corrected for as will be discussed in the following paragraphs. A

special section (3.3) has been dedicated to the treatment of jets originating from HF quarks (i.e. bottom and charm quarks) given their importance in the rest of this thesis. Such HF jets leave distinct signatures in the detector that are exploited by Machine Learning based classification algorithms to distinguish them from light jets.

### Jet clustering

There exist a variety of jet clustering algorithms (see for example Refs. [216, 217] and references therein) with different properties and different required computing times, but the CMS collaboration has opted to use the so–called *anti–$k_T$* algorithm [218]. This is an example of a *sequential recombination algorithm* and is based on a principal of minimal distances between particles. The algorithm starts by defining two types of distances as expressed in equation (3.4).

$$d_{ij} = \min\left(k_{T,i}^{2p}, k_{T,j}^{2p}\right)\frac{\Delta_{ij}^2}{R^2}, \qquad d_{iB} = k_{T,i}^{2p}, \tag{3.4}$$
$$\text{with} \quad \Delta_{ij} = \sqrt{\Delta y_{ij}^2 + \Delta\phi_{ij}^2}.$$

In this notation, $\mathrm{k}_{T,i}$ represents the transverse momentum of particle $i$. $R$ is referred to as the radius parameter of the algorithm that is a measure for the average angular size of a jet and is taken to be 0.4 in CMS for the default jet definition. Due to this choice, the jets are often referred to as AK4 jets, whereas for specific analyses that focus on very boosted objects sometimes a second jet clustering algorithm is used with $R = 0.8$ (resulting in AK8 jets). These wider AK8 jets aim to collect all decay products from the boosted object. Finally, the parameter $p$ regulates the power of the transverse momenta relative to the geometrical separation $\Delta_{ij}$ and is chosen to be $-1$ in the anti-$\mathrm{k}_T$ algorithm.

The algorithm proceeds by comparing for each pair of particles $(i, j)$ the minimum of $d_{iB}$ and $d_{jB}$ to the value of $d_{ij}$. If the distance defined by $d_{ij}$ is smaller than both $d_{iB}$ and $d_{jB}$, the two particles are clustered and combined into one object, referred to as a *pseudo–jet*. This pseudo–jet replaces the two individual particles in the collection and the algorithm can iteratively proceed. If however $d_{ij}$ is not smaller than the other distance measures, the clustering stops and the pseudo–jet is considered a full reconstructed jet. CMS uses the PF algorithm to construct the jet collections, meaning that the objects that are considered for the anti–$\mathrm{k}_T$ clustering are in fact the PF candidates.

The choice of $p = -1$ for the anti-$\mathrm{k}_T$ algorithm ensures that on average soft particles are likely to be clustered around neighboring hard particles, since the smallest values of $d_{ij}$ are obtained by including the hardest particles. It also ensures infrared and collinear safety, in contrast to for example the $\mathrm{k}_T$ algorithm (in which $p$ is taken to be $+1$) which is not collinear safe [219].

### Jet energy scale corrections (JES)

After the jet clustering, a difference in four–momentum is observed between the reconstructed jet and the generated parton. This discrepancy may be due to the presence of pileup particles which are clustered inside the jet, due to (low–energy) particles which were not properly reconstructed or included in the jet, or due to the non–linear

response of the detector itself. In order to correct the energy scale of the reconstructed jets both in data and simulations, CMS follows a factorized approach [220], where at each step the jet four–momentum is corrected for a certain effect and then fed as input to the next correction stage. This procedure is schematically illustrated in Fig. 3.9.

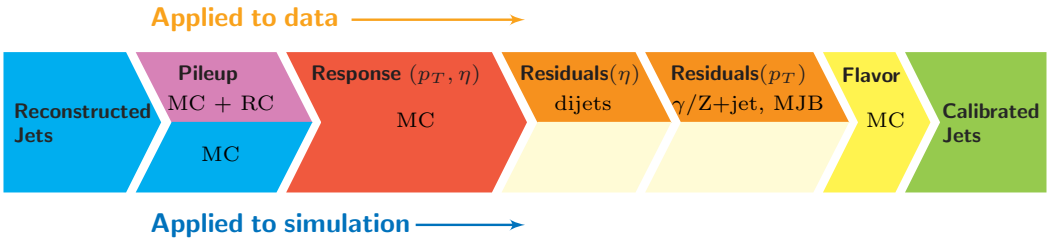


FIGURE 3.9: Different stages in the calculation of the JES correction applied to data (upper row), to simulation (lower row) or both. The final flavor–dependent corrections are optional and are not applied throughout the rest of this thesis. Figure taken from Ref. [220].

The sequential steps in the application of the jet energy scale (JES) corrections are detailed below:

**Pileup:** Additional particles from pileup vertices may end up being clustered inside the jet, resulting in an offset of the jet energy. These offsets are estimated from simulations. Three sources of pileup are considered and corrected for with different treatments. Out–of–time pileup corrections are calculated by varying the bunch spacing and time–integration window used in the calorimeter systems. In–time pileup due to charged particles is subtracted using charged hadron subtraction methods that identify charged hadron PF objects that are associated to pileup vertices. In–time pileup due to neutral particles is estimated using a hybrid jet area method [211, 212, 221], based on the methodology explained for the electron isolation pileup subtraction in Sec. 3.2.3.

**Detector: response** After the pileup offset has been mitigated, there still exist differences between the reconstructed jet $p_T$, and the $p_T$ of the jets which are clustered from generator–level particles (gen–jets). These differences are either caused by hadrons which are not properly reconstructed, or by the non–linear detector response. By using simulated multijet samples, average correction factors are derived which are binned in $p_T$ and $\eta$ to correct the reconstructed jet $p_T$ such that it on average matches the generator–level jet $p_T$.

**Residuals:** Small residual differences between the JES in data and simulation are still observed after the pileup offset and detector response have been taken into account. To this end, residual corrections are applied to data only, by using data–driven measurements. A first residual correction is aimed at measuring the $\eta$–dependent corrections in dijet events. The first jet is required in the central region of the detector ($|\eta| < 1.3$) whereas the second jet has arbitrary $\eta$. By exploiting the transverse momentum balance of the event, the jet response is inferred by comparing the transverse momenta of these two jets. Then Drell–Yan events with

additional jets are used in which leptonically decaying Z bosons and photons are recoiled against one or more jets. Again the transverse momentum balance of the event is used to infer the proper data–to–simulation corrections. Due to the good momentum resolution of the leptons (especially the muons), this method focuses on the $p_T$–dependent residual corrections.

#### Jet energy resolution corrections (JER)

Differences between data and simulations are also observed in the jet energy resolution (JER). To account for a worse resolution observed in data, the simulated jet momentum is artificially smeared. This is done by scaling the jet four–momenta by a smearing factor $c_{\mathrm{JER}}$. The prescribed methodology depends on whether or not a generator–level jet can be matched to the reconstructed jet:

1. If a reconstructed jet can be properly matched to a generator–level jet, $c_{\mathrm{JER}}$ is calculated according to Eq. 3.5, where $p_T^{\mathrm{reco}}$ is the reconstructed jet $p_T$, $p_T^{\mathrm{gen}}$ is the generator–level jet $p_T$ and $s_{\mathrm{JER}}$ is a data–to–simulation scale factor on the measured jet energy resolution [222]. The matching to the generator–level jet is done by requiring $\Delta\mathrm{R}(\mathrm{reco},\mathrm{gen}) < R/2$ and $|p_T^{\mathrm{reco}} - p_T^{\mathrm{gen}}| < 3\sigma_{\mathrm{JER}}\, p_T^{\mathrm{reco}}$, where $R$ is the radius parameter of the jet clustering algorithm (0.4 for AK4 jets) and $\sigma_{\mathrm{JER}}$ is the relative $p_T$ resolution measured from simulations.

2. If no generator–level match can be found, a stochastic smearing can still be applied in which $c_{\mathrm{JER}}$ is calculated according to Eq. 3.6. In this prescription, $\mathcal{N}(0,\sigma_{\mathrm{JER}})$ is a random number sampled from a normal distribution with mean 0 and variance $\sigma^2_{\mathrm{JER}}$.

$$\texttt{matched gen-jet}\text{: } c_{\mathrm{JER}} = 1 + (s_{\mathrm{JER}} - 1) \times \frac{p_T^{\mathrm{reco}} - p_T^{\mathrm{gen}}}{p_T^{\mathrm{reco}}}, \tag{3.5}$$

$$\texttt{no matched gen-jet}\text{: } c_{\mathrm{JER}} = 1 + \mathcal{N}(0,\sigma_{\mathrm{JER}}) \times \sqrt{\max(0, s^2_{\mathrm{JER}} - 1)}. \tag{3.6}$$

### 3.2.6 Missing energy reconstruction

In the initial collisions between the two protons there is zero net momentum in the transverse plane. Conservation of momentum dictates that the vectorial sum of all the outgoing particles should also not have a transverse component. However, not all outgoing particles are detected. A clear example of an undetected particle is the neutrino, that is not expected to interact with the detector material (or only very rarely). Additionally, low–energy particles may not be reconstructed and the reconstruction of the observed particles has a limited precision. All of these effects may enter into a non–zero transverse component of the summed four momentum and is referred to as missing transverse momentum ($\vec{\not{p}}_T$) with the corresponding missing transverse energy ($\not{E}_T$). The missing transverse momentum vector $\vec{\not{p}}_T$ is defined in Eq. (3.7) as the negative vectorial sum of the transverse momentum vectors of all PF candidates. The magnitude of this quantity is the corresponding missing transverse energy.

$$\vec{\not{p}}_T = -\sum_{i=1}^{N_{\mathrm{PF}}} \vec{p}_T^{\,(i)}, \qquad \not{E}_T = \left|\vec{\not{p}}_T\right|. \tag{3.7}$$

The missing energy is also corrected for pileup contributions and also for the effect of the JES and JER corrections discussed in Sec. 3.2.5. More details on the performance of the missing energy can be found in Ref. [223]. A visualization of an event with large $\vec{\not{p}}_T$ is shown in Fig. 3.10.

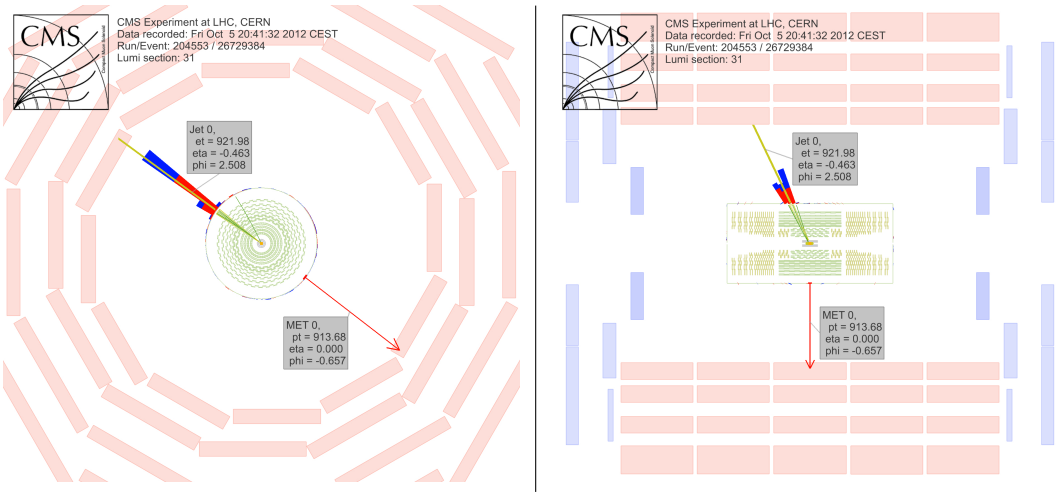


FIGURE 3.10: Event display of a monojet event in the $\rho$–$\phi$ plane on the left and the $\rho$–$z$ plane on the right. In this event a single high–energy jet ($E_T$ = 921.98 GeV) is observed in the very central region of the detector, resulting in a large missing transverse energy vector (red arrow). Figures taken from Ref. [224].

## 3.3 Identification of heavy-flavor jets

A large variety of analyses rely heavily on the identification of jets originating from $b$ and $c$ quarks, referred to as $b$–tagging and $c$–tagging respectively [225–227]. Below a non–exhaustive list of examples is given.

- Due to the fact that the top quark decays almost exclusively to a $b$ quark and a W boson, almost all top quark related analyses make use of these identification algorithms. This thesis is a perfect example, in which top quark pair production with additional HF jets is studied in Chapter 5.
- Many BSM searches rely on decays of heavy new resonances into pairs of $b$ or $c$ quarks.
- The H boson decays predominantly into a pair of $b$ quarks. Even though it is not the cleanest decay channel, it is the one with the largest branching fraction.
- Searches for supersymmetry at the LHC often rely on the production of scalar top, bottom or charm "squarks" that decay into their fermionic SM counterparts (i.e. the known $t$, $b$ and $c$ quarks), resulting in many HF jets.
- Measurements of electroweak boson (W or Z) production with associated HF jet activity are conducted in CMS and ATLAS.

The most recent publication on heavy–flavor tagging in CMS [225] was published in May 2018 and has already over 200 citations while writing this thesis,

whereas its predecessor [228] was cited 861 times[7]. Needless to say that HF identification is of crucial importance in the experimental collaborations operating at the LHC.

This section will cover the properties of HF jets and give an overview of the state–of–the–art $b$– and $c$–tagging algorithms currently used by CMS in the 2017 data–taking period. The calibration measurements will then be discussed, with a focus on the discriminator shape calibration which will be used further along in Chapter 5. The details on the underlying Machine Learning algorithms are described later on in Chapter 4.

### 3.3.1 Flavor definition in CMS

Before discussing the properties of HF jets and the algorithms to identify them, it is important to have a reliable definition of the jet flavor in simulated events. CMS has adopted a jet flavor definition based on the principle of *ghost matching* [211]. Instead of running the jet clustering algorithms using only the reconstructed PF candidates, one can add to the collection of particles also the generator–level $b$ and $c$ hadrons[8] and redo the jet clustering. In order not to affect the reconstructed jet momentum, these $b$ and $c$ hadrons have their momenta scaled down to a negligibly small number (i.e. *ghost hadrons*) such that only the information on their direction is considered in the clustering algorithms. The following jet flavors are then defined:

**$b$ jets:** If at least one $b$ ghost hadron is clustered inside the jet, the jet is labelled a $b$ jet.

**$c$ jets:** If no $b$ ghost hadron and at least one $c$ ghost hadron is clustered inside the jet, the jet is labelled a $c$ jet. Consequently, jets with a clustered $b$ hadron that decays into a $c$ hadron are considered to be $b$ jets.

**light jets:** If no $b$ or $c$ ghost hadrons are clustered inside the jet, it is labelled a light–flavor (LF) jet.

**pileup jets:** If however a jet has no matched generator–level jet (clustered purely from generator–level particles), it is automatically labelled a pileup jet, independent of whether a $b$ or $c$ ghost hadron is found. This is motivated by the fact that during the matrix element generation the pileup is not yet included in the simulation.

The results obtained in the analysis in Chapter 5 are defined using this jet flavor definition. Considering the effect of other flavor definitions would be very interesting. This would however require extensive investigations which are not yet conducted and are left for future studies.

### 3.3.2 Properties of heavy-flavor jets

Heavy–flavor jets originate from $b$ and $c$ hadrons[9]. These hadrons are unstable, though with a significantly long lifetime to travel an observable distance from the interaction point before they decay and produce secondary particles at their decay vertex. The typical lifetime of $b$ hadrons lies around 1.5 picoseconds ($1.5 \times 10^{-12}$ s),

[7] These numbers were extracted from the HEP Inspire database on February 8th, 2019.

[8] In this procedure, only generated $b$ and $c$ hadrons are considered that do themselves not have $b$ or $c$ hadrons as ancestors.

[9] Often the term "D hadrons" is also used to refer to hadrons containing charm quarks.

whereas for $c$ hadrons this average is typically an order of magnitude smaller [40]. Depending on the energy of the outgoing hadrons, these lifetimes give rise to a flight distance from the interaction point of a few mm up to the order of a cm. The decay products of these hadrons therefore do not point back directly to the primary interaction vertex (PV), given that their origin lies at the secondary decay vertex (SV). This results in reconstructed tracks which are displaced with respect to the PV. The impact parameter (IP) defines the distance of closest approach between a track and the PV and is therefore a good measure of this displacement. The average mass of $b$ hadrons lies around 5 GeV, whereas for $c$ hadrons this average lies around 2 GeV [40], still much higher than the typical mass of LF hadrons (such as pions and kaons) which have a typical mass of several 100 MeV. The reconstruction of such a secondary vertex with its corresponding invariant mass can therefore provide crucial information on the jet flavor. The increased (semi)leptonic branching fraction of $b$ and $c$ hadrons compared to light hadrons results in the presence of low–energy (soft) electrons or muons in about 20% (10%) of the $b$ ($c$) jets. Even though this may sound like a small number, it allows for a very pure selection of HF jets and this property can be exploited to distinguish them from LF jets. In Fig. 3.11, all of these properties are illustrated to visualize the distinction between a LF jet and a typical HF jet.

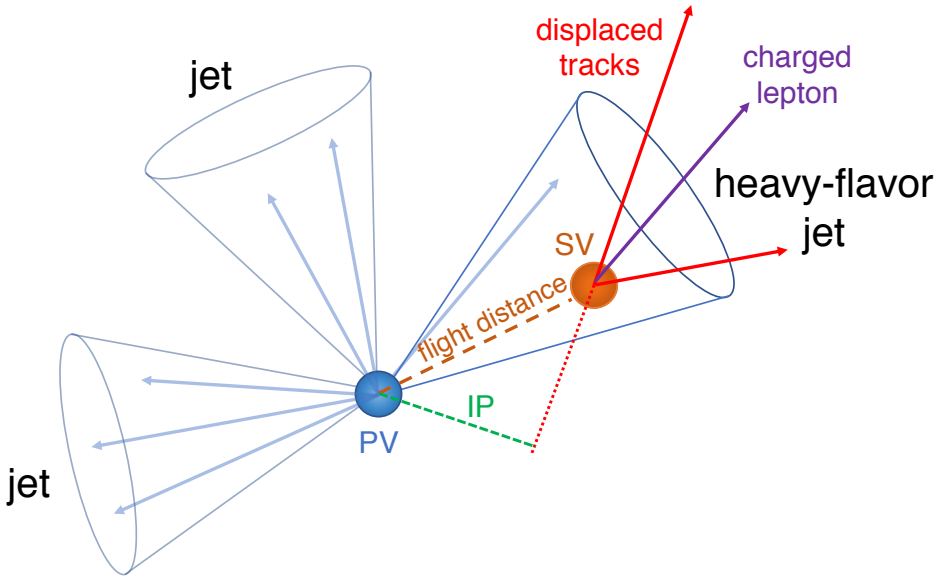


FIGURE 3.11: Illustration of the typical signature caused by a HF jet. The jet contains a secondary vertex (SV) from the decay of a $b$ or $c$ hadron from which a collection of displaced tracks arise, including possible soft leptons. These tracks consequently have a large impact parameter (IP) with respect to the primary vertex (PV). Figure taken from Ref. [225].

### Track preselection

The power of HF jet identification algorithms relies heavily on the tracking performance. When looking at the tracks that are clustered inside HF jets, a priori only a relatively small fraction originates from the $b$ or $c$ hadrons themselves. Fig. 3.12 on the left shows the average track multiplicity inside $b$ jets (top), $c$ jets (middle) and light jets (bottom) for tracks from different origins, as determined from simulations.

The category labelled as "$b$ ($c$) hadron" refers to tracks that have in their decay history a $b$ ($c$) hadron as an ancestor[10]. The category labelled as "*uds* hadron" refers to tracks that have no $b$ or $c$ hadrons, but only light hadrons as ancestors. Finally the labels "Pileup" and "Fake" refer respectively to tracks from pileup vertices or from mis–reconstructed tracks that do not have a true generated charged particle matched to them. From this figure it can be seen that originally the fraction of $b$ ($c$) hadron tracks in $b$ ($c$) jets accounts for only ∼30% (∼15%) of the total track content and a non–negligible fraction of pileup (∼40%) and fake (∼5%) tracks is present. To this end CMS has adopted a preselection on the tracks which are considered for the HF tagging algorithms[11]. Selected tracks must have $p_T > 1$ GeV and the normalized $\chi^2$ value of the fitted trajectory should be below 5, with at least one hit in one of the pixel layers. The track decay length, defined as the distance from the PV to the point of closest approach between the track and the jet axis, should be smaller than 5 cm to remove tracks from long–lived $K^0_S$ or $\Lambda$ hadrons which may fake HF hadron signatures. Tracks from pileup vertices are mostly discarded by requiring the transverse (longitudinal) impact parameter of the track to be smaller than 0.2 (17) cm and the distance between the track and the jet axis at their point of closest approach to be below 0.07 cm. The effect of this track preselection is shown in Fig. 3.12 on the right. It can be seen that by only considering the selected tracks, the fraction of pileup and fake tracks has been reduced to the level of 1–3% and an increased fraction of around 50% (30%) of the tracks in $b$ ($c$) jets come from the decay of the corresponding $b$ ($c$) hadrons. A relatively large fraction of the selected tracks in $b$ and $c$ jets however still originate from light hadrons and result for example from the underlying event or from radiations in the PS that happened before the hadronization.

In the following paragraphs a more detailed overview of the characterizing properties of HF jets will be given. This includes a detailed discussion on the properties (and reconstruction) of SVs, displaced tracks and soft leptons.

### Secondary vertex

The lifetime of $b$ and $c$ hadrons gives rise to displaced secondary vertices at flight distances of a few mm up to the order of a cm from the interaction point. This is still confined within the beam pipe, but the high granularity and precision of the tracking system allows to resolve these distances and to distinguish these secondary vertices from the PV. For a visualization, the event display in Fig. 2.5 nicely illustrates how well these SVs can be resolved next to the PV and additional pileup vertices. It can also happen that a $b$ hadron decays into a $c$ hadron which later decays into light hadrons. Such a decay chain could give rise to a tertiary vertex from the $c$ hadron decay on top of the secondary vertex from the $b$ hadron decay.

Secondary vertices are reconstructed using the Inclusive Vertex Finding algorithm (IVF), which was originally introduced in Ref. [229]. This algorithm is standalone and is not limited to tracks associated to a specific jet. Instead all tracks in the event with $p_T > 0.8$ GeV and a longitudinal IP $< 0.3$ cm are considered. After the vertex reconstruction has been performed, a SV is associated to a jet if the angular distance

[10] Tracks from the subsequent decay of a $b$ hadron into a $c$ hadron are labelled as $b$ hadron tracks.

[11] The most recent DeepFlavor algorithm is an exception, in which the track selection is not applied but rather a complete set of PF candidates are considered. This will be explained in Sec. 3.3.3.

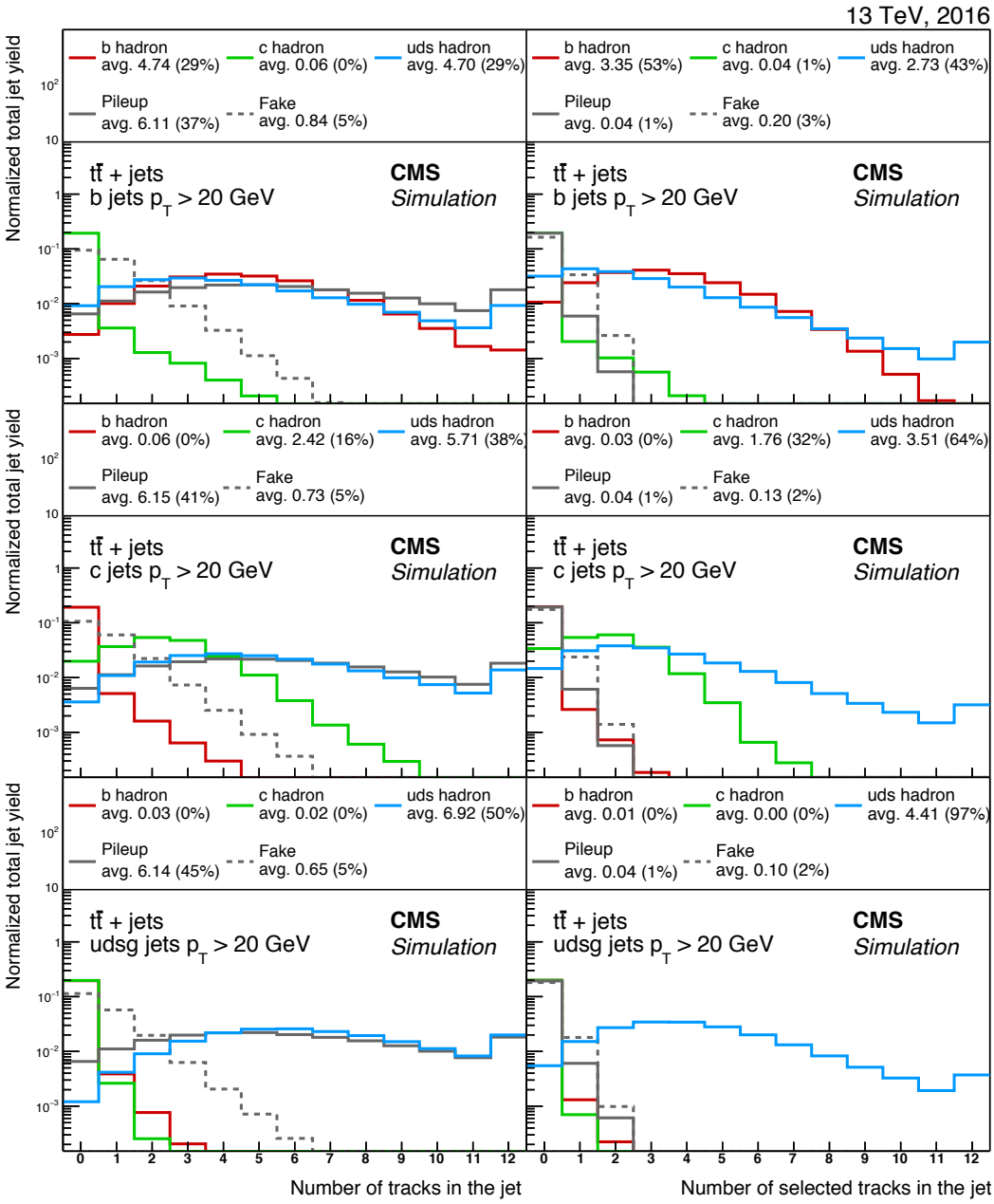


FIGURE 3.12: Fraction of tracks from different origins (see text for more details) before (left) and after (right) applying the track selection requirements for $b$ (upper), $c$ (middle), and light (lower) jets in simulated top quark pair events. The distributions are normalized such that their sum has unit area. The last bin includes the overflow entries. Figure taken from Ref. [225].

between the jet axis and the secondary vertex flight direction[12] satisfies $\Delta$R(jet axis, SV flight direction) $< 0.3$. The IVF reconstruction proceeds through several steps:

1. **Seeding:** The IVF algorithm starts by identifying seeding tracks with a relatively large displacement. Seed tracks are required to have a 3D impact parameter value above 50 µm and a 2D impact parameter significance[13] above 1.2.

[12]The SV flight direction is defined by the vector pointing from the PV to the SV.

[13]The significance of the IP is defined as the IP value divided by its uncertainty. The 2D IP is defined only in the transverse plane.

2. **Clustering:** Then tracks are clustered with the seeds based on a set of specific requirements on their mutual distance at the point of closest approach and the angle between them. A track is only clustered with the seed if the distance between them is smaller than the distance of closest approach between the seed and the PV.

3. **Fitting and cleaning:** The clusters of tracks are then used to fit the vertex position using the AVF algorithm that is also used to fit the PV position. If the fitted SV has a 2D (3D) flight distance significance smaller than 2.5 (0.5) it is removed[14]. Additionally, in order to avoid duplicates, if two reconstructed SVs share more than 70% of their tracks or the flight distance significance between the two SVs is below 2, one of them is removed randomly.

4. **Track arbitration:** During this step, tracks clustered to the SV are removed if they are considered more compatible with the PV. This includes tracks for which $\Delta$R(track, SV flight direction) $> 0.4$ or for which the distance between the SV and the track is larger than the absolute IP value of the track.

5. **Refitting and cleaning:** After the track arbitration, the secondary vertex positions are refitted. At this point duplicates are again removed with more stringent criteria. A SV is discarded if it has at least 20% of its tracks in common with another SV or if the flight distance significance between them is below 10. Finally, all SVs with a 2D flight distance significance less than 2 are discarded from the collections used in HF tagging algorithms.

This procedure results in a collection of reconstructed SVs, tuned for the identification of HF jets. The SV finding efficiency, defined as the number of jets with a reconstructed SV assigned to them divided by the total number of jets, is found to be $\sim$75% for $b$ jets and $\sim$38% for $c$ jets, whereas only around 10% of LF jets has a secondary vertex assigned to them by mistake.

The presence of one or more reconstructed SVs in a jet is therefore a good indication that the jet is likely to originate from a $b$ or $c$ hadron. Not only its presence, but also the properties of a SV can teach us a lot about the origin of a jet. This information allows to disentangle further the HF from the LF jets in the collection of jets that have at least one reconstructed SV. One of the most important observables is the corrected invariant mass of the SV as defined in Eq. (3.8). In this formula, $M_{\mathrm{SV}}$ ($p_{\mathrm{SV}}$) is the invariant mass (momentum) of the summed tracks that are associated to the SV and $\theta$ is the angle between the vector of the SV momentum $\vec{p}_{\mathrm{SV}}$ and the vector spanned between the PV and the SV (i.e. the flight direction of the SV). The motivation behind this corrected mass definition lies in the observed difference between the SV flight direction and its momentum, potentially caused by particles that were not reconstructed or which failed to be associated to the SV. The corrected SV mass artificially corrects for some missing momentum to ensure that the SV flight direction and the momentum associated to it are aligned.

$$M_{\mathrm{SV}}^{\mathrm{corrected}} = \sqrt{M_{\mathrm{SV}}^2 + p_{\mathrm{SV}}^2 \sin(2\theta)} + p \sin(\theta) \tag{3.8}$$

[14] These values are lowered to 1.25 (0.25) when reconstructing SVs for the traditional $c$–tagging algorithms, as will be explained in Sec. 4.3. However, the most recent HF identification algorithms combine $b$– and $c$–tagging in the same algorithm (see Sec. 3.3.3) and use the standard SV collection as explained here.

The normalized distributions of the corrected SV mass for different jet flavors (from simulations) are shown in Fig. 3.13 on the left. It is immediately clear that the distribution peaks at larger values for $b$ jets compared to $c$ jets, as expected by the corresponding hierarchy in the masses of the corresponding hadrons. In Fig. 3.13 on the right also the SV 2D flight distance significance is shown, from which it is again clear that the lifetime of $b$ hadrons is on average significantly larger than that of $c$ jets. The corresponding distributions for LF (udsg) jets are centered at even lower values as they are in fact not expected to have a SV at all.

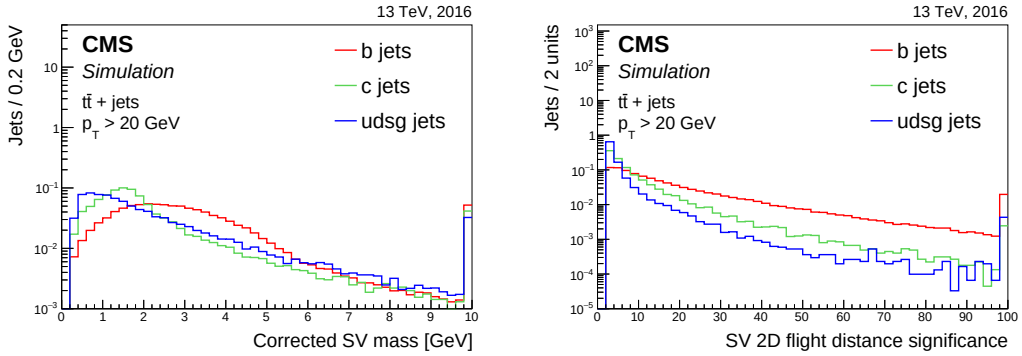


FIGURE 3.13: (left) Corrected SV mass and (right) SV 2D flight distance significance for different jet flavors using jets with $p_T > 20$ GeV from simulated top quark pair events. Figures taken from Ref. [225].

### Displaced tracks

The decay products of $b$ and $c$ hadrons will give rise to a set of reconstructed tracks that have their origin at the SV, resulting in a displacement relative to the position of the PV. This displacement is parametrized by the impact parameter of the track which measures the point of closest approach between the reconstructed tracks and PV. This parameter is defined either in the full three–dimension space (3D), in the transverse plane (2D) or as a one dimensional projection along the beam line (longitudinal). The vector pointing from the PV to the point of closest approach with the track (along the direction in which the IP is measured) is referred to as the IP vector. The IP value can either be positive or negative, depending on whether the angle between the IP vector and the jet axis is smaller or larger than $\pi/2$ respectively. For tracks in LF jets, the IP value is expected to be around zero (mm), although in practice due to resolution effects a spread is observed around that value. For $b$ and $c$ jets however, a much larger positive tail is expected from the truly displaced tracks. This is indeed confirmed by the distributions in Fig. 3.14, showing the 2D IP significance for the first and second most displaced track in a jet for different jet flavors (after the track selection has been applied). This is one of the most important track–related observables in the identification of HF jets. Other useful properties are for example the projected momenta of the tracks parallel or transverse to the jet axis or the angular separation between the track and the jet axis.

Also the track multiplicity in the jet is on average slightly larger for $b$ jets than for $c$ jets than for LF jets. This quantity is shown in Fig. 3.15 as a function of the jet $p_T$ (left) and jet $|\eta|$ (right), both before (open markers) and after (full markers) the track selection. It can be seen that a higher track multiplicity is expected with increasing jet $p_T$ and in the central region of the detector.

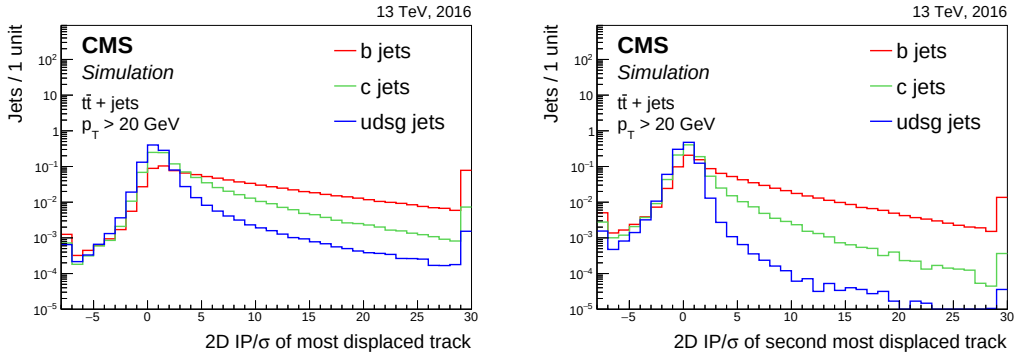


FIGURE 3.14: 2D IP significance of the first (left) and second (right) most displaced track in the jet, for different jet flavors using jets with $p_T > 20$ GeV from simulated top quark pair events. Figures taken from Ref. [225].

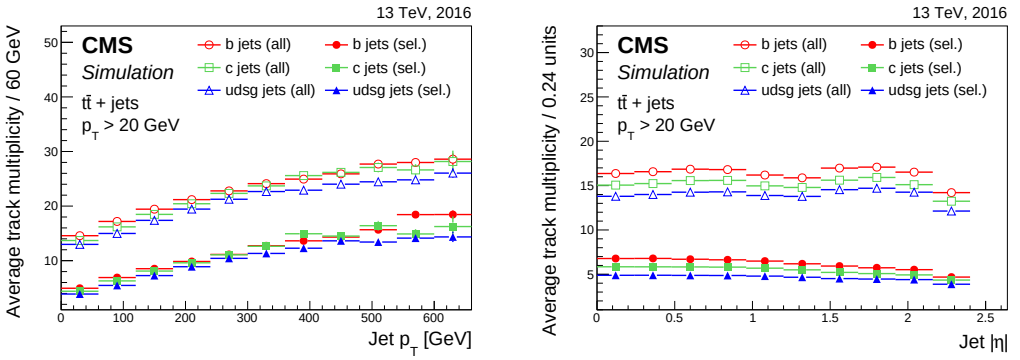


FIGURE 3.15: Average track multiplicity as a function of the jet $p_T$ (left) and $|\eta|$ (right) for jets of different flavors in top quark pair events before (open symbols) and after (filled symbols) applying the track selection requirements. Figures taken from Ref. [225].

The very first HF tagging algorithms used by CMS were based almost exclusively on the track impact parameter measurements [228]. This tagger, known as the "*Jet Probability Tagger*" (JP) is not in use anymore, but still serves its purpose in one of the calibration measurements for the $b$–tagging efficiency (see Sec. 3.3.4).

### Soft leptons

As mentioned in the introduction of this section, the relatively large (semi)leptonic branching fraction of $b$ and $c$ hadrons results in the presence of low–energy (soft) electrons or muons in about 20% (10%) of the $b$ ($c$) jets. This information is often not explicitly used in the HF tagging algorithms, but is exploited in the calibration measurements to select an unbiased sample of events enriched in HF jets (see Sec. 3.3.4). Nevertheless, for illustration purposes, Fig. 3.16 shows the distributions of dedicated soft–electron (SE) and soft–muon (SM) taggers which have been developed in the past [228] and which only use information of soft leptons inside the jets. This shows that indeed the different jet flavors can be distinguished based exclusively on the presence and properties of soft leptons.

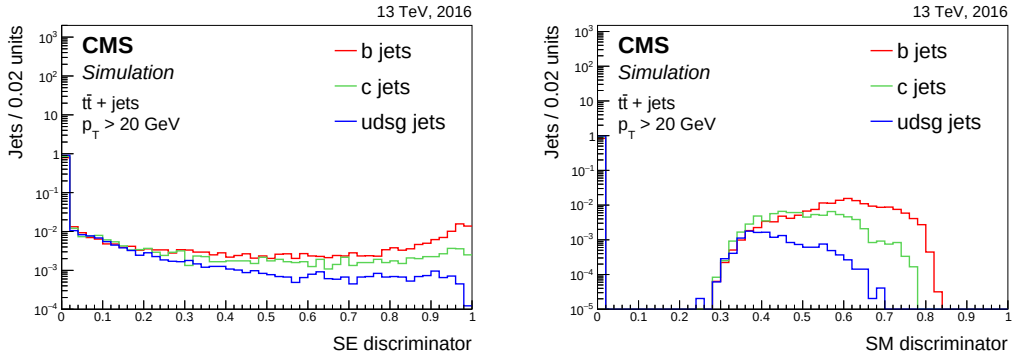


FIGURE 3.16: SE discriminator (left) and SM discriminator (right), for different jet flavors using jets with $p_T > 20$ GeV from simulated top quark pair events. Jets without a soft lepton are assigned a value of 0. Figures taken from Ref. [225].

### 3.3.3 State-of-the-art b- and c-taggers

The main objective of a HF identification algorithm is to provide a jet–based observable that is able to most optimally distinguish between $b$, $c$ and light jets. If the algorithm is used to distinguish $b$ jets from $c$ and light jets it is referred to as a $b$–tagger. If the goal is to identify $c$ jets over a background of $b$ and light jets, the algorithm is referred to as a $c$–tagger. It is clear from Sec. 3.3.2 that this is a multivariate problem that is consequently solved by some kind of a multivariate analysis (MVA) technique. The availability of large–scale simulated datasets makes this classification problem a perfect candidate for Machine Learning (ML) methods. The progress in the field of ML has exploded over the past years, both in terms of hardware and software developments. These developments make it possible to deal with large dimensional input data. The details on state–of–the–art machine learning algorithms and how they are used in HF jet identification are discussed in Chapter 4. For now it is enough to express these ML–based classifiers as a map $f_{\text{tag}}$ that transforms a $\mathcal{D}$–dimensional input space of track– and SV–based observables into a $\mathcal{C}$–dimensional output space of probabilities for a jet to belong to a certain flavor classification label. This is expressed in Eq. (3.9), where $\vec{x}_{\text{track}}$ and $\vec{x}_{\text{SV}}$ are vectors containing a set of observables related to displaced tracks and SVs respectively, whereas $\text{P}(b/c/udsg)$ represents the output observables of the classifier that can be interpreted as if they were a probability for a jet to be a $b$, $c$ or light jet. Their allowed range of values is usually restricted between 0 and 1. These outputs (or linear combinations of them) are often referred to as the *discriminator* values of the tagger. The "$\cdots$" indicate that in principle one is free to add other observables to the input or define more (or less) output classes. The discriminators are themselves new observables that combine information from all the input variables in a clever way[15] such that their distribution separates most optimally the different output classes.

$$f_{\text{tag}}(\text{jet}) : \mathbb{R}^{\mathcal{D}} \rightarrow \mathbb{R}^{\mathcal{C}} : \{\vec{x}_{\text{track}}, \vec{x}_{\text{SV}}, \cdots\} \mapsto \{\text{P}(b), \text{P}(c), \text{P}(udsg), \cdots\} \tag{3.9}$$

The tagger discriminators can then practically be used in analyses in two ways:

- Either a specific selection threshold is defined on the $b$–tag ($c$–tag) discriminator value, above which one considers the jet to be tagged as a $b$ ($c$) jet. Such

[15] The exact underlying structure of $f_{\text{tag}}$ depends on the type of Machine Learning algorithm that is used, see Chapter 4.

a threshold is often referred to as *working point* (WP) and determines the average tagging efficiency of the jet flavor of interest as well as a misidentification probability to tag a jet of another flavor by mistake. With this approach one is then able to make a selection requirement on the number of $b$–tagged (or $c$–tagged) jets in the event.

- Instead of focusing on one discrete working point value, the entire differential shape information of the discriminators can be used for example to perform a fit to data. Another very popular application is to use the flavor–tagging discriminator itself as an input to some other ML–based algorithm that is used to separate the signal from the backgrounds in that specific analysis. This approach will indeed be used later on in Chapter 5 to distinguish the $t\bar{t}c\bar{c}$ from the $t\bar{t}b\bar{b}$ and $t\bar{t}$LF events.

### Heavy–flavor taggers used by CMS during 2017

The recommended tagging algorithm in CMS during the 2017 data–taking period (and the one used throughout Chapter 5) is the *DeepCSV* tagger[16] [225]. This algorithm makes use of a deep neural network (see Chapter 4) to combine information from displaced tracks and SVs inside the jet to define four output classes P($b/bb/c/udsg$). The dedicated P($bb$) output class tries to identify jets with at least two $b$ hadrons inside. This tagging algorithm is a more advanced version of its predecessor *CSVv2* [225, 228]. The most profound differences lie in the elevated number of considered tracks from three to six and the use of one multi–class classifier (with four output classes) instead of a combination of two binary classifiers ($b$ versus $c$ and $b$ versus $light$).
Another more advanced algorithm has also been developed under the name *DeepFlavor* or *DeepJet* (names used interchangeably). Instead of focusing on a set of selected tracks (using the track selection outlined in the previous section), this algorithm uses more low–level inputs directly from all charged and neutral PF candidates that are clustered inside the jet (without additional a priori selections). This information is elevated with jet–based and SV–based observables and fed as input to a more complex neural network architecture that will be discussed in detail in Sec. 4.4. On top of the increased dimensionality of the input space, it also uses a higher–dimensional classification into the following categories: P($b/bb/b_{\text{lep}}/c/uds/g$). It can be seen that this algorithm also tries to disentangle the gluon jets from other light jets (known as gluon–tagging) and it defines an additional $b$ jet category ($b_{\text{lep}}$) in which the $b$ hadron decayed leptonically.

For general $b$ jet identification, a DeepCSV $b$-tagging discriminator is defined as the sum of P($b$) + P($bb$). For the DeepFlavor tagger this discriminator is defined as P($b$) + P($bb$) + P($b_{\text{lep}}$). This is today the most optimal classifier to distinguish $b$ jets from either $c$ or light jets. From the figures in Sec. 3.3.2 it is clear that $b$ jet properties are distributed on the far end of the spectrum, meaning they have on average the largest SV mass, the largest SV flight distance, the largest tracks IPs, etc. This allows to define one discriminator value focusing on a binary classification between $b$ jets and all other jet flavors. The performance of these $b$–taggers depends on the $b$ jet identification efficiency as well as the corresponding $c$ and light jet misidentification probability. This is usually visualized with a so–called ROC[17] curve

[16] DeepCSV = Deep Combined Secondary Vertex.
[17] ROC = Receiver Operating Characteristic.

that shows the $b$ jet efficiency as a function of the $c$ and light jet misidentification (misid.) probability when changing the value of the discriminator threshold (*i.e.* the working point) between 0 and 1 (for more details on ROC curves, see Sec. 4.2). This is illustrated in Fig. 3.17, in which the curves closest to the right lower corner correspond to the best performing tagger. This figure shows in one plot the $b$ versus light jet (full lines) and $b$ versus $c$ jet (dotted lines) identification performance. The green lines show the expected performance (in simulation) of the DeepCSV $b$–tagger using the old (three–layer) pixel detector geometry used in 2016. The red lines show the expected performance of DeepCSV using the upgraded (four–layer) pixel detector geometry, showing a gain of around 7% in $b$ jet identification efficiency for a misid. probability for light jets of 1%. This is a clear proof that HF tagging relies heavily on the tracking performance and that the upgraded pixel detector provides much more accurate information on the SV and IP–related observables. Finally the blue line shows the expected performance for the DeepFlavor $b$–tagging algorithm on the upgraded pixel detector geometry, showing an additional ∼7% gain in the $b$ jet efficiency for a misid. probability for light jets of 1% and perhaps more remarkable a ∼10% improvement in the $b$ versus $c$ jet discrimination over a large range of $c$ misid. probabilities.

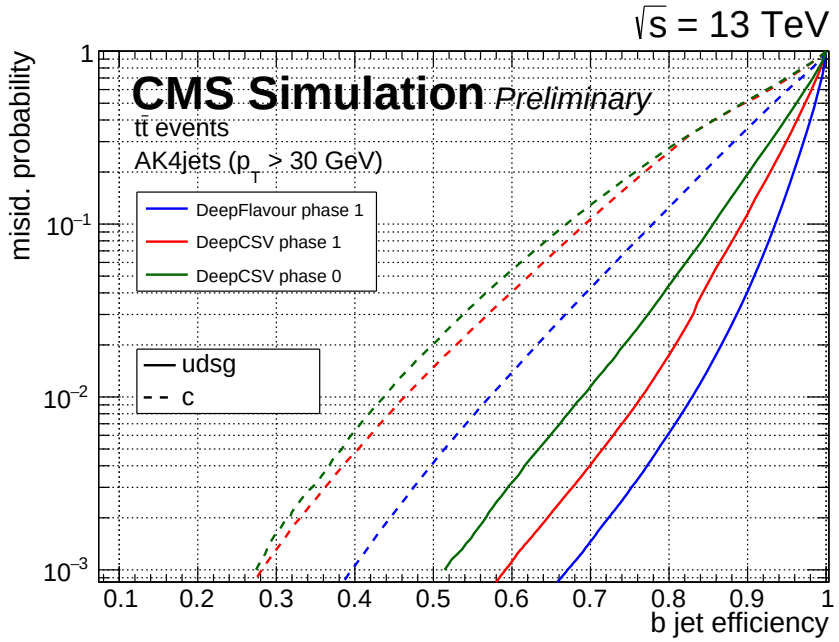


FIGURE 3.17: ROC curves for the expected $b$–tagging performance using jets with $p_T > 30$ GeV in simulated top quark pair events. The full (dotted) lines shown the $b$ jet efficiency as function of the light ($c$) jet misidentification probability. In green (red) the performance of DeepCSV with the old (new) pixel detector geometry is shown. The blue line shows the improved performance of DeepFlavor with the new Phase I pixel geometry. Figure taken from Ref. [230].

The story is slightly more complicated when it comes to $c$ jet identification. The distributions of the discriminating properties in Sec. 3.3.2 for $c$ jets are always centered somewhere in the middle between $b$ jets and light jets. Even though the DeepCSV and DeepFlavor algorithms have a dedicated output to identify $c$ jets, it has been observed that the most optimal discrimination is achieved by constructing

two separate discriminators, one to distinguish $c$ jets from $b$ jets (CvsB) and one to distinguish $c$ jets from light jets (CvsL). The CvsL discriminator for DeepCSV is defined as the normalized probability: $\mathrm{P}(c)/(\mathrm{P}(c)+\mathrm{P}(udsg))$, whereas the CvsB discriminator is defined as $\mathrm{P}(c)/(\mathrm{P}(c)+\mathrm{P}(b)+\mathrm{P}(bb))$. The distribution of $b$, $c$ and light jets in the two–dimensional phase space of the CvsL and CvsB discriminator is shown in Fig. 3.18 on the left, from which it can be seen that light jets are centered in the top left corner, $b$ jets in the lower right corner and $c$ jets towards the upper right corner. Additionally this separation into two binary classifiers allows for a higher flexibility in choosing which background rejection is more important to a specific analysis. This is shown in Fig. 3.18 on the right which is the two–dimensional representation of the ROC curve, which now become ROC contours of constant $c$ jet efficiency, for varying $b$ and light jet misidentification efficiencies. For a given $c$ jet efficiency, there exists now an additional freedom to move along the corresponding line in this graph and either prefer a large $b$ rejection, at the cost of a low light rejection or vice versa. The charm–tagger therefore comes with two discriminators and correspondingly the working point definitions should be comprised of two selection thresholds.

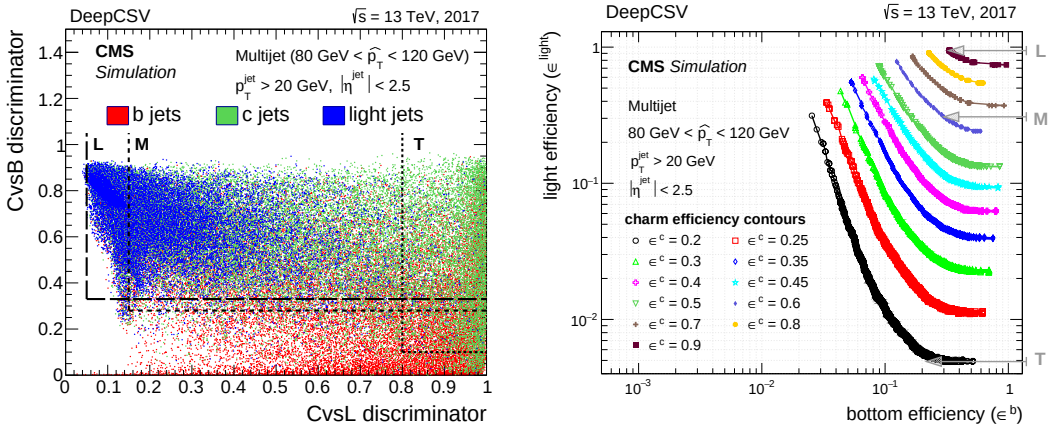


FIGURE 3.18: (left) Scatter plot showing the distribution of $b$, $c$ and light jets in the two–dimensional phase space of the CvsL and CvsB DeepCSV discriminator. (right) The $b$ jet misidentification efficiency as a function of the light jet misid. efficiency for different values of constant $c$ jet identification efficiency. Jets are taken from simulated multijet events and are required to have $p_T > 20$ GeV and $|\eta| < 2.5$.

For DeepFlavor the CvsL discriminator is defined as $\mathrm{P}(c)/(\mathrm{P}(c)+\mathrm{P}(uds)+\mathrm{P}(g))$, and the CvsB discriminator is defined as $\mathrm{P}(c)/(\mathrm{P}(c)+\mathrm{P}(b)+\mathrm{P}(bb)+\mathrm{P}(b_{\mathrm{lep}}))$. An overview of the different discriminator definitions for DeepCSV and DeepFlavor is given in Tab. 3.1.

Finally, it is worth mentioning that dedicated taggers are developed to identify specifically the decay of boosted heavy resonances into pairs of $b$ or $c$ quarks. The boosted nature of these objects often leads to the reconstruction of one single jet (often with an enlarged radius parameter of R = 0.8 in the anti–k$_T$ reconstruction) in which two $b$ or $c$ hadrons are clustered together. The distinct signature of these decays is exploited in these taggers, which are known as *(boosted) double b–tagger and double c–taggers*. See for example Ref. [231] for recent results.

TABLE 3.1: Summary of the HF tagging definitions for both $b$– and $c$–tagging using the DeepCSV and DeepFlavor taggers used by CMS in 2017.

| | **CMS heavy–flavor tagging definitions used in 2017** | | |
|---|---|---|---|
| | $b$ vs $c$/light | $c$ vs $b$ | $c$ vs light |
| `DeepCSV:` | $\mathrm{P}(b)+\mathrm{P}(bb)$ | $\frac{\mathrm{P}(c)}{\mathrm{P}(c)+\mathrm{P}(b)+\mathrm{P}(bb)}$ | $\frac{\mathrm{P}(c)}{\mathrm{P}(c)+\mathrm{P}(udsg)}$ |
| `DeepFlavor:` | $\mathrm{P}(b)+\mathrm{P}(bb)+\mathrm{P}(b_{\mathrm{lep}})$ | $\frac{\mathrm{P}(c)}{\mathrm{P}(c)+\mathrm{P}(b)+\mathrm{P}(bb)+\mathrm{P}(b_{\mathrm{lep}})}$ | $\frac{\mathrm{P}(c)}{\mathrm{P}(c)+\mathrm{P}(uds)+\mathrm{P}(g)}$ |

### 3.3.4 Calibration methods

As will be discussed in detail in Chapter 4, these ML algorithms are trained on simulated events. Also their expected performance is often evaluated through ROC curves which are based purely on simulation such that the truth information on the jet flavor is available. It has however been observed that the performance in data is different from the performance in simulations. To take into account these discrepancies when performing an analysis, dedicated correction factors need to be derived to correct the performance in simulations such that it matches the one observed in data. Two possibilities arise depending on the usage of the HF tagging discriminants. In case a discrete working point is used, the efficiency of that selection in simulation is corrected to match that observed in data through dedicated scale factors. If the full differential discriminator shape is used, the entire shape needs to be corrected to match the shape observed in data.

#### Efficiency corrections for discrete working points

Within the CMS collaboration, efficiency corrections are provided for three $b$–tagging working points and three $c$–tagging working points. The corresponding working points are labelled "Loose (L), Medium (M) and Tight (T)" and Tab. 3.2 summarizes the threshold values as well as the approximate efficiencies corresponding to each of these working points used by the CMS Collaboration in 2017.

For each of these working points, efficiency corrections are derived through dedicated flavor–dependent scale factors (SFs) as a function of jet $p_T$ and $\eta$. These SFs are defined as the efficiency for a given jet flavor $f$ measured in data ($\epsilon_f^{\mathrm{data}}$), divided by the corresponding efficiency observed in the simulation ($\epsilon_f^{\mathrm{MC}}$)

$$\mathrm{SF}_f(p_T, \eta) = \frac{\epsilon_f^{\mathrm{data}}}{\epsilon_f^{\mathrm{MC}}}. \tag{3.10}$$

A variety of calibration measurements exist which measure these scale factors in different topologies and through different methods. A detailed overview is provided in Ref. [225] but a summary is provided here.

First of all, SFs for $b$ jets are measured through several methods applied on QCD multijet or $t\bar{t}$ events. An overview is given below:

**QCD multijet:** A first collection of measurements is performed in a QCD multijet topology which is enriched in $b$ jets by requiring a muon inside one of the jets.

TABLE 3.2: Summary of the $b$– and $c$–tagger working points of the DeepCSV tagger in 2017. The columns labelled "BvsAll, CvsL and CvsB" represent the threshold values of the corresponding working points for the different discriminator definitions in Tab. 3.1. The last three columns show the corresponding efficiencies of $b$, $c$ and light jets derived from a simulated sample of multijet events using jets with $p_T > 20$ GeV and $|\eta| < 2.5$.

| | WP | BvsAll | | $\epsilon_b$[%] | $\epsilon_c$ [%] | $\epsilon_{udsg}$ [%] |
|---|---|---|---|---|---|---|
| | Loose | 0.152 | | 86.4 | 41 | 10.0 |
| ***b*–tag** | Medium | 0.494 | | 71.5 | 12 | 1.0 |
| | Tight | 0.800 | | 54.2 | 2.2 | 0.1 |
| | WP | CvsL | CvsB | $\epsilon_b$[%] | $\epsilon_c$ [%] | $\epsilon_{udsg}$ [%] |
| | Loose | 0.05 | 0.33 | 33 | 89.7 | 94.3 |
| ***c*–tag** | Medium | 0.15 | 0.28 | 30 | 60 | 30.8 |
| | Tight | 0.8 | 0.1 | 21.7 | 19.3 | 0.5 |

- **LTSV**: The *Lifetime–Tagging Secondary Vertex* method is based on a template fitting strategy using templates from the JP tagger as well as from the mass of the SV inside the jet.
- **PtRel**: The *PtRel* method is also based on a template fitting strategy using the transverse component of the momentum of the muon inside the jet with respect to the jet–axis.
- **System8**: In the *System8* method, a set of 8 equations is solved simultaneously. Each of these equations is related to the number of tagged and untagged jets in phase space regions separated by a selection on the transverse component of the momentum of the muon inside the jet with respect to the jet–axis.

**Top quark: pair ($t\bar{t}$)** A second set of measurements is performed in a topology of dileptonic or semileptonic top quark pair events, which are naturally enriched in $b$ jets from the top quark decay.

- **Kin**: The *Kin* method is based on a fit to a kinematic MVA discriminator that uses only kinematic information to separate $b$ jets from the top quark decay from other jets in dileptonic top quark pair events.
- **TagCount**: The *TagCount* method is performed in dileptonic top quark pair events and is based on simply counting the number of events with two $b$–tagged jets.
- **TnP**: The *Tag and Probe* method is performed in semileptonic top quark pair events. Either the leptonic or the hadronic $b$ jet is used as a tag, whereas the tagging efficiency is measured on the other probe jet.

All of these measurements provide results for the $b$ jet scale factors and are eventually combined in a weighted average using the best linear unbiased estimator (BLUE)

method [232] that properly takes into account correlations between the measurements. The results of all these measurements using data collected in 2017 [230], together with the combination are shown in Fig. 3.19 for the Loose WP of the legacy CSVv2 $b$–tagger (left) and the Medium WP of the DeepCSV $b$–tagger (right). The different measurements are consistent and result in a value for $SF_b$ which is slightly below one over the entire jet $p_T$ spectrum.

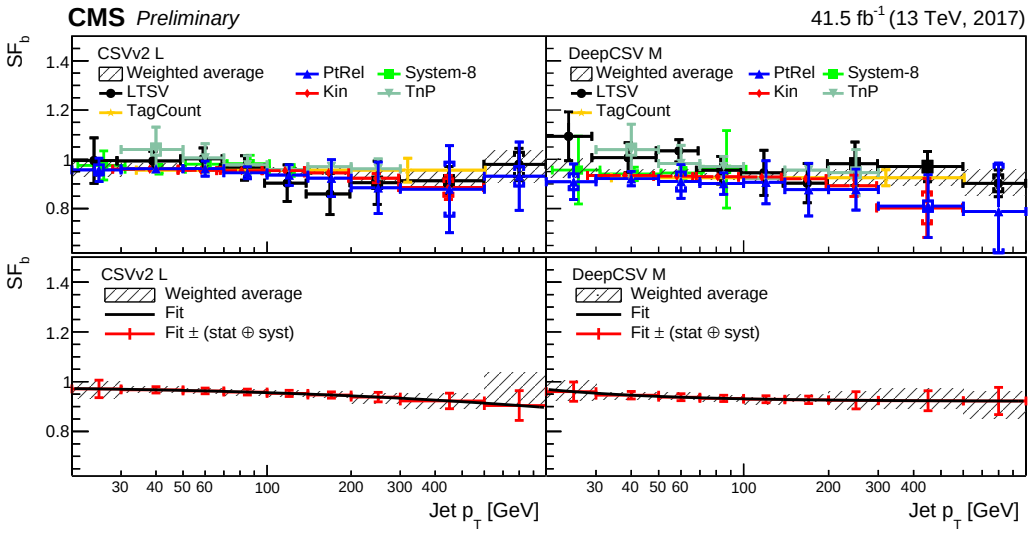


FIGURE 3.19: Summary of the different measurements of the $b$ jet scale factor ($SF_b$) for the (legacy) CSVv2 Loose WP on the left and DeepCSV Medium working point on the right in 2017 as a function of the jet $p_T$. The lower panel shows the combination of all the measurements. Figure taken from Ref. [230].

Secondly, the SFs for light jets are derived in QCD multijet events without any soft lepton requirement, resulting in a light–jet enriched topology. The so–called *NegativeTag* method is used, in which the taggers are evaluated using only tracks with either positive or negative impact parameter values and only SVs with either a positive or a negative flight distance. The outcome of these so–called *positive and negative taggers* are expected to be symmetrical for light jets considering that the non–zero IP values result purely from resolutions effects and not from true physical displacement. The light jet scale factors are derived using the negative tagger output and then projected back onto the full collection of jets. On average the resulting misid. SFs for light jets are found to be slightly larger than 1.

Finally, the derivation of the $c$ jet SFs is overall slightly more challenging given that it is not straightforward to define a large data sample which is very pure in $c$ jets. The only suitable candidate is a topology of W+$c$ events. The key to extracting a pure sample of W+$c$ events lies in the fact that the W boson and the $c$ jet are expected to always have opposite sign (OS) of their electric charge, which is not the case for the relevant backgrounds (top quark pair production, W+$c\bar{c}$/$b\bar{b}$/light), where one expects 50% same sign (SS) and 50% opposite sign electric charge. Subtracting the tagger discriminant distributions from the SS events, from those of the OS events results in a very pure W+$c$ sample that is then used to extract the working points by simply subtracting the small remaining background residuals (estimated from simulations) and counting the number of events passing a given working point of

the considered tagger. This method is validated using a topology of semi–leptonic top quark pair events in which ∼25% of the jets from the hadronically decaying W boson are expected to be $c$ jets. This $t\bar{t}$–based method is based on a binned maximum likelihood fit to a likelihood discriminant that contains information on the kinematic reconstruction of the top quark systems. The results of this measurement usually come with slightly larger uncertainties and are not separated in jet $p_T$ and $\eta$ due to the limited size of the selected data sample. Nevertheless, the two sets of independent measurements of $\mathrm{SF}_c$ are again combined using the BLUE method. The resulting SFs are either used for the $c$ jet efficiency of the $c$–tagger or as the $c$ jet mistag rate for the $b$–taggers and are found to be slightly below 1 (though consistent with 1 within uncertainties).

Overall, the SF results indicate that the performance of the $b$–tagging and $c$–tagging algorithms is slightly overestimated in simulations, compared to what is observed in data. This is illustrated in Fig 3.20, where the ROC curves show the expected performance of the DeepCSV (red) and DeepJet (*i.e.* DeepFlavor, blue) $b$–taggers in simulation, and the triangles illustrate the performance of the three WPs after the application of the corresponding SFs and therefore are representative for the performance observed in data. The degraded performance in data is of the order of ∼ 4% (absolute) lower $b$–tagging efficiency for the same light or $c$ jet misid. probability.

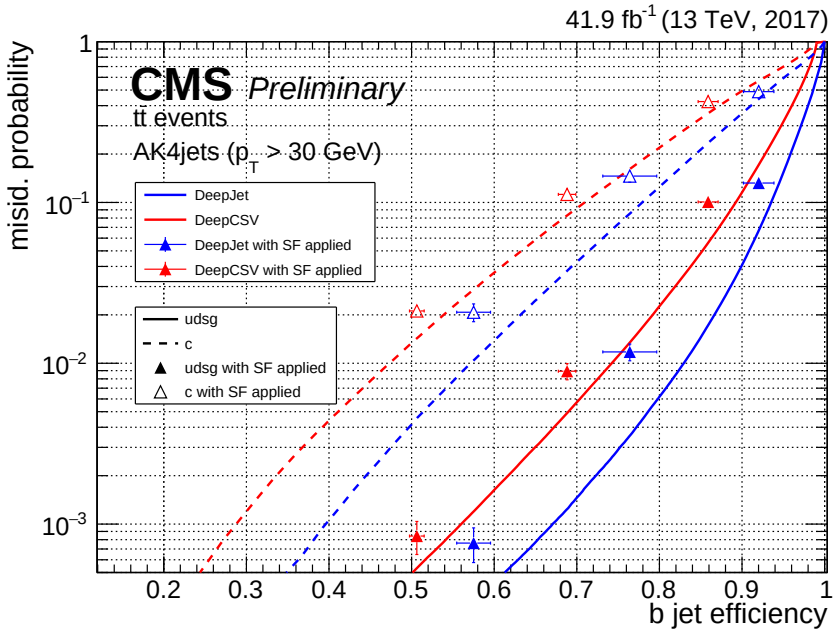


FIGURE 3.20: ROC curves for the expected $b$–tagging performance of DeepCSV (red) and DeepJet (*i.e.* DeepFlavor, blue) using jets with $p_T > 30$ GeV in simulated top quark pair events. The full (dotted) lines shown the $b$ jet efficiency as function of the light ($c$) jet misidentification probability. Additionally, the triangles show the resulting performance of the three WPs after the application of the corresponding SFs and therefore are representative for the performance observed in data. Figure taken from Ref. [233].

The SFs are applied on a per–jet basis. To properly correct the event selection efficiency based on the $b$-tagging multiplicity, an event weight $w_{b-\text{tag}}$ can be constructed as:

$$
\begin{aligned}
w_{b-\text{tag}} &= \frac{P(\text{data})}{P(\text{MC})}, \\
\text{with} \quad P(\text{data}) &= \prod_{\text{i=tagged}} \text{SF}_i \epsilon_i \times \prod_{\text{j=not tagged}} (1 - \text{SF}_j \epsilon_j), \\
P(\text{MC}) &= \prod_{\text{i=tagged}} \epsilon_i \times \prod_{\text{j=not tagged}} (1 - \epsilon_j),
\end{aligned}
\tag{3.11}
$$

where the products run over the jets passing or failing the working point respectively and $\epsilon$ represents the tagging efficiency as a function of jet $p_T$ and $\eta$.

### Shape reweighing for b–tagging discriminators

For analyses that are interested in using the full differential information of the $b$–tagging discriminator rather than just a discrete working point it is important to make sure the full discriminator shape in simulations correctly describes the shape observed in data. To this end, a set of shape reweighing SFs has been derived. These SFs are no longer ratios of efficiencies from data to simulation, but should be interpreted as a per–jet weight that depends on the flavor and the discriminator value of that jet. This method was first introduced in Ref. [234] and is based on an iterative procedure using a tag–and–probe technique to measure the scale factors for both $b$ and light jets simultaneously. A first selection is made, requiring two charged leptons with opposite electric charge and two jets. Further selection requirements are made to divide this sample into two exclusive regions, one to extract the $b$ jet SFs and one to extract the light jet SFs.

- **$\text{SF}_\text{b}$:** A further selection is adopted aimed at extracting dileptonic top quark pair events which are enriched in $b$ jets. One of the jets is required to pass the Medium working point of the considered tagger (the "tag" jet), whereas the other one is used as a probe jet to measure the SFs.

- **$\text{SF}_l$:** A further selection focuses on a light–enriched topology of Z+jets events, by selecting events with same flavor leptons with a combined invariant mass within 10 GeV from the Z boson mass. On one of the jets a Loose $b$–tag veto is applied (the "tag" jet), such that the other jet can be used as a probe to measure the light jet SFs.

The method proceeds iteratively by first measuring $\text{SF}_l$ as a function of the discriminator value in the Z+jets topology and subsequently applying these SFs in the $t\bar{t}$ topology to subtract the light jet component and measure $\text{SF}_b$ as a function of the discriminator value. The method converges to stable SFs after three iterations. The method is not able to measure SFs for $c$ jets and therefore assigns a flat value of 1 with a conservative uncertainty of twice the uncertainty obtained for $b$ jets. The per–jet SFs are also derived in bins of jet $p_T$ and $\eta$ and are combined into an event–based weight:

$$
w_{b-\text{tag}}^{\text{shape}} = \prod_i \text{SF}_i, \tag{3.12}
$$

where the product runs over all the jets of which the $b$–tagging discriminant value is used in the analysis. More details on the event selections can be found in Ref. [225],

chapter 8.5.

This method will be used in Chapter 5 and the SFs which have been derived on 2017 data are shown in Fig. 3.21 for $b$ jets on the left and for light jets on the right. The shaded area represents to the total uncertainty, which is comprised of the statistical uncertainties, uncertainties related to the purity of $b$ and light jets in the two topologies, and uncertainties related to the jet energy scale.

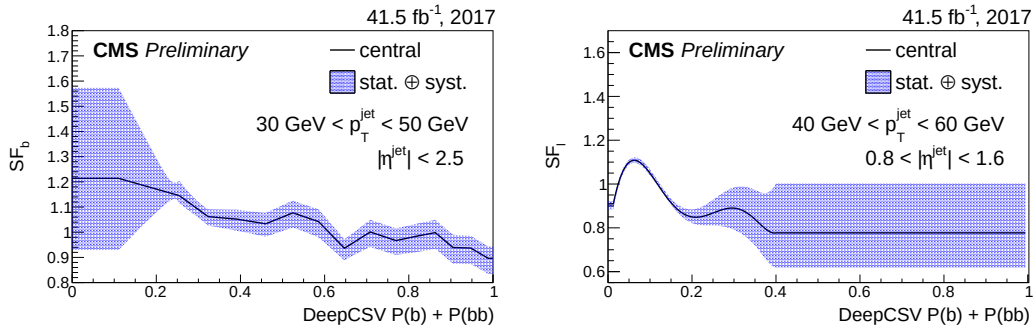


FIGURE 3.21: Shape reweighing SFs derived with the iterative fitting procedure on 2017 data for $b$ jets (left) and light jets (right). The ranges of jet $p_T$ and $\eta$ are indicated on the figures. The shaded area represents the total uncertainty, which is comprised of the statistical uncertainties, uncertainties related to the purity of $b$ and light jets in the two topologies, and uncertainties related to the jet energy scale.

In order to extract also the $t\bar{t}c\bar{c}$ component, the analysis also relies on the use of the full differential information in the $c$–tagger discriminants. Therefore also a shape calibration is needed for the two–dimensional (CvsL–CvsB) $c$–tagger discriminant distributions. A new strategy to derive these $c$–tagging shape reweighing SFs will be presented in Sec. 5.7.

# Chapter 4 Machine Learning for big data analysis

The use of Machine Learning (ML) methods (sometimes referred to as Artificial Intelligence) has gained popularity over the last few decades and stimulated important developments for example in the field of image or speech recognition. The rapidly increasing computational power is, amongst many other reasons, without a doubt responsible for this. Also in the physics community, and particularly in the field of High–Energy Physics (HEP), the advantages of ML techniques have convinced many. Typically, an analysis performed on proton–proton collision data uses a set of observables or features defined for each collision event. One then tries to classify the events as either being signal or background, according to the fundamental interaction happening or being hypothesized in the $pp$ collisions. This binary classification is a benchmark example of a statistical hypothesis test in which one assumes a null–hypothesis ($H_0$, for example the assumption that an event is background) and the conclusion leads either to an acceptance of $H_0$ or a rejection of it. This concept of classification or hypothesis testing is well known and applicable in many analyses.

The applications of ML techniques in HEP have however superseded that of a simple binary classification task. A (far from complete) summary of advanced ML–driven application in HEP is given below:

- The so-called SUSY-AI project uses ML–based predictions to quickly decide whether a point in the SUSY parameter space is excluded by the existing measurements or not [235].
- Enhanced sensitivity to Effective Field Theory models can be achieved by using multi–class ML methods [109] (see Chapter 6), or by constructing specialized loss functions that use information on the event generators [236].
- Auto–encoders may prove to be the key to purely data–driven detection of BSM effects [237].
- The dependence of an analysis on specific sources of systematic uncertainties may be reduced by using adversarial neural networks [238].
- Recently, a collaboration between CMS, ATLAS and LHCb has launched an official ML competition under the name TrackML [239]. The goal is to search both within the HEP community, as well as the ML community for a solution to the very large pileup tracking scenario which will present itself at the HL–LHC.
- Fast tracking–based triggers for the HL–LHC era could become reality by implementing ML–based methods on hardware components such as FPGAs [240].

- State–of–the–art HF tagging algorithms [233] rely on advanced deep neutral network structures to obtain optimal classification performance, as will be discussed in detail in Sec. 4.4.
- ...

Even though the list could go on for many more pages, it already provides a convincing argument that ML application in HEP are multifold. The applications presented in this thesis are mostly aimed at multi–class classification tasks. After summarizing the basic ideas behind ML algorithms in Sec. 4.1, the problem of multivariate classification will first be formally introduced in Sec. 4.2. Then in Sec. 4.3, an overview will be given of the available ML methods, focusing mainly on the (deep) neural networks as they are used throughout the rest of this thesis. Finally, Sec. 4.4 is dedicated to a detailed discussion on the structure of the state–of–the–art HF tagging algorithms, which were introduced before in Sec. 3.3. Many of the ideas in this chapter are inspired by Ref. [241], chapter 5 and Ref. [242], chapters 6 – 10.

## 4.1 Basics of Machine Learning

The most fundamental idea behind ML is the concept of training an algorithm on a (preferably large) set of example data from which it can learn the desired patterns. Once the algorithm is trained, it can be used to make predictions on unseen data. This is a very generic definition and the precise application of an ML algorithm depends on the nature of the desired "pattern recognition" that one has in mind. Some examples are listed below.

- A *classification algorithm* should be trained to assign the correct class labels to a given set of examples, based on the input properties or *features* that are available. Image recognition is an example of a classification task in which the algorithm is trained to recognize what is displayed in a given picture.
- A *regression algorithm* is trained to predict a continuous output as a function of the provided inputs, without a given analytical expression. It can be seen as a generic interpolation or fitting algorithm that does not rely on the knowledge of the underlying functional dependence between inputs and outputs.
- The goal of a *generative algorithm* is to be able to produce new examples from random inputs, resembling as close as possible the structure of the data it was trained on. Such algorithms are usually referred to as GANs (Generative Adversarial Networks).
- An *auto–encoder* is trained to compress the input data into a lower–dimensional latent space and to reconstruct from that latent space the original information as accurately as possible.

The training process can either be *supervised*, meaning that the example training dataset comes with a set of true labels or outputs which are known a priori. This is the most common type of ML algorithms. There exist however also *unsupervised* learning techniques in which the algorithm does not have access to the true desired outcome during its training phase (often related to clustering tasks). A third category is called *reinforced* learning in which the algorithm does not have access to the true desired output, but instead its predicted outcome is either penalized or

rewarded after which is can update its predictions for a next training iteration[1].

The training phase proceeds iteratively, starting out with an algorithm that outputs random predictions and progressively adapting its predictions to be closer to the desired output. This raises the need of a *loss function* that gives a quantitative assessment of how well a predicted output matches the desired true output. In principle the functional form of the loss function can be chosen freely as long as it possesses a minimum at the desired output predictions. A standard loss function used for multivariate classification is the categorial cross–entropy, defined in Eq. (4.1), which reduces to Eq. (4.2) in the case of binary classification.

$$L_{\text{cross–entropy}}(y, \hat{y} \,|\, \vec{\theta}) = -\sum_{i=1}^{\mathcal{C}} y_i \log(\hat{y}_i) \tag{4.1}$$

$$\overset{\mathcal{C}=2}{=} -\left(y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right) \tag{4.2}$$

Here $\mathcal{C}$ denotes the number of classes, $y_i$ is the true label assigned to the class $i$ and $\hat{y}_i$ is the predicted output by the ML classifier. By minimizing the loss function, the ML algorithm will adapt its internal parametrization ($\vec{\theta}$) such that its predictions become closer to the desired outputs. This minimization is performed through a procedure known as back–propagation [243]. A popular minimization algorithm is the stochastic gradient descent, of which the foundations were laid already in 1951 [244]. After a certain amount of training data have passed through the algorithm, the gradient of the loss function with respect to the internal parametrization of the algorithm is determined. The internal parametrization is then updated in the direction of the negative gradient, resulting in a minimization of the loss function and therefore an improved prediction. By progressively improving its predictions over many iterations, also called *epochs*, the ML algorithm is able to converge towards its most optimal performance.

In order to make accurate predictions on unseen examples over a large range of the considered phase space, ML algorithms rely heavily on the availability of very large training datasets. The production of large–scale simulations (with truth–level information) for HEP analyses provides the perfect environment to train very complex algorithms. Additionally, the training of complex algorithms with thousands or millions of internal parameters on a set of millions of training data requires efficient and fast computing resources. The developments in Graphics Processing Units (GPUs) have proven to be an indispensable ingredient to allow for the most powerful algorithms, sometimes referred to as *Deep Learning* (DL) algorithms, to be trained on reasonable time scales.

## 4.2 Multivariate classification

The goal of this section is to introduce formally the concept of multivariate classification using ML classifiers. A mathematical notation will be introduced with a focus on the statistical interpretation of the performance and the analogy to hypothesis testing. A general classification task has an arbitrary number of classes that can be considered. However, by combining the predicted probabilities

[1] An example of reinforced learning can be found in Artificial Intelligence for games, in which a strategic move can either be evaluated as positive (leading to a possible victory) or negative (resulting in defeat).

for the different classes in an appropriate way, such a multi–class classification can always be interpreted as a collection of binary classifiers. For simplicity and visualization purposes we will therefore introduce ML classifiers mostly through binary classification.

Consider a dataset containing $\mathcal{N}$ *events* (for example *pp* collisions or individual jets). Each of those events is described by a set of $\mathcal{D}$ *features* (for example the track and SV related observables in HF taggers). We also call $\mathcal{D}$ the *dimensionality of the feature–space*. Each event can now be represented by a *feature–vector* $\vec{x}$ of size $\mathcal{D}$

$$\vec{x} = \{x_1, x_2, ..., x_{\mathcal{D}}\}.$$

The $j$-th event has a feature–vector denoted by $\vec{x}^j$ and the $i$-th feature of event $j$ is denoted by $x_i^j$. Furthermore each event also carries a label ($y$) which denotes its category. Maintaining the link to HEP and the restriction to binary classification tasks, each event is either labelled signal ($y = \mathcal{S}$) or background ($y = \mathcal{B}$). This label is needed for supervised learning techniques, however it is obsolete when unsupervised or reinforced learning is considered. Each of the $\mathcal{D}$ features has a probability distribution (PDF) for signal and for background:

$$\underline{\texttt{Signal}} : \text{PDF}(x_i|y = \mathcal{S})\ i \in \{1, 2, ..., \mathcal{D}\},$$
$$\underline{\texttt{Background}} : \text{PDF}(x_i|y = \mathcal{B})\ i \in \{1, 2, ..., \mathcal{D}\}.$$

However these underlying distributions are a priori unknown and are usually estimated from simulations in the form of binned histograms.

In order to separate the events into different classes, the classifier is supposed to construct a *separation boundary* between signal and background events in the $\mathcal{D}$-dimensional feature–space. For $\mathcal{D} = 2$ an example is shown in Fig. 4.1 on the left. This can be generalized to $\mathcal{D}$ dimensions where the decision boundary is represented by a $\mathcal{D} - 1$ dimensional hyperspace. The way this boundary is formed depends on the classifier of choice (see Section 4.3), but the procedure is always based on a training phase during which the most optimal boundary is constructed for a set of labelled training events. The performance of the trained classifier can then be tested by making predictions on unseen labelled events (independent from the training events). This is called the validation phase. Finally once (or if) the desired performance is achieved, the trained classifier can be used to make predictions on unlabelled data.

Each of the available classification methods has its own mathematical description, containing an internal parametrization of this decision boundary. These *internal parameters* are usually a set of weights that are updated during the training phase in a way such that they minimize the loss function, yielding the best possible separation between signal and background events. Each classifier has however also a set of *external parameters* that can be tuned by the user. These parameters typically dictate the allowed complexity of the decision function and its speed of convergence to an optimal solution. The final values of the internal parameters depend on the chosen values of the external ones. When the set of external parameters is chosen such that the determination of the decision boundary does not have a lot of degrees of freedom, typically the easiest boundaries are constructed. The most simple example is a linear boundary (a line in two dimensions or a plane in three dimensions). Such simple boundaries will follow the main large features of the underlying PDFs of the input

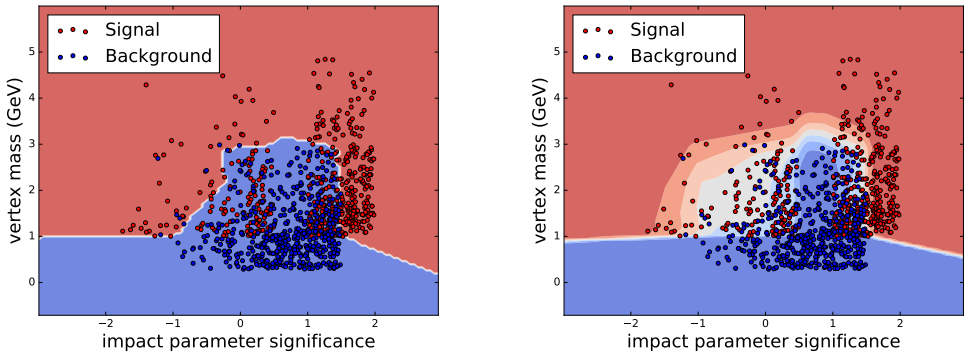


FIGURE 4.1: Illustration of a decision boundary between signal and background events, constructed by a shallow neural network. On the left the hard decision boundary in drawn, whereas on the right the contours show the "soft" output scores of the classifier. The dataset was artificially constructed from $b$ jets (signal) and light jets (background) within certain predefined and partially overlapping ranges of SV mass and impact parameter significance.

features, but they will fail to extract more subtle details. In this case, the classifier is said to have a small *variance*, meaning that the determination of the boundary on a different dataset with the same underlying PDFs will lead to very similar results. They are also said to have a large *bias*, meaning they will systematically fail to see small features and therefore fail to achieve the absolute best separation between signal and background. On the other extreme, algorithms which allow for a large complexity in the decision function will also consider the finest features in the underlying PDFs and will typically have much smaller bias, although they will have a larger variance. In this case it is very important that the algorithm does not learn features that are not desired (like statistical fluctuations) and becomes too specific to the training dataset[2]. This is called *overtraining* and should be avoided at any time. For this purpose, regularization methods exist that restrict the allowed complexity of the decision function. A natural way of reducing the overtraining is to use large enough training datasets to suppress statistical fluctuations or parts of the phase space which are uncovered during training. Overtraining can be detected by monitoring the value of the loss function or the classification accuracy on the training data as well as on a small, statistically independent validation sample, over different training iterations. If both the training and validation data show a continuously falling (rising) value of the loss (accuracy), the classifier is not being overtrained.

The construction of a decision boundary between signal and background events offers a visual description of the output of the classifier. However, in a more general picture it is better to interpret a classifier as a function or map ($f_{\text{clf}}$) from a $\mathcal{D}$–dimensional feature–space to a $\mathcal{C}$–dimensional output space of probabilities for the event to belong to one out of $\mathcal{C}$ classes. This was already expressed in Eq. (3.9) in the case of HF tagging, and can be written in a more general form as in Eq. (4.3). This boils down to Eq. (4.4) in the case of binary classification.

[2]It is said that in this case the algorithm does not generalize well enough.

$$f_{\text{clf}}^{\text{multi-class}} : \mathbb{R}^{\mathcal{D}} \to \mathbb{R}^{\mathcal{C}} : \vec{x} = \{x_1, x_2, ..., x_{\mathcal{D}}\} \mapsto \{\text{P}_1, \text{P}_2, \cdots, \text{P}_{\mathcal{C}}\} \tag{4.3}$$

$$f_{\text{clf}}^{\text{binary}} : \mathbb{R}^{\mathcal{D}} \to \mathbb{R}^{2} : \vec{x} = \{x_1, x_2, ..., x_{\mathcal{D}}\} \mapsto \{\text{P}, (1 - \text{P})\} \tag{4.4}$$

In this notation, the output discriminators are denoted $\text{P}_{(i)}$. As mentioned before, these are continuous outputs bounded by an interval (typically $[0, 1]$ or $[-1, 1]$) that are in some way related to the probability of an event belonging to one of the classes. Usually the discriminators sum up to 1 (or the width of the allowed interval), such that in the case of $\mathcal{C}$ classes, there exist only $\mathcal{C} - 1$ independent outputs. A binary classification problem can therefore be characterized by one discriminator value P, which represents the probability to belong for example to the signal class, whereas the complementary output probability to belong to the background class is then known to be $1 - \text{P}$. These continuous outputs are sometimes called *soft scorer functions*. As opposed to the discrete hard decision boundary, one can now draw contours of constant soft scores as illustrated in Fig. 4.1 on the right.

After the training phase, the probability distribution of the discriminator for signal events $\text{P}(y = \mathcal{S})$ is most optimally separated from that of background events $\text{P}(y = \mathcal{B})$. A toy–example of such a discriminator distribution is shown in Figure 4.2 on the left, where the background is distributed more towards 0 and the signal towards 1. By making a selection on this discriminator value, one can choose how much signal and background is selected. When we denote this selection threshold by $\kappa \in [\text{P}_{\text{min}}, \text{P}_{\text{max}}]$, all events for which $\text{P} \geqslant \kappa$ are classified as being signal. The resulting part of the phase–space in which events are classified as signal will be denoted as $\omega$. All events for which $\text{P} < \kappa$ are classified as being background and the resulting phase–space is denoted as $\overline{\omega}$ (since it is the complement of $\omega$). These regions are also indicated in Fig. 4.2 on the left. The hard separation of events into these two sub–spaces results in so–called *hard labels* (in contrast to soft scorer values P) for the events, which we will denote by $\hat{y}_{\mathcal{S}}$ for events in $\omega$ and by $\hat{y}_{\mathcal{B}}$ for events in $\overline{\omega}$. It is important to note that $\hat{y}_{\mathcal{S}/\mathcal{B}}$ denotes the final decision of the classifier (dependent on the choice of $\kappa$) and is not necessarily equal to the true label ($y$) of the event. The decision can be right or wrong, which is described by the signal selection efficiency $\epsilon$ and the background acceptance $1 - r$ (where $r$ is called the background rejection). The former will be defined as the fraction of the selected true signal ($\mathcal{S}_\omega$) events to the sum of the selected true signal events and the rejected false background ($\mathcal{S}_{\overline{\omega}}$) events. In other words $\epsilon$ is the fraction of true signal events that are correctly classified as being signal (given the selection threshold $\kappa$). The background acceptance $1 - r$ is defined as the fraction of false signal[3] events ($\mathcal{B}_\omega$) to the sum of the false signal events and the rejected true background ($\mathcal{B}_{\overline{\omega}}$) events. Again this is more easily expressed by the fraction of true background events that were misidentified as being signal for the given value of $\kappa$. Symbolic definitions are shown in Eqs. (4.5) – (4.10).

$$\mathcal{S}_\omega \sim \mathcal{S} \cup \omega \tag{4.5}$$

$$\mathcal{S}_{\overline{\omega}} \sim \mathcal{S} \cup \overline{\omega} \tag{4.6}$$

$$\mathcal{B}_\omega \sim \mathcal{B} \cup \omega \tag{4.7}$$

$$\mathcal{B}_{\overline{\omega}} \sim \mathcal{B} \cup \overline{\omega} \tag{4.8}$$

[3]Background events which are mistakenly identified as signal events are referred to as *false signal events.*

$$\epsilon = \frac{\mathcal{S}_\omega}{\mathcal{S}_\omega + \mathcal{S}_{\overline{\omega}}} \qquad (4.9) \qquad\qquad 1 - r = \frac{\mathcal{B}_\omega}{\mathcal{B}_\omega + \mathcal{B}_{\overline{\omega}}} \qquad (4.10)$$

Considering a background event as signal by mistake ($\mathcal{B}_\omega$) is called a type–I error and similarly considering a signal event as background by mistake ($\mathcal{S}_{\overline{\omega}}$) is called a type–II error. This is summarized in Table 4.1. These concepts are known from the theory of hypothesis testing [241, 242].

TABLE 4.1: Possible right or wrong decisions when defining a decision threshold on a discriminator distribution with corresponding probabilities.

| | $\omega$ | $\overline{\omega}$ |
|---|---|---|
| $\mathcal{S}$ | Correct decision with probability $\epsilon$ | Wrong decision: type-II error with probability 1-$\epsilon$ |
| $\mathcal{B}$ | Wrong decision: type-I error with probability 1-$r$ | Correct decision with probability $r$ |

For each value of $\kappa$ one can calculate $\epsilon$ and $1 - r$ and plot that point in what is called a Receiver-Operating-Characteristic (ROC) curve. An example is shown in Fig. 4.2 on the right where the point obtained from the chosen value of $\kappa$ on the left side of Fig. 4.2 is indicated with a star. ROC curves give a more global view of the performance scanning over different selection thresholds. The closer the ROC curve is to the right lower corner: $(\epsilon, 1 - r) = (1,0)$, the better. Therefore, a popular figure of merit for the performance of a classifier is the area under the ROC curve (AUC), which in this notation should be minimized for optimal classification performance. Nevertheless, the AUC score only gives information on the global performance, and it could very well be that some classification algorithms perform better in the high purity range, whereas others perform better in the high efficiency range. The worst possible performance results from a classifier that randomly assigns the signal or background label to an event and corresponds to a diagonal ROC curve (i.e. completely identical discriminator distributions for signal and background events).

## 4.3 Machine Learning methods

The internal parametrization of the classifier in Eq. (4.3) depends on which type of ML algorithm is used. There exist a variety of methods, each with their advantages and shortcomings. This section will start with an overview of the more traditional ML algorithms that are available. These algorithms are typically well suited for small–scale problems, but mostly suffer from the "*curse of dimensionality*", which refers to the fact that their computational time rises (or their predictive power drops) drastically as the dimensionality of the input space becomes too large. These algorithms are therefore not preferred in HEP applications. After that Boosted Decision Trees (BDTs) will be briefly discussed given their frequent use for HEP problems. The main focus of this section however lies on a detailed explanation of (deep) neural networks,

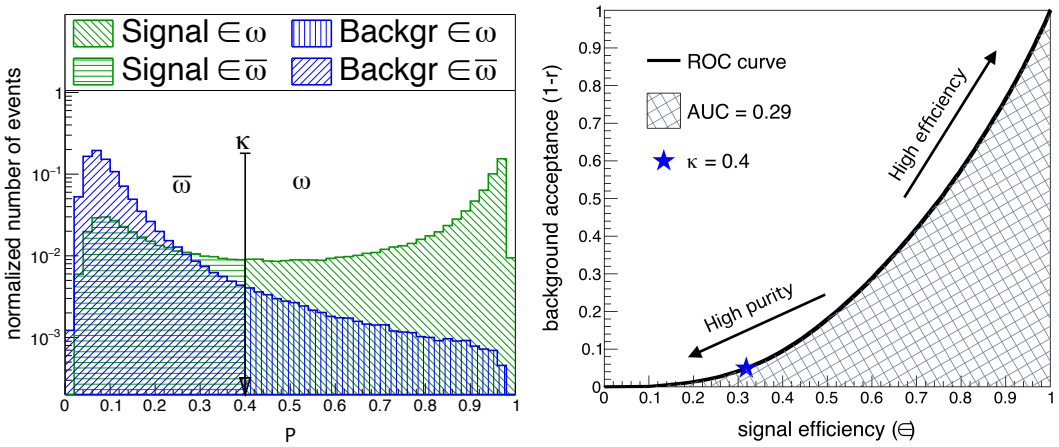


FIGURE 4.2: (left) Toy–example of the discriminator distribution for signal events (green) and background events (blue). A selection threshold is indicated at $P = \kappa$, separating the output space into two sub–spaces ($\omega$ and $\overline{\omega}$).
(right) Example of the ROC curve constructed from the discriminator distributions on the left. The star indicates the point on the curve that corresponds to the example selection threshold $\kappa$. The shaded area corresponds to the area under the ROC curve (AUC) that should be minimized for an optimal performance.

which have shown to be the most powerful algorithms for large–scale problems and are used extensively throughout the rest of this thesis.

### Classical Machine Learning methods

Below, a brief summary of some classical ML methods will be provided. These methods are however not very well suited for classification tasks with large input dimensions and are computationally not efficient enough to be trained on millions of training examples. Even though they have been used in HEP in the past, their performance is surpassed by that of BDTs and (deep) neural networks. Nevertheless, these examples often obtain a very intuitive view on solving a classification task, making it very useful to list their basic ideas here.

- **Support Vector Machines (SVM)** [245]: These algorithms aim to construct a decision boundary with the largest possible "*gap*" or separation between elements of the different classes. Kernels are used to elevate the data into a high–dimensional representation in which this separation can be more efficiently achieved. The idea originates from fully separable classes. However also when classes can not be fully separated, there exists extensions that make use of "penalty" terms in the loss function for mislabelled events.

- **Naive Bayes or Likelihood Ratios (LR)** [246]: The Neyman–Pearson lemma [247] states that the ratio of likelihoods under different hypotheses is the strongest possible test statistic to be used. It requires however an a–priori knowledge on the functional form of the underlying PDFs under the different hypotheses. This is often not the case and ML algorithms can approximate the likelihood ratio for example through the use of binned histograms of simulated events. There is a large effort within the ML community to estimate more accurately the likelihood ratios in problems where they are impossible

to retract[4] [248]. These problems are known under the name *likelihood–free inference*.

- **k–Nearest Neighbors (kNN)** [249]: The basic principle behind this type of classifiers relies on the definition of a distance metric in the input feature space. The classification is based on the class labels of the $k$ nearest (with respect to the chosen distance metric) neighboring training examples, where $k$ is a tunable parameter. This is a very intuitive type of classification: if an event is surrounded by many signal events, it is likely to be a signal event itself.

- **Decision trees and random forests (RF)** [250]: A decision tree is defined by a set of consecutive selections that each split the sample of events into two *branches*. Branches can themselves be split into other branches until the final *leaves* are reached, which are supposed to be more signal– or more background–like. This is illustrated in Fig. 4.3, where the features are denoted $x_{i,j,k,l}$ with corresponding selection thresholds $c_{1,2,3,4}$. Weights are associated to the branches that are optimized during the training phase such that the classification results in the highest possible purity in the final leaves. A random forest is an ensemble of decision trees (usually each with a limited complexity), where the discriminator output of the forest is an average taken over the individual outputs of each tree. Each individual tree uses randomly a subset of features. The use of ensemble classifiers (i.e. the average decision of multiple simple classifiers) is used to reduce the possibility of overtraining.

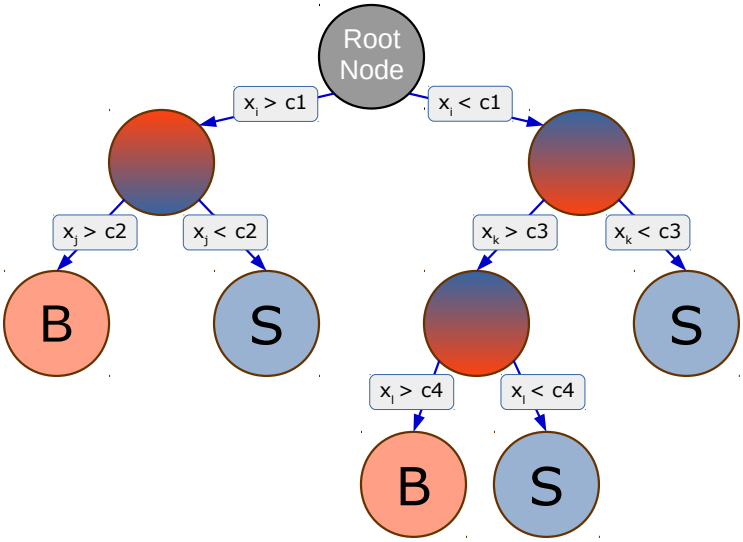


FIGURE 4.3: Illustration of the structure of a decision tree. Starting from the root node, consecutive decisions are made based on the value of the input features $x_{i,j,k,l}$, splitting the sample into different branches. The training objective is to create output leaves with the highest possible purity in either signal or background events.

This is a non–exhaustive list of classical ML algorithms and each of the algorithms mentioned above comes in different variants (based on the choice of kernel, distance

[4]For example the full simulation chain used to simulate proton–proton collisions (as outlined in Sec. 3.1) makes it completely impossible to write down a functional form of the PDF of a given observable as a function of a theoretical input parameter that entered the matrix element calculation.

metric, randomization, loss function, etc.). Below we discuss two ML classification methods that are more popular in HEP applications.

### Boosted decision trees (BDTs)

Boosted decision trees are a special application of the principle of *gradient boosting* [251]. A boosting algorithm is an ensemble technique that uses many *weak*[5] base–classifiers (often called *weak learners*) to construct a strong classifier with a reduced chance of overtraining, similar to the idea behind a random forest. However, instead of training many weak learners at the same time, boosting is an iterative procedure. The algorithm starts by training the first base–classifier which yields a certain accuracy of the classification performance. After the first base–classifier has been trained, its output is assigned a weight related to its accuracy. Additionally, the training events are also reweighed by assigning higher weights to misclassified events. This will make sure that the misclassified events become more important in the training of the next base–classifier. The algorithm proceeds iteratively, updating the event–weights such that the events which are hardest to classify correctly are given a higher and higher importance throughout the training. The first iterations are able to detect the rough features that separate signal from background, whereas the last iterations should focus on more delicate differences between the features of the different categories. The final ensemble output is a weighed average of the outputs of each base–classifier, with the weights determined by the accuracy of each individual weak learner.

A boosted decision tree is therefore simply a boosting algorithm that uses shallow decision trees as weak base–learners. The depth of each individual decision tree is usually restricted to be less than $\sim$5, and dictates the amount of correlation between the input features that can be taken into account. If the depth of each decision tree is restricted to one (meaning only one branch split is allowed), the correlation between two or more input features can never be taken into account. Given that correlations are often present and important to distinguish signal from background, the depth usually varies between two and five. If the allowed depth is chosen too large, the algorithm becomes prone to overtraining and the required computational time also increases considerably. There exist many ways to reduce the overtraining of BDTs, ranging from the use of regularization terms in the loss function, to the use of bagging techniques that randomly use a subset of the training data or only consider a random subset of the input features in each iteration.

`Application in legacy c-tagger:` The first available $c$ jet identification algorithm in CMS [252] was based on a combination of two BDTs, one to distinguish $c$ jets from light jets and one to distinguish $c$ jets from $b$ jets. The corresponding CvsL and CvsB discriminator distributions of this legacy tagger are shown on the left side in Fig. 4.4 on the top and bottom respectively. First of all it can be observed that these distributions show a very peaked behavior. This is due to the fact that the tagger substitutes default values for track–, SL– or SV–related properties in case not enough tracks pass the track selection or in case no SV or no SL was found. These unphysical peaks in the distribution however lead to unnatural discontinuities in the performance of the tagger for different WP thresholds. The one–dimensional ROC curves (considering one discriminator at a time) are shown on the right of Fig. 4.4. It can be seen that this original $c$–tagger was optimized for $c$ jet versus light jet discrimination, showing

[5] Weak refers to the restricted allowed complexity of the individual base classifier.

indeed a superior performance compared to the existing $b$–tagging algorithms at that time (CSVv2 and CMVAv2). This is shown with the blue lines. However, the CvsB discrimination was not yet optimized and indeed shows a worse performance to dedicated and more evolved $b$–tagging algorithms, as shown by the red lines.

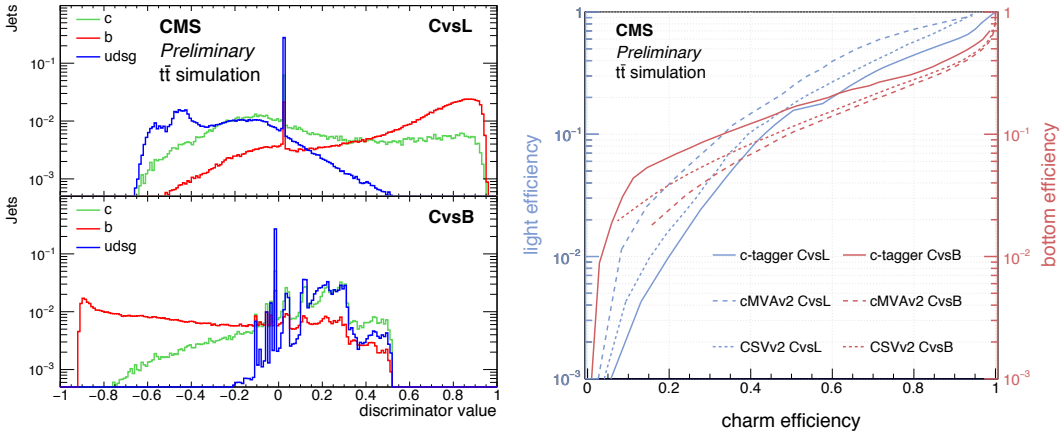


FIGURE 4.4: (left) Normalized discriminator distributions of the legacy BDT–based $c$–tagger for CvsL (top) and CvsB (bottom) discrimination. The distributions are separated by jet flavor and drawn from simulated $t\bar{t}$ events. (right) Corresponding one–dimensional ROC curves for the individual $c$–tagger discriminants compared to the performance of supported $b$–tagging algorithms at that time (CSVv2 and CMVAv2). Figures taken from Ref. [252].

All of these problems are currently circumvented in the state–of–the–art taggers by constructing both $b$– and $c$–tagging discriminators from the same multi–class tagger as explained in Sec. 3.3.3. These taggers are based on neural networks, which will be outlined in more detail in the next paragraph. The legacy BDT–based $c$–tagging algorithm is not supported anymore for 2017 data processing.

### Neural networks

The today perhaps most popular ML methods rely on the use of (artificial) neural networks (NN). These methods find their foundations already in the 1940s[6], inspired by the structure of the nervous system in the human brain. Based on the ideas pioneered by Warren S. McCulloch and Walter Pitts [254], it was Frank Rosenblatt who first came up with the concept of a *perceptron*, which is a theoretical model trying to explain how information is transferred through the neurons in the nervous system to the brain [255]. Such a model connects the input nodes to subsequent layers of multiple neurons each. The neurons of subsequent layers are fully interconnected and finally result in output nodes that can be used to make predictions on the desired pattern recognition. Due to the lack of computational power and efficient numerical optimization algorithms at that time, such NN algorithms were not very popular to be used practically. However in the last decades, the development of highly efficient back–propagation algorithms to optimize the loss function in combination with the explosion in computational power and the availability of GPUs has given neural

[6] A pleasant read on the historical aspects of the origin of chaos theory to model complex systems using artificial intelligence can be found in Ref. [253]. This book nicely explains the interactions between scientists at the Santa Fe institute, where the foundations of modern artificial intelligence were laid.

networks a prominent position in the ML landscape. In fact these developments have triggered the so–called field of *Deep Learning*, in which very large neural network architectures (tens of layers with hundreds of neurons in each layer) are used to learn patterns from a high–dimensional, low–level feature space [256].

The basic structure of a simple feed–forward and fully–connected neural network is schematically illustrated in Fig. 4.5. In this example, a $\mathcal{D}$–dimensional input features space is connected to two internal layers with multiple neurons, which are then connected to three output nodes, representing the discriminator probabilities in case of a classification task. On the right side of this figure, a closer look at one of the neurons is depicted. The basic idea behind a NN is relatively simple and starts by assigning weights ($w_i$) to each of the connections between two neurons (considering input and output nodes as neurons as well). Each neuron calculates the weighed average of its inputs ($x_i$) that originate from neurons in the previous layer that it is connected to. A bias term $b$ is added to this weighed sum, of which the purpose becomes clear in a few lines. At this point, each neuron outputs a value that is a simple linear combination of its inputs and the entire interconnected network can be seen as a linear transformation from a $\mathcal{D}$–dimensional input vector to a $\mathcal{C}$–dimensional output vector. In terms of pattern recognition, this would allow only to derive linear decision boundaries, or equivalently to derive only linear relations between input and output. The non–linearity of the network is ensured by adding a so–called *activation function* ($\sigma$) on top of the weighed sum (with bias). Such an activation function usually takes the form of a threshold function of which the strength of its response is proportional to the weighed sum of its inputs. If the weighed sum of the inputs falls below a given threshold, the response of the neuron is highly suppressed (or even taken to be 0). Once the threshold is surpassed, the signal is passed on to the next layers of the network. The bias term added to the weighed sum allows to shift that specific threshold and is therefore crucial. This results in the response of a neuron as expressed in Eq. (4.11).

$$\underline{\texttt{Response of a neuron:}} \quad \sigma\left(\sum_{i=1}^{\mathcal{D}}(x_i \cdot w_i) + b\right) \tag{4.11}$$

When training the neural network, the weights and biases are updated after each iteration by back–propagation of the derivative of the loss function such that it can be minimized. This results in a network of which certain regions will be activated in response to a certain pattern in the inputs. For a clear and detailed explanation on efficient back–propagation and optimization of the NN, the reader is referred to Ref. [243] (see specifically lecture 4).

A large diversity of activation functions is available with different properties. One important aspect of an activation function is that is has an easy analytical expression for its derivative that can be efficiently computed during the back–propagation and optimization of the network. Popular ones are the rectified linear unit (ReLu) which outputs 0 for inputs below 0 and scales linearly with the input for values larger than 0. Other popular choices are sigmoid or logistic activation functions ($\sigma(x) = [1 + e^{-x}]^{-1}$), hyperbolic tangent functions ($\sigma(x) = (e^x - e^{-x})/(e^x + e^{-x})$) or inverse tangent activations ($\sigma(x) = \tan^{-1}(x)$). A special activation is usually considered on the output layers, which are not a function of a single input, but rather combine information of all the output nodes to define the final discriminator. A popular choice for classification is the use of a *softmax* activation on the final output layer. This

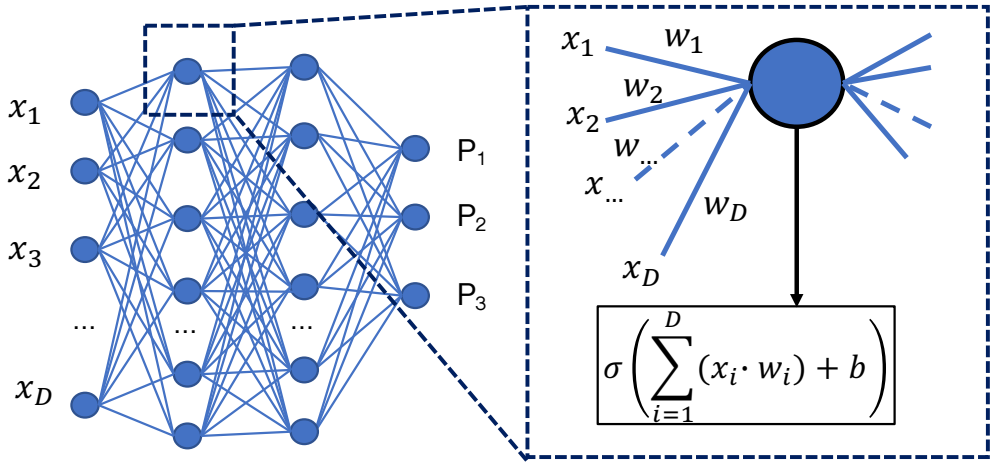


FIGURE 4.5: Schematic representation of a fully–connected, feed–forward neural network on the left, with a more detailed look into the structure and response of a single neuron on the right.

activation is defined in Eq. (4.12) and outputs a number between 0 and 1 that can roughly be interpreted as the probability to belong to any of the $\mathcal{C}$ output classes.

$$\underline{\texttt{Softmax activation:}} \quad \sigma_i(\{x_1, \cdots, x_\mathcal{C}\}) = \frac{e^{x_i}}{\sum_{j=1}^{\mathcal{C}} e^{x_j}}, \quad i \in \{1, \cdots, \mathcal{C}\} \tag{4.12}$$

Different types of network layers have been developed for specific purposes. An overview of the three most popular types of layers is given below. In practice the most advanced network architectures usually combine different types of layers to achieve the most optimal performance.

- **Fully–connected or dense layers**: This is the most basic type of layer already discussed above and consists of an ensemble of neurons that are connected to all neurons in the previous layer and to all neurons of the next layer.

- **Convolutional layers**: Convolutional neural networks (CNNs) are used mostly in the field of image recognition. It is based on the idea that the input has some spatial structure that connects neighboring features with each other. The analogy is most easily made with the pixels of an image, where neighboring pixels are related to each other in order to create larger–scale features such as edges, squares, circles or other basic shapes. Images are three–dimensional data structures with a given width, height and a "depth" that represents the base colors which are mixed in each pixel. Image recognition algorithms rely on two–dimensional convolutional layers that consist of a set of two–dimensional filters. These filters "slide" over the entire image, searching for specific patterns in the image. The first convolutional layers in a CNN therefore return a set of filters that are capable of detecting basic shapes. Consecutive layers will then combine these low–level features into high–level characteristics such as eyes, a nose and a mouth in case of a facial recognition algorithm. Typically the convolutional layers are followed by pooling layers that downsample the dimensionality of the input data into a lower–dimensional representation (for example by only taking the pixel with the maximum intensity in a window of $n \times m$ pixels, which is

known as *max pooling*). The final layers in a CNN are usually filled by simple fully–connected layers that perform the final classification based on the activated filters in the first part of the network. Even though these networks are mostly used in image–processing software, the general description of a convolutional layer is not restricted to a set of two–dimensional filters that process images. The dimensionality of the filters can be chosen freely and in a more general interpretation, the filters of a convolutional neural network are capable of building higher–level features that characterize different classes of events. Therefore one–dimensional convolutional layers are often used for *feature engineering* purposes, constructing (abstract) more powerful characterizing features from a set of low–level inputs that are provided to the network. An example will be given in the structure of the DeepFlavor $b$–tagging algorithm in Sec. 4.4.

- **Recurrent layers**: Recurrent neural networks (RNNs) contain layers which are specialized in processing input data with some type of sequential ordering. This sequential property can for example be related to chronological ordering in time, spatial ordering or a mathematical ordering from large to small values of a defining property. RNNs are typically used in language processing algorithms that rely on the order in which words appear in a sentence. The layers have an internal memory that allows the network to compare the features of a given element to the previous element in the list. Recurrent layers exist in many forms, but one of the most popular ones is the so–called *long short–term memory* (LSTM) layer. An application of this type of layers in the DeepFlavor $b$–tagging algorithm will again be given in Sec. 4.4.

Finally, some important notes are in order to keep in mind when training a neural network. First of all, it is much easier for a network to process input data with features that are distributed in a similar range. For this reason it is often advisable to apply some data pre–processing in which the mean values of each feature are shifted to zero, and the variance of the distributions of each input feature is rescaled to a value of one[7]. These scaling and shifting procedures should not affect the separating power between the feature PDFs of different classes, but simply fix the order–of–magnitude of the range in which the features are defined. Secondly, in order to avoid overtraining two things need to be considered. One has to make sure that enough training data are available and additionally the use of regularization procedures is often recommended. A popular regularization method for neural networks is known as *dropout* and consists of randomly freezing a fraction of the neurons in each layer for a given iteration, such that their weights can not be updated. Finally one should find an appropriate value for the *batch size* and the *learning rate* of the optimization algorithm that is used. The batch size defines the amount of training data that are passed through the network before it performs an update of its weights. Larger batch sizes will give more accurate estimations of the gradient of the loss function, but also lead to a large increase in computational time. The learning rate defines the size of a step in the direction that minimizes the loss function. If taken too large, the algorithm might fail to converge to an absolute minimum in the loss function, but tiny learning rates may result in a very slow learning speed, such that the absolute minimum of the loss function is not reached before all iterations are done. A decaying learning rate is often preferred which initially allows large steps in the direction of the negative gradient, but progressively lowers the learning rate over multiple iterations.

[7]There exist also *batch–normalization layers* that automatically perform this task during the training.

## 4.4 Machine Learning for heavy-flavor identification

Heavy–flavor tagging algorithms are a prime example of the use of ML methods for classification in HEP. The underlying physics principles have been discussed already thoroughly in Sec. 3.3, and this section focuses on the ML algorithms used to perform jet flavor classification in CMS. The original $b$–taggers used in Run I of the LHC were mostly based on the use of a likelihood ratio classifier [228]. At the start of Run II, the standard $b$–tagger (CSVv2) was based on the use of an ensemble of shallow neural networks, trained specifically to distinguish $b$ jets from either light or $c$ jets in the presence or absence of a reconstructed secondary vertex. These individual shallow neural networks were then combined in a likelihood ratio to obtain the final discriminator value. With the rise of Deep Learning, studies have shown that deep neural networks would be able to provide much better performance for HF identification and would allow to use a higher–dimensional and low–level input feature space [257]. In 2017 the CSVv2 algorithm was converted into a deep neural network architecture to increase its performance, resulting in the DeepCSV algorithm. Recently an even more involved network architecture (based on different types of layers) was developed under the name DeepFlavor or DeepJet. These two deep–learning classifiers are the recommended ones in CMS for data–taking from 2017 onwards and will be discussed in detail below.

### DeepCSV architecture

The DeepCSV algorithm is based on the same input observables that were used in its predecessor, CSVv2[8]. This includes seven track–related variables, 5 observables related to the SV, and seven global jet–based variables. The DeepCSV algorithm uses up to six selected tracks (ordered according to 2D IP significance) as opposed to only three in the CSVv2 algorithm. Furthermore, DeepCSV considers only the SV with the smallest uncertainty on its flight distance. This total of 54 input features are then served to a dense fully–connected NN composed of 4 layers with 100 nodes each. As discussed before it then outputs four probabilities, P($b$/$bb$/$c$/$udsg$) to belong to one of the jet flavor classes. This structure is schematically shown in Fig. 4.6. The input observables are scaled to have a zero mean and a variance of one. In case some input observables are not defined (for example when less than 6 tracks are selected or no SV was reconstructed), they are assigned a value of zero (*zero–padding*) after this scaling has been performed. The algorithm is trained on a mixture of QCD multijet events and top quark pair events to avoid a possible dependence of the tagger output on the underlying production mechanism of the jets. The tagger is trained using the Keras deep learning library [258], interfaced with TensorFlow [259] as a backend.

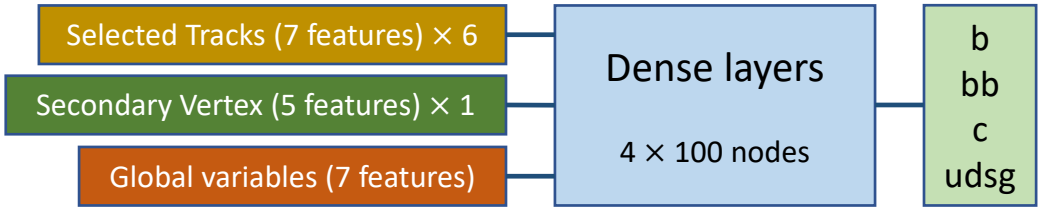


Figure 4.6: Schematic overview of the network architecture of the DeepCSV algorithm.

[8] A detailed list of these observables can be found in Ref. [225], Chapter 5.1.2.

### DeepFlavor/DeepJet architecture

The recently developed DeepFlavor algorithm [233] has a much richer architecture that results in an improved performance with respect to DeepCSV. The first striking difference with respect to DeepCSV is related to the chosen set of input features. DeepFlavor does not impose any a priori selection criteria on the tracks. In fact it uses directly all charged and neutral PF candidates that are associated to the jet. This results in the use of up to 25 charged PF candidates (again ordered according to 2D IP significance) of which 16 features are considered for each candidate. Additionally this algorithm uses for the first time the information included in 8 features of up to 25 neutral PF candidates as well. This information is extended with 12 SV–related properties of up to four reconstructed SVs in the jet and finally a set of 15 global jet–related features are added to the list. This clearly results in a much higher dimensionality of the input space in which more low–level information is used.
The second striking difference is related to the network architecture of DeepFlavor that combines different layer types to extract different kinds of information before performing the final classification. For convenience, the full structure is schematically illustrated in Fig. 4.7. For each collection of charged/neutral particles and vertices a set of $1 \times 1$ convolutional layers is trained. Charged candidates as well as the SVs are exposed to four hidden layers with 64, 32, 32 and 8 convolutional filters respectively. Neutral candidates are exposed to three hidden layers with 32, 16 and 4 filters respectively. These $1 \times 1$ filters act on each PF candidate or vertex individually and their goal is to perform some kind of feature engineering to encapsulate the information from the original 16, 8 and 12 features (for charged, neutral and SV candidates) into a set of 8, 4 and 8 abstract but lower–dimensional input features respectively. After the convolutional layers, the network is still represented by 25 charged and neutral candidates and 4 SVs, but the original input features are replaced by these abstract lower–dimensional representations in the hidden layers.
Next, these collections are fed into three separate recurrent LSTM networks with 150, 50 and 50 output nodes for charged and neutral candidates and SVs respectively. It is important that the objects in each of the three collections are ordered with respect to some sensitive observable, such that they can be interpreted as a sequentially ordered list. The goal of the recurrent layers is to detect correlations between the list of ordered objects. If for example a highly displaced charged candidate is found at the start of the list, it is very likely that another highly displaced candidate is found resulting from the decay of a $b$ or $c$ hadron. Such correlations can be exploited to learn about the properties of HF jets.
Finally the output nodes of the recurrent layers are recombined with the global jet–based features that have so far not been considered and these are collectively fed into a deep fully–connected dense network. This dense network starts with one layer composed of 200 neurons, followed by five more layers with 100 neurons each. Through a softmax activation, the final output nodes result in the discriminator probabilities $\mathrm{P}(b/bb/b_{\mathrm{lep}}/c/uds/g)$ as discussed in Sec. 3.3.

It was already shown for example in Fig. 3.17 that the architecture of DeepFlavor leads to a superior performance compared to DeepCSV. In order to motivate the architecture of DeepFlavor outlined above, dedicated studies have been performed to investigate separately the gain in performance from the increased number of inputs (without any a priori track selections), the convolutional layers and the recurrent layers. Each of these changes has shown a significant increase in performance

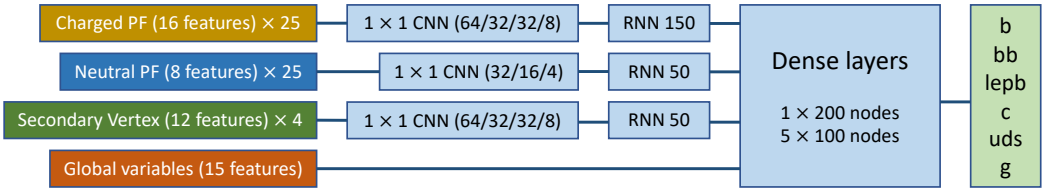


FIGURE 4.7: Schematic overview of the network architecture of the DeepFlavor/DeepJet algorithm. Figure adapted from Ref. [233].

compared to the DeepCSV inputs and architecture. As an example, the need for convolutional layers with a large dimensionality of the input feature space is demonstrated in Fig. 4.8. From the ROC curves in this figure it can be seen that the performance of DeepFlavor (red lines) degrades drastically when the algorithm is trained without the convolutional layers (blue lines) and becomes even worse than the performance of DeepCSV (green lines). Similarly, a degraded performance of DeepFlavor has been observed when the algorithm was trained without the recurrent layers (but with the convolutional layers).

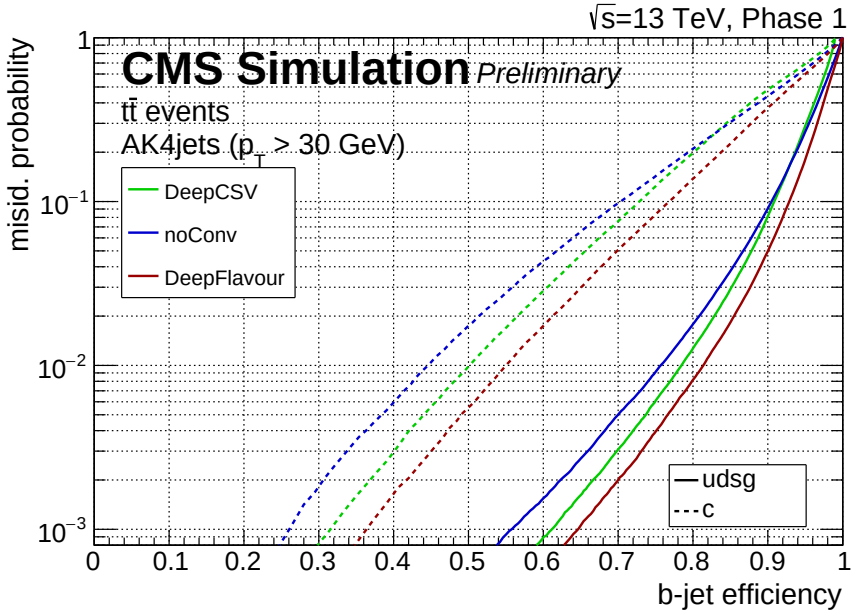


FIGURE 4.8: ROC curves for the expected $b$–tagging performance using jets with $p_T > 30$ GeV in simulated top quark pair events using the four–layer pixel geometry during Phase I. The full (dotted) lines shown the $b$ jet efficiency as function of the light ($c$) jet misidentification probability. In green (dark red) the performance of DeepCSV (DeepFlavor) is shown. For comparison, the blue line shows the degraded performance of DeepFlavor, trained without the convolutional layers. Figure taken from Ref. [260].

Additional checks were performed for the DeepFlavor algorithm to study its dependence of the random initialization of the network weights before training, as well as the size of the used training dataset. In Fig. 4.9 on the left the blue (red) band shows the $1\sigma$ variation of the DeepFlavor $b$ versus light ($c$) ROC curve for different initializations of the network weights before training. Differences are limited to

the level of 1–2%, which is a motivation for the robust character of this algorithm[9]. The complex architecture of this algorithm also dictates the need for a large training dataset in order to avoid overtraining and to be able to learn the most delicate features relevant for jet–flavor identification. The right side of Fig. 4.9 illustrates the dependence of the DeepFlavor performance on the size of the training dataset, showing that a minimum of around 100 million jets are needed to achieve an absolute optimal performance. This number perfectly justifies the need for large data storage and fast computational processing units.

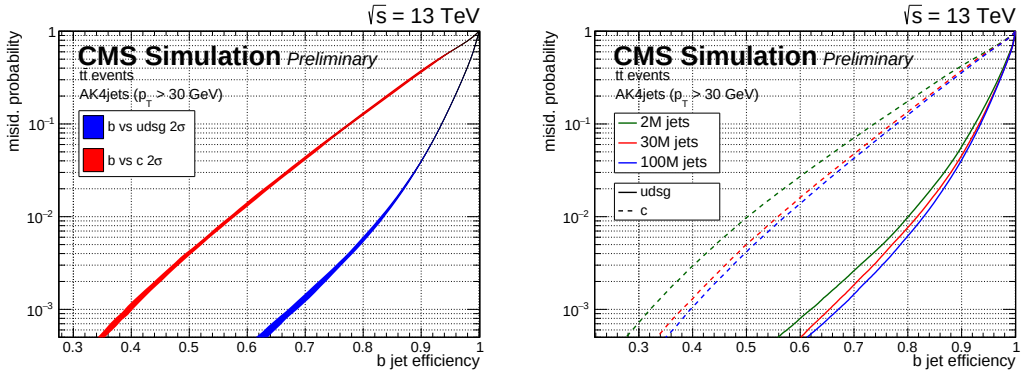


FIGURE 4.9: (left) The blue (red) band shows the $1\sigma$ variation of the DeepFlavor $b$ versus light ($c$) ROC–curve for different initializations of the network weights before training.
(right) Dependence of the DeepFlavor performance on the training sample size. Performances are evaluated on jets with $p_T > 30$ GeV from simulated top quark pair events. Figures taken from Ref. [233].

[9]The performance of DeepFlavor does not change significantly when the training is performed multiple times with different initial conditions.

# Chapter 5

# A measurement of $t\bar{t}$ production with additional charm jets

This chapter will present for the first time a measurement of top quark pair production with additional charm jets ($t\bar{t}c\bar{c}$) with the CMS detector at the LHC. Such a measurement would provide an improved background estimation in the top–Higgs sector and would complement the existing $t\bar{t}b\bar{b}$ measurements to obtain a global picture of the SM production of top quark pairs with additional HF jets.

After giving a clear motivation for this measurement in Sec. 5.1, the simulated datasets as well as the recorded datasets of proton–proton collisions in 2017 with the corresponding triggers will be outlined in Sec. 5.2 and 5.3 respectively. Then a robust signal definition will be provided to be able to classify events into categories related to the HF content of the additional jets (not from the top quark decay) in Sec. 5.4. This includes a definition in a fiducial (or "visible") phase space of the detector, as well as the full phase space. A detailed event selection is outlined in Sec. 5.5 to end up with a dataset of selected events that is very pure in dileptonic top quark pair events with at least two additional jets. After this event selection, a matching is performed between the jets and the expected generator–level partons based on a neural network trained to correctly perform this matching. This is discussed in Sec. 5.6 and provides a way to identify as accurately as possible the two additional jets which did not originate from the decay of the top quark. A novel method has been developed to calibrate the differential shape of the $c$–tagger discriminators in analogy to the existing shape reweighing procedure for $b$–tagging which was outlined in Sec. 3.3.4. This first implementation of the $c$–tagger shape calibration is performed in a control region of semileptonic top quark pair events (*i.e.* the single lepton control region) and is discussed in Sec. 5.7. Finally, an event–based neural network discriminator is constructed that combines information on the $c$–tagging discriminators of the additional jets with information on the event kinematics to extract for the first time the cross section of the $t\bar{t}c\bar{c}$ process. The presented analysis strategy is based on a template fitting method and allows for a simultaneous extraction of the $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$ + two light jet ($t\bar{t}$LF) cross sections as well as the ratios of the $t\bar{t}c\bar{c}$ and $t\bar{t}b\bar{b}$ cross section to the inclusive $t\bar{t}$ + two jet ($t\bar{t}jj$) cross section. This strategy is presented in detail in Sec. 5.8, whereas a detailed discussion on the corrections and systematic uncertainties is provided in Sec. 5.9. Finally, the results are summarized in Sec. 5.10.

## 5.1 Motivation

In 2012, the CMS [4, 5] and ATLAS [6] experiments at the LHC discovered a new scalar boson with a mass of around 125 GeV. Since then, measurements of its properties have been conducted to confirm its consistency with the Standard Model Higgs boson [75, 76]. We are therefore entering a precision era in the field of Higgs physics. In the exploration for signs of Beyond the Standard Model physics phenomena, measuring the coupling of the Higgs boson to other (massive) SM particles is of crucial importance. The interaction strength between the Higgs boson and other SM particles is predicted to be proportional to the mass of these particles (see Sec. 1.4). Therefore the Higgs boson is expected to interact the strongest with the heaviest known SM particle, namely the top quark.

The strength of the top quark Yukawa coupling is indirectly constrained by means of a global fit of the Higgs interactions [75], but recently also a direct measurement of the top quark Yukawa strength was performed by measuring the production of a Higgs boson in association with a top quark pair ($t\bar{t}H$) [9, 10]. This measurement combines different Higgs boson decay channels, of which the decay of the Higgs boson into bottom quarks has the largest branching fraction. However, this channel deals with an irreducible SM background of $t\bar{t}b\bar{b}$ and $t\bar{t}c\bar{c}$ production, for which both precise measurements and accurate theory predictions are needed in order to reliably estimate the contribution of these background processes. Measurements of the $t\bar{t}b\bar{b}$ cross section have previously been conducted by the CMS and ATLAS collaborations at center–of–mass energies of 7, 8 and 13 TeV [11–15] (see Tab. 1.4). Theory calculations for this cross section exist to NLO accuracy in QCD [170–172], but suffer from large uncertainties due to the choice of factorization ($\mu_F$) and renormalization ($\mu_R$) scales in this process which inherently posses very different energy scales (from the top quark mass to the relatively soft additional jets, resulting mostly from gluon splitting into $b\bar{b}$ and $c\bar{c}$ pairs).

At least equally important is the measurement of the $t\bar{t}c\bar{c}$ process, because it can contribute as a background to the $t\bar{t}H$ ($H \to b\bar{b}$) measurement due to misidentified charm quarks, and to a potential future measurement of the $pp \to t\bar{t}H$ ($H \to c\bar{c}$) process. In the existing analyses, this process has been estimated from simulations with a very conservative uncertainty of 50% on its rate [80,81]. With the development of charm–jet identification algorithms [225, 252], this final–state can be disentangled from the $t\bar{t}b\bar{b}$ and $t\bar{t}$LF components. In this chapter a measurement is presented that allows for the simultaneous extraction of the $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$LF cross sections, together with the ratio of the $t\bar{t}c\bar{c}$ and $t\bar{t}b\bar{b}$ cross section to the inclusive $t\bar{t}jj$ cross section. This provides a fully consistent description of the $t\bar{t}$ final–state with two additional jets and comprises the first measurement of the $t\bar{t}c\bar{c}$ cross section. This measurement is performed in the dilepton decay channel of the top quark pair and uses a data sample of $pp$ collisions collected in 2017 with the CMS detector at a center–of–mass energy of 13 TeV, corresponding to an integrated luminosity of 41.5 fb$^{-1}$.

## 5.2 Simulated datasets

Samples of simulated events are generated for the top quark pair signal, as well as for other backgrounds that may pass the event selections outlined in Sec. 5.5. The used samples with the corresponding cross sections are summarized in Tab. 5.1. In

this table also the ME generators are mentioned and the final column represents the order to which the cross section was calculated (not necessarily the order used for the ME generation).

The matrix element generation of the $t\bar{t}$ signal samples has been performed with POWHEG (v2) [163–165, 261] at next–to–leading–order (NLO), followed by a simulation of the parton shower with PYTHIA8 [167], using the CP5 underlying event tune [262] with the NNPDF3.1 [263] parton distribution function set. A top quark mass of 172.5 GeV is used in the sample generation. Additionally, a $t\bar{t}$ sample is used for which the ME generation was performed using MADGRAPH5_AMC@NLO (MG5_AMC@NLO) [166], followed by parton shower simulations using PYTHIA8 with FXFX merging [177]. This sample is used for cross checks and for training the matching neural network described in Sec. 5.6.

The backgrounds considered in this analysis consist of Drell–Yan, single top quark, W+jets, diboson, triboson, $t\bar{t}$+Z, $t\bar{t}$+W and $t\bar{t}$+H production. For all these samples, the PS is simulated using PYTHIA8. The ME generation of the W+jets and Drell-Yan processes is performed with MG5_AMC@NLO at NLO, with cross sections normalized to NNLO calculations [264]. The same accounts for the triboson, $t\bar{t}$+Z and $t\bar{t}$+W processes with the cross section determined at NLO precision. The production of $t\bar{t}$+H was performed with POWHEG. The single top quark production in the s–channel was also simulated using MG5_AMC@NLO in the four–flavor scheme, whereas the single top quark t–channel was simulated using POWHEG using also the four–flavor scheme and finally the tW channel was simulated using POWHEG in the five–flavor scheme [265, 266], with its cross section normalized to NNLO calculations [267]. The diboson samples are simulated at leading order (LO) using PYTHIA8 for the ME generation. Their cross section is normalized to that calculated at NLO (WZ and ZZ) [268] and NNLO (WW) [269].

## 5.3 Proton–proton collision datasets and triggers

This analysis focuses on the dileptonic decay channel of the top quark pairs and therefore uses the double electron (DOUBLEEG), double muon (DOUBLEMUON) and muon–electron (MUONEG) primary datasets collected by the CMS collaboration in 2017. This data–taking period was subdivided into several run periods of which runs B to F were used in this analysis. The integrated luminosities[1] corresponding to each of these runs are summarized in Tab. 5.2, resulting in a total integrated luminosity of 41.5 $fb^{-1}$.

### Triggers

A set of unprescaled[2] dilepton trigger paths were used to select dilepton events, as summarized in Tab. 5.3. In each channel, a logical OR of the trigger paths is used. The triggers used in data are also applied to the simulated events. The name of the trigger path defines the minimal $p_T$ thresholds that are imposed on the leptons to pass the trigger. As an example, the

[1] CMS summarizes all of the validated runs which are of high quality for physics analyses in a so–called "*golden JSON file*" This analysis is based on the follow golden JSON: `Cert_294927-306462_13TeV_EOY2017ReReco_Collisions17_JSON_v1`.

[2] Prescaled trigger paths systematically disregard some of the events that pass the trigger to keep the rate manageable and comparable to other trigger paths.

TABLE 5.1: List of simulated samples with the corresponding ME generator and cross section ($\sigma$). The final column represents the order to which the cross section was calculated, not the order used for the ME generation.

| **Simulated samples** | | | | |
|---|---|---|---|---|
| | Channel | ME generator | $\sigma$ [pb] | Order |
| **$t\bar{t}$ + jets** | dilepton | POWHEG | 88.29 | NNLO |
| | single lepton | POWHEG | 365.34 | NNLO |
| | fully hadronic | POWHEG | 377.96 | NNLO |
| | inclusive | MG5_aMC@NLO | 831.59 | NNLO |
| **DY + jets** | $m_{\ell\ell} > 50$ GeV | MG5_aMC@NLO | 6529 | NNLO |
| | 10 GeV $< m_{\ell\ell} <$ 50 GeV | MG5_aMC@NLO | 16270 | NLO |
| **Single top** | s–channel ($t + \bar{t}$) | MG5_aMC@NLO | 10.32 | NLO |
| | t–channel ($t$) | POWHEG | 136.02 | NLO |
| | t–channel ($\bar{t}$) | POWHEG | 80.95 | NLO |
| | tW–channel ($t$) | POWHEG | 34.91 | NNLO |
| | tW–channel ($\bar{t}$) | POWHEG | 34.91 | NNLO |
| **W + jets** | $W \to \ell\nu_\ell$ | MG5_aMC@NLO | 52940 | NNLO |
| **Diboson** | ZZ | PYTHIA8 | 12.14 | NLO |
| | WZ | PYTHIA8 | 27.6 | NLO |
| | WW | PYTHIA8 | 118.7 | NNLO |
| **Triboson** | ZZZ | MG5_aMC@NLO | 0.01398 | NLO |
| | WZZ | MG5_aMC@NLO | 0.0557 | NLO |
| | WWZ | MG5_aMC@NLO | 0.4651 | NLO |
| | WWW | MG5_aMC@NLO | 0.2086 | NLO |
| **$t\bar{t}$Z + jets** | $Z \to \ell^+\ell^-/\nu_\ell\overline{\nu}_\ell$ | MG5_aMC@NLO | 0.2432 | NLO |
| | $Z \to q\bar{q}$ | MG5_aMC@NLO | 0.5104 | NLO |
| **$t\bar{t}$W + jets** | $W \to \ell\nu_\ell$ | MG5_aMC@NLO | 0.2198 | NLO |
| | $W \to q\overline{q'}$ | MG5_aMC@NLO | 0.5269 | NLO |
| **$t\bar{t}$H + jets** | $H \to b\bar{b}$ | POWHEG | 0.5269 | NLO |
| | other H decays | POWHEG | 0.5638 | NLO |

`HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ_v*` trigger path requires the presence of at least one muon with a $p_T > 23$ GeV and at least one electron with $p_T > 12$ GeV.

TABLE 5.2: List of used data samples with the corresponding run periods and their integrated luminosity ($\mathcal{L}^{\text{int}}$).

| **Primary datasets** | **Runs** | $\mathcal{L}^{\text{int}}$ **[fb$^{-1}$]** |
|---|---|---|
| DoubleEG | Run2017 B | 4.8 |
| | Run2017 C | 9.7 |
| DoubleMuon | Run2017 D | 4.2 |
| MuonEG | Run2017 E | 9.3 |
| | Run2017 F | 13.5 |
| | **Total** | 41.5 |

A trigger logic is applied to select all events that appear in any of the trigger paths, without double–counting any of the events. Events from the DoubleEG primary dataset are selected if they pass any of the double electron triggers. Events from the DoubleMuon primary dataset are selected if they pass any of the double muon triggers but not any of the double electron triggers. Finally, events from the MuonEG primary dataset are selected if they pass any of the electron–muon triggers but not any of the double electron or the double muon triggers.

TABLE 5.3: List of used triggers in each of the dilepton channels.

| **Channel** | **Trigger paths** |
|---|---|
| double electron | `HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ_v*` |
| | `HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_v*` |
| double muon | `HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8_v*` |
| | `HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_v*` |
| electron–muon | `HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ_v*` |
| | `HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_v*` |
| | `HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*` |
| | `HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*` |

## 5.4 Signal definitions

The main objective of the analysis is to measure for the first time the cross sections of top quark pair production with two additional $c$ jets ($t\bar{t}c\bar{c}$), using the dileptonic decay channel of the top quarks. Therefore, the $t\bar{t}c\bar{c}$ process will need to be disentangled from the $t\bar{t}b\bar{b}$ and $t\bar{t}$LF final–states. A clear definition (at generator level) of these processes is needed to allow for a transparent interpretation of the result. The measurement will be reported for two different phase space definitions, the *visible phase space* and the *full phase space*, as defined below.

A typical Feynman diagram describing the production of the dileptonic $t\bar{t}c\bar{c}$ final–state is shown in Fig. 5.1. Such a final–state gives rise to two oppositely charged leptons and two $b$ jets from the top quark decays, together with two additional $c$ jets

from the gluon splitting. The neutrinos from the leptonic W boson decays result in a significant amount of missing transverse energy in the detector.

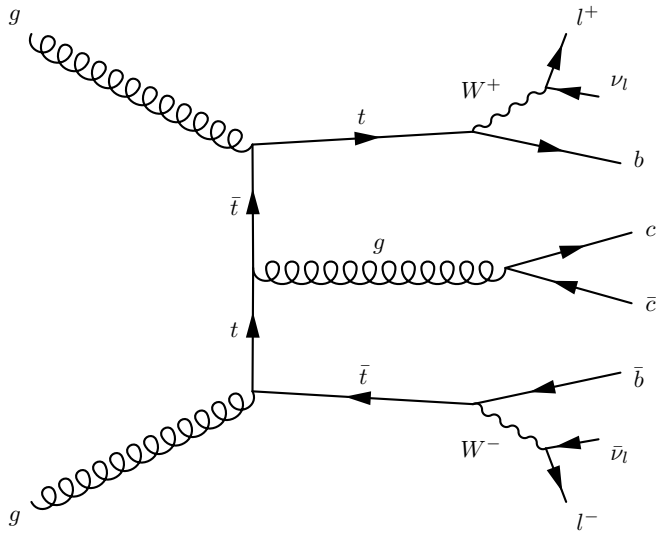


FIGURE 5.1: Example of a Feynman diagram describing the dileptonic decay channel of a top quark pair with two additional $c$ quarks produced via gluon splitting.

### Visible phase space

In the definition of the visible phase space, all of the final–state particles (except for the neutrinos) resulting from the decay chain: $pp \to t\bar{t}jj \to \ell^+\overline{\nu_\ell}b\ \ell^-\nu_\ell\bar{b}\ jj$ are requested to be within the region of the detector in which these objects can be properly reconstructed. This implies the presence of two oppositely charged leptons (including electrons, muons and tau leptons[3]) at generator level with $p_T > 25$ GeV and $|\eta| < 2.4$. Furthermore, these leptons are requested to originate from the decay of a W boson, which in turn resulted from a top quark decay. Two particle–level[4] $b$ jets from the top quark decay are requested to be present with $p_T > 20$ GeV and $|\eta| < 2.4$. Additionally, two particle–level jets are requested on top of these $b$ jets with the same kinematical requirements imposed. The above requirements define the inclusive $t\bar{t}jj$ signal, which is then further subdivided based on the HF content of the additional jets (not from the top–quark decay). Note that the jet flavor definitions were already discussed in Sec. 3.3.1. The further subdivision is defined as follows:

**$t\bar{t}b\bar{b}$:** At least two additional $b$ jets are present, each containing at least one $b$ hadron.

**$t\bar{t}bL$:** Only one additional $b$ jet is present, containing at least one $b$ hadron. This category does not contain true physical processes, but rather results from events in which two $b$ jets are merged into one (resulting in

[3]As will be outlined in Sec. 5.5, the event selections used in the analysis will only consider events with two leptons, being either electrons or muons. Tau leptons are only included in this selection if they decay leptonically into electrons or muons. Nevertheless the measured cross section will be unfolded to the phase space definition in which tau leptons are included in the W boson decays, using appropriate efficiency and/or acceptance factors calculated from simulations.

[4]Particle–level jets are clustered from final–state particles (except for neutrinos) at generator level using the same clustering algorithm as for reconstructed jets.

two clustered $b$ hadrons in one jet), or events in which one of the two additional $b$ jets is lost due to acceptance requirements.

**$t\bar{t}c\bar{c}$:** No additional $b$ jets are present, and at least two additional $c$ jets are found, each containing at least one $c$ hadron.

**$t\bar{t}cL$:** No additional $b$ jets are present, and only one additional $c$ jet is found, containing at least one $c$ hadron. This category does not contain true physical processes either, but similarly results from events in which two $c$ jets are merged into one (resulting in two clustered $c$ hadrons in one jet), or events in which one of the two additional $c$ jets is lost due to acceptance requirements.

**$t\bar{t}LF$:** No additional $b$ or $c$ jets are present.

**$t\bar{t}$+other:** Events that do not fit in any of the above categories, because they do not fulfill the acceptance requirements described in the definition of the visible phase space. These could for example be semileptonic top quark pair events with one additional fake lepton or events in which the $b$ quarks from the top quark decay are not within the fiducial detector volume.

Jets which have both $b$ and $c$ hadrons clustered inside are labelled as $b$ jets by the jet flavor definition. In the event categorization outlined above these $c$ hadrons, which are accompanied by at least one $b$ hadron in the same jet, are consequently ignored, since they most likely originate from a subsequent decay of a $b$ hadron into a $c$ hadron.

### Full phase space

The measurement in the full phase space is provided as well for easier comparison to theoretical calculations. In this phase space the presence of a top quark and a top antiquark is requested, with two particle–level $b$ jets ($p_T > 20$ GeV) that originate from the decays of the top quarks. Further requirements depend on the presence of generator–level charged leptons (including electrons, muons and tau leptons) with $p_T > 25$ GeV that result from the decay of the top quark. If none such charged leptons are found (*i.e.* the fully hadronic decay channel), at least six more particle–level jets are required with $p_T > 20$ GeV. In case exactly one such charged lepton is found (*i.e.* the semileptonic decay channel), at least four more particle–level jets are required with $p_T > 20$ GeV. Finally if two such charged leptons are found (*i.e.* the dileptonic decay channel), at least two more particle–level jets are required with $p_T > 20$ GeV. No requirement is made on the pseudorapidity of these objects. This phase space includes all decay channels of the top quark pair and is therefore not restricted to the dilepton channel. The further flavor–based categorization happens exactly the same as defined in the previous paragraph.

## 5.5 Event selection

A cut–based event selection has been developed in order to select a subset of events which consists almost exclusively of dileptonic top quark pair events with at least two additional jets. This subsample will later on be used in a fitting procedure to extract components with a different flavor content of the additional jets, as described in Sec. 5.8. This section describes the different requirements on the reconstructed objects and observables that result in this inclusive dileptonic $t\bar{t}$ final–state with at least two additional jets.

**Lepton and (b) jet multiplicity**

The topology of interest results from the following decay chain: $pp \to t\bar{t}jj \to \ell^+\overline{\nu_\ell}b\ \ell^-\nu_\ell\bar{b}\ jj$. Therefore, exactly two reconstructed oppositely–charged leptons (either electrons or muons) are required to be present. We only consider the presence of reconstructed electrons and muons and consequently no requirements are imposed on reconstructed tau leptons. Nevertheless it is possible that a tau lepton decays leptonically into electrons or muons, resulting in the inclusion of these events if the final–state electrons or muons pass the kinematic requirements outlined in this section. Motivated by the different background contributions, a distinction is made between the di–electron ($ee$), di–muon ($\mu\mu$) and electron–muon ($e\mu$) channels during the event selection. These three channels are however combined in the final extraction of the cross sections. At least four reconstructed jets are required in the event. Later on, as described in Sec 5.6, an assignment of the jets to the parton–level objects (*i.e.* the quarks) is made to identify $b$ jets from the top quark decay and jets originating from additional radiation. The two jets assigned to the $b$ quarks from the top quark decay are required to be $b$–tagged using the Medium working point of the DeepCSV $b$–tagging algorithm (see Tab. 3.2).

**Muons**

From the collection of muons reconstructed with the PF algorithm, further requirements are imposed on the selected muons in the analysis. Muons are required to have $p_T > 25$ GeV and $|\eta| < 2.4$. Additional quality requirements on the reconstruction of the muon are imposed according to the Tight muon ID [141, 214] (see Sec. 3.2.4). Muons are required to be isolated from other objects in the event, by imposing an upper threshold on the relative isolation as defined in Eq. (3.3).

**Electrons**

Electrons are required to have $p_T > 25$ GeV and $|\eta| < 2.4$. Additional quality requirements on the reconstruction of the electrons are imposed according to the Medium cut–based electron ID [209, 210] (see Sec. 3.2.3). Electrons are required to be isolated from other objects in the event, by imposing an upper threshold on the relative isolation as defined in Eq. (3.2).

**Jets**

All jets are required to have $p_T > 20$ GeV and $|\eta| < 2.4$. On the jets that are matched to the $b$ quarks from the top quark decay, a more stringent requirement is imposed of $p_T > 30$ GeV. The reason for this is that the $p_T$ spectrum of the additional jets is expected to be softer than that of the $b$ jets from the top quark decay. The additional jets are of crucial importance in order to distinguish between the $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ and $t\bar{t}$LF categories. In order to keep as much of the additional jets within the acceptance, the threshold on the $p_T$ was lowered for these jets compared to those matched to the top quark decay products. Additional quality requirements on the reconstruction of the jets are imposed according to the Tight jet ID [270]. The jets are additionally required to have meaningful values for the $b$–tagging and $c$–tagging discriminators[5]. Jets that overlap with isolated electrons or muons ($\ell$) within $\Delta\mathrm{R}(\ell, \mathrm{jet}) < 0.5$ are removed from the collection.

[5]These values sometimes are missing due to limited available information, for example when none of the tracks in the jet pass the selection criteria (see Sec. 3.3.2) to evaluate the $b$– and $c$–tagging discriminators.

**Missing transverse momentum**

The neutrinos resulting from leptonically decaying W bosons are not detected by the CMS detector, and will result in missing transverse momentum. The magnitude of this vector ($\not\!\!E_T$) is required to be above 30 GeV in the $ee$ and $\mu\mu$ channels only, in order to reduce contributions from Drell–Yan events. In the $e\mu$ channel the purity is already very high (due to a much smaller contamination from Drell–Yan processes) and no selection criterion is imposed on the $\not\!\!E_T$.

**Additional selections to reduce backgrounds**

In order to further reduce the contribution from Drell–Yan events in the $ee$ and $\mu\mu$ channels, the invariant mass of the two leptons is required to be outside of the Z boson mass window, defined by:

$$m_{\text{inv}}(\ell^+, \ell^-) \notin [m_Z - 15 \text{ GeV}, m_Z + 15 \text{ GeV}], \tag{5.1}$$

with $m_Z = 91.2$ GeV [40]. In all lepton channels, the dilepton invariant mass is also bounded from below to be larger than 12 GeV.

**Control distributions**

After this initial event selection more than 95% of the selected events originate from top quark pair events. Some control distributions are displayed in Fig. 5.2 to assess how well the data are described by the simulations. Distributions are shown for the angular separation ΔR between the two leptons (top left), the combined invariant mass of the two leptons (top right), the angular separation ΔR between the two additional jets, assigned via the matching described in Sec. 5.6, (middle left) and their combined invariant mass (middle right) and the $\not\!\!E_T$ (bottom). The different simulated contributions are shown as filled histograms and are stacked on top of each other with the data superimposed. The top quark pair contributions are split by their event category (see Sec. 5.4). The dominant backgrounds after the event selection are single top quark and Z + jets (Drell–Yan) events, which are separately shown. All other backgrounds have negligible contributions to this signal region and are added together under the label "Rare". The overflow is added to the last bin. The shaded uncertainty band includes only the statistical uncertainty due to the limited amount of simulated data. The bottom panels each time show the ratio of the data over the simulated yields. An overall good agreement is observed between the data and the simulations.

## 5.6 Jet–parton matching

The possibility to distinguish among the $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ and $t\bar{t}$LF categories relies on the correct identification of the additional jets, not coming from the decay of the top quarks. Assuming a 100% branching ratio for the decay $t \to bW$, and focusing on the dileptonic decay channel, two $b$ jets are expected from the top quark decays and at least two additional jets are required by the event selection criteria. In practice, not all the $b$ jets from the top quark decays will be reconstructed within the acceptance of the detector and if additional HF jets are present, also those do not necessarily pass the reconstruction and selection criteria. In this section, a matching procedure is described to achieve the most accurate correspondence between the jets and the expected partons (*i.e.* quarks) in the final–state. This is done by considering all

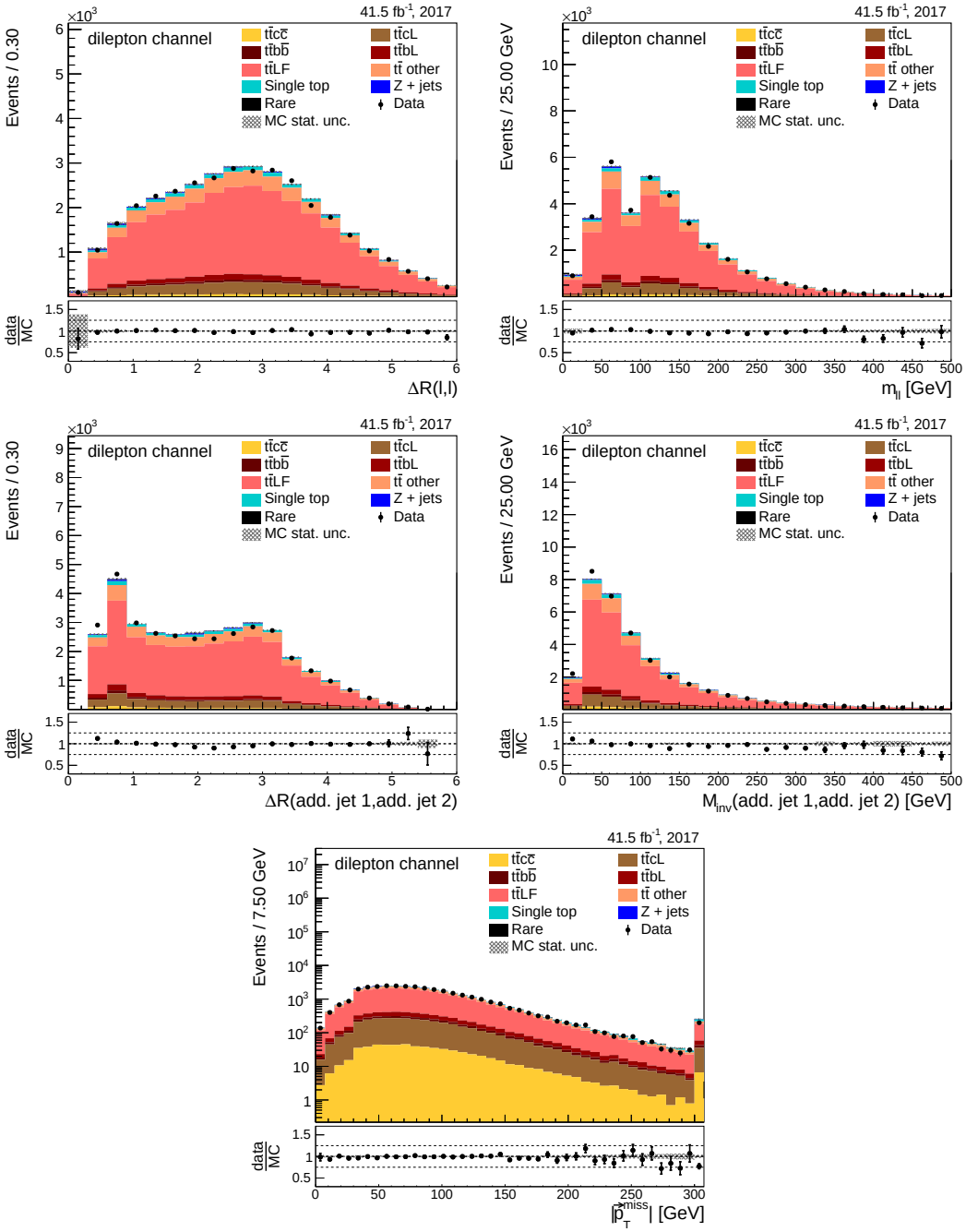


FIGURE 5.2: Control distributions in the signal region after applying the event selection outlined in Sec. 5.5. Distributions are shown for the angular separation $\Delta$R between the two leptons (top left), combined invariant mass of the two leptons (top right), angular separation $\Delta$R between the two additional jets, assigned via the matching described in Sec. 5.6, (middle left) and their combined invariant mass (middle right) and the $\not\!\!E_T$ (bottom). See text for more details.

possible permutations of four jets in the collection of jets passing the selection criteria described in Sec. 5.5 and training a neural network to identify correct and wrong permutations of jet–parton assignments.

### Input variables

Whether or not a given permutation indeed corresponds to a correct jet–parton assignment, can be inferred from different quantities such as the jet kinematics, $b$– and $c$–tagging discriminators, angular separation and invariant masses between combinations of jets (or between jets and leptons). A list of all the input variables to the neural network is shown in Tab. 5.4 and the normalized distributions are shown in Figs. 5.3 and 5.4. Before feeding these properties into the neural network, their values are rescaled such that they have a mean of zero and a variance of one over all training samples.

TABLE 5.4: Variables used in the matching neural network. The notation is as follows: $b_t$ ($b_{\bar{t}}$) denotes the jet matched to the $b$ quark from the (anti)top quark whereas $j_1$ and $j_2$ represent the first and second additional jet, ranked according to the DeepCSV $b$–tag discriminator. Finally, $\ell^{\pm}$ denotes the positively or negatively charged lepton.

| jet $p_T$ | jet $\eta$ | $b$–tag | CvsL $c$–tag | CvsB $c$–tag | $m_{inv}$ | $\Delta$R |
|---|---|---|---|---|---|---|
| $p_T(b_t)$ | $\eta(b_t)$ | BvsAll($b_t$) | CvsL($b_t$) | CvsB($b_t$) | $m_{inv}(b_t, \ell^+)$ | $\Delta$R($b_t, \ell^+$) |
| $p_T(b_{\bar{t}})$ | $\eta(b_{\bar{t}})$ | BvsAll($b_{\bar{t}}$) | CvsL($b_{\bar{t}}$) | CvsB($b_{\bar{t}}$) | $m_{inv}(b_{\bar{t}}, \ell^-)$ | $\Delta$R($b_{\bar{t}}, \ell^-$) |
| $p_T(j_1)$ | $\eta(j_1)$ | BvsAll($j_1$) | CvsL($j_1$) | CvsB($j_1$) | $m_{inv}(j_1, j_2)$ | $\Delta$R($j_1, j_2$) |
| $p_T(j_2)$ | $\eta(j_2)$ | BvsAll($j_2$) | CvsL($j_2$) | CvsB($j_2$) | | |

### Definition of correct matching

The main focus lies on correctly identifying additional HF jets in the event. In the correct assignment of the $b$ jets from the top quark decays, it does not matter which $b$ jet was matched to the top quark or to the top antiquark. If at least one additional HF ($b$ or $c$) jet is present, a correct permutation has to correctly identify these as the additional jets, in favor of other light jets in the event. It is important to note that the neural network was trained only on events for which the two generator–level $b$ quarks from the top quark decays were found to lie within $\Delta R < 0.3$ of a reconstructed $b$ jet. The neural network architecture exhibits three output nodes corresponding to the following three classes:

1. **Correct match**: the $b$ quarks from the top quark decays are correctly matched to their corresponding $b$ jets and if one or more additional HF jets were present in the event they are correctly identified as the additional jets.

2. **Flipped match**: same as the *correctly matched* category, but the $b$ jet from the top quark decay was matched to the $b$ antiquark from the top antiquark decay and vice versa.

3. **Wrong match**: Either at least one of the $b$ jets from the top quark decays was not correctly matched, or an additional HF jet was found but was not identified as one of the two additional jets.

The matching also prefers a solution in which the first additional jet has a larger DeepCSV $b$–tag discriminator value compared to the second additional jet. This is relevant to identify events from the $t\bar{t}b$L and $t\bar{t}c$L categories.

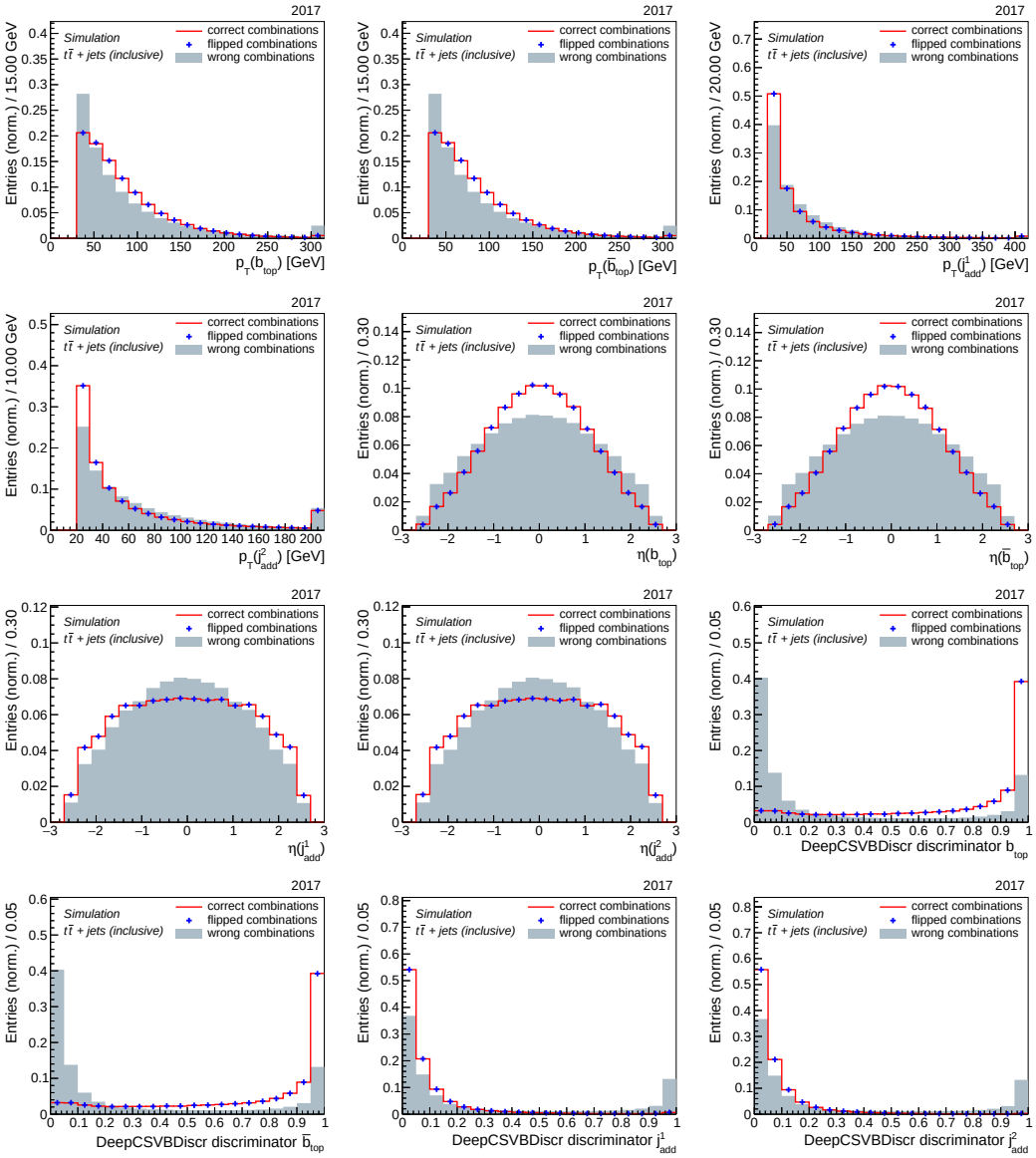


FIGURE 5.3: Input variables to the matching neural network for correct permutations (red line), flipped permutations (blue crosses) and wrong permutations (gray area). See text for definitions of the correct, flipped and wrong matching.

## Training of the neural network

The matching neural network was trained with the KERAS deep learning library [258], interfaced with TENSORFLOW [259] as a backend. The training was performed on the inclusive simulated $t\bar{t}$ dataset (with the MG5_aMC@NLO ME generator) from Tab. 5.1, which is not used for the final extraction of the results later on. The 26 input nodes (corresponding to the input variables listed in Tab. 5.4) are linked to two fully–connected hidden layers with each 50 neurons and with a ReLu activation. A dropout layer is added after each hidden layer which randomly freezes 10% of the neurons in these inner layers in every training batch to avoid overfitting. This layer is connected to the 3 outputs with a softmax activation such that the outputs sum up to one. A categorical cross–entropy loss function is used and the minimization of this loss function is performed with a stochastic gradient descent set to an initial learning rate of 0.001 and a decay of $2 \times 10^{-6}$. The training is performed in batches of 128

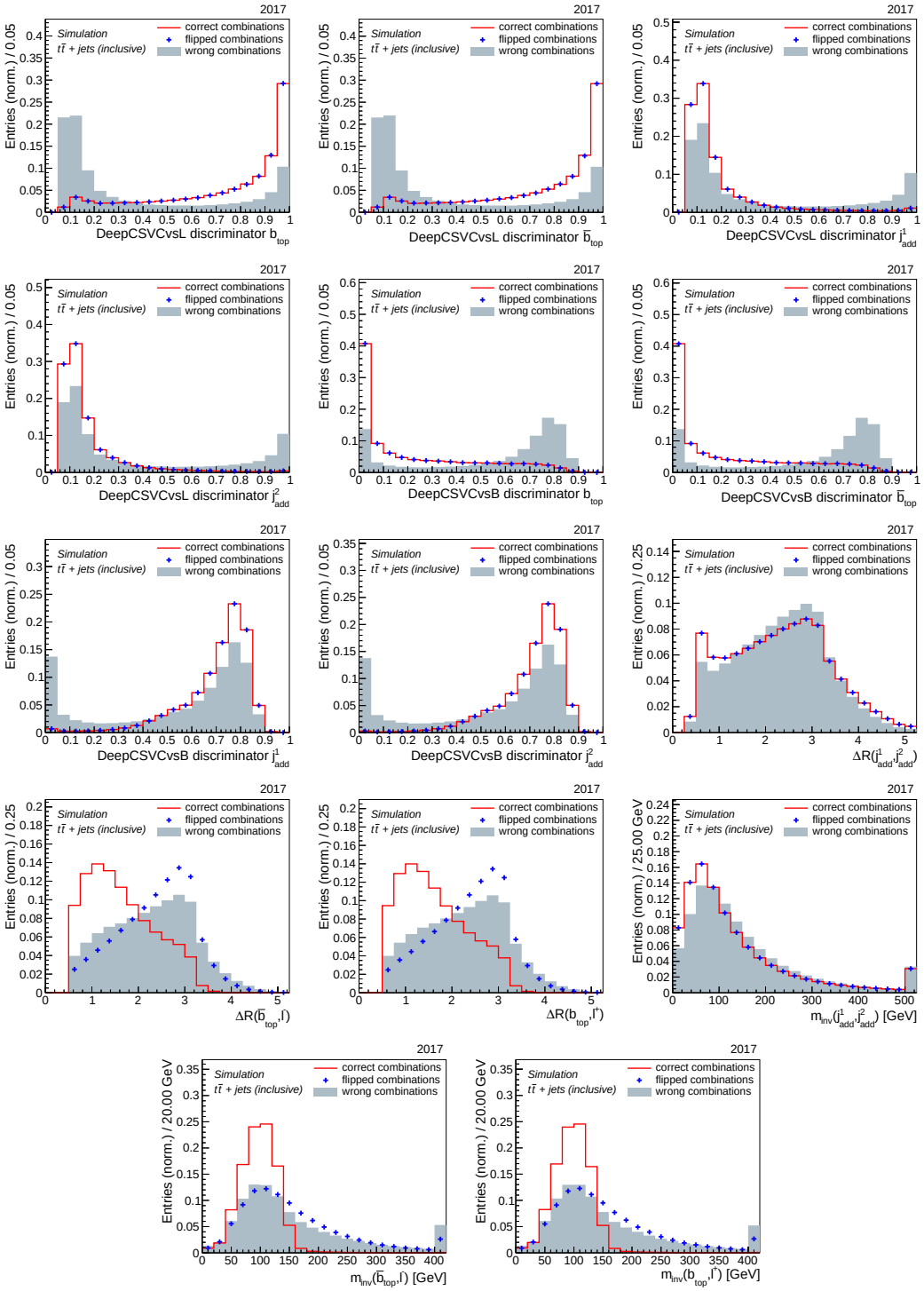


FIGURE 5.4: continuation of Fig. 5.3

events and is stopped after 100 epochs. During the first 50 epochs, events are given weights according to the event weights used in the analysis (see Sec. 5.9), including weights to correct for $b$–tagging and $c$–tagging calibrations, electron identification and reconstruction, muon identification and isolation and pileup corrections. After 50 epochs, different event categories (defined by the flavor of the additional jets) are given different additional weights such that the network does not favor the most

abundant category, but rather treats all event categories on the same footing[6]. These weights are summarized in Tab. 5.5. The training curve is shown in Fig. 5.5, showing a convergence of the loss function and the accuracy to a plateau both for training (red) and validation (blue) datasets. The top panel shows the accuracy whereas the bottom panel shows the value of the loss function.

TABLE 5.5: Additional event weights applied to the different categories during the last 50 epochs of the training of the matching neural network.

| **Training Weights (epoch > 50)** | $t\bar{t}b\bar{b}$ | $t\bar{t}b$L | $t\bar{t}c\bar{c}$ | $t\bar{t}c$L | $t\bar{t}$LF |
|---|---|---|---|---|---|
| | 20 | 10 | 20 | 5 | 1 |

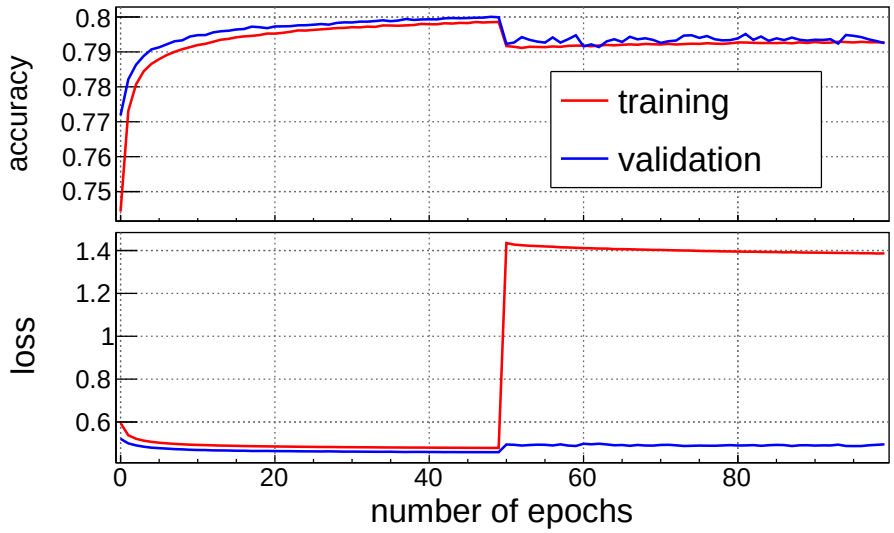


FIGURE 5.5: Training curves of the matching neural network training, displaying the evolution of the accuracy (top) and value of the loss function (bottom) for increasing number of epochs. These curves are shown both for the training (red) and for an independent validation data set (blue). The jump in the loss and accuracy after 50 epochs is a consequence of the application of the weights in Tab. 5.5 in the training sample.

### Evaluation of the performance

The performance of the matching is evaluated on an independent $t\bar{t}$ + jets sample using again only events for which the two generator–level $b$ quarks from the top quark decays were found to lie within $\Delta R < 0.3$ of a reconstructed $b$ jet. This constitutes $\sim$76% of all selected events before the matching[7], meaning that for one event in four a proper identification of the additional HF jets is very unlikely independent of how well the NN performs. A comparison is made between the multivariate approach using the matching neural network, and a simple approach in which the jets are

[6]The values of these weights are roughly estimated from simulations, based on the abundance of events in each of the categories.

[7]The selected events before matching are required to pass all selections outlined in Sec. 5.5 except for the $b$–tagging requirement on the two jets matched to the $b$ quarks from the top quark decay (which are not yet identified at this stage) as well as the elevated $p_T$ threshold of 30 GeV on these jets.

ranked according to decreasing b–tag discriminator value[8].

The trained neural network will provide for each permutation of jets three output values corresponding to the *Correct* ($\mathrm{P}_C$), *Flipped* ($\mathrm{P}_F$) and *Wrong* ($\mathrm{P}_W$) categories. As already discussed before, both the *Correct* and *Flipped* assignments are considered appropriate for the goal of this analysis. Therefore the choice was made to pick the permutation with the highest value of the quantity

$$\max\left(\frac{\mathrm{P}_C}{\mathrm{P}_C+\mathrm{P}_W}, \frac{\mathrm{P}_F}{\mathrm{P}_F+\mathrm{P}_W}\right). \tag{5.2}$$

The final performance of the matching is shown in Fig. 5.6, using the naive $b$–tag discriminator ranking on left and using the matching NN on the right. The quality of the matching from the chosen permutation is further subdivided in 5 categories:

1. **correct**: The $b$ quarks from the top quark decays are correctly matched to their corresponding $b$ jets and if one or more additional HF jets were present in the event they are correctly identified as the additional jets.

2. **flipped**: Same as the *correctly matched* category, but the $b$ jet from the top quark decay was matched to the $b$ antiquark from the top antiquark decay and vice versa.

3. **one/two**: The $b$ jets from the top quark decays are either correctly matched or flipped, but two or more additional HF jets were present of which only one was correctly identified as one of the two additional jets. The other HF jet was by mistake replaced by an additional light jet in the event.

4. **only top**: The $b$ jets from the top quark decays are either correctly matched or flipped, but one or more additional HF jets were present of which none were correctly identified as being one of the two additional jets. All additional HF jets were replaced by light jets in the event in the assignment.

5. **wrong**: At least one of the $b$ jets from the top quark decays was not correctly matched.

The fraction of jets belonging to either of these categories is depicted in Fig. 5.6 in different colors for each of the event categories (on the x–axis). It can be seen that the $t\bar{t}$LF category does not benefit much from the neural network matching, since the $b$–tagging information is already the dominating factor to choose the proper assignment in these events. However, for the categories that have at least one additional HF jet, the neural network matching has a clear increased efficiency in choosing the correct permutation. The matching neural network shows an absolute improvement of up to $\sim$ 15% in matching efficiency in some categories. It can for example be seen that for around 50% of the $t\bar{t}c\bar{c}$ events, a correct (or flipped) permutation is chosen and indeed the two additional $c$ jets are correctly identified. Additionally in around 20% of the cases one out of two additional $c$ jets is correctly identified while the other is lost. This is a clear improvement to the $b$–tag ranking procedure, in which only 40% is correctly matched (or flipped) and only an additional $\sim$ 10% recovers only one out of two additional $c$ jets. This will be important for an improved discrimination during the fitting procedure outlined in Sec. 5.8.

[8]This approach was used in previous $t\bar{t}b\bar{b}$ analyses. With this approach one considers only one permutation, and assigns the jets with the two highest b–tag discriminator values to the top quark decays and the remaining ones to the additional jets.

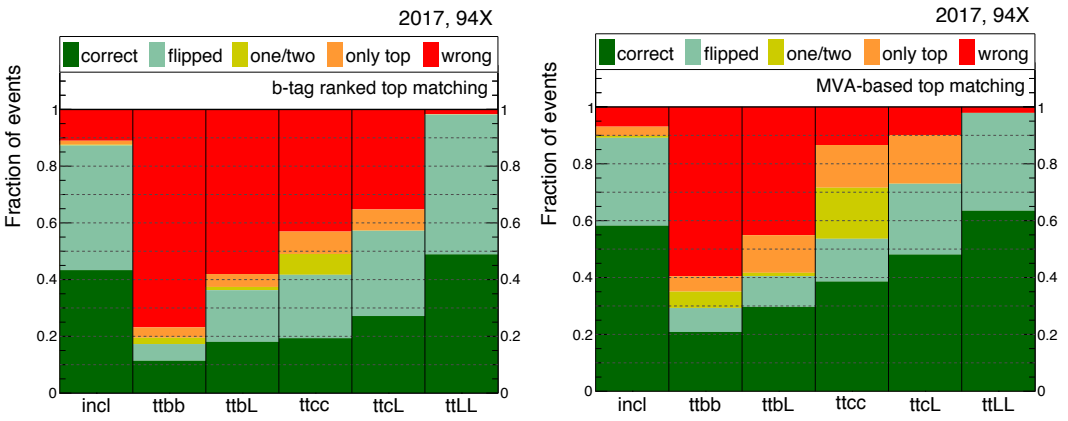


FIGURE 5.6: Performance of the jet–parton matching, using a top–down ranking of the jets according to their $b$–tag discriminator value (left) and using the matching neural network (right). The performance is evaluated on an independent simulated $t\bar{t}$ + jets sample using only events for which the two generator–level $b$ quarks from the top quark decays were found to lie within $\Delta R < 0.3$ of a reconstructed $b$ jet. See text for definitions of the different colors.

Finally, the data–to–simulation comparison of the NN output of the best permutation in each event is shown in Fig. 5.7. It can be seen that the shape of this output in data is well described by the simulation.

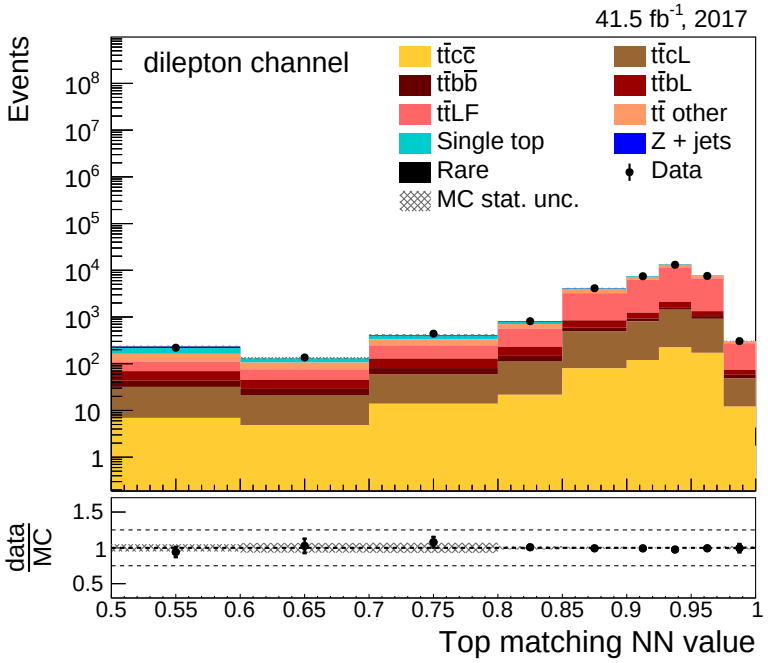


FIGURE 5.7: Data–to–simulation comparison for the distribution of the matching NN output defined in Eq. (5.2) for the best permutation found in each event.

## 5.7 Charm–tagger calibration

Similarly as for $b$–tagging, also $c$–tagging discriminators require calibration to data given that these algorithms are trained on simulated events that do not necessarily

reproduce exactly the observations in real data. This is illustrated in Fig. 5.8, showing that the DeepCSV CvsL and CvsB $c$–tagger distributions for the two additional jets in the dilepton signal region are not perfectly described by the simulations. The simulated yields are normalized to the yields observed in data to focus on the shape difference. These discrepancies show a clear slope in the data–to–simulation ratio and can be as large as 25%. Considering that these $c$–tagger discriminators will be used later on in a fitting procedure (see Sec. 5.8), it is not sufficient to calibrate the selection efficiencies for a single working point. Instead the entire differential shape of the two–dimensional (CvsL – CvsB) distribution needs to be corrected to reproduce the observed shape in data. In this section, a new shape calibration method is proposed, using a single lepton control region enriched in semileptonic top quark pair events in which one top quark decays leptonically, while the other decays hadronically. By making use of the $b$ jets from the top quark decays and the $c$ and light jets from the hadronic W boson decay, shape corrections are derived such that they can later on be applied in the dileptonic $t\bar{t}$ + two jets signal region of the $t\bar{t}$+HF analysis.

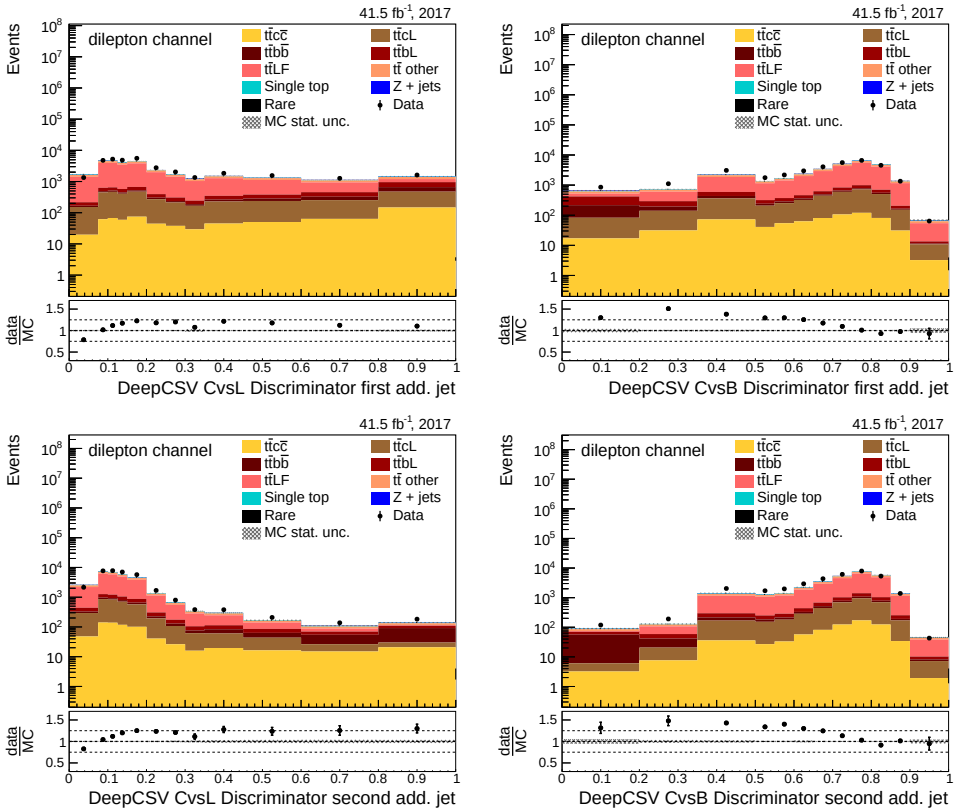


FIGURE 5.8: Data–to–simulation comparisons for the CvsL (left) and CvsB (right) $c$–tagging discriminator distributions for the first (top) and second (bottom) additional jet in the dilepton signal region. The total yield in the simulation is normalized to the data yield.

### 5.7.1 Single lepton control region

The $c$–tagger calibration is performed in a single lepton control region. In this section the details of this control region are discussed. The single electron and single muon primary datasets are used with corresponding single lepton triggers. The datasets

and trigger paths are summarized in Tab. 5.6 and Tab. 5.7 respectively.

TABLE 5.6: List of used data samples in the single lepton control region with the corresponding run periods and their integrated luminosity ($\mathcal{L}^{\text{int}}$).

| **Primary datasets** | **Runs** | $\mathcal{L}^{\text{int}}$ **[fb$^{-1}$]** |
|---|---|---|
| SINGLE MUON | Run2017 B | 4.8 |
| | Run2017 C | 9.7 |
| | Run2017 D | 4.2 |
| SINGLE ELECTRON | Run2017 E | 9.3 |
| | Run2017 F | 13.5 |
| | **Total** | 41.5 |

TABLE 5.7: List of used trigger paths that are applied in the single lepton control region.

| **Channel** | **Trigger paths** |
|---|---|
| Single electron | `HLT_Ele35_WPTight_Gsf_v*` |
| Single muon | `HLT_IsoMu27_v*` |

Furthermore, a cut–based event selection is defined below that aims to extract a sample enriched in semileptonic $t\bar{t}$ events. This event selection leads to dominant background contributions of single top quark and W+jets events. The presence of these backgrounds does not affect the calibration method which is based on the flavor composition of the jets, as explained below.

**Lepton and ($b$) jet multiplicity.**

The topology of interest results from the following decay chain: $pp \to t\bar{t} \to \ell^{\pm}\overset{(-)}{\nu}_{\ell}\, q\overline{q'}\, b\bar{b}$. Therefore, exactly one charged lepton (either an electron or muon) is required to be present. Leptonically decaying $\tau$ leptons into an electron or muon are again included in this selection if the corresponding final–state electrons or muons pass the kinematic selections outlined in this section. Additionally, exactly four jets are requested, and they are ranked according to a decreasing value of the DeepCSV $b$–tagging discriminator. The one with the highest $b$–tagging discriminator is required to be $b$–tagged using the Tight DeepCSV working point. The next three jets in the ranked list (excluding the tightly $b$–tagged one) will be used further along in the calibration method. No additional $b$–tagging requirements are imposed on any of these jets in order not to bias the measurement.

**Muons**

Muons are required to have $p_T > 35$ GeV and $|\eta| < 2.4$. Additional quality requirements on the reconstruction of the muon are imposed according to the Tight ID [141, 214]. The same isolation requirements are imposed as defined in Eq. (3.3). If a muon is found with the above mentioned kinematical selections, no additional leptons are allowed with $p_T > 25$ GeV and $|\eta| < 2.4$ in order to avoid overlap with the dilepton signal region.

**Electrons**

Electrons are required to have $p_T > 35$ GeV and $|\eta| < 2.4$. Additional quality requirements on the reconstruction of the electron are imposed according to the Medium cut–based ID [209, 210]. The same isolation requirements are imposed as defined in Eq. (3.2). If an electron is found with the above mentioned kinematical selections, no additional leptons are allowed with $p_T > 25$ GeV and $|\eta| < 2.4$ in order to avoid overlap with the dilepton signal region.

**Jets**

All jets are required to have $p_T > 30$ GeV and $|\eta| < 2.4$. Jets have to pass the Tight jet ID [270] and are additionally required to have meaningful values for the $b$–tagging and $c$–tagging discriminators. Jets that overlap with isolated electrons or muons within $\Delta R(\ell, \text{jet}) < 0.5$ are removed from the collection.

**Missing transverse momentum**

The missing transverse momentum is required to be above 20 GeV.

**Additional selections to reduce backgrounds**

The combination of the cut on the missing momentum, the presence of an isolated lepton and tight $b$–tagging requirements already reduces the contributions from QCD multijet events to a negligible level. The dominant remaining backgrounds are W+jets, single top quark production and potentially dileptonic top quark pair events of which one lepton was not reconstructed. Nevertheless these backgrounds stay below 10% and do not compromise the methodology outlined in the next section. It should also be noted that after this event selection there may still be some events from the signal region (*i.e.* dileptonic top quark pair events with two additional jets) end up in this selected sample of events. Dedicated checks confirmed that this fraction of events stays below 5%.

The data–to–simulation comparisons for the CvsL and CvsB discriminator distributions of the jet with the second, third and fourth highest $b$–tagging discriminator value in the single lepton control region are shown in Fig. 5.9. These distributions show similar discrepancies in their shapes as those observed for the distributions in the dilepton signal region (see Fig. 5.8).

### 5.7.2 Methodology: iterative fit

The semileptonic $t\bar{t}$ topology exhibits a mixture of jet flavors, including $b$ jets from top quark decays, together with $c$ and light jets from hadronic W boson decays[9] or from extra radiation. The methodology to calibrate the shape of the $c$–tagging distributions relies on this mixture of jet flavors. The goal is to select a subsample of jets enriched in $b$ jets, another exclusive subset enriched in light jets and finally a third subset enriched in $c$ jets.

For each event that passes the event selection outlined in Sec. 5.7.1, the jets are ranked according to decreasing values of their DeepCSV $b$–tagging discriminator. As already mentioned, the first one in the list (the one with the highest $b$–tagging discriminator value) is required to be $b$–tagged using the Tight DeepCSV working

[9] Around 25% of jets from hadronic W boson decays are expected to be $c$ jets.

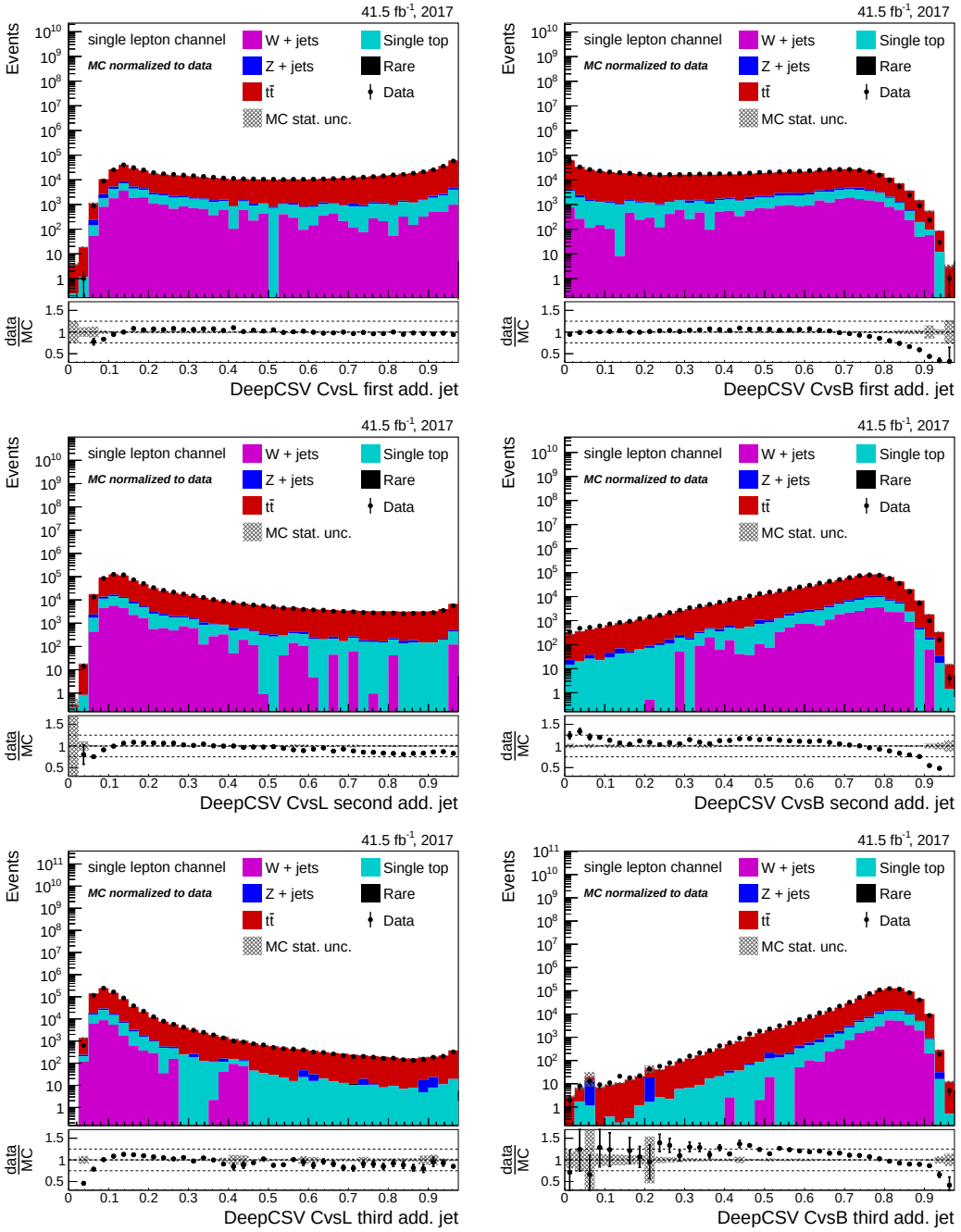


FIGURE 5.9: Data–to–simulation comparisons for the CvsL (left) and CvsB (right) $c$–tagging discriminator distributions for the jet with the second (top), third (middle) and fourth (bottom) highest $b$–tag discriminator value in the events in the single lepton control region. The total yield in the simulation is normalized to the data yield.

point and is not considered anymore in what follows. In order to make sure that the shape of the $b$–tagging discriminator is correctly described in the simulation, a priori $b$–tagging shape correction factors are applied to the events based on the procedure described Sec. 3.3.4. This will also ensure a proper selection efficiency of the Tight $b$–tagging requirement. The subsequent calibration of the $c$–tagger shapes will be derived on top of this $b$–tagging shape calibration. The three

consequent jets in the list are referred to as the first, second and third additional jet of the event. The first additional jet has a larger $b$–tagging discriminator value than the second additional jet, which has a larger $b$–tagging discriminator value than the third additional jet. With this assignment, the first additional jet has a large probability to be one of the $b$ jets from the top quark decay and is therefore largely enriched in $b$ jets. The third additional jet on the other hand is almost exclusively composed of light jets. The second additional jet has a relative large component in $c$ jets from hadronically decaying W bosons. This enrichment is illustrated in Fig. 5.10, where the CvsL and CvsB $c$–tagging discriminator distributions are shown for the first, second and third additional jet in the event, comparing predictions from simulated events (subdivided in jet flavor) to data.

The goal is now to mitigate the observed data–to–simulation discrepancies by applying bin–per–bin and flavor–dependent scale factors to the jets. To this end, two–dimensional CvsL–CvsB distributions are constructed for the first, second and third additional jet. A proper binning was chosen to make sure each bin contains enough jets of each flavor. As shown in Fig. 5.11, $b$ jets are dominantly expected in the lower right corner (of the first additional jet) of this two–dimensional phase space, whereas $c$ jets have a large fraction in the upper right corner (for the second additional jet) and light jets mostly occupy the top left corner (for the third additional jet).

The methodology is based on an iterative fit approach, in which on a bin–per–bin basis a flavor–dependent scale factor is extracted that can be applied to the simulated events to rescale the distributions and match those observed in data. These scale factors should alter only the shape of the distributions, without modifying the overall yield of events. Therefore the integrated yield in the simulation is normalized to match that observed in data. Any discrepancy in the overall normalization is therefore not taken into account with this method, and instead the focus lies on correcting the differential shape. In each bin $i$, a set of three scale factors (SFs) is determined, one to be applied to $b$ jets ($\mathrm{SF}_b^i$), one for $c$ jets ($\mathrm{SF}_c^i$) and one for light jets ($\mathrm{SF}_\ell^i$). Introducing some notation: given a certain bin $i$, let $N_{b_k}^{\mathrm{MC},i}$, $N_{c_k}^{\mathrm{MC},i}$ and $N_{\ell_k}^{\mathrm{MC},i}$ ($k \in 1,2,3$) be the number of predicted $b$, $c$ and light jets respectively in simulation for additional jet $k$. Furthermore, let $N_{j_k}^{\mathrm{Data},i}$ be the total number of observed jets (for the considered bin) in data for additional jet $k$. An initial guess for each of the three SFs is taken as the overall normalization of data yield over simulated yield in that bin. Then, each iteration proceeds in four steps:

**Step 1:** The third additional jet is highly enriched in light jets. The small contamination from $b$ and $c$ jets is subtracted using the best estimation for $\mathrm{SF}_b^i$ and $\mathrm{SF}_c^i$ at that stage and the light–flavor SF, $\mathrm{SF}_\ell^i$, is determined by minimizing the $\chi^2_\ell$ defined in Eq. 5.3 with respect to $\mathrm{SF}_\ell^i$, keeping $\mathrm{SF}_b^i$ and $\mathrm{SF}_c^i$ frozen at this stage. The uncertainty in the denominator, $\delta N_{\ell_3}^i$, is taken to be the statistical uncertainty on the observed number of jets in data, after subtraction of the $b$ and $c$ jet contamination.

$$\chi^2_\ell(\mathrm{SF}_\ell^i) = \frac{\left(\mathrm{SF}_\ell^i \cdot N_{\ell_3}^{\mathrm{MC},i} - \left[N_{j_3}^{\mathrm{Data},i} - \mathrm{SF}_b^i \cdot N_{b_3}^{\mathrm{MC},i} - \mathrm{SF}_c^i \cdot N_{c_3}^{\mathrm{MC},i}\right]\right)^2}{\left(\delta N_{\ell_3}^i\right)^2} \tag{5.3}$$

**Step 2:** The first additional jet is highly enriched in $b$ jets. The contamination from $c$ and light jets is subtracted using the best estimation for $\mathrm{SF}_c^i$ and $\mathrm{SF}_\ell^i$ at

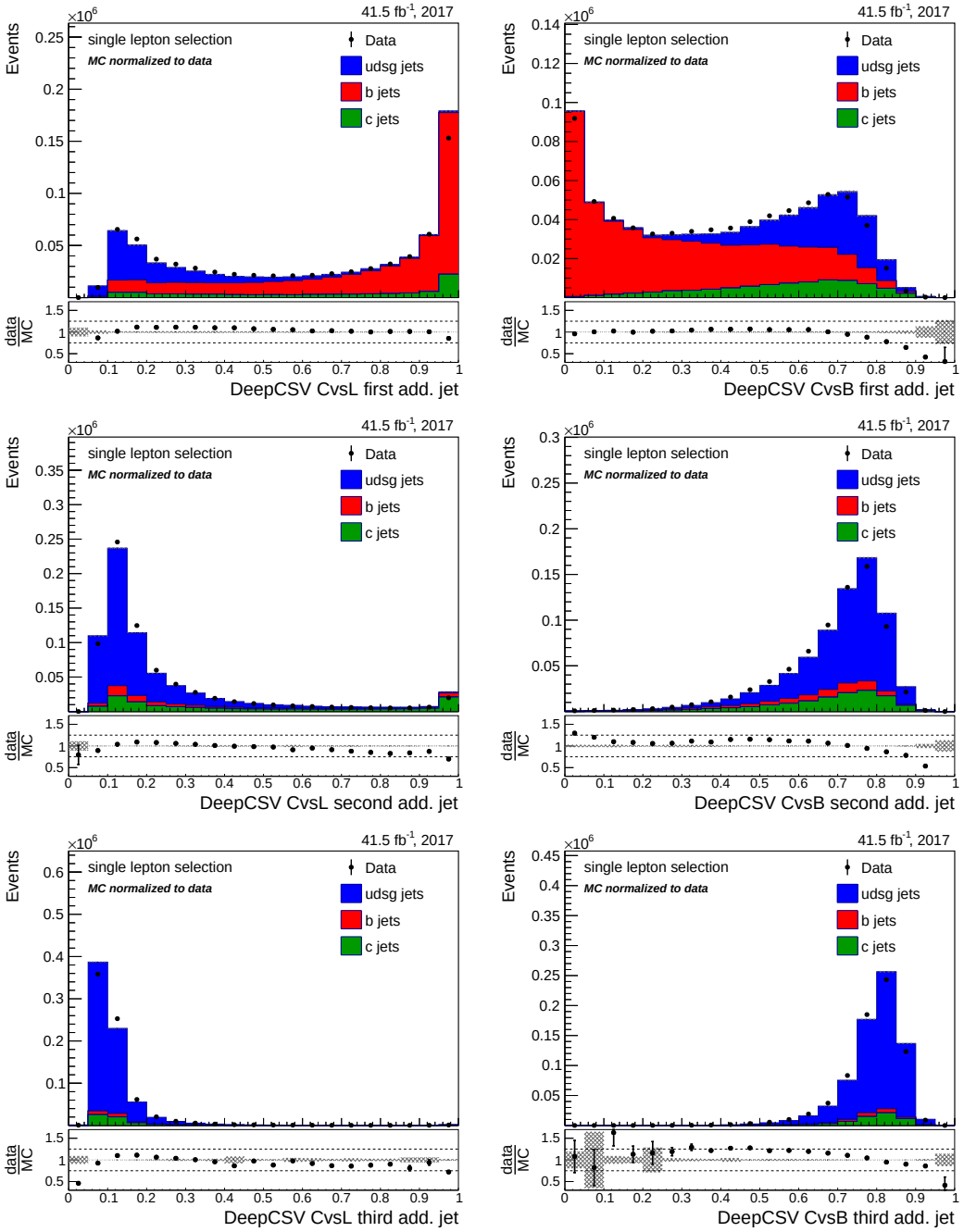


FIGURE 5.10: Data–to–simulation comparisons for the CvsL (left) and CvsB (right) $c$–tagging discriminator distributions for the first (top), second (middle) and third (bottom) additional jet in the events in the single lepton control region. The simulated events are subdivided by jet flavor to show the enrichment of specific flavors for each of the additional jets. The total yield in the simulation is normalized to the data yield.

that stage and the $b$ jet SF, $\mathrm{SF}_b^i$, is determined by minimizing the $\chi_b^2$ defined in Eq. 5.4 with respect to $\mathrm{SF}_b^i$, keeping $\mathrm{SF}_c^i$ and $\mathrm{SF}_\ell^i$ frozen at this stage. The uncertainty in the denominator, $\delta N_{b_1}^i$, is taken to be the statistical uncertainty on the observed number of jets in data, after subtraction of the

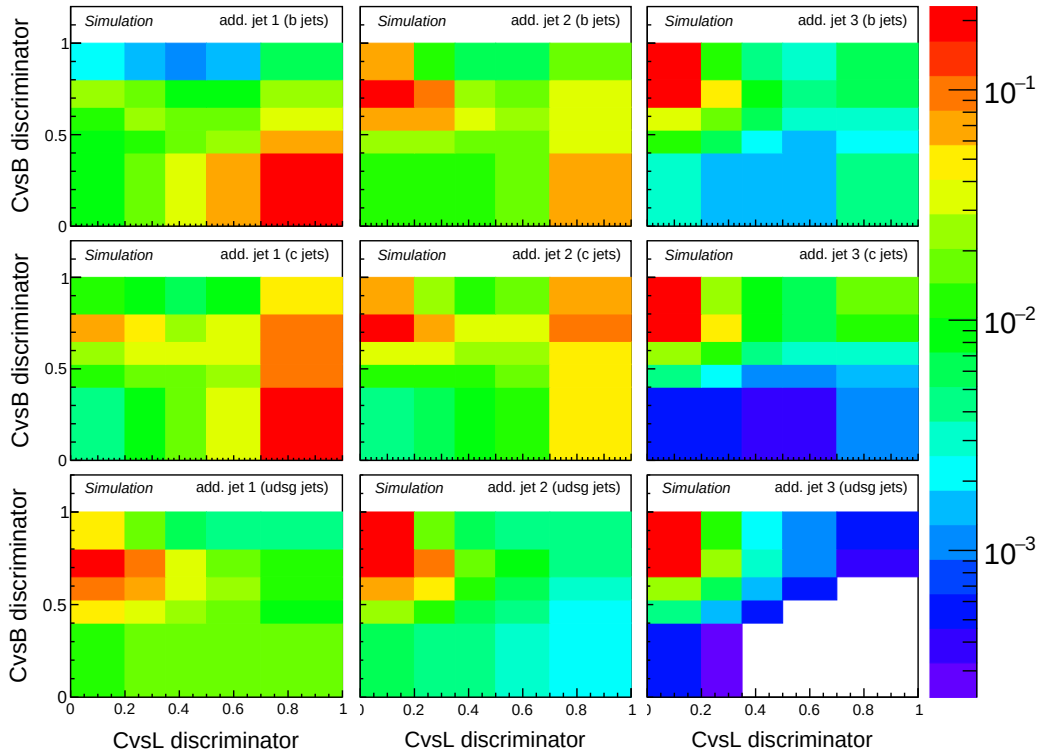


FIGURE 5.11: Normalized two–dimensional CvsL–CvsB distributions of the first (left), second (middle) and third (right) additional jet in the single lepton control region, constructed from simulated $t\bar{t}$ events for $b$ jets (top), $c$ jets (middle) and light jets (bottom).

$c$ and light jet contamination.

$$\chi^2_b(\mathrm{SF}^i_b) = \frac{\left(\mathrm{SF}^i_b \cdot N^{\mathrm{MC},i}_{b_1} - \left[N^{\mathrm{Data},i}_{j_1} - \mathrm{SF}^i_\ell \cdot N^{\mathrm{MC},i}_{\ell_1} - \mathrm{SF}^i_c \cdot N^{\mathrm{MC},i}_{c_1}\right]\right)^2}{\left(\delta N^i_{b_1}\right)^2} \quad (5.4)$$

**Step 3:** Finally, the second additional jet has a significant fraction of $c$ jets. The contamination from $b$ and light jets is subtracted using the best estimation for $\mathrm{SF}^i_b$ and $\mathrm{SF}^i_\ell$ at that stage and the $c$ jet SF, $\mathrm{SF}^i_c$, is determined by minimizing the $\chi^2_c$ defined in Eq. 5.5 with respect to $\mathrm{SF}^i_c$, keeping $\mathrm{SF}^i_b$ and $\mathrm{SF}^i_\ell$ frozen at this stage. The uncertainty in the denominator, $\delta N^i_{c_2}$, is taken to be the statistical uncertainty on the observed number of jets in data, after subtraction of the $b$ and light jet contamination.

$$\chi^2_c(\mathrm{SF}^i_c) = \frac{\left(\mathrm{SF}^i_c \cdot N^{\mathrm{MC},i}_{c_2} - \left[N^{\mathrm{Data},i}_{j_2} - \mathrm{SF}^i_\ell \cdot N^{\mathrm{MC},i}_{\ell_2} - \mathrm{SF}^i_b \cdot N^{\mathrm{MC},i}_{b_2}\right]\right)^2}{\left(\delta N^i_{c_2}\right)^2} \quad (5.5)$$

**Step 4:** In order to converge to an appropriate solution, each update of the SF values in steps 1 to 3 is only allowed to change the previous SF value by at most 0.01. After each update, a check is performed whether the total (summed) $\chi^2 = \chi^2_\ell + \chi^2_b + \chi^2_c$ also decreased. If this is not the case, the updated value of the SF made the data–to–simulation agreement worse in another additional jet and therefore the update is undone and the SF value

is restored to its value from the previous iteration. If neither $\mathrm{SF}_\ell^i$ nor $\mathrm{SF}_b^i$ nor $\mathrm{SF}_c^i$ were updated in a given iteration, the procedure is finished and the final SFs are fixed.

As mentioned, this procedure is applied to each bin individually. An example of the convergence of the SFs to their final value for one of the chosen bins is shown in Fig. 5.12.

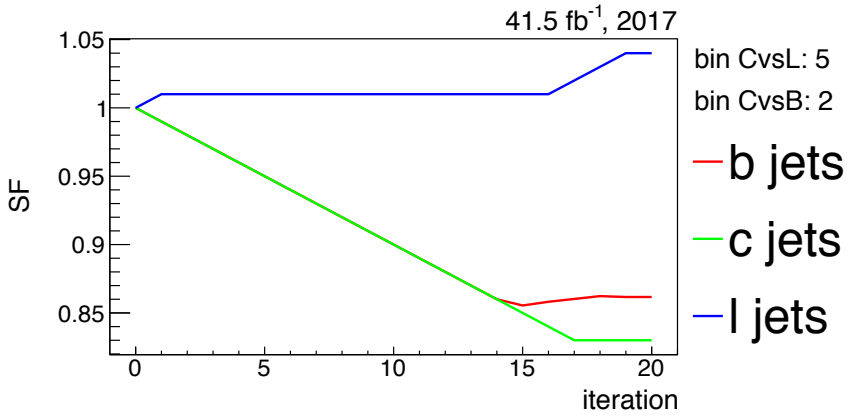


FIGURE 5.12: Convergence of the scale factors for the different jet flavors in a given bin of the two–dimensional CvsL–CvsB phase space.

Finally, in order to apply these corrections to simulated events, one should iterate over all the selected jets in the analysis (of which the $c$–tagging discriminator is used) and multiply their SFs (corresponding to their flavor $f$) to obtain a final event–based weight $w_{c\text{–tag}}$ that can be applied to the simulated events. This weight is defined in Eq. (5.6).

$$w_{c\text{–tag}} = \prod_{i=0}^{\#\text{ jets}} \mathrm{SF}_{(f)}^i \tag{5.6}$$

### Smoothing via interpolation

The SFs are derived with a two–dimensional binning that has to be sufficiently broad in order to have enough events in each bin. In order to obtain a smooth SF as a function of the CvsL and CvsB $c$–tagging discriminators, the obtained result has been interpolated between the centers of each bin. The interpolation is performed with a linear bivariate spline interpolation.

## 5.7.3 Corrections and systematic uncertainties

This section summarizes the systematic uncertainties related to the derivation of the $c$–tagging SFs, as well as corrections applied to the simulated events to account for differences with respect to the data. These effects are subdivided in experimental and theoretical sources of uncertainty.

### A. Experimental uncertainties

**Jet energy scale and resolution.** Systematic uncertainties due to up and down variation of the JER smearing within uncertainties are taken into account in the calculation of the $c$–tagging SFs. Similarly, corrections and uncertainties due to observed differences in the jet energy scale are applied.

These corrections are evaluated and applied in different regions of the jet $p_T$ and $\eta$. See Sec. 3.2.5 for more information on the JES and JER corrections.

**Lepton identification, isolation and reconstruction.** Observed differences in muon and electron identification, isolation and reconstruction between data and simulation are taken into account by $p_T$ and $\eta$ dependent event weights [271, 272]. The corresponding uncertainties on these weights are evaluated by varying them within the $\pm 1$ $\sigma$ uncertainty interval.

**Trigger efficiencies.** The single lepton triggers used to select the events for the SF measurement also show slightly different efficiencies in data and simulation. Again $p_T$ and $\eta$ dependent event weights correct for these differences and the corresponding uncertainties are taken into account [271,272].

**Pileup.** The pileup profile in simulated events is reweighed to match the one observed in data, using an inelastic proton–proton cross section of 69.2 mb [146]. An uncertainty related to this correction is applied by varying this inelastic cross section by $\pm$ 4.6%.

**b–tagging shape.** After ranking the jets according to decreasing value of the $b$–tagging discriminator, the first one in the list is required to be $b$–tagged using the Tight DeepCSV working point. Instead of applying the usual correction for the efficiency of this selection, the dedicated shape correction is applied as discussed in Sec. 3.3.4. Uncertainties corresponding to this calibration are included as systematic uncertainties when evaluating the $c$–tagger calibration.

**B. Theoretical uncertainties**

**Factorization ($\mu_F$) and renormalization ($\mu_R$) scales.** In the matrix element calculation and the parton shower simulation, the choice of the renormalization and factorization scales may have an impact on the kinematical distributions of the final–state objects (see Sec. 3.1). Uncertainties on these scales are taken into account by varying $\mu_F$ and $\mu_R$ independently by a factor of 0.5 and 2. These variations are included as event weights in the simulated samples.

**Matching between the ME and the PS.** The matching between the ME and the PS (see Sec. 3.1.2) is governed by a parameter called *hdamp*. The value of this parameter is varied according to hdamp $= 1.58^{+0.66}_{-0.59}$ $m_{top}$ [273] and included as a systematic uncertainty.

**Underlying event.** The remnants of the proton collisions that do not take part in the hard scatter are referred to as the underlying event and are tuned in the generators to match observed distributions in data (see Sec. 3.1.4). The uncertainties resulting from the variations of these tunes are propagated to the measurement of the $c$–tagging SFs.

Many of the above mentioned uncertainties will be in common with the uncertainties relevant in the final extraction of the $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$LF cross sections. The uncertainties in the $c$–tagging calibration will be correlated to those in the final measurement, as will be discussed in Sec. 5.9.

### 5.7.4 Results

The resulting $c$–tagger SFs for $c$ jets, $b$ jets and light jets are summarized in Fig. 5.13. The uncertainties that are shown on this figure are only statistical uncertainties. The

text and dotted lines indicate the result of the binned calibration, whereas the color scale underneath shows the result of the interpolation. The total systematical and statistical uncertainty on the SFs is shown in Fig. 5.14. The largest uncertainties are observed in regions of the CvsL–CvsB phase space where the given jet flavor is least abundant. These large uncertainties will therefore only have a limited impact on the actual measurement in which they are applied, since it is expected that only very few jets of that flavor are present in that part of the phase space.

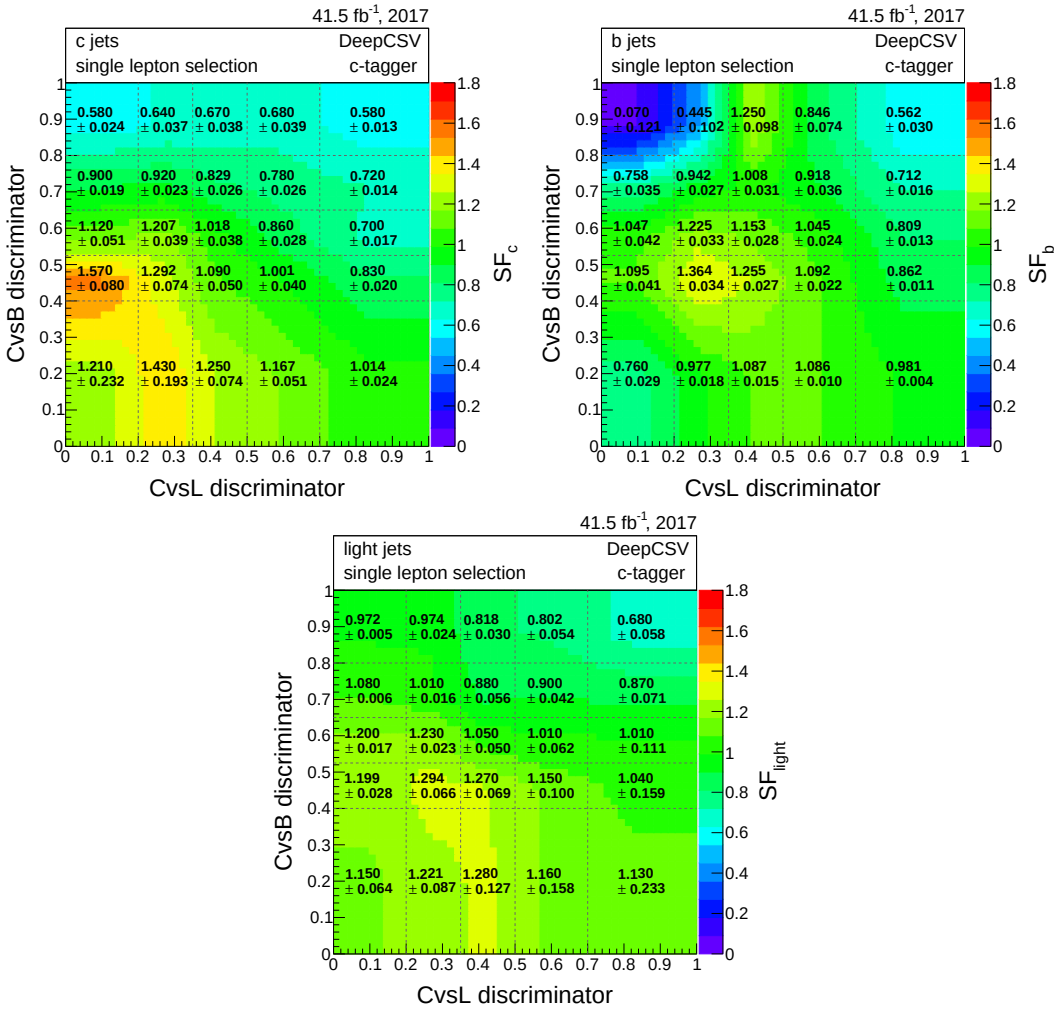


FIGURE 5.13: Scale factors for the $c$–tagger calibration as a function of the DeepCSV CvsL and CvsB discriminator value for $c$ jets (top left), $b$ jets (top right) and light jets (bottom). Uncertainties that are quoted are statistical uncertainties only. The text and dotted lines indicate the result of the binned calibration, whereas the color scale underneath shows the result of the linear interpolation.

### 5.7.5 Validation

A first check is performed by applying the obtained $c$–tagging SFs to the events in the single lepton control region. The results are summarized in Fig. 5.15 and should be compared to Fig. 5.9. It is clear that a much better data–to–simulation agreement is observed and the obtained uncertainty covers well the remaining discrepancies.

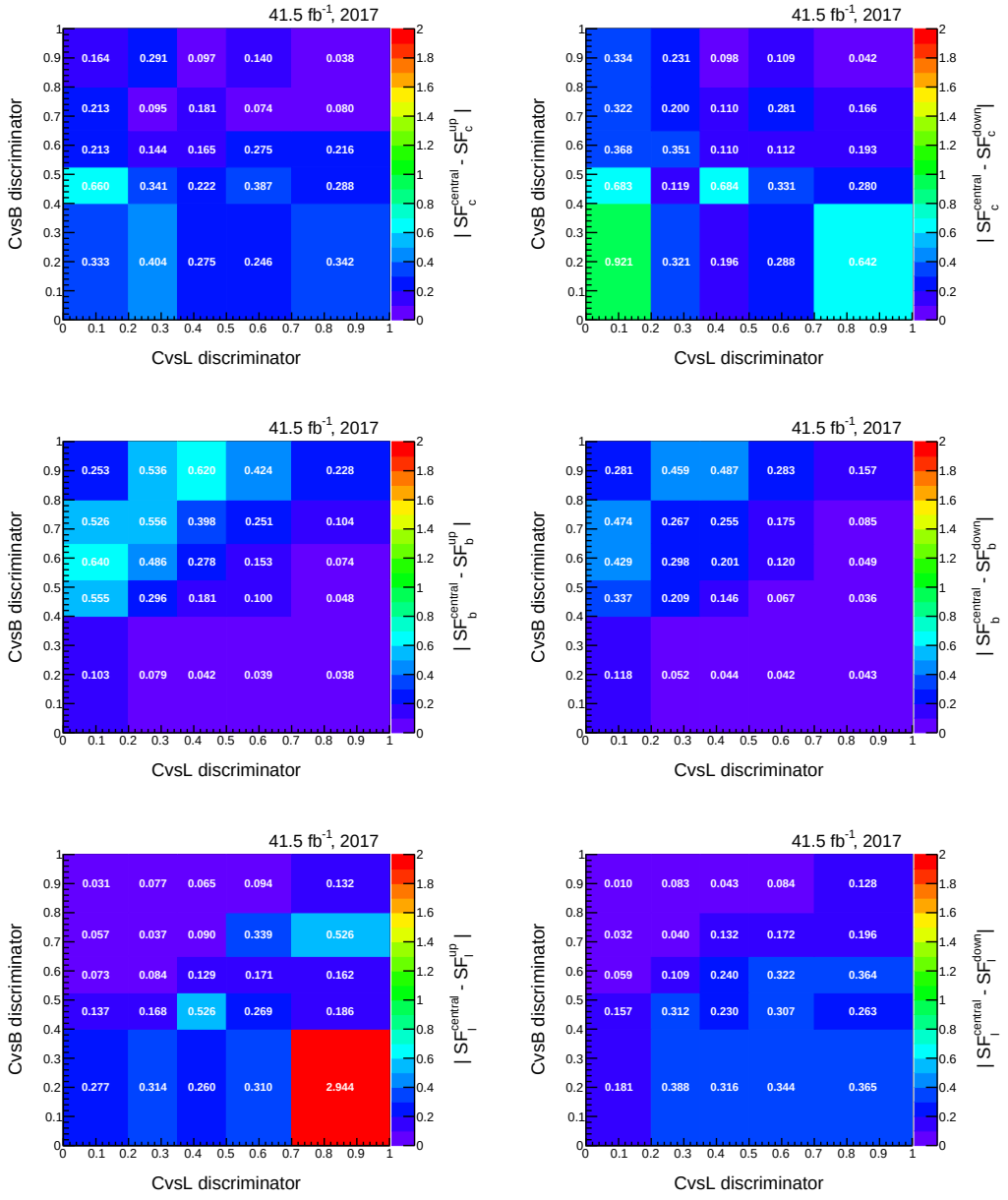


FIGURE 5.14: Total up–variation uncertainty (left) and down–variation uncertainty (right) on the scale factors for the $c$–tagger calibration as a function of the DeepCSV CvsL and CvsB discriminator value for $c$ jets (top), $b$ jets (middle) and light jets (down).

The derived set of SFs can now be applied in the dilepton $t\bar{t}$ + two jets signal region to look at the effect on the DeepCSV $c$–tagging distributions. These distributions are shown in Fig. 5.16 for the first (top) and second (bottom) additional jet and can be compared to the distributions in Fig. 5.8 before the calibration. A clear improvement in the overall data–to–simulation agreement is observed for all distributions. These distributions are before any fitting procedure is applied (which will be explained in Sec. 5.8).

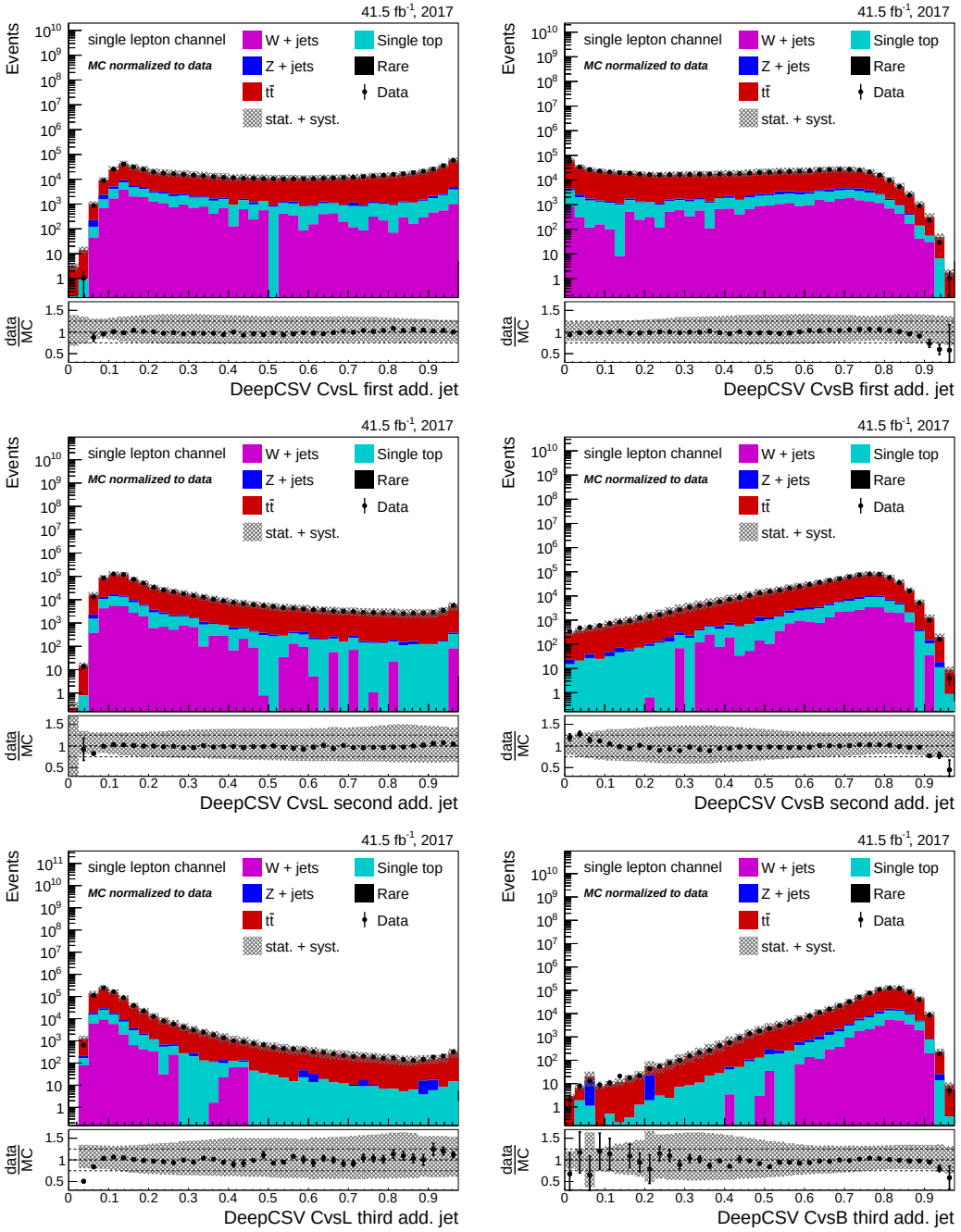


FIGURE 5.15: Data–to–simulation comparisons after application of the $c$–tagging SFs, for the CvsL (left) and CvsB (right) $c$–tagging discriminator distributions for the jet with the second (top), third (middle) and fourth (bottom) highest $b$–tagging discriminator in the events in the single lepton control region. The uncertainty shown with the grey band includes the statistical uncertainty as well as the systematical uncertainty due to the $c$–tagging calibration only. The total yield in the simulation is normalized to the data yield to focus on the comparison of the differential shape.

## Consistency with c–tagging working points

Within the CMS collaboration, SFs are provided by the $b$–tagging and vertexing group for discrete working points of the $c$–tagger. These are depicted in Fig. 3.18 and

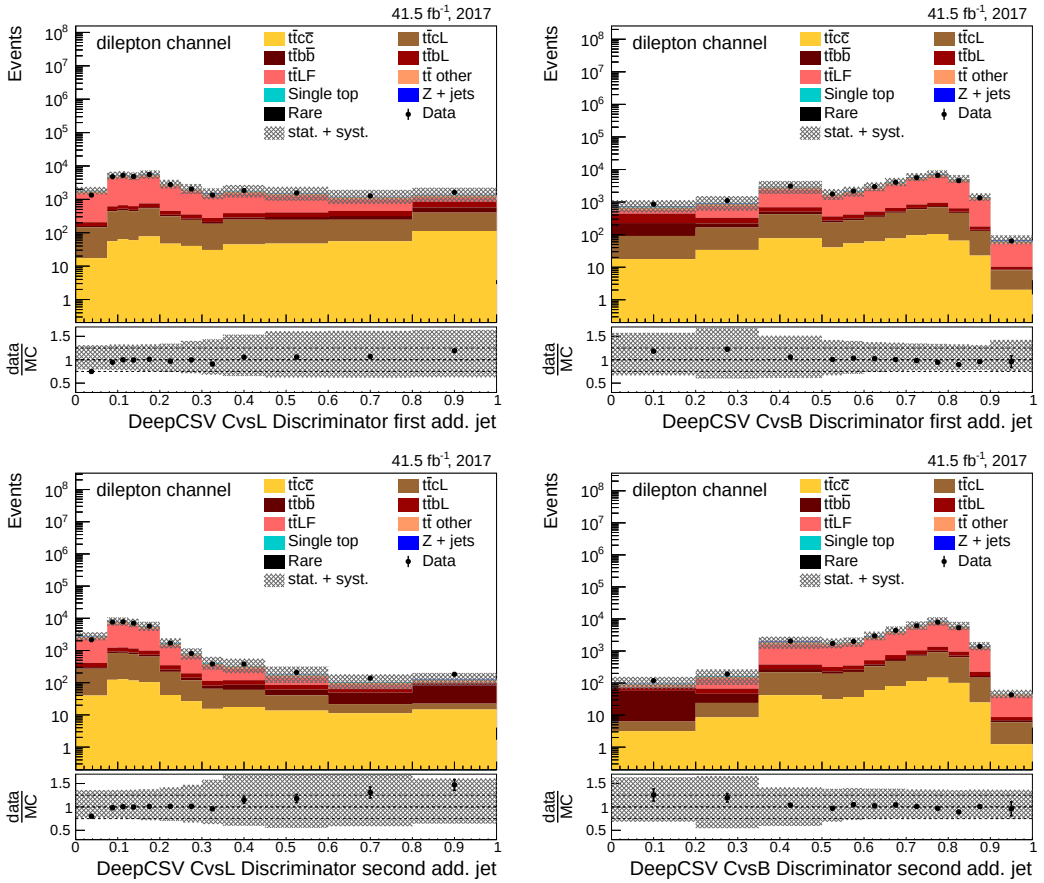


FIGURE 5.16: Data–to–simulation comparisons for the DeepCSV CvsL (left) and CvsB (right) discriminator distributions for the first (top) and second (bottom) additional jet in the dilepton $t\bar{t}$ + two jets signal region after applying the $c$–tagging calibration. The total yield in the simulation is normalized to the data yield to compare the shapes of the distributions. The systematical uncertainties due to the $c$–tagging calibration are added to the statistical uncertainty in quadrature, as displayed by the gray uncertainty band.

summarized in Tab. 3.2. For these measurements to be consistent with the results obtained from the $c$–tagger shape calibration one should be able to recover these SFs (within uncertainties) by an appropriate choice of binning. Indeed, by subdividing the CvsL–CvsB plane in four bins, such that the upper right bin corresponds to a working point in Fig. 3.18, one can compare results obtained by the two methods.

This cross–check was indeed performed and results are shown in Fig. 5.17 for different jet flavors and different working points. It should be noted that the working points obtained by the iterative fit are not binned in $p_T$ and $\eta$ due to limited statistics in some bins. The obtained results are indeed consistent between the two methods, though often the uncertainties from the iterative fitting procedure are relatively large, especially for the $c$ jets. This is expected, given that the $c$ jet working point SFs are derived in a W+c topology that is much more enriched in $c$ jets.

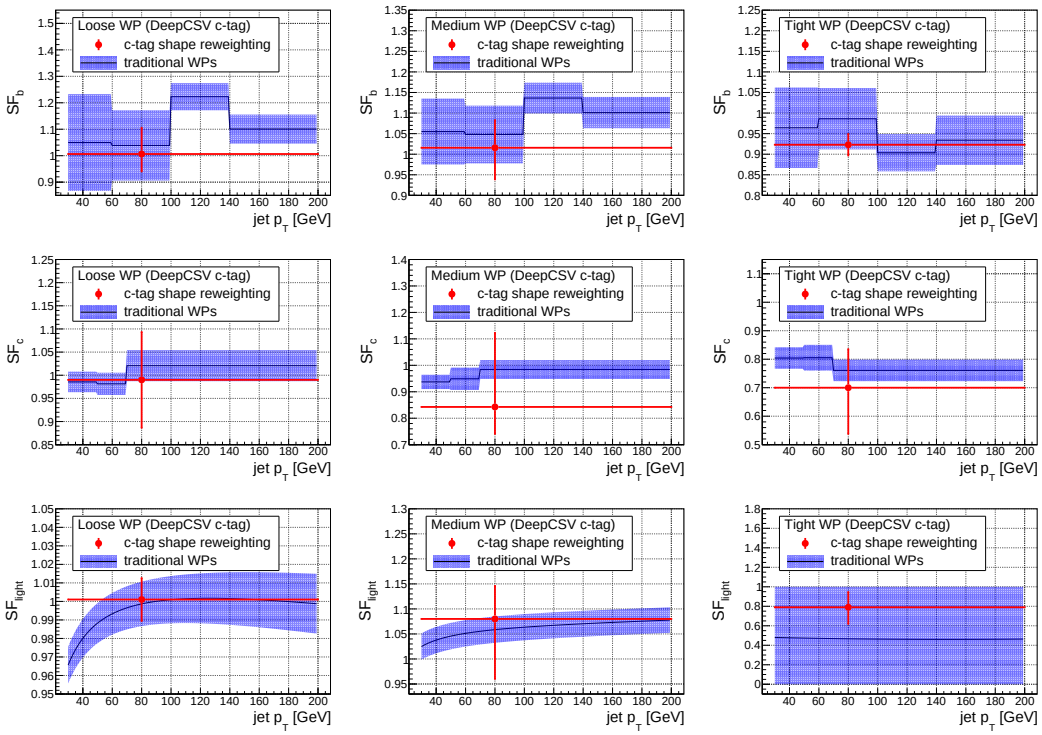


FIGURE 5.17: Comparison between the scale factors obtained for the Loose (left), Medium (middle) and Tight (right) working points of the DeepCSV $c$–tagger and full shape reweighing procedure presented in this work. The $b$ jet SFs are shown on the top, $c$ jet SFs in the middle and light jet SFs on the bottom.

## 5.8 Neural network based event categorization

The extraction of the $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ and $t\bar{t}$LF cross section will proceed through a global fit of an observable that is able to distinguish among these different signals. The $c$–tagging discriminators of the two additional jets indeed possess the needed information to do so, as shown in Fig. 5.18. The CvsL discriminator will be able to differentiate the $t\bar{t}c\bar{c}$ and $t\bar{t}b\bar{b}$ from the $t\bar{t}$LF component, whereas the CvsB discriminator will provide additional information that can be used to distinguish the $t\bar{t}c\bar{c}$ from the $t\bar{t}b\bar{b}$ final–state.

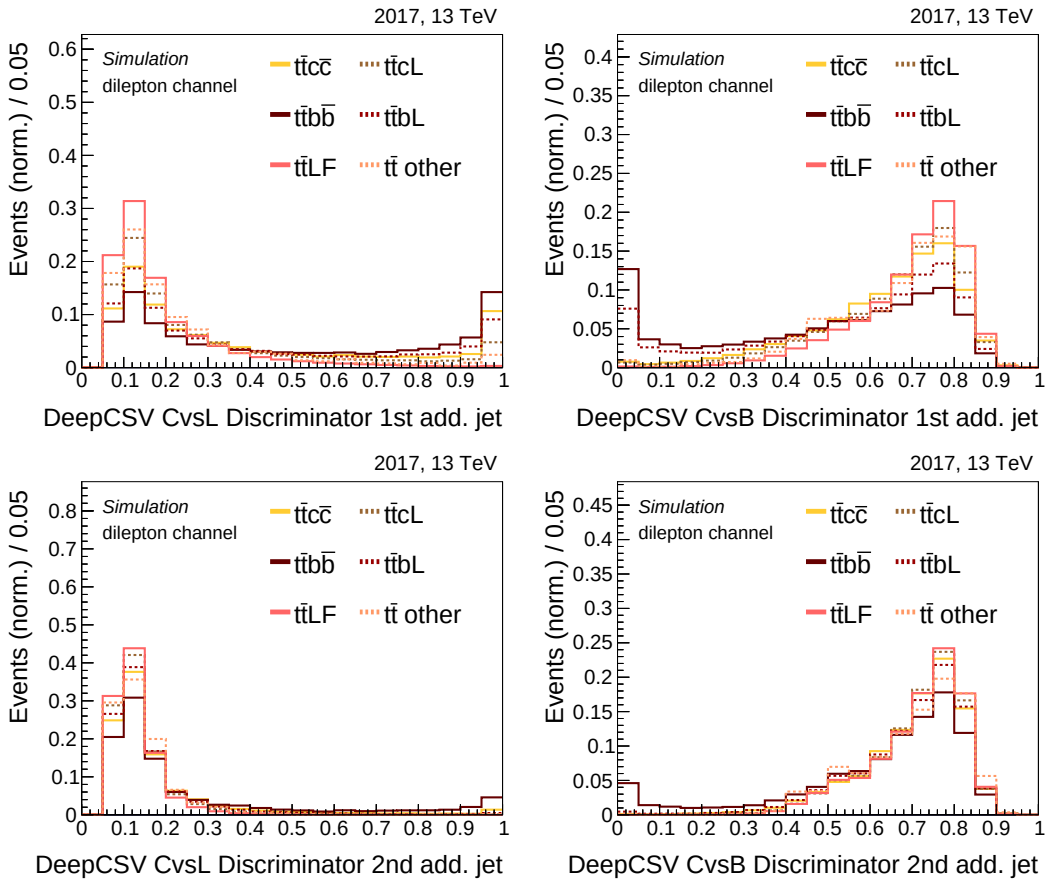


FIGURE 5.18: Normalized distributions of the DeepCSV CvsL and CvsB discriminators for the first and second additional jet in simulated top quarks pair events after the event selection.

Even though these four observables (two discriminators for each additional jet) will possess the most powerful discriminating power, there are other observables that may add additional information, based on the event kinematics rather than the jet flavor. Examples are the separation in $\Delta$R between the two additional jets and also the matching NN output of the best permutation is expected to be larger on average for $t\bar{t}$LF events, for which the matching is easier based on the $b$–tagging information that is used. Examples of these variables are shown in Fig. 5.19. In order to encapsulate as much discriminating information as possible, another neural network was trained to distinguish $t\bar{t}c\bar{c}$ events from $t\bar{t}b\bar{b}$ and from $t\bar{t}$LF events, using the $c$–tag discriminator values of the additional jets, together with the additional discriminating properties depicted in Fig. 5.19. Before feeding these properties into the neural

network, their values are rescaled such that they have a mean of zero and a variance of one over all training samples. Given the relatively small number of input features, a network architecture was chosen with one hidden layer that comprises 30 neurons with ReLu activation functions and a 10% dropout. The network was designed to have five outputs corresponding to the probabilities: $\mathrm{P}(t\bar{t}c\bar{c})$, $\mathrm{P}(t\bar{t}cL)$, $\mathrm{P}(t\bar{t}b\bar{b})$, $\mathrm{P}(t\bar{t}bL)$ and $\mathrm{P}(t\bar{t}\mathrm{LF})$ of an event belonging to the corresponding event category. These network outputs can then be used to project all of the useful information onto a two–dimensional phase space spanned by the two discriminators defined in Eqs. (5.7) and (5.8).

$$\Delta_b^c = \frac{\mathrm{P}(t\bar{t}c\bar{c})}{\mathrm{P}(t\bar{t}c\bar{c}) + \mathrm{P}(t\bar{t}b\bar{b})} \tag{5.7}$$

$$\Delta_L^c = \frac{\mathrm{P}(t\bar{t}c\bar{c})}{\mathrm{P}(t\bar{t}c\bar{c}) + \mathrm{P}(t\bar{t}\mathrm{LF})} \tag{5.8}$$

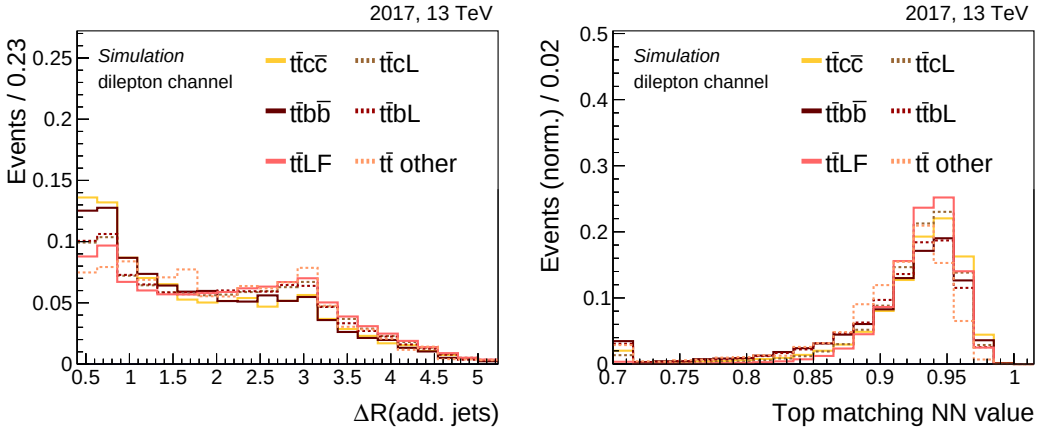


FIGURE 5.19: Normalized distributions of the angular separation between the additional jets (left) and the matching NN output of the best permutation (right) in simulated top quark pair events after the event selection.

One could interpret these discriminators as extended, topology–specific $c$–tagger discriminators which augment information on the jet flavor of the two additional jets with additional information on the event kinematics to most optimally distinguish different event categories. The distributions of these discriminators in data and simulation are shown in Fig. 5.20. Also the normalized distributions from simulations for the different event categories are shown in Fig. 5.21, whereas the two-dimensional normalized distributions are shown in Fig. 5.22. Finally, in order to demonstrate that these NN discriminators have an optimal separating power between the $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$LF classes, Fig. 5.23 shows the ROC curves for the individual input observables to the final NN discriminators, together with the ROC curves of the discriminators themselves. The top panel shows how well the $t\bar{t}c\bar{c}$ events can be distinguished from the $t\bar{t}b\bar{b}$ events, whereas the bottom panel shows the discriminating power between $t\bar{t}c\bar{c}$ and $t\bar{t}$LF events. It is clear that a combination of the $\Delta_b^c$ and $\Delta_L^c$ discriminators provides the most powerful discrimination between the three categories.

The final fit will be performed on a two–dimensional distribution of the $\Delta_b^c$ and $\Delta_L^c$ discriminators using templates from simulation. A first fit will be performed to

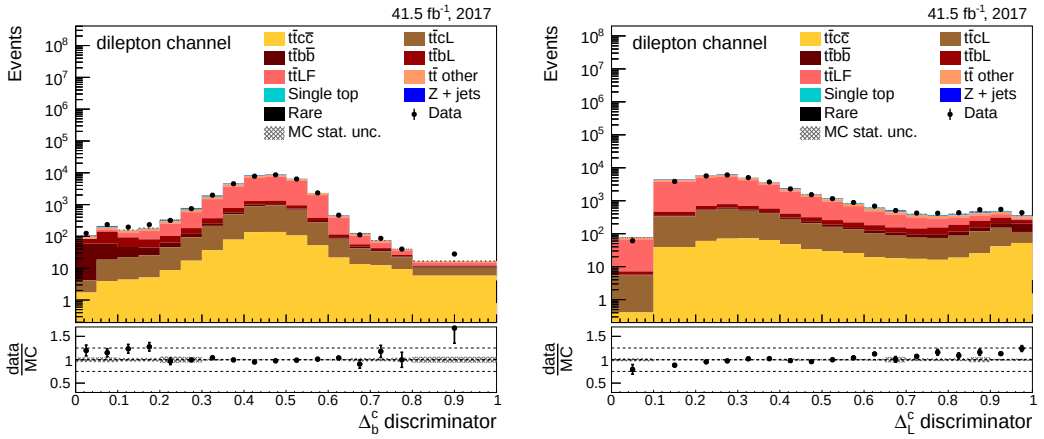


FIGURE 5.20: Distributions of $\Delta_b^c$ (left) and $\Delta_L^c$ (right) in data and simulation before the fit.

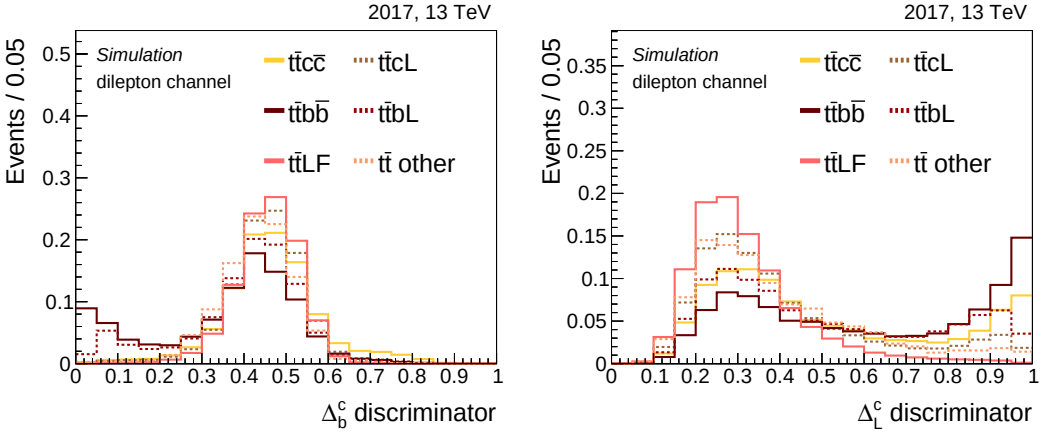


FIGURE 5.21: Normalized distributions of $\Delta_b^c$ (left) and $\Delta_L^c$ (right) in simulated top quark pair events after the event selection.

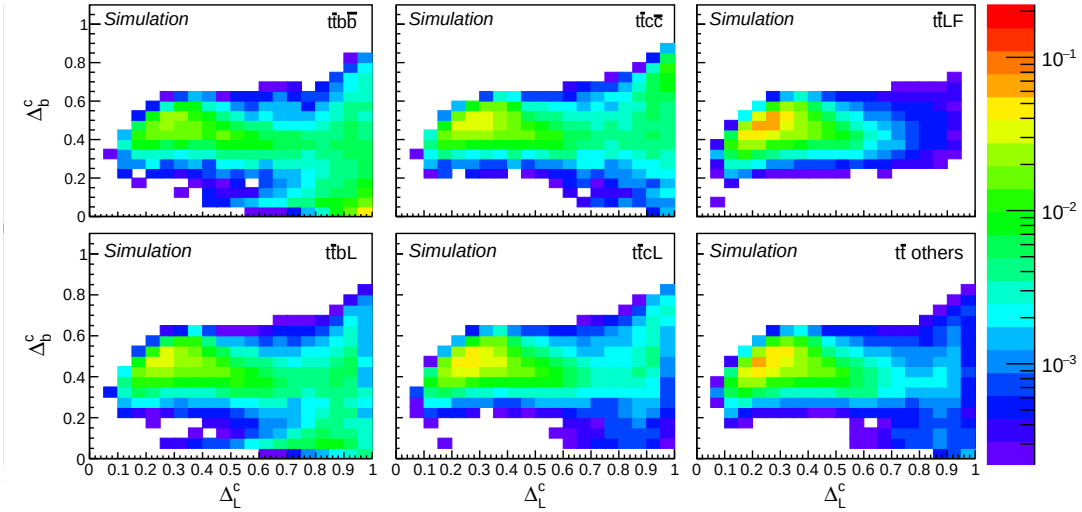


FIGURE 5.22: Normalized two–dimensional distributions of $\Delta_b^c$ on the y–axis and $\Delta_L^c$ on the x–axis in simulated top quark pair events after the event selection.

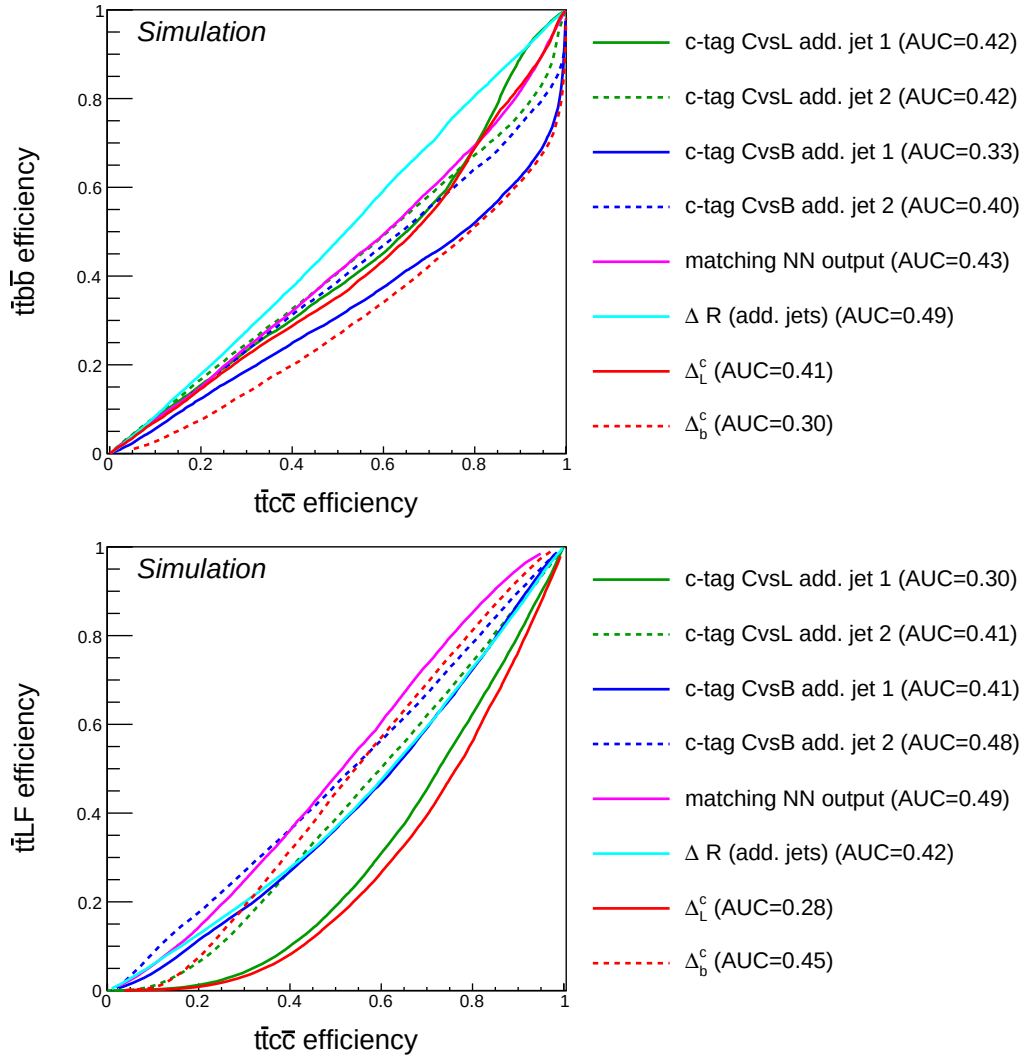


FIGURE 5.23: ROC curves for the individual input observables to the final NN discriminators, and for the discriminators themselves. The top panel shows how well the $t\bar{t}c\bar{c}$ events can be distinguished from the $t\bar{t}b\bar{b}$ events, whereas the bottom panel shows the discriminating power between $t\bar{t}c\bar{c}$ and $t\bar{t}$LF events.

simultaneously extract the absolute cross section of the $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$LF events in the visible and in the full phase space, as defined in Sec. 5.4. As explained before, the $t\bar{t}cL$ and $t\bar{t}bL$ categories do not contain true physical processes, but rather results from events in which two $b/c$ jets are merged into one (resulting in two clustered $b/c$ hadrons in one jet), or events in which one of the two additional $b/c$ jets is lost due to acceptance requirements. Therefore these processes are scaled with the same factor as the $t\bar{t}c\bar{c}$ and $t\bar{t}b\bar{b}$ templates, and the relative yield of $t\bar{t}cL$ ($t\bar{t}bL$) with respect to $t\bar{t}c\bar{c}$ ($t\bar{t}b\bar{b}$) is fixed by that predicted in the simulation.

A second fit will then extract the ratio of the $t\bar{t}c\bar{c}$ and $t\bar{t}b\bar{b}$ to the overall inclusive $t\bar{t}jj$ cross section also in the visible and full phase space. These ratios will further be referred to as $R_c$ and $R_b$ respectively. The advantage of calculating these ratios lies in the cancellation of some common systematic uncertainties, resulting in a more accurate prediction of the ratio compared to the absolute cross sections. These two cases will be discussed separately in Sec. 5.8.2 and Sec. 5.8.3. First a formal description of

the template fitting procedure is given in Sec. 5.8.1.

### 5.8.1 Statistical procedure: template fitting

The statistical procedure is based on a binned maximum likelihood fit to a set of predefined (normalized) templates (for a comprehensive overview, see Ref. [241], chapter 10.4). These templates are typically extracted from simulations (or sometimes from control regions in data) in the form of binned histograms that approximate the probability density $f_p(x)$ for an observable $x$. Templates are provided for each of the signal and background processes ($p$) under consideration[10]. Consider a total of $N_p$ processes with the corresponding template histograms ($\hat{f}_p$) that have $N_{\text{bins}}$ bins with the central value of each bin denoted as $\{x_1, x_2, \cdots, x_i, \cdots, x_{N_{\text{bins}}}\}$. The total expected number of events ($\nu_i$) in bin $i$ can be expressed as the sum of the expected total contributions from each process, $\nu^p$, multiplied by the probability that an event of process $p$ ends up in bin $i$

$$\nu_i = \sum_{p=1}^{N_p} \nu^p \cdot \int_{\Delta x_i} f_p(x)\, dx \tag{5.9}$$

$$\approx \sum_{p=1}^{N_p} \nu^p \cdot \hat{f}_p(x_i)\, \Delta x_i. \tag{5.10}$$

The second line follows by approximating the integral of the probability density $f_p(x)$ over bin $i$ by the product of the value of the binned template histogram $\hat{f}_p$ at the bin center $x_i$ and the width of the bin $\Delta x_i$. The value of $\hat{f}_p(x_i)$ is exactly the bin content of the normalized template histogram that is provided.

Now assume one measures in each bin $i$ an event yield $n_i$. The objective of the template fit is to determine which values of $\vec{\boldsymbol{\nu}}^{\mathbf{p}} = \{\nu^{p=1}, \cdots, \nu^{p=j}, \cdots, \nu^{p=N_p}\}$ will yield the best agreement between the expected yields ($\nu_i$) and the observed yields $\vec{\mathbf{n}} = \{n_1, \cdots, n_i, \cdots, n_{N_{bins}}\}$. The observed yields in each bin, $n_i$, are expected to be random numbers drawn from a Poisson distribution with an expectation value $\nu_i$. Therefore a likelihood function $\mathcal{L}(\vec{\mathbf{n}}|\vec{\boldsymbol{\nu}}^{\mathbf{p}})$ can be constructed that expresses the likelihood for the observation $\vec{\mathbf{n}}$ given that the yields of each of the processes is given by $\vec{\boldsymbol{\nu}}^{\mathbf{p}}$, as expressed in Eq. (5.11).

$$\mathcal{L}(\vec{\mathbf{n}}|\vec{\boldsymbol{\nu}}^{\mathbf{p}}) = \prod_{i=1}^{N_{\text{bins}}} \frac{(\nu_i)^{n_i}}{n_i!} \cdot e^{-\nu_i} \tag{5.11}$$

The maximum likelihood estimators for $\vec{\boldsymbol{\nu}}^{\mathbf{p}}$ are then obtained by minimizing $-\ln\left[\mathcal{L}(\vec{\mathbf{n}}|\vec{\boldsymbol{\nu}}^{\mathbf{p}})\right]$ with respect to $\vec{\boldsymbol{\nu}}^{\mathbf{p}}$.

Systematic uncertainties (discussed further in Sec. 5.9) are treated as nuisance parameters and can therefore be additionally constrained by the fit. Practically this means that the likelihood in Eq. (5.11) is expanded with a set of nuisance parameters $\vec{\boldsymbol{\theta}}$ which are typically constrained by a gaussian with a mean of zero and a standard deviation of one. For a simple normalization uncertainty (for example on the total

[10] Sometimes templates are merged together if they have a very similar shape or if the fit is not supposed to extract the yields separately for those processes. This will be the case for all non–$t\bar{t}$ processes in this analysis, which will be merged into one single template in the fitting procedure.

integrated luminosity), the negative log–likelihood can then be expressed as

$$-\ln\left[\mathcal{L}(\vec{\mathbf{n}}|\vec{\boldsymbol{\nu}}^{\mathbf{p}},\theta)\right] = \sum_{i=1}^{N_{\text{bins}}}\left[-n_i\ln(\nu_i)+\ln(n_i!)+\nu_i\right]+\frac{(\theta-\mu_\theta)^2}{2\sigma_\theta^2}, \tag{5.12}$$

where the additional last term denotes the probability density function of the nuisance parameter $\theta$. This is taken to be a gaussian with mean $\mu_\theta$ and standard deviation $\sigma_\theta$ that can alter the overall yield due to a normalization uncertainty. The minimization of the negative log–likelihood now takes into account this additional nuisance parameter, expanding the allowed range of values for the parameters $\vec{\boldsymbol{\nu}}^{\mathbf{p}}$ and therefore reducing the sensitivity of the measurement. In case a given source of uncertainty changes the shape of the templates rather than the normalization, more advanced techniques such as *template morphing* are considered [274].

### 5.8.2 Extraction of the absolute cross sections

The absolute cross sections in the visible phase space, $\sigma^{\text{vis}}_{t\bar{t}c\bar{c}}$, $\sigma^{\text{vis}}_{t\bar{t}b\bar{b}}$ and $\sigma^{\text{vis}}_{t\bar{t}\text{LF}}$, are extracted by fitting the templates from simulations to the observed data according to Eq. (5.13). In this notation, $\text{H}^{\text{norm}}$ represents the normalized two–dimensional $\Delta_b^c$ – $\Delta_L^c$ distribution (*i.e.* the templates) corresponding to each of the processes[11]. The signal categories were previously defined in Sec. 5.4. Other background abbreviations are DY (Drell–Yan), ST (single top quark) and Rare (collectively describing diboson, triboson, $\text{t}\bar{\text{t}}$Z/W/H and W+jets). The total integrated luminosity of 41.5 $\text{fb}^{-1}$ is denoted as $\mathcal{L}^{int}$. In order to unfold the fitted cross sections to the visible phase space, efficiencies ($\epsilon$) for the $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$LF categories are included in the formula. These efficiencies are calculated from the simulation and their values are summarized in Tab. 5.8. The ratio of $t\bar{t}c\bar{c}$ ($t\bar{t}b\bar{b}$) events with respect to $t\bar{t}cL$ ($t\bar{t}bL$) events is fixed at the value observed in simulations (denoted by the ratio $N^{\text{MC}}_{t\bar{t}\text{c}\text{L}(t\bar{t}\text{b}\text{L})}/N^{\text{MC}}_{t\bar{t}c\bar{c}(t\bar{t}b\bar{b})}$) and both components are scaled with the same factor. Also the $t\bar{t}+\text{Other}$ component is scaled with the same factor as the $t\bar{t}$LF component, motivated by their very similar distributions as can be seen in Fig. 5.21.

$$\begin{aligned} f\left(\sigma^{\text{vis}}_{t\bar{t}c\bar{c}},\sigma^{\text{vis}}_{t\bar{t}b\bar{b}},\sigma^{\text{vis}}_{t\bar{t}\text{LF}}\right) = \mathcal{L}^{int}\cdot\Bigg\{ & \sigma^{\text{vis}}_{t\bar{t}c\bar{c}}\cdot\epsilon_{t\bar{t}c\bar{c}}\cdot\left(\text{H}^{\text{norm}}_{t\bar{t}c\bar{c}}+\frac{N^{\text{MC}}_{t\bar{t}\text{c}\text{L}}}{N^{\text{MC}}_{t\bar{t}c\bar{c}}}\cdot\text{H}^{\text{norm}}_{t\bar{t}cL}\right) \\ & +\sigma^{\text{vis}}_{t\bar{t}b\bar{b}}\cdot\epsilon_{t\bar{t}b\bar{b}}\cdot\left(\text{H}^{\text{norm}}_{t\bar{t}b\bar{b}}+\frac{N^{\text{MC}}_{t\bar{t}\text{b}\text{L}}}{N^{\text{MC}}_{t\bar{t}b\bar{b}}}\cdot\text{H}^{\text{norm}}_{t\bar{t}bL}\right) \\ & +\sigma^{\text{vis}}_{t\bar{t}\text{LF}}\cdot\epsilon_{t\bar{t}\text{LF}}\cdot\left(\text{H}^{\text{norm}}_{t\bar{t}\text{LF}}+\frac{N^{\text{MC}}_{t\bar{t}+\text{Other}}}{N^{\text{MC}}_{t\bar{t}LF}}\cdot\text{H}^{\text{norm}}_{t\bar{t}+\text{Other}}\right) \\ & +\sigma_{\text{DY}}\cdot\text{H}^{\text{norm}}_{\text{DY}}+\sigma_{\text{ST}}\cdot\text{H}^{\text{norm}}_{\text{ST}}+\sigma_{\text{Rare}}\cdot\text{H}^{\text{norm}}_{\text{Rare}}\Bigg\} \end{aligned} \tag{5.13}$$

To extract the result in the full phase space, an additional acceptance factor, $\mathcal{A}$, is included to account for the difference in acceptance between the full and the visible phase space. These factors are also summarized in Tab. 5.8.

[11] Practically, this two–dimensional distribution is unrolled into a one–dimensional histogram that can be handled by the software framework. See for example Fig. 5.25.

TABLE 5.8: Selection efficiencies and acceptance factors for events in different signal categories. These values were derived from simulations.

| **Category** | $t\bar{t}b\bar{b}$ | $t\bar{t}c\bar{c}$ | $t\bar{t}$LF |
|---|---|---|---|
| **Efficiency** $\epsilon$ (%) | 11.9 | 4.7 | 2.7 |
| **Acceptance** $\mathcal{A}$ (%) | 6.0 | 5.2 | 10.9 |

### 5.8.3 Extraction of the ratios $R_c$ and $R_b$

The extraction of the ratios $R_c$ and $R_b$, accompanied by the measurement of the full inclusive $\sigma^{\text{vis}}_{t\bar{t}jj}$ cross section can in principle be derived from the results in Sec. 5.8.2. However, the choice was made to perform a separate fit according to Eq. (5.14), and the results can serve as a cross check of the ones obtained from fitting Eq. (5.13).

$$
\begin{aligned}
f\left(\sigma^{\text{vis}}_{t\bar{t}jj}, \mathrm{R_c}, \mathrm{R_b}\right) = & \mathcal{L}^{int} \cdot \sigma^{\text{vis}}_{t\bar{t}jj} \cdot \Bigg\{ \mathrm{R_c} \cdot \epsilon_{t\bar{t}c\bar{c}} \cdot \left( \mathrm{H}^{\text{norm}}_{t\bar{t}c\bar{c}} + \frac{N^{\text{MC}}_{t\bar{t}c\text{L}}}{N^{\text{MC}}_{t\bar{t}c\bar{c}}} \cdot \mathrm{H}^{\text{norm}}_{t\bar{t}cL} \right) \\
& + \mathrm{R_b} \cdot \epsilon_{t\bar{t}b\bar{b}} \cdot \left( \mathrm{H}^{\text{norm}}_{t\bar{t}b\bar{b}} + \frac{N^{\text{MC}}_{t\bar{t}b\text{L}}}{N^{\text{MC}}_{t\bar{t}b\bar{b}}} \cdot \mathrm{H}^{\text{norm}}_{t\bar{t}bL} \right) \\
& + (1 - \mathrm{R_c} - \mathrm{R_b}) \cdot \epsilon_{t\bar{t}\text{LF}} \cdot \left( \mathrm{H}^{\text{norm}}_{t\bar{t}\text{LF}} + \frac{N^{\text{MC}}_{t\bar{t}+\text{Other}}}{N^{\text{MC}}_{t\bar{t}LF}} \cdot \mathrm{H}^{\text{norm}}_{t\bar{t}+\text{Other}} \right) \Bigg\} \\
& + \mathcal{L}^{int} \cdot \{ \sigma_{\text{DY}} \cdot \mathrm{H}^{\text{norm}}_{\text{DY}} + \sigma_{\text{ST}} \cdot \mathrm{H}^{\text{norm}}_{\text{ST}} + \sigma_{\text{Rare}} \cdot \mathrm{H}^{\text{norm}}_{\text{Rare}} \}
\end{aligned}
\tag{5.14}
$$

## 5.9 Corrections and systematic uncertainties

This section summarizes the systematic uncertainties related to the extraction of the $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ and $t\bar{t}$LF cross sections (and the ratios $R_c$ and $R_b$), as well as corrections applied to the simulated events to account for differences with respect to the data. These effects are subdivided in experimental and theoretical sources of uncertainty.

**A. Experimental uncertainties**

**Jet energy scale and resolution.** Systematic uncertainties due to up and down variations of the JER smearing within uncertainties are taken into account in the determination of the cross sections and their ratios. Similarly, corrections due to observed differences in the jet energy scale are applied and uncertainties are taken into account. These corrections are evaluated and applied in different regions of the jet $p_T$ and $\eta$ (see Sec. 3.2.5).

**Lepton identification, isolation and reconstruction.** Observed differences in muon and electron identification, isolation and reconstruction between data and simulation are taken into account by $p_T$ and $\eta$ dependent event weights [271, 272]. The corresponding uncertainties on these weights are taken into account by varying them within the $\pm 1$ $\sigma$ uncertainty interval.

**Pileup.** The pileup profile in simulated events is reweighed to match the one observed in data, using an inelastic proton–proton cross section of 69.2 mb [146]. An uncertainty related to this correction is applied by varying this inelastic cross section by $\pm$ 4.6%. The effect of this reweighing procedure is illustrated in Fig. 5.24.

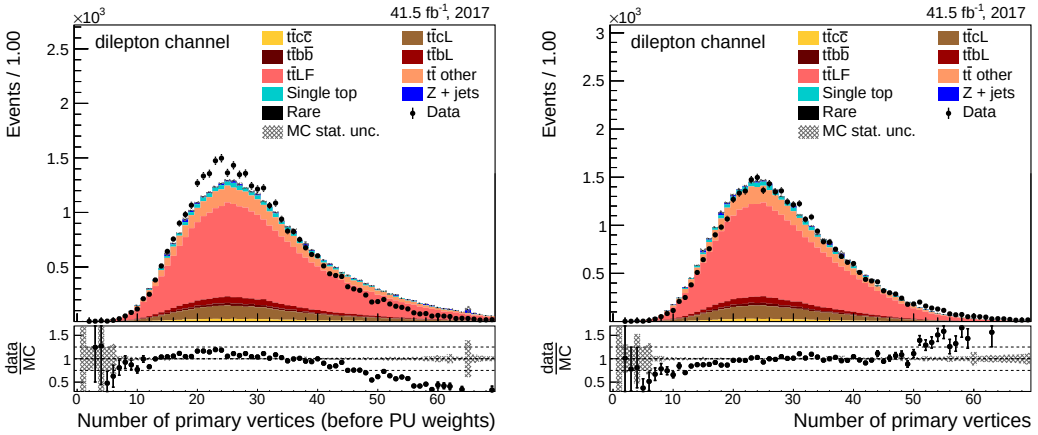


FIGURE 5.24: Distribution of the number of primary vertices compared between data and simulation. On the left the distribution is shown before reweighing the pileup profile and on the right after the corrections are applied.

**Luminosity.** An uncertainty of 2.3% [130] on the total integrated luminosity of 41.5 fb$^{-1}$ accumulated by the CMS experiment in 2017 is taken into account in the measurement.

**b–tagging calibration.** After matching the jets to the partons based on the permutation with the largest output of the matching NN (see Sec. 5.6), the two jets matched to the $b$ quarks from the top quark decays are required to be $b$–tagged using the Medium DeepCSV working point. Dedicated shape corrections have been applied that match the shape of the simulated $b$–tagging discriminator to that observed in data (see Sec. 3.3.4). This way the efficiency of this $b$–tagging requirement in the simulation is ensured to match the efficiency observed in data. Uncertainties related to this shape calibration are taken into account in the final measurement.

**c–tagging calibration.** The event weights from the $c$–tagging calibration, which was in detail explained in Sec. 5.7.2, have been applied. Many of the theoretical and experimental sources of uncertainties are in common between the single lepton control region in which the $c$–tagger calibration was derived and the dilepton signal region considered here. In order not to double–count any systematic uncertainties, the up and down variations of each of these common uncertainties are considered fully correlated and applied simultaneously. This means for example that when the JES uncertainties are evaluated, separate samples are used in which the JES is scaled up and down corresponding to the $\pm 1\sigma$ variation and the corresponding $c$–tagger calibration is derived using these dedicated JES up and down samples. Only the purely statistical uncertainty of the $c$–tagging calibration is applied as an independent source of uncertainty which is uncorrelated to any of the systematic uncertainties.

## B. Theoretical uncertainties

**Factorization ($\mu_F$) and renormalization ($\mu_R$) scales.** As discussed in Sec. 3.1.1, the choice of the renormalization and factorization scale in the matrix element calculation and the parton shower simulation may have an impact on the kinematical distributions of the final–state objects. This is especially relevant in the simulation of the $t\bar{t}b\bar{b}$ and $t\bar{t}c\bar{c}$ final–state, which

inherently posses very different scales (from the top quark mass to the relatively soft additional jets, resulting mostly from gluon splitting into $b\bar{b}/c\bar{c}$ pairs). Uncertainties on these scales are taken into account by varying $\mu_F$ and $\mu_R$ independently by a factor of 0.5 and 2. These variations are included as event weights in the simulated samples.

**Matching between the ME and the PS.** The matching between the ME and the PS is governed by a parameter called *hdamp*. The value of this parameter is varied according to hdamp $= 1.58^{+0.66}_{-0.59}$ $m_{top}$ [273] and included as a systematic uncertainty.

**Underlying event.** The remnants of the proton collisions that do not take part in the hard scatter are referred to as the underlying event and are tuned in the generators to match observed distributions in data (see Sec. 3.1.4). The uncertainties obtained from the variations of these tunes are propagated to the measurement of the cross sections and the corresponding ratios.

**Simulated event sample size.** Only a limited number of simulated events can be produced, resulting in a statistical uncertainty on the bin contents for the templates to which the data are fitted. The statistical uncertainty due to the limited size of the simulated samples is taken into account in the fit.

**Background normalization.** A conservative uncertainty of 30% is added to the total cross section of all background processes that are considered (Drell–Yan, single top quark, diboson, triboson, ttZ/W/H and W+jets). Given that the total background contamination is only $\sim 5\%$, the impact of this uncertainty should in principle stay in the order of $\sim 1-2\%$.

For the integrated luminosity, the simulated sample size and the background normalization only the effect on the overall normalization is taken into account, whereas for all other systematic uncertainties also the effect on the shape of the templates is considered. Finally, we summarize the different sources of systematic uncertainties with their individual impact on the different parameters of interest in Tab. 5.9 for the visible phase space and in Tab. 5.10 for the full phase space. The quoted numbers represent the relative impact on the measured values of the different cross sections ($\Delta\sigma_i$) or ratios ($\Delta \mathrm{R}_i$) when only that systematic source is taken into account. The impact is taken to be the maximum of the up or down variation of the considered systematic uncertainty. The dominant uncertainties to the $t\bar{t}$+HF cross sections are due to the $b$– and $c$–tagging calibrations as well as the choice of the renormalization scale. The theoretical uncertainties are highly suppressed for the measurement of the ratios as expected. Details on the pull, constraint and impact of each nuisance parameter on each of the parameters of interest are summarized in App. A.

## 5.10 Results

Fig. 5.25 shows the data–to–simulation agreement of the unrolled histogram corresponding to the two–dimensional distribution of the $\Delta_b^c$ and $\Delta_L^c$ discriminators. The left panel shows the agreement before the fit, whereas the right panel shows the agreement after the fit. The results of the fit are translated into three scale factors $\alpha_c$, $\alpha_b$ and $\alpha_l$, that scale the simulated templates up or down to obtain the best match with the observed data. The $\alpha_c$ factor scales the $t\bar{t}c\bar{c}$ and $t\bar{t}cL$ components, $\alpha_b$ scales the $t\bar{t}b\bar{b}$ and $t\bar{t}bL$ components and $\alpha_l$ scales the $t\bar{t}LF$ and $t\bar{t}$ + Other components. A clear overall improvement is observed in the agreement with the observed data,

TABLE 5.9: Summary of the individual impacts of the nuisance parameters on the different parameters of interest in the visible phase space.

| Source | $\Delta\sigma_{t\bar{t}c\bar{c}}$[%] | $\Delta\sigma_{t\bar{t}b\bar{b}}$[%] | $\Delta\sigma_{t\bar{t}\mathrm{LF}}$[%] | $\Delta\mathrm{R}_c$[%] | $\Delta\mathrm{R}_b$[%] |
|---|---|---|---|---|---|
| Experimental systematic uncertainties | | | | | |
| Jet Energy Scale | 5.5 | 5.1 | 7.2 | 2.2 | 2.2 |
| Jet Energy Resolution | 1.1 | 0.7 | 0.6 | 1.7 | 1.3 |
| Electron ID | 2.6 | 3.0 | 2.4 | 0.3 | 0.5 |
| Electron reconstruction | 0.5 | 0.5 | 0.5 | $< 0.1$ | 0.1 |
| Muon ID/reconstruction | 0.4 | 0.5 | 0.5 | 0.1 | 0.1 |
| Muon isolation | $< 0.1$ | $< 0.1$ | $< 0.1$ | $< 0.1$ | $< 0.1$ |
| Pileup | 0.5 | 1.2 | 0.6 | 1.2 | 0.7 |
| Total integrated luminosity | 2.3 | 2.5 | 2.4 | 0.1 | 0.1 |
| b-tag calibration | 7.9 | 5.8 | 3.1 | 9.5 | 5.9 |
| c-tag calibration (stat.) | 7.7 | 2.4 | 0.8 | 8.4 | 1.8 |
| Theoretical systematic uncertainties | | | | | |
| Factorization scale ($\mu_F$) | 5.1 | 3.4 | 3.7 | 1.4 | 0.4 |
| Renormalization scale ($\mu_R$) | 10.0 | 7.9 | 9.0 | 0.9 | 1.5 |
| Matching ME-PS (hdamp) | 3.8 | 2.9 | 2.9 | 1.4 | 0.6 |
| Underlying event | 5.5 | 1.7 | 3.6 | 2.3 | 4.7 |
| Background normalization | 0.2 | 2.1 | 0.9 | 1.1 | 1.3 |
| Simulated data sample size | 4.9 | 4.1 | 3.9 | 3.7 | 3.9 |

TABLE 5.10: Summary of the individual impacts of the nuisance parameters on the different parameters of interest in the full phase space.

| Source | $\Delta\sigma_{t\bar{t}c\bar{c}}$[%] | $\Delta\sigma_{t\bar{t}b\bar{b}}$[%] | $\Delta\sigma_{t\bar{t}\mathrm{LF}}$[%] | $\Delta\mathrm{R}_c$[%] | $\Delta\mathrm{R}_b$[%] |
|---|---|---|---|---|---|
| Experimental systematic uncertainties | | | | | |
| Jet Energy Scale | 6.1 | 5.4 | 8.2 | 1.9 | 2.5 |
| Jet Energy Resolution | 0.4 | 0.6 | 0.9 | 0.6 | 1.5 |
| Electron ID | 2.7 | 2.9 | 2.4 | 0.3 | 0.4 |
| Electron reconstruction | 0.5 | 0.5 | 0.5 | $< 0.1$ | $< 0.1$ |
| Muon ID/reconstruction | 0.5 | 0.5 | 0.5 | $< 0.1$ | 0.1 |
| Muon isolation | $< 0.1$ | $< 0.1$ | $< 0.1$ | $< 0.1$ | $< 0.1$ |
| Pileup | 2.0 | 1.5 | 0.5 | 2.5 | 1.1 |
| Total integrated luminosity | 2.2 | 2.4 | 2.4 | 0.1 | 0.1 |
| b-tag calibration | 10.1 | 7.0 | 1.8 | 10.7 | 6.9 |
| c-tag calibration (stat.) | 7.1 | 4.1 | 1.5 | 8.4 | 2.9 |
| Theoretical systematic uncertainties | | | | | |
| Factorization scale ($\mu_F$) | 5.2 | 2.7 | 3.1 | 2.2 | 0.5 |
| Renormalization scale ($\mu_R$) | 10.2 | 8.0 | 9.3 | 1.2 | 1.4 |
| Matching ME-PS (hdamp) | 6.1 | 6.0 | 5.7 | 4.7 | 0.7 |
| Underlying event | 1.2 | 4.1 | 0.6 | 1.0 | 4.3 |
| Background normalization | 0.4 | 2.0 | 0.8 | 0.7 | 1.12 |
| Simulated data sample size | 5.4 | 3.9 | 3.3 | 3.6 | 4.7 |

consistent within uncertainties. This is also quantified by the improved $\chi^2/_{\mathrm{ndof}}$ as defined in Eq. (5.15). Here the number of bins is denoted as $N_{\mathrm{bins}}$, whereas $N_i^{\mathrm{data}}$ and $N_i^{\mathrm{MC}}$ denote the number of events in data and simulations in bin $i$ respectively. The

total uncertainty on the yield in bin $i$, corresponding to the grey band in Fig. 5.25 is denoted as $\delta N_i$.

$$\chi^2/_{\text{ndof}} = \frac{1}{N_{\text{bins}}} \sum_{i=1}^{N_{\text{bins}}} \left( \frac{N_i^{\text{data}} - N_i^{\text{MC}}}{\delta N_i} \right)^2 \tag{5.15}$$

The binning of the two–dimensional distribution of the $\Delta_b^c$ and $\Delta_L^c$ discriminators was chosen to be

$$\Delta_L^c \otimes \Delta_b^c : [0, 0.25, 0.4, 0.6, 0.9, 1.0] \otimes [0, 0.3, 0.45, 0.5, 0.55, 1.0].$$

This results in a total of 25 bins (as shown in Fig. 5.25) with different compositions of the different event categories.

The final results for the cross sections and their ratios both in the visible and full phase space are summarized in Tab. 5.11. These results are also visualized in Figs. 5.26 and 5.27 for the visible and full phase space respectively. The systematical uncertainties are conservatively taken to be the maximum of the total up and down variation. The last two columns of Tab. 5.11 display the theoretical predictions from the simulated top quark pair samples using either POWHEG or MG5_aMC@NLO as a matrix element generator. These are also visually depicted in Figs. 5.26 and 5.27 as the blue and red lines. In these figures the $t\bar{t}$LF cross section is scaled down by a factor 100 (50) in the visible (full) phase space to more clearly display the results on the same scale.

TABLE 5.11: Results on the parameters of interest in the visible and full phase space with uncertainties. The systematical uncertainties are conservatively taken to be the maximum of the total up and down variation. The last two columns display the theoretical predictions from the simulated top quark pair samples using either POWHEG or MG5_aMC@NLO as a matrix element generator.

| | Result | Unc. (stat. + syst.) | POWHEG | MG5_aMC@NLO |
|---|---|---|---|---|
| Visible Phase Space | | | | |
| $\sigma_{t\bar{t}c\bar{c}}$ [pb] | 0.278 | ± 0.028 (stat.) ± 0.049 (syst.) | 0.303 | 0.303 |
| $\sigma_{t\bar{t}b\bar{b}}$[pb] | 0.195 | ± 0.011 (stat.) ± 0.023 (syst.) | 0.149 | 0.158 |
| $\sigma_{t\bar{t}LF}$ [pb] | 19.1 | ± 0.25 (stat.) ± 2.32 (syst.) | 22.6 | 24.1 |
| $R_c$ [%] | 1.42 | ± 0.14 (stat.) ± 0.21 (syst.) | 1.31 | 1.26 |
| $R_b$ [%] | 1.00 | ± 0.06 (stat.) ± 0.09 (syst.) | 0.64 | 0.65 |
| Full Phase Space | | | | |
| $\sigma_{t\bar{t}c\bar{c}}$ [pb] | 5.86 | ± 0.48 (stat.) ± 1.03 (syst.) | 6.71 | 5.87 |
| $\sigma_{t\bar{t}b\bar{b}}$ [pb] | 3.53 | ± 0.20 (stat.) ± 0.45 (syst.) | 2.87 | 2.57 |
| $\sigma_{t\bar{t}LF}$ [pb] | 176.9 | ± 2.6 (stat.) ± 22.5 (syst.) | 221 | 183 |
| $R_c$ [%] | 3.15 | ± 0.29 (stat.) ± 0.48 (syst.) | 2.92 | 3.22 |
| $R_b$ [%] | 1.90 | ± 0.10 (stat.) ± 0.15 (syst.) | 1.25 | 1.41 |

This analysis provides a first measurement of the $t\bar{t}c\bar{c}$ cross section of $\sigma_{t\bar{t}c\bar{c}}^{\text{vis}} = 0.278 \pm 0.028\,(\text{stat.}) \pm 0.049\,(\text{syst.})$ pb in the visible phase space and $\sigma_{t\bar{t}c\bar{c}}^{\text{full}} = 5.86 \pm 0.48\,(\text{stat.}) \pm 1.03\,(\text{syst.})$ pb in the full phase space. The measurement reaches a precision of around 20% which will provide a much more accurate estimation of this background in the $t\bar{t}H$ analyses. The ratio of the $t\bar{t}c\bar{c}$ to the inclusive $t\bar{t}jj$ cross section is found to be $R_c^{\text{vis}} = 1.42 \pm 0.14\,(\text{stat.}) \pm 0.21\,(\text{syst.})$ %

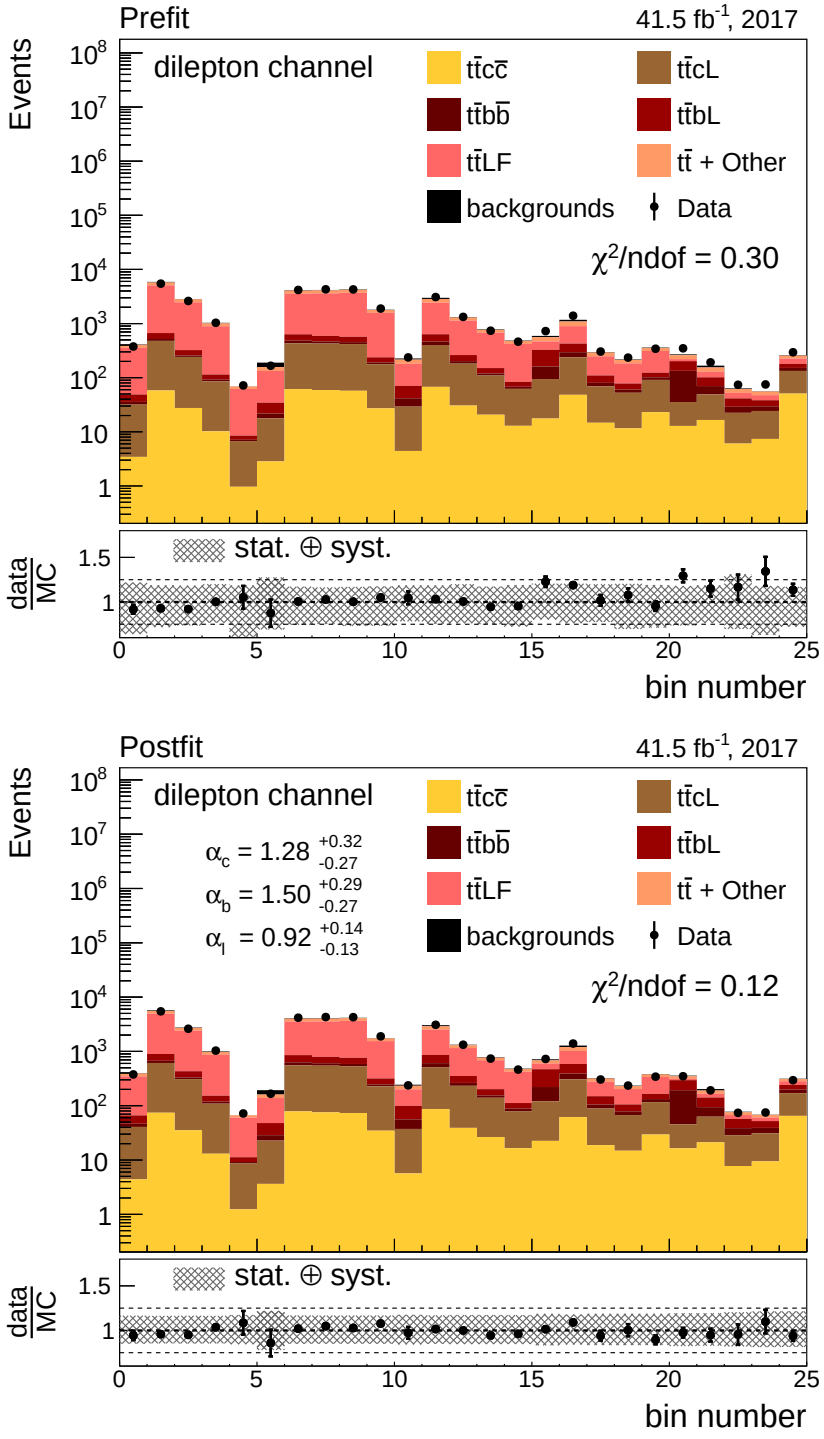


FIGURE 5.25: Unrolled distribution of the two–dimensional distribution of the $\Delta_b^c$ and $\Delta_L^c$ discriminators onto a one–dimensional histogram. The top figure shows the data–to–simulation agreement before the fit. The bottom figure shows the results after the fit, where the scale factors $\alpha_c$, $\alpha_b$ and $\alpha_l$ are also shown with their uncertainties.

in the visible phase space and $R_c^{\text{full}} = 3.15 \pm 0.29\,(\text{stat.}) \pm 0.48\,(\text{syst.})$ % in the full phase space. The ratios are obtained with a slightly better precision of around 18% compared to the cross sections. This is related to some highly suppressed theoretical

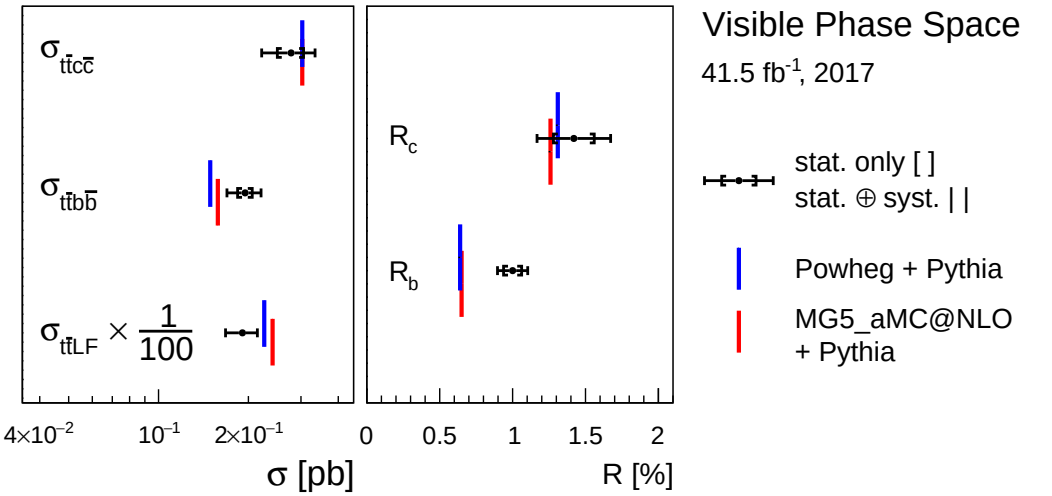


FIGURE 5.26: Summary of the results in Tab. 5.11 for the visible phase space. The theoretical predictions using either the POWHEG or MG5_aMC@NLO matrix element generators are shown in blue and red respectively.

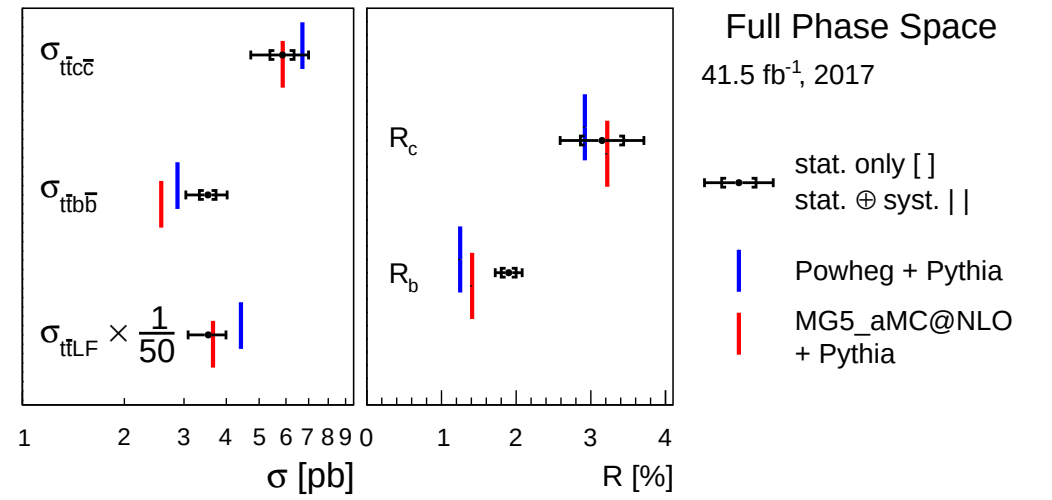


FIGURE 5.27: Summary of the results in Tab. 5.11 for the full phase space. The theoretical predictions using either the POWHEG or MG5_aMC@NLO matrix element generators are shown in blue and red respectively.

uncertainties (see Tabs. 5.9 and 5.10), though the dominant uncertainties come from the $b$– and $c$–tagging calibrations which clearly do not vanish in these ratios. The measured values for $\sigma_{t\bar{t}c\bar{c}}$ and $R_c$ coincide well with the theoretical expectations.

An overall good agreement is observed also for the other fitted parameters with the theoretical predictions from the generators, although the $t\bar{t}b\bar{b}$ cross section and its ratio to the inclusive $t\bar{t}jj$ production seem to be underestimated in the simulations. This trend is also observed in previous dedicated $t\bar{t}b\bar{b}$ analyses (using the dilepton decay channel) in CMS at center–of–mass energies of 8 and 13 TeV. The $t\bar{t}b\bar{b}$ analysis at 8 TeV using 19.6 fb$^{-1}$ [11] measured a value of $R_b = 2.2 \pm 0.4$ (stat.) $\pm$ 0.5 (syst.)%, in the full phase space with a minimum $p_T$ of the jets above 40 GeV, to be compared to the corresponding NLO prediction of 1.1%. Due to the relatively large total uncertainty, this result is still consistent with the theoretical prediction

within two standard deviations. The $t\bar{t}b\bar{b}$ analysis at 13 TeV using 2.3 fb$^{-1}$ [12] measured a value of $R_b = 2.2 \pm 0.3$ (stat.) $\pm 0.6$ (syst.)%, in the full phase space with a minimum $p_T$ of the jets above 20 GeV, to be compared to the corresponding NLO prediction from POWHEG simulations of 1.2%. Again the result is almost twice as large as the predicted value, with a relatively large uncertainty. The analysis presented in this manuscript indeed observes a similar underestimation of $R_b$ in the simulations. Using 41.5 fb$^{-1}$ of proton–proton collisions at 13 TeV, we obtain a value of $R_b = 1.9 \pm 0.10$ (stat.) $\pm 0.15$ (syst.)% in the full phase space with a minimum $p_T$ of the jets above 20 GeV. Compared to the predicted values by POWHEG and MG5_aMC@NLO of 1.25 % and 1.41 % respectively, this result deviates from the predictions to the level of around three standard deviations. This consistent discrepancy motivates a profound investigation of the modeling of this complex final–state. Efforts from the theory community on this front should go hand–in–hand with further explorations of this result in the dedicated $t\bar{t}b\bar{b}$ analyses. Whereas this analysis is focused on the measurement of the $t\bar{t}c\bar{c}$ process, it is expected that updated analyses dedicated to measuring the $t\bar{t}b\bar{b}$ process will soon provide more accurate results. Also the results from ongoing analyses in the semileptonic and fully hadronic decay channel of the top quark pairs will provide indispensable information to either confirm or counter the observed differences in the dileptonic decay channel. Extending these analyses to measure also differential cross sections as a function of for example the transverse momenta of the additional jets could provide interesting information. Such differential measurements could indicate whether the observed differences are enhanced or suppressed in specific kinematic phase space regions. To this end it could be advisable to apply additional $b$–tagging requirements on the additional jets to enhance the purity of the final–state in $t\bar{t}b\bar{b}$ events.
The theoretical modeling of this final–state could perhaps require a more adequate choice of the renormalization and factorization scales or perhaps need a calculation to higher orders in perturbation theory (NNLO). Whether to model the additional HF jets directly in the ME calculation or rather in the parton showering is also a subject under study [172]. Alternatively, the larger observed $t\bar{t}b\bar{b}$ cross section could potentially be the result of a new physics phenomenon, such as the decays of heavy vector–like quarks into top and bottom quarks [38] or a heavy resonance (such as a Z' or W' vector boson) that couples to top and/or bottom quarks [275]. In Chapter 6 a novel ML–based method will be proposed and demonstrated using the $t\bar{t}c\bar{c}$ process, that can be used in the future to interpret as well this $t\bar{t}b\bar{b}$ result in the framework of the SMEFT. It will therefore be interesting to follow up with future results from the $t\bar{t}b\bar{b}$ final–state both from the experimental collaborations and the theory community.

The values of $R_c$ and $R_b$ are significantly larger in the full phase space, compared to the visible phase space. This is potentially explained by the larger expected production of $b$ and $c$ hadrons in the forward direction (at large pseudorapidities), compared to light jets. The visible phase space is limited to the coverage of the tracker ($|\eta| < 2.4$), whereas the full phase space does not have any restriction on the pseudorapidity of the jets and leptons.

# Chapter 6 Probing new physics in the SMEFT

The Standard Model Effective Field Theory has already been introduced in Sec. 1.6 and provides a model–independent framework to interpret measurements at the LHC. The SMEFT parametrizes BSM interactions that are mediated by new particles with masses beyond the currently accessible energy reach. The effective description of the SMEFT operators does not reveal much about the potentially undiscovered mediators themselves, but may still provide information on the underlying Lorentz and gauge structure of the interactions. The lack of signs of new physics at the LHC motivates the use of such a generic interpretation framework given that there exist no clear clues on where new physics phenomena may be hiding. The ultimate objective would be to search within the vast amount of analyses at the LHC for any kind of deviation from the SM using the effective new physics operators of the SMEFT. This would allow either to constrain the allowed values of the corresponding Wilson coefficients, or perhaps to discover new physics. In the latter case, the Lorentz and gauge structure of the responsible operators may hint towards possible UV completions that could provide viable models for the yet undiscovered BSM physics phenomena. However, for analyses that are sensitive to multiple SMEFT operators, it is crucial to disentangle the effect of different operators. This way any possible deviation can be linked to a given type of interaction and would therefore provide a clear direction in which to look for plausible BSM models. In this chapter a novel method will be introduced in which multi–class Machine Learning classifiers are used to increase the sensitivity to the SMEFT operators. Additionally, this method allows to disentangle contributions from different types of SMEFT operators to be able to pinpoint the origin of any deviation from the SM predictions.

This chapter will be divided into two parts. The first part, described in Sec. 6.1, will describe a phenomenological study in which this novel ML–based method to interpret analyses in the SMEFT will be introduced. By making use of the $t\bar{t}b\bar{b}$ signature at the LHC as an interesting case study, it will be demonstrated how the four–heavy–quark operators summarized in Tab. 1.5 can be distinguished and optimally constrained using ML classifiers.
In the second part, a demonstration of this novel method will be provided using the results from the measurement of the $t\bar{t}c\bar{c}$ cross section, described in Chapter 5. The $t\bar{t}$+HF analysis will be extended to allow for a SMEFT interpretation with an increased sensitivity of the $t\bar{t}c\bar{c}$ final–state to the two–heavy–two–light operators from Tab. 1.5.

## 6.1 Learning to pinpoint effective operators at the LHC

In order to interpret results from physics analyses in terms of BSM models, it is crucial to provide an exact definition of the phase space in which the result is presented. This allows the members of the theoretical (or phenomenological) HEP community to make predictions for their favorite models in that phase space and compare those predictions to the measurements. For this reason the inclusive and differential measurements conducted by experimental collaborations are usually unfolded to some fiducial phase space. Practically, this means that the results are corrected by estimating and subtracting effects from detector efficiencies and acceptances to allow for an easier comparison to theoretical predictions. Unfortunately this often provides a bottleneck towards an optimal sensitivity to new physics. The effects from SMEFT operators are typically characterized by a heightened energy–dependence of some reconstructed observables. Examples of such observables are invariant masses, transverse momenta, transverse masses, *etc.* and the sensitivity to the SMEFT operators resides in the high–energy tails of these distributions. Unfolding procedures will often reduce the overall sensitivity and it is clearly preferred to perform the interpretation in the most sensitive phase space region. In this chapter it will be shown how Machine Learning classifiers are capable of finding the most sensitive phase space regions for specific types of operators such that they can be most stringently constrained. Information from the final–state kinematics is combined into a neural network classifier that is trained to separate the SM signatures from those of SMEFT operators with different chiral structures. As a case–study it will be shown how a set of four–heavy–quark operators affects the $t\bar{t}b\bar{b}$ process differently for operators with a top quark of either left–handed (LH) or right–handed (RH) chirality. The neural network classifier is trained to distinguish those LH or RH top–quark operators and can therefore identify the phase–space regions in which the most stringent constraints on these LH or RH operators can be obtained.

### 6.1.1 The $t\bar{t}b\bar{b}$ process in the SMEFT

The $t\bar{t}b\bar{b}$ signature has been thoroughly investigated at different center–of–mass energies and in different top–quark decay channels [11–15]. The motivation behind these measurements lies in the top–Higgs sector, but within the experimental collaborations there has not yet been a lot of attention for interpreting these measurements in terms of new physics signatures. In the theory community however, the sensitivity of this final–state to SMEFT operators was first discussed in Ref. [276]. In the last year, the LHC top working group (LHCTopWG) has put together a set of guidelines in how to interpret top–quark related measurements at the LHC in the SMEFT framework [103]. It is therefore clear that also the experimental collaborations aim to include directly in the performed analyses an interpretation of the results in a model–independent[1] framework such as the SMEFT. This section will discuss the virtues of the $t\bar{t}b\bar{b}$ final–state at the LHC when it comes to constraining SMEFT operators.

One of the most profound advantages of the $t\bar{t}b\bar{b}$ process is the fact that it is sensitive to a set of four–quark contact operators involving third–generation quarks,

[1] Even though it was already mentioned several times that the SMEFT provides a model–independent interpretation framework, it is good to keep in mind that there are always some underlying assumptions, for example on the energy scale of the new physics and on the flavor–structure of the operators (see Sec. 1.6).

as summarized in the top rows (4H) of Tab. 1.5. The dominant Feynman diagrams involving a single insertion of such a four–heavy–quark vertex are shown in Fig. 6.1. Some of these operators are also constrained by four–top–quark processes as will be discussed below. However, the set of operators that do not involve four top quarks, but only give rise to vertices with two top quarks and two bottom quarks have not yet been directly constrained[2] and obtain sensitivity in the $t\bar{t}b\bar{b}$ final–state. This includes the operators $O^1_{tb}$, $O^8_{tb}$, $O^1_{Qb}$, $O^8_{Qb}$, $O^1_{QtQb}$ and $O^8_{QtQb}$.

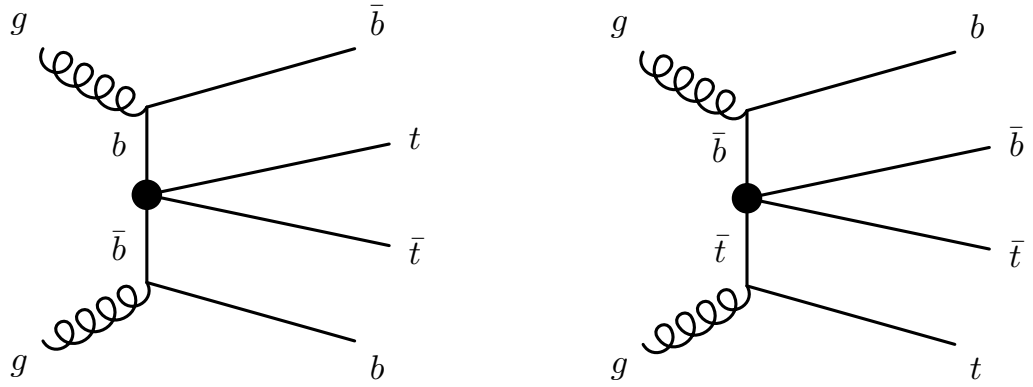


FIGURE 6.1: Dominant EFT contributions to $t\bar{t}b\bar{b}$ production.

### Complementarity to four–top–quark production

Four–heavy–quark operators involving four top quarks can additionally be constrained using measurements (or limits on the signal–strength) of four–top–quark production [107, 108, 278]. The fully right–handed four–top–quark operator, $O^1_{tt}$, can only be constrained from these measurements. However, $O^1_{QQ}$, $O^8_{QQ}$, $O^1_{Qt}$ and $O^8_{Qt}$ can be constrained both by the $t\bar{t}t\bar{t}$ and the $t\bar{t}b\bar{b}$ processes. In fact, the four–top–quark process is only sensitive to one specific linear combination of the first two operators $O^1_{QQ}$ and $O^8_{QQ}$. This can be seen by writing out the corresponding interaction terms:

$$O^1_{QQ} = \frac{1}{2}\left[(\bar{t}_L\gamma^\mu t_L)\,(\bar{t}_L\gamma_\mu t_L) + (\bar{b}_L\gamma^\mu b_L)\,(\bar{b}_L\gamma_\mu b_L)\right] + (\bar{t}_L\gamma^\mu t_L)\,(\bar{b}_L\gamma_\mu b_L)\,,$$
$$O^8_{QQ} = \frac{1}{6}\left[(\bar{t}_L\gamma^\mu t_L)\,(\bar{t}_L\gamma_\mu t_L) + (\bar{b}_L\gamma^\mu b_L)\,(\bar{b}_L\gamma_\mu b_L)\right] + \left(\bar{t}_L\gamma^\mu T^A t_L\right)\left(\bar{b}_L\gamma_\mu T^A b_L\right),$$

where Fierz identities [279] have been used in the first term in $O^8_{QQ}$ to convert this part into a color–singlet structure. Consequently, in four–top–quark production only the linear combination

$$C^{(+)}_{QQ} = \frac{1}{2}C^1_{QQ} + \frac{1}{6}C^8_{QQ}, \tag{6.1}$$

is probed. In contrast, both degrees of freedom are probed independently in $t\bar{t}b\bar{b}$ production, because of the different color structures in the $t\bar{t}b\bar{b}$ terms in $O^1_{QQ}$ and in $O^8_{QQ}$. This lifts the flat direction $C^8_{QQ} = -3C^1_{QQ}$ in the $C^1_{QQ}$– $C^8_{QQ}$ plane, which is present in the SMEFT interpretations of four–top–quark measurements.

For comparison, the best projected sensitivities on $C^1_{QQ}$, $C^8_{QQ}$, $C^1_{Qt}$ and $C^8_{Qt}$ assuming 300 fb$^{-1}$ of integrated luminosity, obtained from the $t\bar{t}b\bar{b}$ study presented

[2] A subset of these operators has however been indirectly constrained through renormalization group induced contributions to electroweak precision observables [277].

in this chapter are compared to other existing bounds from four–top–quark measurements in Tab. 6.1. The first column shows individual 95% CL intervals (for projections to 300 fb$^{-1}$) from four–top–quark production [108] assuming a value of $M_{\text{cut}}$ at 4 TeV. These constraints assume an upper limit of the four–top–quark signal strength, $\mu < 1.87$, obtainable at the LHC with 300 fb$^{-1}$ at 13 TeV [280]. The second column quotes results from the four–top–quark measurement by CMS in the single lepton and opposite–sign dilepton final–states [107], using 35.8 fb$^{-1}$ at 13 TeV. No upper threshold on the allowed energy scales has been applied in this analysis. The third column shows the bounds from a global fit of multiple SMEFT operators to top–quark related measurements at center–of–mass energies of 8 and 13 TeV [105]. These bounds result mostly from the CMS measurement of four–top–quark production in the same–sign dilepton and multilepton final–states [73] at 13 TeV using 35.9 fb$^{-1}$. The final column represents the best sensitivities obtained from the $t\bar{t}b\bar{b}$ study presented in this chapter for comparison.

TABLE 6.1: Comparison between the sensitivity of $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ production to the mutual SMEFT operators. The first column shows individual 95% CL intervals (for projections to 300 fb$^{-1}$) from four–top–quark production [108] at 13 TeV, assuming a value of $M_{\text{cut}}$ at 4 TeV. The second column quotes 95% CL intervals from the four–top–quark measurement by CMS in the single lepton and opposite–sign dilepton final–states [107], using 35.8 fb$^{-1}$ at 13 TeV. The third column shows the bounds from a global fit of multiple SMEFT operators to top–quark related measurements at center–of–mass energies of 8 and 13 TeV [73, 105]. The last column compares these intervals to the best constraints from $t\bar{t}b\bar{b}$ production (assuming 300 fb$^{-1}$ at 13 TeV) obtained in this work.

| | 4-top (300 fb$^{-1}$) ($M_{\text{cut}} = 4$ TeV) | 4-top (35.8 fb$^{-1}$) (no $M_{\text{cut}}$) | global fit (no $M_{\text{cut}}$) | $t\bar{t}b\bar{b}$ (300 fb$^{-1}$) ($M_{\text{cut}} = 2$ TeV) |
|---|---|---|---|---|
| $C_{QQ}^{1}$ | $[-2.8, 2.5]$ | $[-2.2, 2.0]$ | $[-5.4, 5.2]$ | $[-2.1, 2.3]$ |
| $C_{QQ}^{8}$ | $[-8.4, 7.4]$ | n.a. | $[-21, 16]$ | $[-4.5, 3.1]$ |
| $C_{Qt}^{1}$ | $[-2.2, 2.3]$ | $[-3.5, 3.5]$ | $[-4.9, 4.9]$ | $[-2.1, 2.3]$ |
| $C_{Qt}^{8}$ | $[-5.1, 4.1]$ | $[-7.9, 6.6]$ | $[-11, 8.7]$ | $[-3.9, 3.8]$ |

An additional difference between the two processes is the comparative rarity of four–top–quark production. The four–top–quark cross section in 13 TeV $pp$ collisions is expected to be of the order of 9 fb [71], whereas the cross section for the $t\bar{t}b\bar{b}$ process is predicted to be in the order of 2–3 pb (see Tab. 5.11). This results in the limited size of the available datasets to search for four–top–quark production and means that it will only ever be measured at the inclusive level with the available dataset of 300 fb$^{-1}$ that will be collected by the end of Run–3 of the LHC. On the contrary, $t\bar{t}b\bar{b}$ production does not suffer from this low cross section and therefore differential measurements are now already feasible for this process [15]. The methods developed in the rest of this chapter rely on the sufficiently large datasets available to measure differential properties of $t\bar{t}b\bar{b}$ production, assuming a projected integrated luminosity of 300 fb$^{-1}$.

In summary, the EFT interpretation of $t\bar{t}b\bar{b}$ measurements at the LHC presents the following advantages:

- A sufficiently large inclusive cross section that allows for the use of differential information after 300 fb$^{-1}$ of integrated luminosity.
- It directly constrains six four–heavy–quark operators for the first time.
- It breaks the degeneracy in a blind direction of the parameter space spanned by $O^1_{QQ}$ and $O^8_{QQ}$ with respect to four–top–quark measurements.

These arguments make the $t\bar{t}b\bar{b}$ process an indispensable component in a future global fit of top–quark interactions in the SMEFT [105].

### 6.1.2 Validity of the effective field theory

The interpretation of a measurement in terms of an effective field theory is only valid if one ensured that the measurement probes scales below the mass scale of the new physics. This scale, which we will denote as $\Lambda_{\text{NP}}$ is a priori unknown and therefore this ansatz is in practice not verifiable. Instead, an upper limit on the allowed energy of the process under consideration is introduced, denoted by $M_{\text{cut}}$. When performing a measurement, all energy scales in the events are required to be below $M_{\text{cut}}$ for the events to be considered in the interpretation. This way one has a handle on the probed energy scales and the interpretation of the measurement is valid for all models with

$$\Lambda_{\text{NP}} > M_{\text{cut}}. \tag{6.2}$$

Ideally the acquired limits on the Wilson coefficients could be quoted as a function of the value of $M_{\text{cut}}$. The higher the value of $M_{\text{cut}}$ is chosen, the better the constraints on the Wilson coefficients will be but at the cost of a smaller landscape of viable UV models that are consistent with the obtained limits.

An additional difficulty arises from the fact that the actual scale of new physics is degenerate with the value of the Wilson coefficient as shown in Eq. (1.27). It is only possible to constrain the ratio $C_i/\Lambda^{d-4}$, where $d$ is the dimensionality of the operator. Typically one either fixes the value of $\Lambda$ and puts limits on the value of the Wilson coefficients, or the limits are quoted on the allowed new physics energy scales, assuming a value of the Wilson coefficient of $\pm 1$. In what follows, the value of $\Lambda$ will be fixed to 1 TeV and will for simplicity be absorbed in the definition of the Wilson coefficient $C_i$, which will be expressed in units of TeV$^{-2}$ for dimension six operators. Because of this ambiguity, quantitive statements on the validity of the EFT require some assumptions on the power counting scheme of the EFT and therefore on the nature of the UV completions. Following the conventions of the SILH description on the Higgs EFT [281], we assume that there is only one single BSM coupling $g_*$ corresponding to $\Lambda_{\text{NP}}$. An arbitrary EFT operator can then be built using the following building blocks:

$$\mathcal{L}_{\text{EFT}} = \frac{\Lambda^4_{NP}}{g^2_*}\mathcal{L}\left(\frac{D_\mu}{\Lambda_{NP}}, \frac{g_* H}{\Lambda_{NP}}, \frac{g_* f_{L,R}}{\Lambda^{3/2}_{NP}}, \frac{g F_{\mu\nu}}{\Lambda^2_{NP}}\right). \tag{6.3}$$

To clarify this notation, it can be observed that derivatives do not introduce any powers of the BSM coupling $g_*$ and are of dimension one. When an operator includes scalar fields ($H$) or fermionic fields ($f_{L,R}$), each insertion of such a field introduces

one power of $g_*$. Scalar fields are of dimension one, whereas fermionic fields are of dimension $3/2$. Finally, if a gauge vector field is introduced in the operator, it is assumed that the coupling of this gauge field corresponds to its own gauge coupling ($g$) instead of the BSM coupling $g_*$, which is a natural assumption. The factor $\Lambda_{\rm NP}^4/g_*^2$ in front normalizes the Lagrangian to be of the right dimension and order in $g_*$. As an example, the four–quark operator $O_{Qb}^1$ is constructed from two LH fermionic fields and two RH fermionic fields, which according to Eq. (6.3) scales as

$$\left(\frac{\Lambda_{\rm NP}^4}{g_*^2}\right)\left(\frac{g_* f_L}{\Lambda_{NP}^{3/2}}\right)^2\left(\frac{g_* f_R}{\Lambda_{NP}^{3/2}}\right)^2 \quad\sim\quad \frac{g_*^2}{\Lambda_{NP}^2} f_L f_L f_R f_R.$$

This shows that four–quark operators come with a coefficient of the order of $C = g_*^2/\Lambda_{NP}^2$. Combining this with the validity requirement in Eq. (6.2), this can be rewritten as $|C_i|\, M_{\rm cut}^2 \lesssim g_*^2$. The perturbative description of the EFT requires $g_* \lesssim 4\pi$, and therefore

$$|C_i|\, M_{\rm cut}^2 \lesssim (4\pi)^2 . \tag{6.4}$$

This expression was already introduced in Eq. (1.29) (after absorbing $\Lambda$ in the definition of the Wilson coefficient) and will be used later on to determine a proper value for $M_{\rm cut}$ that ensures the validity of the obtained limits on $C_i$.

Finally this power counting scheme allows to determine the relative contributions from dimension six interference (i.e. linear in the Wilson coefficient) or quadratic terms, as expressed in Eq. (1.28) and to motivate the omission of dimension eight operators[3] in the production of the $t\bar{t}b\bar{b}$ final–state. Dimension six interference and quadratic contributions to the $gg \to t\bar{t}b\bar{b}$ cross section are of the order of

$$\underline{\texttt{dim-6 interference}}\text{:}\ \frac{g_s^6 g_*^2 E^2}{\Lambda_{NP}^2}, \tag{6.5}$$

$$\underline{\texttt{dim-6 quadratic}}\text{:}\ \frac{g_s^4 g_*^4 E^4}{\Lambda_{NP}^4}. \tag{6.6}$$

Here $E$ is the largest energy scale in the $gg \to t\bar{t}b\bar{b}$ process and is limited from above by $M_{\rm cut}$. Even though the quadratic contribution is suppressed by a higher power of $\Lambda_{\rm NP}$, it is possible for the quadratic term to become dominant over the interference term if $g_*$ is sufficiently large such that $(g_*/g_s)^2 E^2/\Lambda_{NP}^2 \gtrsim 1$. It will be shown later on that in this phenomenological analysis the obtained limits indeed lie within this regime and therefore both the interference and the quadratic terms should be included.

The relative size of the dimension eight interference with respect to the dimension six quadratic contribution is now assessed to motivate the omission of dimension eight operators. A four–fermion operator of dimension eight would have the schematic form $(f_{L,R})^4 D^2$ and according to Eq. (6.3) a coefficient of order $g_*/\Lambda_{\rm NP}^4$. This would lead to the following interference contributions to the $gg \to t\bar{t}b\bar{b}$ cross section

$$\underline{\texttt{dim-8 interference}}\text{:}\ \frac{g_s^6 g_*^2 E^4}{\Lambda_{NP}^4} \tag{6.7}$$

[3]Remember that operators of odd–numbered dimensions can only generate baryon or lepton number violating processes and are not considered here.

and is for sure subleading to the dimension six interference as long as Eq. (6.2) is satisfied. It is also subleading to the dimension six quadratic contribution if one assumes $g_* > g_s$, i.e. the UV completion of the EFT is a strongly coupled theory. Also other dimension eight operators of the schematic form $(f_{L,R})^4 G_{\mu\nu}$ or $(f_{L,R})^4 D_\mu D_\nu$ could contribute to the $t\bar{t}b\bar{b}$ process. These would generate $gt\bar{t}b\bar{b}$ and $ggt\bar{t}b\bar{b}$ contact interactions. The gluonic field $G_{\mu\nu}$ would come with its own gauge coupling, namely the strong coupling $g_s$. Therefore the coefficients of $(f_{L,R})^4 G_{\mu\nu}$ and $(f_{L,R})^4 D_\mu D_\nu$ are of the order $^{(g_*^2 g_s)}/_{\Lambda^4_{NP}}$ and $^{(g_*^2)}/_{\Lambda^4_{NP}}$ respectively. The interference contributions to the SM $gg \to t\bar{t}b\bar{b}$ amplitude from these $gt\bar{t}b\bar{b}$ and $ggt\bar{t}b\bar{b}$ vertices would be of the same order as in Eq. (6.7) and would also be subleading. Using this (model–dependent) power counting scheme we have thus motivated the omission of any dimension eight operators.

In summary, we will ensure that the validity requirement in Eq. (6.4) is fulfilled and include dimension six interference and quadratic contributions, while neglecting higher order operators. The latter is motivated by the arguments given above in case of a strongly coupled UV completion[4].

### 6.1.3 Strategy of the phenomenological analysis

The most straightforward way to interpret an analysis in terms of a set of SMEFT operators is to assess the effect an operator has on the cross section of the process under consideration, using Eq. (1.28). The Wilson coefficients can then be constrained from a cross section measurement in a predefined phase space. Here we present a new methodology that uses Machine Learning classifiers to search for the most sensitive phase space region of each coefficient (or group thereof) and to optimize the constraints that can be obtained on the Wilson coefficients. This method will gradually be introduced by considering first one operator at a time and proceeding through the following steps:

1. First the latest CMS measurement of the $t\bar{t}b\bar{b}$ cross section at 13 TeV [12] will be used to derive individual constraints on the Wilson coefficients of the four–heavy–quark operators in Tab. 1.5. This measurement was conducted using the dilepton decay channel and uses only 2.3 fb$^{-1}$ of integrated luminosity. This resulted in a measurement of $\sigma_{t\bar{t}b\bar{b},\text{CMS}} = 0.088 \pm 0.012(\text{stat.}) \pm 0.029(\text{syst.})$ pb in a specific fiducial phase space (see below), i.e. with a total uncertainty of $\sim 35\%$. This will provide some very first individual constraints on some of these operators that have not yet been directly constrained in the past.

2. Projections to a total integrated luminosity of 300 fb$^{-1}$ at 13 TeV will be made, scaling the statistical uncertainty accordingly and making some assumptions on the achievable level of the systematic uncertainties by the end of Run–3 of the LHC.

3. Using the projections for 300 fb$^{-1}$ as a benchmark, the sensitivity will be increased by applying event selections and moving towards a reconstructed phase space. It will be shown that by imposing a minimal threshold on the invariant mass of the four $b$ jets in the $t\bar{t}b\bar{b}$ events one obtains more stringent limits on the Wilson coefficients.

4. Instead of focusing on one sensitive kinematical observable, a combination of all the available kinematic information in the final–state into a neural network

[4] An example of such a strongly coupled UV completion is given in Appendix A of Ref. [109].

classifier will be shown to increase significantly the sensitivity. The multi–class neural network discriminator allows to select a phase space enriched in SMEFT contributions, but will additionally provide a way of distinguishing different types of SMEFT operators. As a case–study it will be demonstrated how the network learns to distinguish between operators with LH top–quark currents and those with RH top–quark currents. This benefits the sensitivity to the SMEFT operators by selecting even more tailored phase space regions that are enriched in contributions from either LH or RH top quark operators. Also template–fitting methods to the discriminator outputs are explored to further optimize the sensitivity.

Finally the strength of this multi–class neural network architecture will be demonstrated in the case where more than one operator is allowed to have a non–zero Wilson coefficient. The distinction between LH and RH top–quark operators leads to the most stringent constraints under the SM–only hypothesis, and allows to pinpoint most accurately the origin of a hypothetical excess in the data.

In the following paragraphs, the details on the event simulation and reconstruction will be outlined in more detail, together with a discussion on the statistical procedures that were used and the assumptions behind the projections towards 300 $\text{fb}^{-1}$ of integrated luminosity.

### Simulation

The SMEFT model including the four–heavy–quark operators is generated using the UFO model[5] DIM6TOP [103] that includes both the SM and the SMEFT operators of Tab. 1.5. The ME generation of the dileptonic $t\bar{t}b\bar{b}$ final–state is simulated at LO in the four–flavor scheme[6] from 13 TeV $pp$ collisions using MG5_aMC@NLO 2.6.0 [166]. The ME generation is employed in a phase space that mimics as closely as possible the "visible phase space" definition outlined in the CMS measurement of the $t\bar{t}b\bar{b}$ cross section. Events are generated with two oppositely charged leptons (electrons or muons) with $p_T > 20$ GeV and $|\eta| < 2.4$. Additionally at least four particle–level $b$ jets are required with $p_T > 20$ GeV and $|\eta| < 2.5$. The angular separation in $\Delta$R between any two objects is required to be larger than 0.5. To simulate the reconstructed phase space, parton showering and hadronization processes are simulated with PYTHIA8 [167] and the DELPHES [282] detector simulation software is used with the default CMS card to model the detector effects and the final object reconstruction.

### Event reconstruction

The effect of the SMEFT operators on the $t\bar{t}b\bar{b}$ cross section can be evaluated using the fiducial phase space definition at the level of the ME generation which was outlined above. The CMS measurement is unfolded to this phase space and therefore it can be interpreted using the predictions from MG5_aMC@NLO. Also the projections for 300 $\text{fb}^{-1}$ can be estimated from this phase space. However, to further increase the sensitivity, one has to step away from the unfolded result corresponding to the

[5]The event generation is also validated with an independent (private) implementation of the same operators using the FEYNRULES package [168].

[6]This choice was motivated by the recent studies on simulating $t\bar{t}$ + $b$ jet production at the LHC [172], suggesting the most accurate predictions at LO with the four–flavor scheme compared to the five–flavor scheme.

fiducial detector volume, and instead impose further selection requirements on the reconstructed objects. These object selections again try to mimic as closely as possible those in the CMS analysis. Events are selected with two reconstructed, isolated leptons (electrons or muons) with $p_T > 20$ GeV and $|\eta| < 2.4$. A minimal threshold on the missing transverse energy is chosen to be $\not{E}_T > 30$ GeV. At least four jets must be present with $p_T > 30$ GeV and $|\eta| < 2.5$. In the CMS analysis the jets are ranked according to decreasing values of the $b$–tagging discriminator and the first two in this list are required to be $b$–tagged and are assigned to the top–quark decays. The $b$–tagging discriminator information is not available in DELPHES, which merely uses a parametrized $b$–tagging efficiency for a fixed working point. Instead, the four jets with the highest $p_T$ are identified and the one closest in $\Delta$R to the leptons are identified as those coming from the top–quark decays. The two remaining jets are ordered by decreasing value of their transverse momentum and are identified as the *additional jets*. Finally, the upper threshold on the allowed energy scales ($M_{\text{cut}}$) as discussed in Sec. 6.1.2, is imposed on all combinations of invariant masses of final–state particles (including the total invariant mass of all final–state objects) as well as the scalar sum of transverse momenta, $H_T$. The chosen value of $M_{\text{cut}}$ is taken at 2 TeV and is motivated later in Sec. 6.1.4.

### Sensitivity analysis

In order to evaluate the sensitivity to the individual SMEFT operators, we start by calculating the functional dependence of an observable $O$ on each of the Wilson coefficients, one at a time. This corresponds to Eq. (1.28) in the case that only one operator is considered and leads to

$$O_{\text{fit}} = O_{SM}\left(1 + p_1 \cdot C_i + p_2 \cdot C_i^2\right), \tag{6.8}$$

where $O_{\text{fit}}$ is the total observed value of the observable, $O_{SM}$ is its SM prediction, $C_i$ is the value of the Wilson coefficient and $p_i$ ($i \in \{1, 2\}$) are parameters to be determined. In this notation, $p_1$ represents the fractional importance of the interference of the SMEFT operator with the SM and $p_2$ signifies the squared contribution to the observable at quadratic order in the Wilson coefficient. The observable, $O$, will be either a cross section or a number of events in a certain phase space, corresponding to a given integrated luminosity or extracted from a template–fitting method. It should be noted that the choice of observable is not restricted to a cross section and can be interpreted more generally as any observable quantity that has some dependence on the SMEFT operator (for example the bin contents of a binned differential distribution). Once this functional dependence is obtained, it can be used to interpret an experimentally measured value $O_{\text{obs}}$. This observed value is either taken from the CMS measurement or it is chosen to be the theoretical prediction obtained from the simulation when future projections are used. The latter allows to compare expected limits at 13 TeV with 300 fb$^{-1}$ for the different approaches that are investigated in this study. By combining the statistical and systematical uncertainties in quadrature, a total uncertainty $\delta O$ is derived on the observed quantity and a $\Delta\chi^2$ is constructed according to

$$\begin{aligned}\Delta\chi^2(C_i|p_1, p_2) &= \chi^2(C_i|p_1, p_2) - \chi^2_{min} \\ &= \frac{\left(O_{\text{fit}}(C_i|p_1, p_2) - O_{\text{obs}}\right)^2}{\delta O^2} - \chi^2_{min},\end{aligned} \tag{6.9}$$

where $\chi^2_{min}$ is the minimum value of the $\chi^2$ function in the EFT parameter space. Limits are quoted using the 95% CL sensitivity interval on the individual Wilson coefficients $C_i$, determined by the region in which the $\Delta\chi^2$ value is lower than 3.84, corresponding to a p–value of 0.05 for a $\chi^2$ distribution with one degree of freedom in the Gaussian limit. The only exception to this resides in Section 6.1.5, when two operators are allowed to vary simultaneously. In this case the corresponding number of degrees of freedom is augmented to two, with a corresponding threshold of 5.991 for the same p–value.

#### Future projections after Run–3

The projected sensitivities for 300 fb$^{-1}$ after Run–3 of the LHC require an adequate assessment of the uncertainties. Statistical uncertainties can be scaled accordingly, but some assumptions are needed on the possible improvements for systematical uncertainties when a much larger dataset is available. The choice was made to quote all results from future projections with a 10% relative systematic uncertainty. This choice is motivated by the fact that the 35% systematic uncertainty of the CMS $t\bar{t}b\bar{b}$ measurement [12] is dominated by the $b$–tagging scale factors. An improved precision by a factor of ∼4 has been observed from $b$–tagging calibration studies performed with 2.6 fb$^{-1}$ [283] and 36.1 fb$^{-1}$ [225]. The corresponding scale factors remain stable up to very large jet transverse momenta (up to ∼1 TeV), which will constitute the sensitive region in this analysis. This causes uncertainties from theoretical modeling to become of similar (or higher) importance once this level of integrated luminosity is reached. This is also observed in the $t\bar{t}$+HF analysis which was presented in Chapter 5. In the CMS $t\bar{t}b\bar{b}$ analysis with 2.3 fb$^{-1}$, this uncertainty was found to be of the order of 17%, consisting mainly of MC generator and parton shower scale variations, whereas in the $t\bar{t}$+HF analysis presented in this manuscript this uncertainty is reduced to around ∼ 10% (see for example Tab. 5.9). Based on the importance of this topology in the Higgs sector, it can be expected that the theoretical modeling of this final–state may improve over the coming years [172, 284] and we deem the estimate of a 10% overall systematic uncertainty reasonable.

### 6.1.4 Constraining individual operators

In this section only one operator will be allowed to have a non–zero value of the Wilson coefficient at a given time. This provides interesting information on the sensitivity of the $t\bar{t}b\bar{b}$ final–state to each of the SMEFT operators and is therefore referred to as a sensitivity study. It should be noted that the obtained limits can be optimistic, given that the sensitivity can be reduced by other operators with similar contributions that can not be disentangled. Nevertheless the emphasis of this section lies on the relative improvement of the limits when subsequently including more and more kinematical information, as outlined in the beginning of Sec. 6.1.3. We start from an interpretation of the unfolded cross section measurement of CMS and the corresponding future projections at the end of Run–3. After this the validity of the EFT is quantified by choosing an appropriate value for $M_{\text{cut}}$. Going beyond the unfolded cross section by applying an event selection on the reconstructed objects, we progressively include more kinematical information to improve the sensitivity. First a selection on the invariant mass of the four $b$ jets in the event is considered. Then a multi–class neural network classifier is trained and the sensitivity is further improved by making a selection on its discriminator outputs, or by means of a template–fitting method to the differential discriminator distributions.

### Cross section interpretation in the fiducial detector volume

First the parametric dependence of the unfolded cross section on the value of the Wilson coefficients is determined for the 10 four–heavy–quark (4H) operators that are listed in Tab. 1.5. The values of the parameters $p_1$ and $p_2$ in Eq. (6.8) are visualized in Fig. 6.2 for each of these operators and immediately show some trends.

- The color singlet operators show a comparatively small interference with the SM and show a larger relative contribution from the squared order contributions. On the contrary, the color octet operators exhibit a relatively large interference compared to the squared order contributions. This can be explained by the fact that the SM $t\bar{t}b\bar{b}$ process is predominantly mediated by QCD processes involving gluons.
- The quadratic contribution of the color octet operators is suppressed with respect to the quadratic contributions of the color singlet operators. This is indeed expected given that a color factor of $^2/_9$ shows up in the color–octet EFT vertices, which is not there for the color singlet operators. The relative suppression is indeed consistent with this factor of $^2/_9$.
- For all operators, the interference contributions would only dominate over the squared order contributions when the Wilson coefficients are below roughly 3.3 ($\text{TeV}^{-2}$). It will be shown later on that this falls below the sensitivity that can be achieved with the $t\bar{t}b\bar{b}$ measurement and hence is out of reach. Squared order contributions are thus dominant and can not be neglected in this analysis.

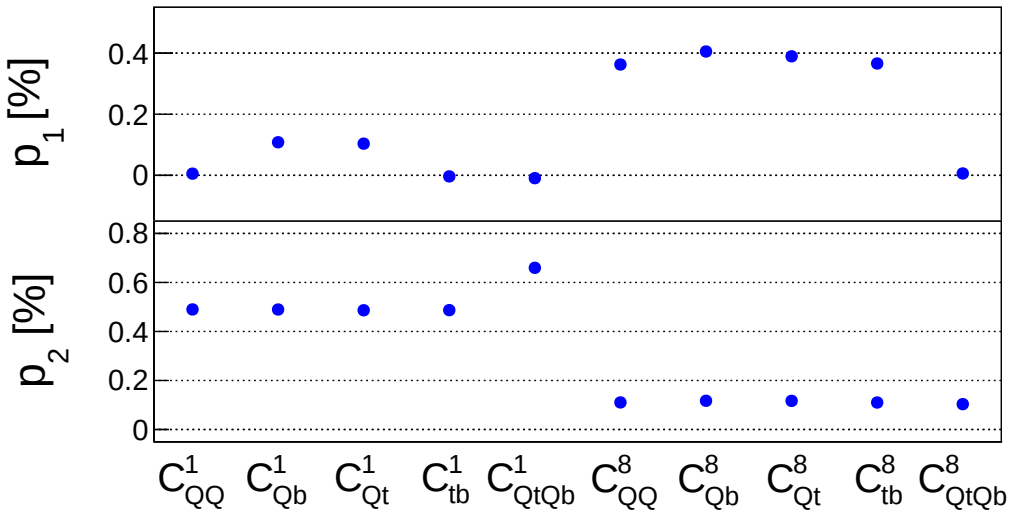


FIGURE 6.2: Coefficients of the fit to the cross section in the fiducial detector volume, for different four–heavy–quark SMEFT operators turned on one by one. The fit function has the form $\sigma_{fit} = \sigma_{SM}\,(1 + p_1 \cdot C_i + p_2 \cdot C_i^2)$. In this notation $\sigma_{SM}$ represents the SM cross section, $p_1$ signifies the fractional importance of the interference of the EFT with the SM and $p_2$ represents the fractional pure EFT contribution to the cross section at quadratic order in the coupling strength of the EFT operator.

Next, the result of the CMS $t\bar{t}b\bar{b}$ measurement in the "visible phase space", $\sigma_{t\bar{t}b\bar{b},\text{CMS}} = 0.088 \pm 0.012(\text{stat.}) \pm 0.029(\text{syst.})$ pb, is compared to the LO

MG5_aMC@NLO prediction of 78 fb and is then interpreted in terms of possible SMEFT contributions. In the next paragraphs we will use one color–singlet operator, $C^1_{Qb}$, and one color–octet operator, $C^8_{Qb}$, as benchmarks to compare the sensitivity of the different methods. A summary for all operators using different methods is shown at the end of this section in Fig. 6.15. Indicative results for the sensitivity to these benchmark operators are shown in Fig. 6.3. In the top panel the red band shows the fitted cross section to the generated sample points with uncertainties (the function is also quoted in red on top of the figure). The CMS measurement with the corresponding 95% CL region is indicated with the light brown band. In the bottom panel the full line is the resulting $\chi^2$ as a function of the Wilson coefficient and the light brown band shows the corresponding 95% CL interval. The minima of the $\chi^2$ are not exactly centered at 0, indicating the fact that the LO theory prediction of the cross section is slightly below the measured value by CMS, but still well within the uncertainty of the measured value.

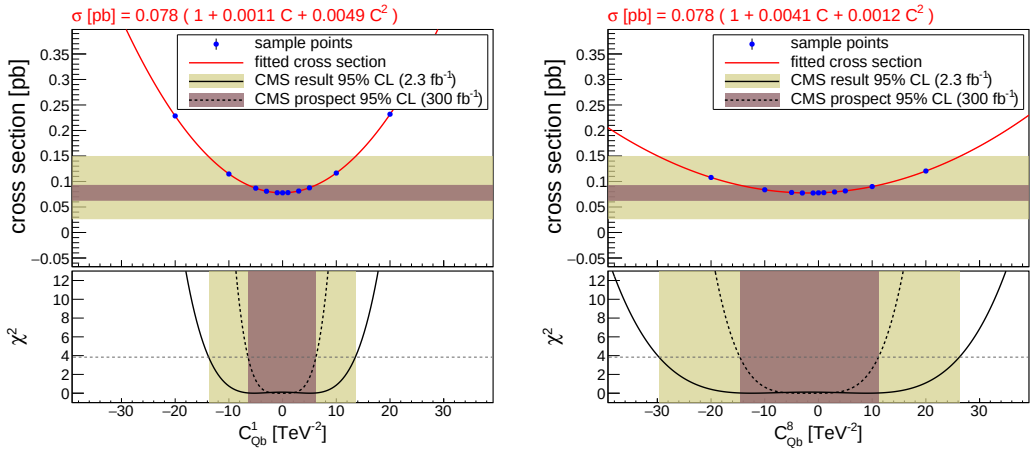


FIGURE 6.3: Limits at 95% CL on $C^1_{Qb}$ (left) and $C^8_{Qb}$ (right) using the CMS result with 2.3 fb$^{-1}$ (full line, light brown area) and prospects for 300 fb$^{-1}$ (dashed line, dark brown area).

Overall it is observed for all operators that both the linear and quadratic contributions stay below the percent level. Limits of the order of $[-14, 14]$ (TeV$^{-2}$) and $[-30, 26]$ (TeV$^{-2}$) are obtained for the benchmark color singlet and octet operators respectively. When assessing the projected sensitivities for 300 fb$^{-1}$, these limits improve to roughly $[-6, 6]$ (TeV$^{-2}$) and $[-14, 11]$ (TeV$^{-2}$) respectively. Note that in this case it is assumed that the observed cross section lies exactly at the SM prediction from the simulation, resulting in expected rather than observed limits. This should be kept in mind when comparing the results for 2.3 fb$^{-1}$ and for 300 fb$^{-1}$. These projections for the benchmark operators are indicated by the dark brown bands and the dotted lines in Fig. 6.3. The limits on all Wilson coefficients are summarized in the left panel of Fig. 6.15 (black and red lines for 2.3 fb$^{-1}$ and 300 fb$^{-1}$ respectively) at the end of this section.

### Quantifying the validity of the EFT

The limits obtained above were derived using a value of $M_{\text{cut}} = 2$ TeV. This choice will now be motivated through a proper assessment of the validity criterium in Eq. (6.4). The dependence of the limits on the value of $M_{\text{cut}}$ is shown in Fig. 6.4 for the two benchmark operators. The full black lines represent this dependence for

the CMS measurement, whereas the dotted lines show the dependence for the future projections. A higher upper threshold for $M_{\text{cut}}$ results in the inclusion of more events in the high–energy regime where the SMEFT contributions are expected to be more abundant. Consequently higher values for $M_{\text{cut}}$ result in better constraints on the operators. However, the validity criterium expressed in Eq. (6.4) is also depicted in Fig. 6.4 as the light pink shaded region. Any point within the light pink shaded region can not be given a meaningful interpretation using an effective description of the interactions and therefore this criterium limits the appropriate choices for the value of $M_{\text{cut}}$ from above[7]. It can be observed that the limits are almost insensitive to the value of $M_{\text{cut}}$ down to ∼1.5 TeV and we therefore fix it to 2 TeV throughout the rest of this study. This ensures that all limits are within the valid regime and we hardly sacrifice any sensitivity. To illustrate this, Fig. 6.5, shows the normalized distributions of the scalar sum of the transverse momentum of all visible objects, $H_T$, in the final–state, comparing the SM contributions (black), with those of the $O^1_{Qb}$ operator with $C^1_{Qb}$ fixed at 10 TeV$^{-2}$ (blue) and 20 TeV$^{-2}$ (red) respectively. This is a representative variable for the typical energy scale of the $t\bar{t}b\bar{b}$ events and indeed it can be seen that only a small fraction of the events are present above $H_T = 2$ TeV.

The dark pink shaded region in Fig. 6.4 represents a more restricted validity region corresponding to the criterium $\frac{|C_i|M^2_{\text{cut}}}{(4\pi)^2} < \kappa^2$, where the specific value of $\kappa^2$ is chosen such that the edge of the new valid region intersects the projected upper limit for $M_{\text{cut}} = 2$ TeV (at 300 fb$^{-1}$). The corresponding value of $\kappa$ provides an estimate of the lower bound on the new physics coupling strength ($g_*$) that would still ensure perturbativity of the EFT description with the obtained limits.

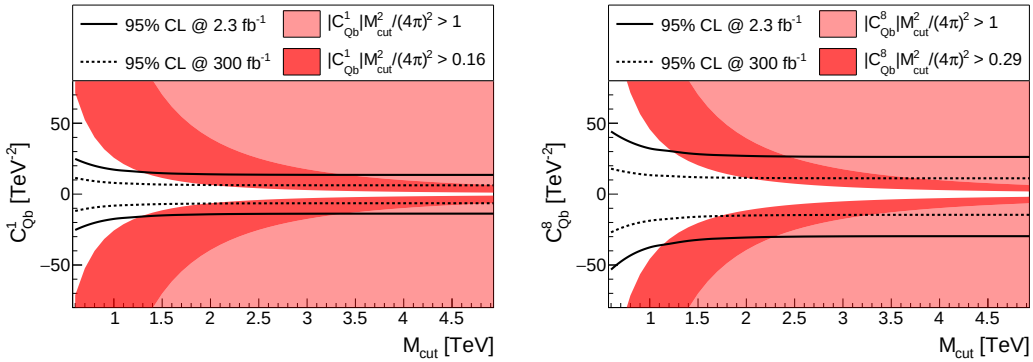


FIGURE 6.4: Limits at 95% CL on $C^1_{Qb}$ (left) and $C^8_{Qb}$ (right) as a function of the mass cut $M_{\text{cut}}$ for an integrated luminosity of 2.3 fb$^{-1}$ (full line) and projections to 300 fb$^{-1}$ (dashed line). The non–perturbative regime of the EFT in which $|C_i|M^2_{\text{cut}} > (4\pi)^2$ is indicated with the light pink shaded region. The darker red region represents a more stringent perturbativity requirement for which the upper limit on the Wilson coefficient (at 300 fb$^{-1}$) intersects the perturbativity threshold at $M_{\text{cut}} = 2$ TeV.

### Selection on the invariant mass of the four b jets

The sensitivity obtained from the cross section measurement in the fiducial phase space of the detector results in relatively weak limits. Additionally, it can only be

[7] If $M_{\text{cut}}$ is chosen too large the obtained limits fall outside of the valid EFT region and have no meaningful interpretation.

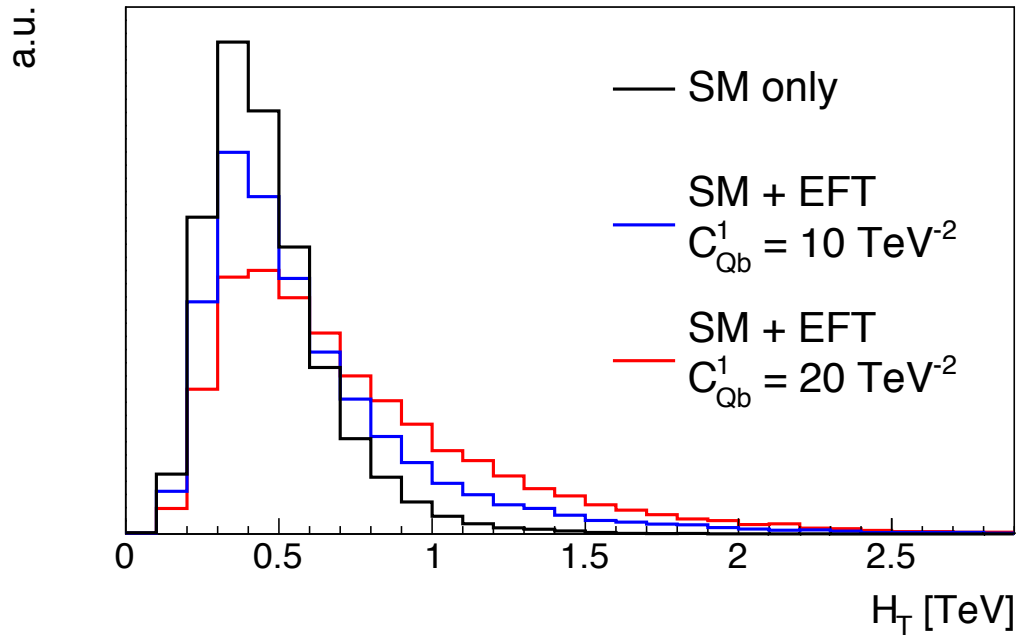


FIGURE 6.5: Scalar sum of transverse momentum of all final–state physics objects in the event ($H_T$).

improved by collecting more data and by trying to reduce as much as possible the systematical uncertainties. A further improvement in the obtained sensitivity can be achieved by moving away from the fiducial phase space and by applying additional event selections to increase the dependence of the observed cross section to the SMEFT operators. After applying the event reconstruction and selections outlined in Sec. 6.1.3, several kinematical variables have been investigated to look for those that are clearly sensitive to the presence of the SMEFT operators. These variables are summarized in Tab. 6.2 and constitute transverse momenta of the leptons and the jets, angular separations in $\Delta$R between combinations of two of these objects and invariant masses of combinations of two, four or even all six observed final–state objects. The separating strength of each variable is calculated by an ANOVA (analysis of variance) F–statistic as defined in Ref. [285]. This value scales with the absolute difference between the mean values of the SM distribution and the EFT distribution for a given variable. At the same time it is inversely proportional to the average variance of the individual SM and EFT distributions. Qualitatively, the F–value thus describes both the overlap between two distributions and the distance between their mean values, thereby providing information on which observables have strong separating power between SM and EFT contributions. The strongest separating power was observed for the invariant mass of the four $b$ jets in the event, $M_{4b}$. Its normalized distribution is shown in Fig. 6.6, comparing the shape of the SM (black) prediction to that of the $O^1_{Qb}$ operator with the Wilson coefficient fixed at 10 TeV$^{-2}$ (blue) and 20 TeV$^{-2}$ (red). This observable is sensitive to the heightened energy dependence of the SMEFT vertices. From the example Feynman diagrams in Fig. 6.1 it can be seen that each time two $b$ jets reside from the EFT vertex (either directly from the production of a $b$ quark or from the decay of a top quark). The calculation of the $M_{4b}$ variable always includes those two $b$ jets with a larger average momentum.

The high–energy tails of the $M_{4b}$ distribution clearly show an increased relative

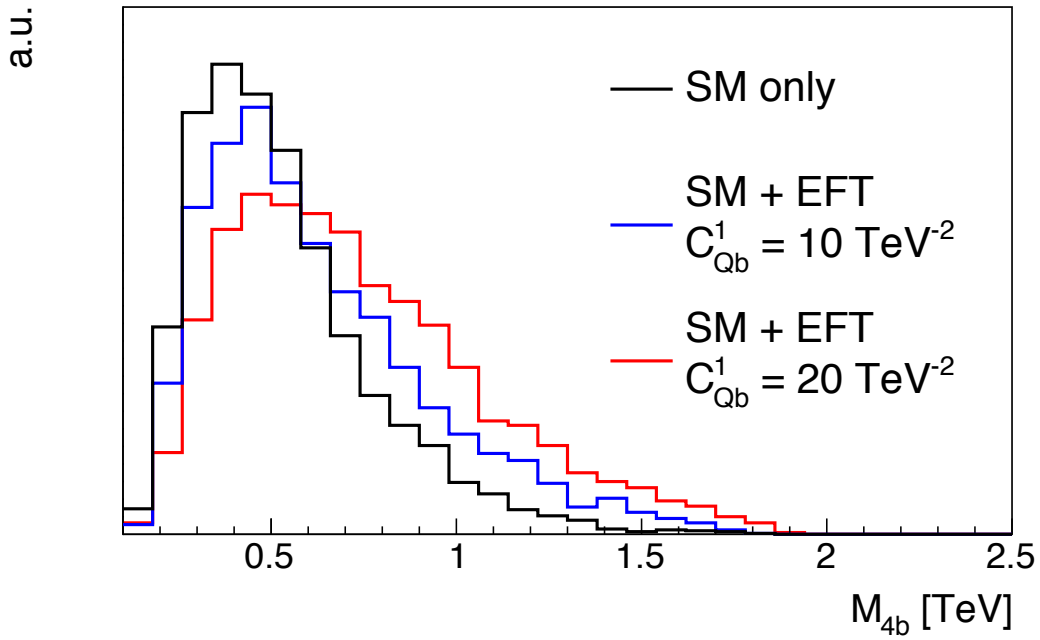


FIGURE 6.6: Invariant mass of the four leading jets in the event ($M_{4b}$) with and without the presence of EFT operators and before cutting on the maximal mass scale $M_{\text{cut}}$.

abundance of SMEFT contributions with respect to the SM. The phase space beyond $M_{4b} = 2$ TeV is already excluded from the measurement by the validity requirement on $M_{\text{cut}}$, but it can once again be seen that a negligible fraction of events is expected beyond that threshold. We will now determine a lower threshold on $M_{4b}$ that maximizes the sensitivity to the SMEFT operators. The higher this threshold is chosen, the higher the relative abundance of SMEFT contributions compared to the SM background, but at the cost of a larger statistical uncertainty due to the reduced size of the selected dataset. To find the optimal selection, Fig. 6.7 shows the dependence of the obtained limits on the lower threshold applied to the invariant mass of the four $b$ jets in the event ($M_{4b}^{sel}$) for the two benchmark operators. The bottom panel in each of these figures shows the drop in visible (*i.e.* observed) cross section with larger values of $M_{4b}^{sel}$. These results motivate a lower threshold of $M_{4b} > M_{4b}^{sel} = 1.1$ TeV.

Finally, after requiring $M_{4b} > M_{4b}^{sel} = 1.1$ TeV, the dependence of the cross section on the Wilson coefficients of the operators in this selected phase space can be once again determined. This is shown by the red band in Fig. 6.8, where these predictions are compared to the expected observation (taken to be the predicted cross section from MG5_aMC@NLO with uncertainties as outlined in Sec. 6.1.3) for the future projections of 300 fb$^{-1}$. The values for the coefficients $p_{1,2}$ lead to a stronger dependence of the cross section on $C_i$ than for the unfolded cross section in Fig. 6.3. The limits on the Wilson coefficients consequently improve, resulting in 95% CL limits between $[-3, 3]$ (TeV$^{-2}$) and between $[-6, 7]$ (TeV$^{-2}$) for the benchmark color singlet and octet operators respectively. The results for all other operators are summarized in Fig. 6.15 (blue) on the right. An improvement in the limits of a factor of $\sim$2 is observed when compared to the sensitivity from the unfolded cross section observable.

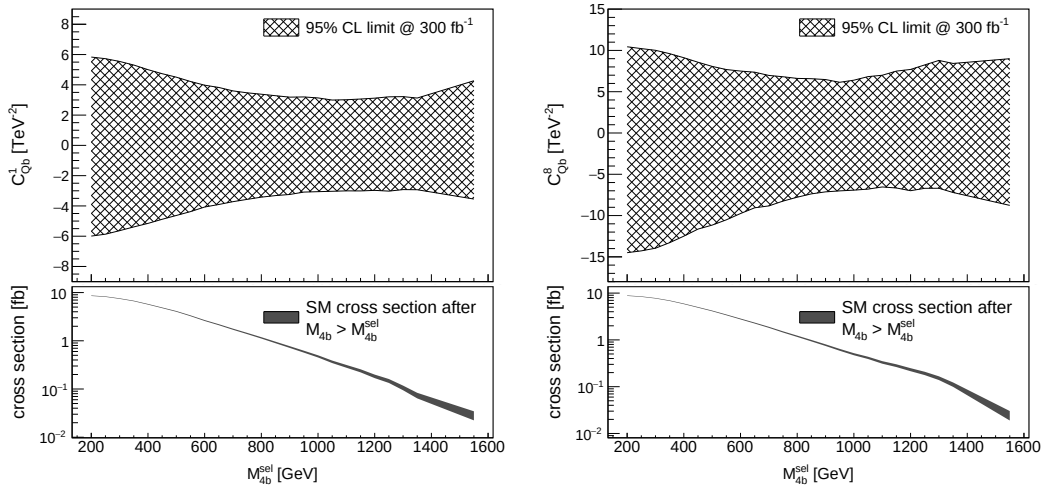


FIGURE 6.7: Individual limits at 95% CL on $C^1_{Qb}$ and $C^8_{Qb}$ as a function of the threshold on the invariant mass of the 4 $b$ jets in the event ($M^{sel}_{4b}$). In the bottom panel the predicted SM cross section as a function of $M^{sel}_{4b}$ is shown, with the corresponding statistical uncertainty shown as a grey band.

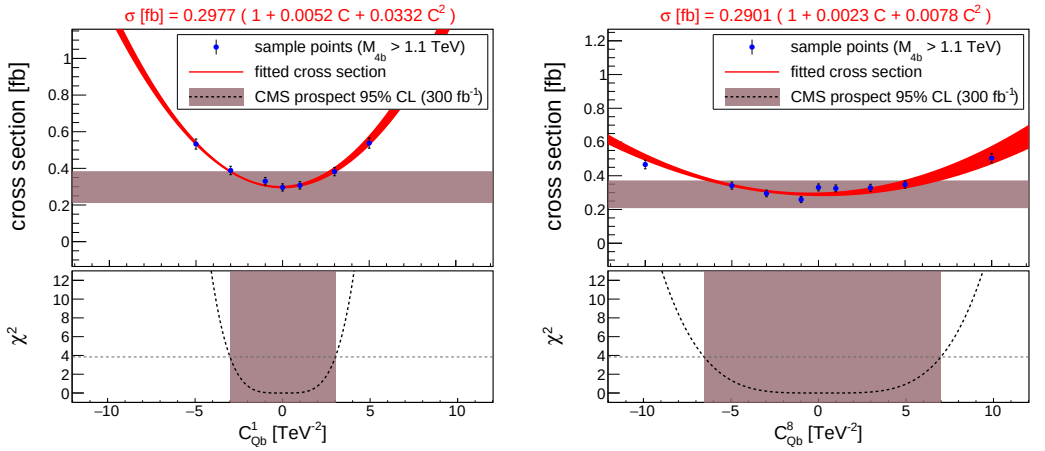


FIGURE 6.8: Limits at 95% CL on $C^1_{Qb}$ (left) and $C^8_{Qb}$ (right) after applying a cut on $M_{4b} > 1.1$ TeV, assuming an integrated luminosity of 300 fb$^{-1}$.

## Neural network classification

An EFT enriched phase space can thus be selected by imposing thresholds on kinematical observables like $M_{4b}$. The most optimal phase space selection, resulting in the best sensitivity to SMEFT operators, is obtained by combining all kinematical information in the final–state. The solution to this multivariate problem lies once again in the use of Machine Learning classifiers. Therefore, all kinematical information summarized in Tab. 6.2 is served as input to a multi–class neural network classifier that is trained to distinguish SM events from SMEFT contributions. However, many analyses are sensitive to multiple operators, which may cause a degeneracy in the interpretation when multiple of these operators are allowed to vary simultaneously. The novelty of the approach presented in this chapter resides in the multi–class output structure of the neural network that is trained to disentangle SM and SMEFT effects, but additionally tries to distinguish between different types of operators. As a case–study it will be shown that a shallow neural network can be

trained to disentangle effects from SMEFT operators with LH top quark currents from those with RH top quark currents in the $t\bar{t}b\bar{b}$ final–state. For now we restrict ourselves to the case of varying only one operator at a time, but the true power of this method stands out when multiple operators are allowed to vary simultaneously, as will be demonstrated in Sec. 6.1.5.

| $\Delta R$ | F-value | $m_{inv}$ | F-value | $p_T$ | F-value |
|---|---|---|---|---|---|
| $\Delta R(\ell_1,\ell_2)$ | 274 | $m_{inv}(\ell_1,\ell_2)$ | 312 | $p_T(\ell_1)$ | 580 |
| $\Delta R(b_1,b_2)$ | 12 | $m_{inv}(b_1,b_2)$ | 8455 | $p_T(\ell_2)$ | 10 |
| $\Delta R(b_1,\ell_1)$ | 1493 | $m_{inv}(b_1,\ell_1)$ | 1505 | $p_T(b_1)$ | 8500 |
| $\Delta R(b_2,\ell_2)$ | 714 | $m_{inv}(b_2,\ell_2)$ | 1673 | $p_T(b_2)$ | 8434 |
| $\Delta R(add_1,add_2)$ | 309 | $m_{inv}(add_1,add_2)$ | 6589 | $p_T(add_1)$ | 9664 |
| | | $m_{inv}(b_1,b_2,add_1,add_2)$ | 14805 | $p_T(add_2)$ | 5081 |
| | | $m_{inv}(\ell_1,\ell_2,b_1,b_2,add_1,add_2)$ | 12895 | | |

TABLE 6.2: Kinematical variables used in the SMEFT neural network. The F–values denote the results of the ANOVA (analysis of variances) F–statistics. The notation is as follows: $b_{1,2}$ denote the $b$ jets assigned to the top quark decays, $add_{1,2}$ refers to the two additional jets, ranked according to decreasing $p_T$ and $\ell_{1,2}$ denote the two charged leptons (arbitrarily assigned label 1 or 2).

A neural network classifier was trained using the 18 observables defined in Tab. 6.2 as inputs. Again a preprocessing of the input features happens to ensure that they have a mean of zero and a variance of one over all training samples. This NN has three output nodes, corresponding to the probability (P) to be a pure SM event, an event from a SMEFT operator with a LH top quark current ($t_L$) and an event from a SMEFT operator with a RH top quark current ($t_R$). The normalized distributions of these observables are shown in Figs. 6.9 and 6.10 for each of the three output classes. In order to unambiguously identify the class labels, the NN was trained using only the squared order contributions from the SMEFT operators. This leads to the clear advantage that the shape of the output discriminators is independent of the value of the Wilson coefficient. It however also means that the network is not capable of learning interference effects. To properly take into account interference effects during the training, one needs to adopt a parametrized learning approach in which the differential shapes of the distributions depend on the value of the Wilson coefficients[8]. We leave this interesting possibility for future studies. The network is expected to learn the differences between the $t_L$ and $t_R$ operators based on the expected kinematic differences from the decay products of LH and RH top quarks. One expects for example a harder leptonic $p_T$ spectrum from RH top–quark decays compared to LH top–quark decays [286]. The training was once again performed using the KERAS deep learning library [258], interfaced with TENSORFLOW [259] as a backend. The 18 input nodes are linked to a fully–connected dense layer with 50 neurons with a rectified linear unit activation. A dropout layer is added which randomly freezes 10% of the neurons in this inner layer. This layer is connected to the 3 ouput nodes with a softmax activation such that the outputs sum up to one. A categorical cross–entropy loss function is used and the minimization of this loss function is performed with a stochastic gradient descent set to an initial learning rate of 0.005 and a decay of $10^{-6}$. The training is performed in mini–batches

[8] In such parametrized networks, the value of the Wilson coefficient would be an additional input variable such that the network is able to learn how the shape of the observables (and consequently also the shape of the output discriminator) depends on the Wilson coefficient.

of 128 events and is stopped after 100 epochs. The training curve is shown in Fig. 6.11, showing a convergence to a plateau for the accuracy (top) and the loss function (bottom), both for the training (red) and an independent testing (blue) dataset.

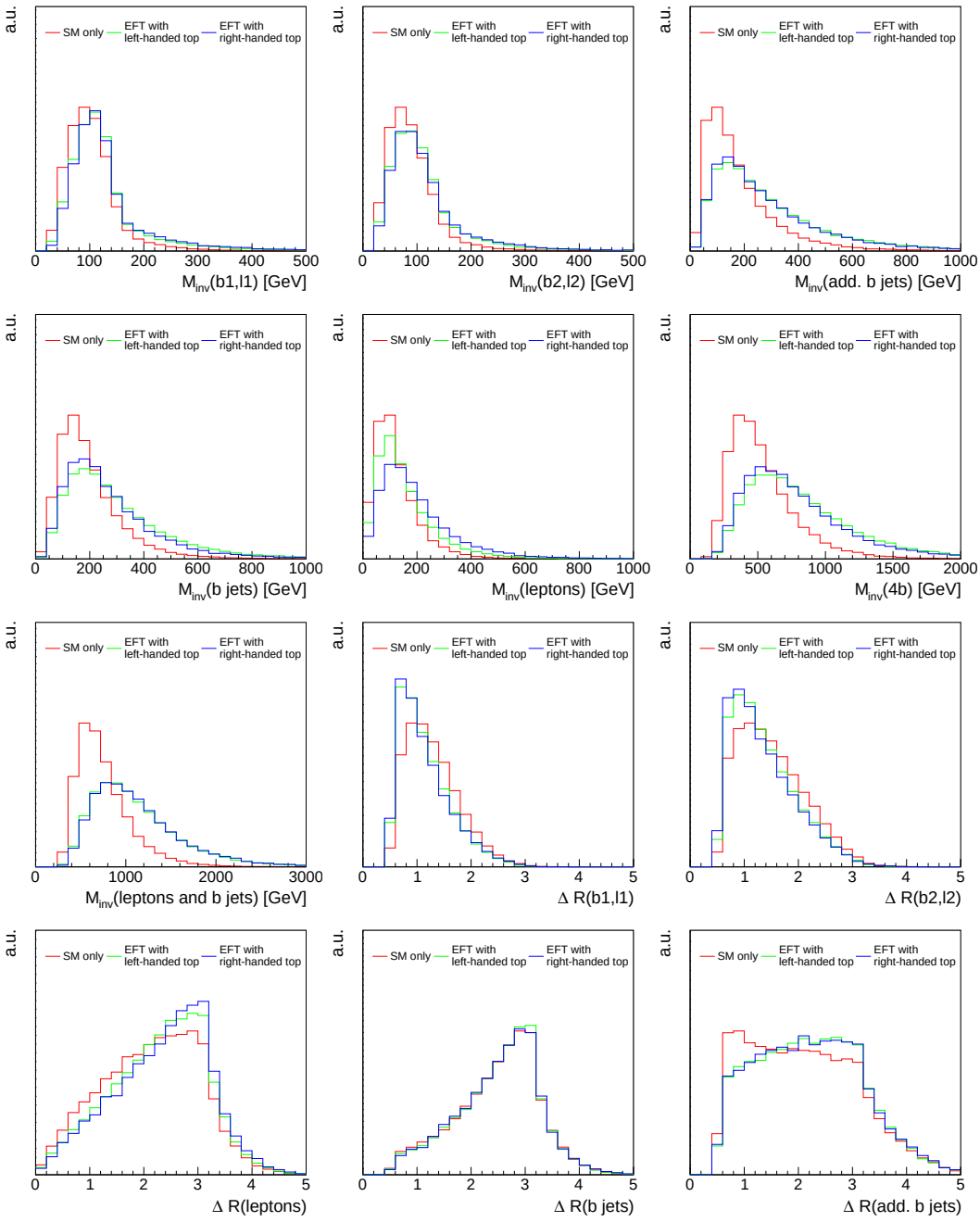


FIGURE 6.9: Normalized distributions of the SMEFT neural network input variables for SM events (red), LH top–quark operators (green) and RH top–quark operators (blue).

By choosing appropriate combinations of the output probabilities, different observables can be constructed for different desired types of discrimination. When only one operator is considered at a time, specialized outputs are constructed that try to distinguish the type of operator (either $t_L$ or $t_R$) from the pure SM contributions.

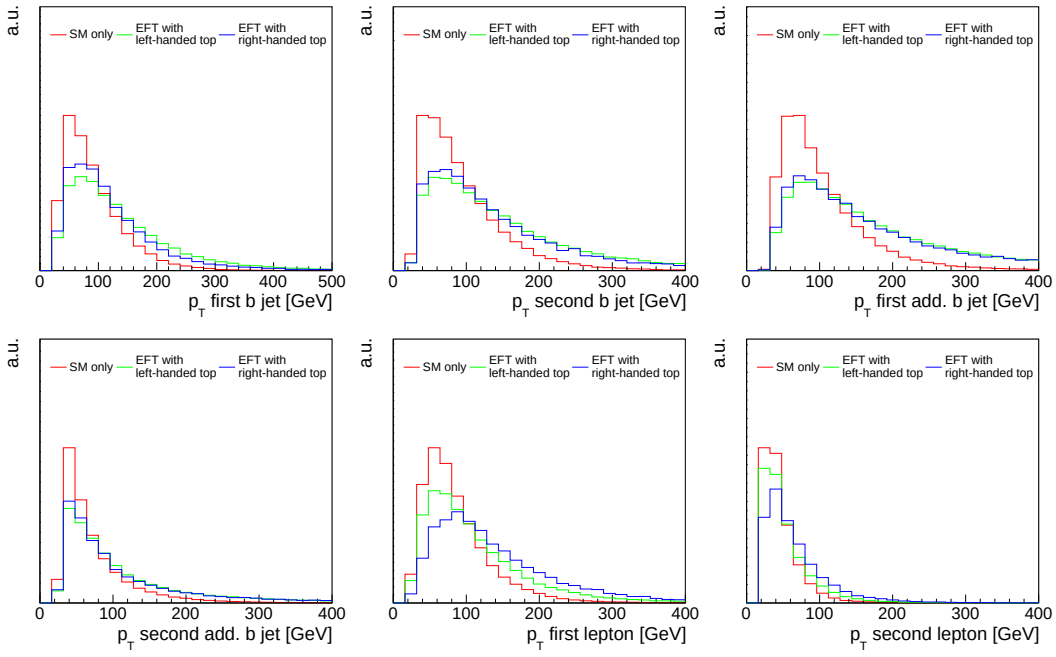


FIGURE 6.10: Continuation of Fig. 6.9.

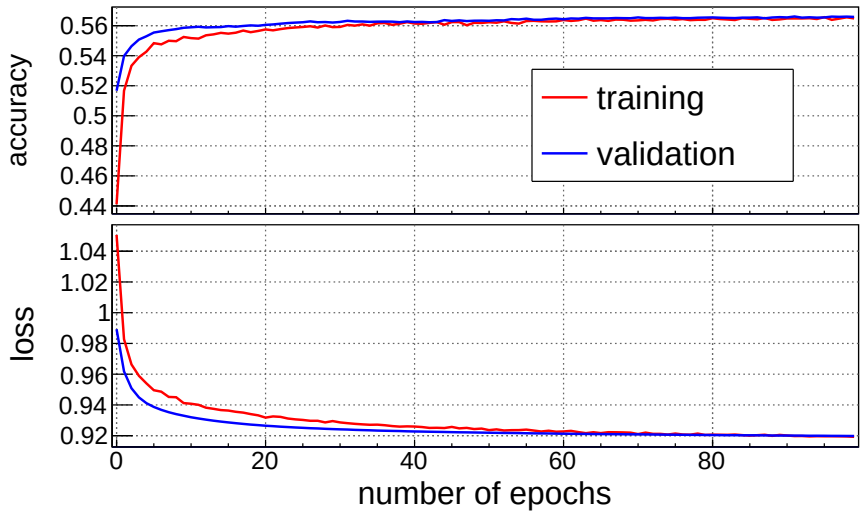


FIGURE 6.11: Training curves of the SMEFT neural network, displaying the evolution of the accuracy (top) and value of the loss function (bottom) for increasing number of epochs. These curves are shown both for the training (red) and for an independent validation data set (blue) and reach a plateau after about 100 epochs.

These discriminators are defined in the two top rows of Tab. 6.3. They will be used below to derive individual limits on the operators. However, when multiple operators with different top quark chirality are considered, it is of interest to construct two discriminators, one to distinguish the SM effects from any kind of SMEFT operator (EFT vs SM) and one to distinguish between the effects caused by either the $t_L$ or $t_R$ operators (ignoring the pure SM effects). These two discriminators are defined in the bottom row of Tab. 6.3. Indeed the sum $P(t_L) + P(t_R)$ allows to disentangle

SM and EFT contributions, as can be seen from the normalized distributions and corresponding ROC curve in the top of Fig. 6.12. The normalized discriminator distribution of the $t_L$ versus $t_R$ discriminator, which is defined as $P(t_L)/[P(t_L)+P(t_R)]$, is shown in the bottom panels of Fig. 6.12 and clearly demonstrates that the NN has learned to distinguish between the $t_L$ and $t_R$ classes.

| | **Desired Discrimination** | **Combined NN Output used for limits** |
|---|---|---|
| only $t_L$ operator | SM vs $t_L$ | $\frac{P(t_L)}{P(t_L)+P(SM)}$ |
| only $t_R$ operator | SM vs $t_R$ | $\frac{P(t_R)}{P(t_R)+P(SM)}$ |
| including both $t_L$ and $t_R$ operators | EFT vs SM | $P(t_L)+P(t_R)$ |
| | $t_L$ vs $t_R$ | $\frac{P(t_L)}{P(t_L)+P(t_R)}$ |

TABLE 6.3: Definitions of the combined NN outputs used for deriving limits in different situations.

Constraints on the individual Wilson coefficients can now be derived using the combined NN outputs defined in the first two rows of Tab. 6.3 (depending on the chirality of the top–quark current[9]). A selection is made on these discriminators to select the most sensitive phase space region for a given SMEFT operator. This selection was again optimized as shown in Fig. 6.13 for the two benchmark operators. The threshold on the NN output was taken to be $\mathrm{NN} > \mathrm{NN}^{\mathrm{sel}} = 0.83$ for all operators. After this selection, the limits on the individual operators are derived using the same method as before, resulting in a further improvement of the sensitivities on our example color singlet and octet operators to $[-2.1, 2.3]$ ($\mathrm{TeV}^{-2}$) and $[-5, 4.5]$ ($\mathrm{TeV}^{-2}$) at 95% CL, respectively. This is illustrated in Fig. 6.14, whereas the limits for all operators are once again summarized in Fig. 6.15 (green lines). A consistent improvement with respect to the $M_{4b}$ selection is seen for all operators.

### Template fits to the NN discriminators

By making a selection on the network discriminators some information is disregarded to obtain a better purity of SMEFT contributions. A more effective approach might be to consider the full differential shape of the discriminator distributions and apply a template fitting technique as outlined in Sec. 5.8.1. To illustrate this, template

[9]For the scalar operators $O^1_{QtQb}$ and $O^8_{QtQb}$, involving both LH and RH top quark currents, the choice was made to assign them to the $t_R$ category. This was motivated by the fact that the distributions of the kinematical variables show more similarity to this category.

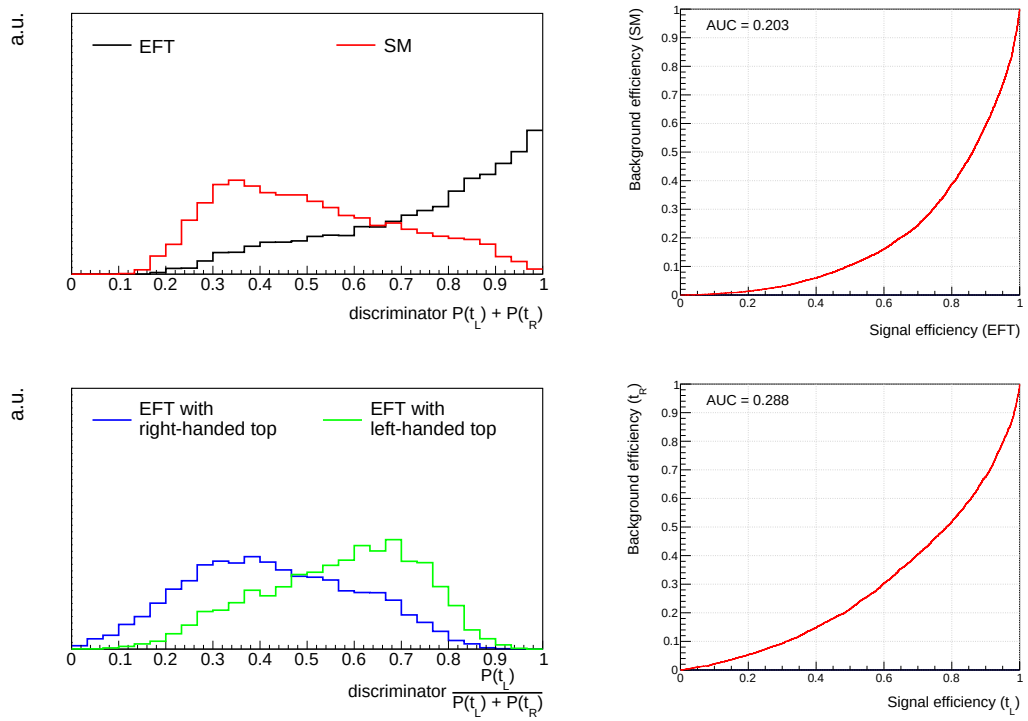


FIGURE 6.12: Discriminator distributions of combined outputs of the neural network that discriminate the SM processes from the EFT processes (top left) and the EFT operators with a $t_L$ current from the ones with a $t_R$ current (bottom left). The ROC curves corresponding to each of these distributions are shown on the right and the area under the ROC curve (AUC) is displayed.

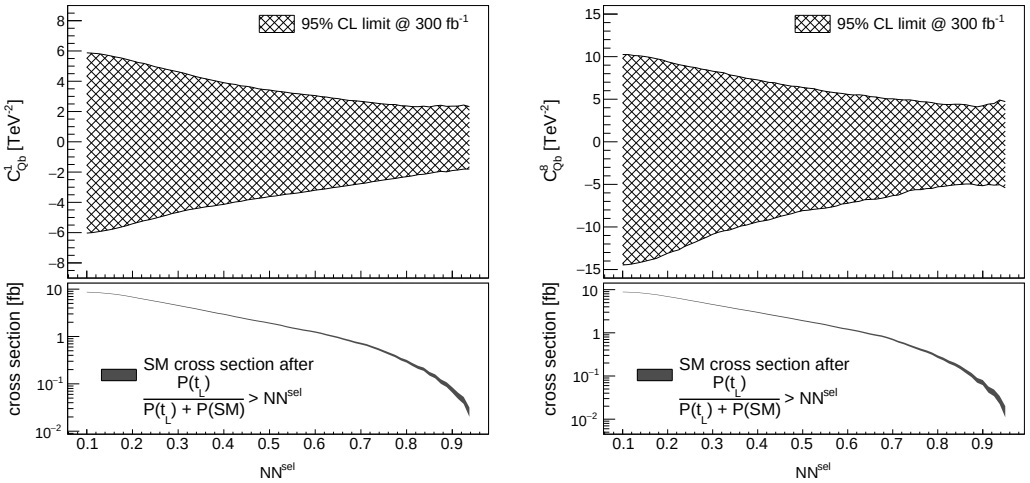


FIGURE 6.13: Limits at 95% CL on $C^1_{Qb}$ (left) and $C^8_{Qb}$ (right) as a function of the threshold on the network output. In the bottom panels, the effective (visible) SM cross section as a function of the NN output threshold is shown, with the corresponding statistical uncertainty shown as a grey band.

histograms ($T^{1D}$) have been derived for the discriminators (first two rows of Tab. 6.3) for each of the three output classes. These templates are constructed purely from the squared order contributions to preserve the independence of their shape on the

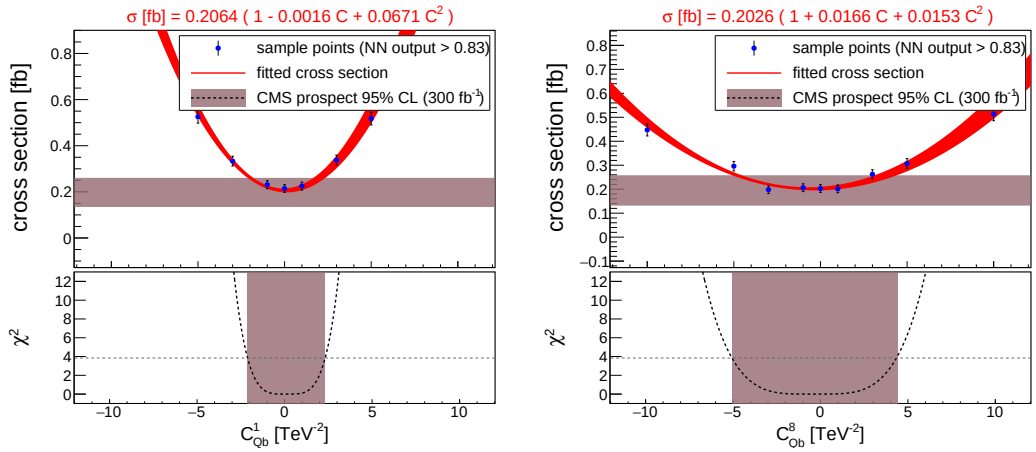


FIGURE 6.14: Limits at 95% CL on $C^1_{Qb}$ (left) and $C^8_{Qb}$ (right) after requiring the network output to be above 0.83 and assuming an integrated luminosity of 300 fb$^{-1}$.

value of the Wilson coefficient. This allows to parametrize the discriminators for any point in the SMEFT parameter space as linear combination of the templates scaled by the event yields in each category ($N_{SM}$ and $N_L$ or $N_R$). This is shown in Eqs. (6.10) and (6.11) for $t_L$ and $t_R$ operators respectively.

$$f_L\left(N_{SM}, N_L\right) = N_{SM} \cdot T^{1D}_{SM} + N_L \cdot T^{1D}_L \tag{6.10}$$

$$f_R\left(N_{SM}, N_R\right) = N_{SM} \cdot T^{1D}_{SM} + N_R \cdot T^{1D}_R \tag{6.11}$$

The yields then normalize each of the templates and can thus be extracted by fitting $f_L$ or $f_R$ to data. The RooFit package [287] incorporated in the ROOT data analysis framework [288] was used to perform the fit of pseudo–data to Eqs. (6.10) and (6.11), generated for different values of the Wilson coefficients assuming 300 fb$^{-1}$ of integrated luminosity. The obtained fitted yields can then be used in a $\Delta\chi^2$ as described in Eq. (6.9) to derive limits on the Wilson coefficients, which are again summarized for all operators in Fig. 6.15 (brown lines). No systematic uncertainties are considered on the templates, but a 10% systematic uncertainty is added when evaluating this $\Delta\chi^2$ to obtain a more fair comparison to the previous strategies. The obtained sensitivity from the template–fitting method is similar to that of the selection on the NN discriminators. The true benefit of these template–fitting procedures will be demonstrated in Sec. 6.1.5, where multiple operators with different top–quark chiralities are allowed to vary simultaneously.

### 6.1.5 Multiple operators: pinpointing the EFT

This section will demonstrate the strength of the multi–class output structure of the neural network, which becomes apparent when multiple EFT operators with both $t_L$ and $t_R$ currents are given non–zero values of their Wilson coefficients simultaneously. For illustration purposes, we will consider an example in which only two operators are allowed to have non–zero Wilson coefficients, keeping all other operators at a value of zero (TeV$^{-2}$). One LH operator ($O^1_{Qb}$) and one RH operator ($O^1_{tb}$) have been chosen to demonstrate the methodology. As mentioned before, the two discriminators in the bottom row of Tab. 6.3 are used to disentangle the different event classes. The distribution of each of these classes in the two–dimensional phase space of these discriminators is shown in Fig. 6.16. The x–axis represents the

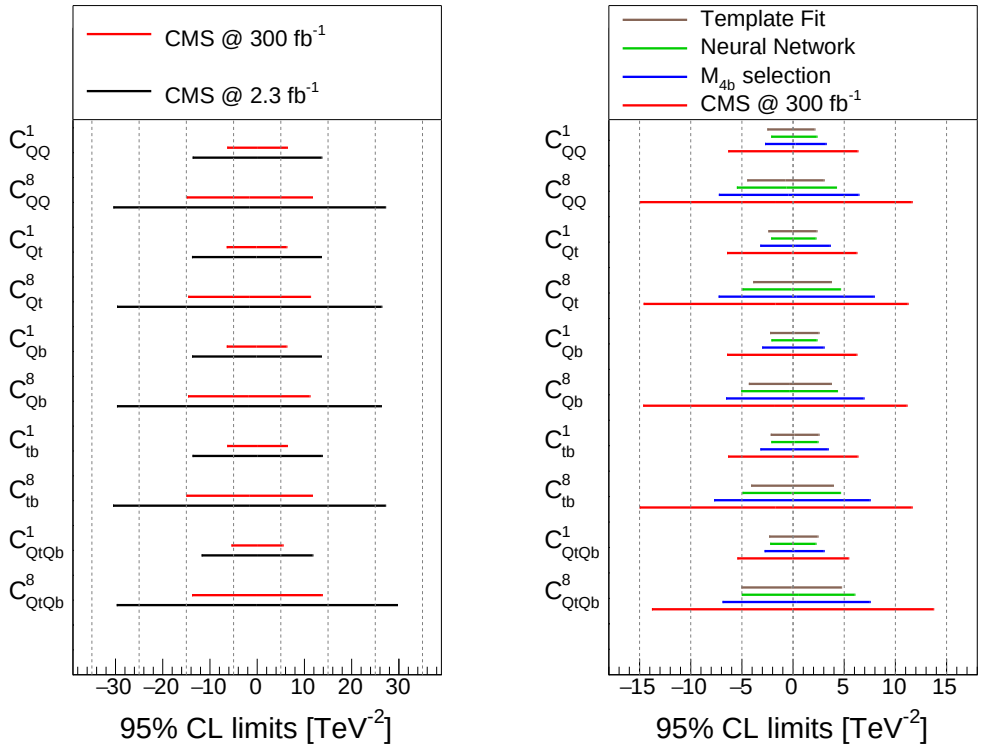


FIGURE 6.15: (left) Summary of the individual limits at 95% CL on all the Wilson coefficients, from the 13 TeV inclusive $t\bar{t}b\bar{b}$ cross section measurement of CMS with 2.3 fb$^{-1}$ (black), as well as projections for 300 fb$^{-1}$ (red). (right) Corresponding limits with 300 fb$^{-1}$ obtained in this work: by making a selection on $M_{4b}$ (blue) and on the neural network output (green) and by applying template fitting techniques to the network outputs (brown). For comparison we also include in red the limits from the 300 fb$^{-1}$ projection shown on the left plot. An upper cut on every energy scale of the process of $M_{\text{cut}} = 2$ TeV has been applied throughout.

summed probability $P(t_L) + P(t_R)$ that is able to separate the SM events (red) from events that are generated with any of the considered EFT operators. The y–axis displays the normalized probability $\frac{P(t_L)}{P(t_L)+P(t_R)}$, allowing for the distinction between the $t_L$ (green) and the $t_R$ (blue) categories. In this phase space, SM events are concentrated to the left, whereas the $t_L$ and $t_R$ contributions dominantly populate the upper and lower right hand corners, respectively.

The strategy remains the same as before, with the only difference that the cross section dependence in Eq. (1.28) reduces to

$$\sigma_{fit} = \sigma_{SM} \{ 1 + p_A \cdot C_A + p_B \cdot C_B + p_{AA} \cdot C_A^2 + p_{BB} \cdot C_B^2 + p_{AB} \cdot C_A C_B \}, \tag{6.12}$$

when two operators (generically denoted $A$ and $B$) are considered simultaneously. We first consider the hypothesis of observing a measurement consistent with the SM prediction. One could first apply a one–dimensional selection on the SM vs EFT discriminator (x–axis in Fig. 6.16) to select an EFT enriched phase space in analogy to what is done in Sec. 6.1.4. Applying a lower threshold on $P(t_L) + P(t_R)$,

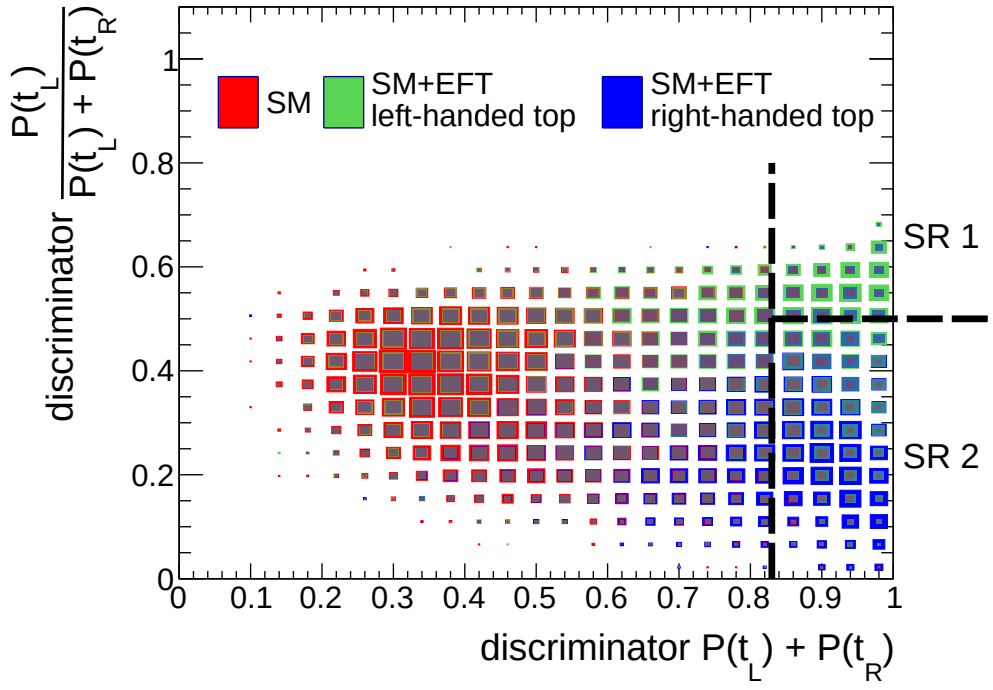


FIGURE 6.16: Normalized distributions of the combined NN outputs for events corresponding to SM (red), SM+EFT ($t_L$ operators) (green) and SM+EFT ($t_R$ operators) (blue). The Wilson coefficients are set to 20 (TeV$^{-2}$). The size of each box is proportional to the abundance of events of the corresponding sample. The discriminators on the x– and y–axis are defined in Tab. 6.3. The dashed lines define SR 1 and SR 2. See text for more details.

asking this value to be larger than 0.83, one obtains 95% CL contours in the $C^1_{Qb}$– $C^1_{tb}$ phase space. This is shown by the full red contour in Fig. 6.17 on the left. Applying an additional selection on the $t_L$ versus $t_R$ discriminator allows to define two signal regions, denoted SR 1 and SR 2 in Fig. 6.16, which are enriched in EFT contributions from either the $t_L$ or the $t_R$ categories, respectively. When limits are derived in SR 1 (SR 2), the sensitivity to $O^1_{Qb}$ ($O^1_{tb}$) increases as indicated by the green (blue) contours. A combination of the limits in these two signal regions results in the dashed red contour which shows a more stringent constraint on the operators compared to the one–dimensional selection on $P(t_L) + P(t_R)$.

The observations become even more interesting in case of a potential discovery of new physics. Following the hypothesis that an EFT signal is indeed observed it will be demonstrated how the previously outlined strategy allows to identify more accurately which operators are responsible for this possible excess. To illustrate this, an artificial benchmark signal is injected into the pseudo–data with $C^1_{Qb} = 5$ TeV$^{-2}$ and $C^1_{tb} = 3$ TeV$^{-2}$. Again the different steps from the previous paragraph can be followed to determine the 95% CL intervals on the allowed values of the Wilson coefficients. The one–dimensional selection on $P(t_L) + P(t_R)$ results in the toroidal contour shown by the full red line in the right panel of Fig. 6.17. The phase–space point ($C^1_{Qb}$,$C^1_{tb}$) = (0,0) is excluded, but the contour shows a symmetry around this central point, indicating the degeneracy between the operators. This approach does not allow to determine the individual strength of each of the operators that contributed to the

excess. A combination of CL intervals from SR 1 and SR 2 however partially lifts the degeneracy, as shown by the dashed red contour. Indeed, a value of zero ($\text{TeV}^{-2}$) for $C^1_{Qb}$ is now also excluded at 95% CL. This was not possible with the one–dimensional selection.

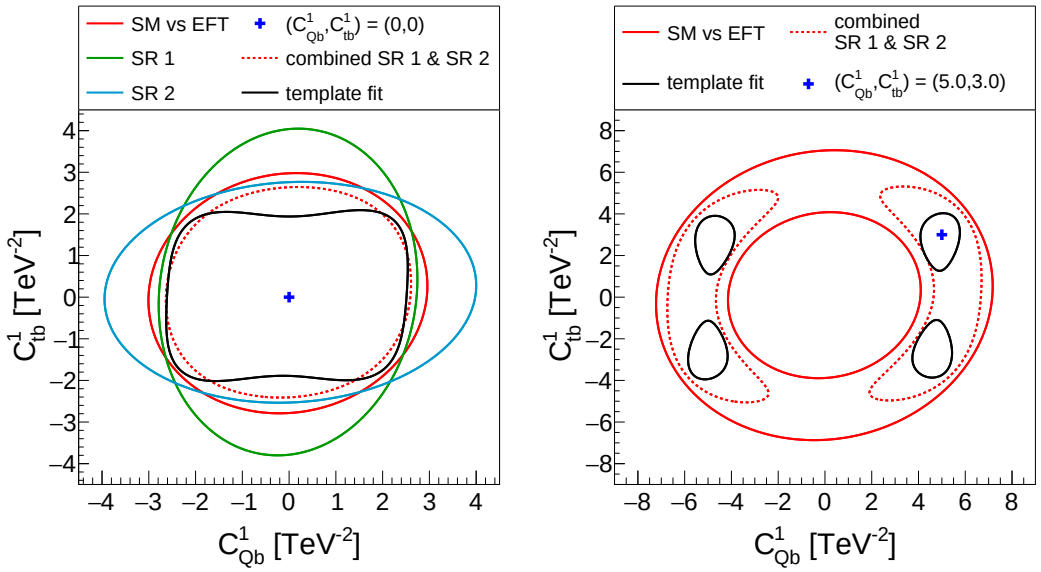


FIGURE 6.17: (left) Two-dimensional limits at 95% CL assuming a measurement consistent with the SM–only hypothesis (blue cross) and allowing two couplings, $C^1_{Qb}$ and $C^1_{tb}$ to vary simultaneously: (red) one dimensional cut on $P(t_L) + P(t_R)$ output; (green) SR1; (blue) SR2; (red dashed) combination of SR1 and SR2; (black) two–dimensional template fit. (right) Same as on the left plot, but for the EFT signal injection hypothesis. See text and Fig. 6.16 for more details.

The application of template–fitting methods shows its true potential in this two–dimensional example. We perform a two–dimensional maximum likelihood fit to predefined templates ($T^{2D}$) according to

$$f_{2D}\left(N_{SM}, N_L, N_R\right) = N_{SM} \cdot T^{2D}_{SM} + N_L \cdot T^{2D}_{L} + N_R \cdot T^{2D}_{R}, \tag{6.13}$$

to pseudo–data built from the SM–only hypothesis, or by injecting the benchmark signal mentioned above. The corresponding 95% CL contours that are derived from the yields ($N_{SM}$ and $N_L$ or $N_R$), are shown in black in Fig. 6.17 on the left (SM–only hypothesis) and the right (possible observation of a signal due to EFT operators). Under the SM hypothesis this approach leads to the most stringent constraints in some parts of the phase space. In the case of an observed signal, the template–fitting strategy is clearly able to pinpoint with more accuracy the origin of the observed signal. A value of 0 ($\text{TeV}^{-2}$) for $C^1_{tb}$ is now also excluded at 95% CL, which was not the case for combined limits in SR 1 and SR 2. This method therefore helps to identify the operators responsible for an observed excess, which could point into the direction in which to search for UV completions that may describe the corresponding new physics. There however remains a degeneracy in the sign of the Wilson coefficient, which can not be resolved with this method[10]. Finally, Fig. 6.18 shows the projected distributions of the fitted templates for $P(t_L) + P(t_R)$ on the left and for $\frac{P(t_L)}{P(t_L)+P(t_R)}$ on the right. This provides a

[10]The negligibly small interference of these color singlet operators with the SM results in a symmetric dependence of the cross section around a value of zero for the Wilson coefficients.

visual illustration on how the templates are scaled to match the inserted pseudo–data.

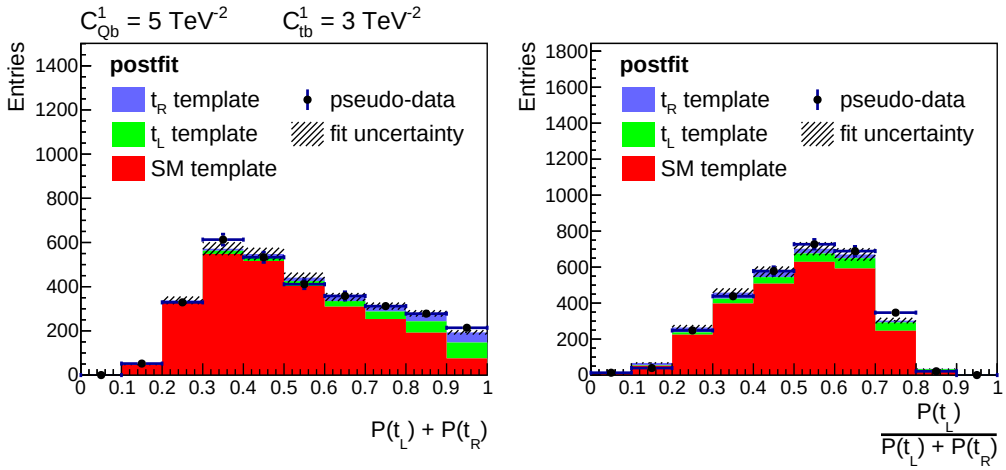


FIGURE 6.18: Projected distributions of the fitted templates and one of the generated pseudo–datasets onto the $P(t_L) + P(t_R)$ axis (left) and onto the $\frac{P(t_L)}{P(t_L)+P(t_R)}$ axis (right). The pseudo–experiments are generated from a sample with the Wilson coefficients $C^1_{Qb} = 5$ TeV$^{-2}$ and $C^1_{tb} = 3$ TeV$^{-2}$, and assuming an integrated luminosity of 300 fb$^{-1}$.

We have demonstrated how multi–class neural network classifiers are capable of distinguishing between different types of operators. This allows to select phase space regions in which these operator types are most abundant. By selecting these specific regions, or by applying template–fitting methods to the full differential shape of the discriminators, the most stringent limits on the SMEFT operators can be obtained or the origin of a potential excess can be pinpointed most accurately. In Sec. 6.2, this methodology will be demonstrated with real data using the measurement of the $t\bar{t}c\bar{c}$ process with the CMS detector, as described in Chapter 5.

This method can be extended by using more advanced network architectures in combination with more optimal input variables and larger training datasets to further exploit the power of these Machine Learning algorithms to constrain the SMEFT [236]. We have demonstrated this novel approach by using the kinematics of top–quark decay products to distinguish between operators with LH and RH top quark currents. An interesting extension would be to include interference effects in the training through a parametrized learning approach, which would potentially allow to discriminate also color octet from color singlet operators. It would be extremely interesting to investigate how far this discrimination can be pushed and whether or not it is possible to construct a neural network that is capable of distinguishing individual SMEFT operators.

## 6.2 Interpreting $t\bar{t}$+HF in the SMEFT

The novel method to constrain the Wilson coefficients of SMEFT operators will now be demonstrated using the measurement of the $t\bar{t}c\bar{c}$ cross section from Chapter 5. Using the full CMS simulation and reconstruction chain, this demonstration will serve as a proof–of–principle of the concept using actual collision data from $pp$ collisions at the LHC. A full interpretation of the $t\bar{t}$+HF final–state in terms of SMEFT operators is a highly involved task and is left for the future when the full dataset collected from 13 TeV collisions during Run–2 is analyzed. For now, we restrict ourselves to the set of two–heavy–two–light operators from Tab. 1.5, involving top quarks and charm quarks only, and focus on their impact on the $t\bar{t}c\bar{c}$ cross section. For future reference, the best individual observed limits obtained in this work from $t\bar{t}c\bar{c}$ production using 41.5 fb$^{-1}$ are compared to other existing bounds in Tab. 6.4. The first set of bounds comes from projected (expected) limits (for 300 fb$^{-1}$) on the mutual operators from four–top–quark production [108], assuming a value of $M_{\text{cut}} = 4$ TeV. These constraints assume an upper limit of the four–top–quark signal strength, $\mu < 1.87$, obtainable at the LHC with 300 fb$^{-1}$ at 13 TeV [280]. The other set of bounds results from a global fit of multiple SMEFT operators involving top quarks, to a variety of top–quark related measurements at center–of–mass energies of 8 and 13 TeV [105]. This preview already demonstrates the competitive (or even improved) constraints on most operators from $t\bar{t}c\bar{c}$ production using the novel ML–based methodology, as compared to those obtained from four–top–quark production. The individual bounds from the global fit are stronger since these operators are more stringently constrained by measurements of pure top–quark pair production, which are included in this global fit.

Following for a large part the strategy from Sec. 6.1, an assessment of the sensitivity is made by considering only one non–zero Wilson coefficient at a time. A neural network will be trained to distinguish SM events from SMEFT operators with LH top–quark currents and those with RH top–quark currents. In the end a pair of LH and RH operators will be allowed to vary simultaneously and the most stringent constraints will be derived using the NN discriminators.

### 6.2.1 Simulation of the SMEFT events

The 2H2L four–quark operators from Tab. 1.5 are included in the DIM6TOP [103] UFO model provided by the LHCTopWG. Events from the $t\bar{t}c\bar{c}$ final–state are generated at LO using the MG5_aMC@NLO [166] ME generator, followed by a showering and hadronization using Pythia8 [167]. This final–state is simulated only for dileptonic decays of the top–quark pair into electrons or muons. Different values of the Wilson coefficients are included in a single sample by adding event weights to the samples [289]. By applying these weights, both the cross section and the differential distributions of the simulated event sample change to match those expected for a specific value of the Wilson coefficients. One sample has been created that includes only the squared order SMEFT contributions (no SM) and will be used to train the NN and to create templates for the template–fitting methods. Another sample is created that includes effects of both the SM and the SMEFT operators (including properly the interference) and will be used to derive the dependence of the cross section on the value of the Wilson coefficients and to constrain their allowed range.

TABLE 6.4: Comparison between the sensitivity of $t\bar{t}t\bar{t}$ and $t\bar{t}c\bar{c}$ production, as well as from a global fit of top–quark related measurements to the mutual two–heavy–two–light SMEFT operators. The first column shows individual 95% confidence intervals (for projections to 300 fb$^{-1}$ at 13 TeV) from four–top–quark production [108] assuming a value of $M_{\text{cut}} = 4$ TeV. The middle column shows the results from a global fit of multiple SMEFT operators to a variety of top–quark related measurements [105] at 8 and 13 TeV. The last column compares these intervals to the best individual constraints obtained in this work from $t\bar{t}c\bar{c}$ production using 41.5 fb$^{-1}$ at 13 TeV. These numbers correspond to the blue lines in Fig. 6.29.

| | 4-top (300 fb$^{-1}$) ($M_{\text{cut}} = 4$ TeV) | global fit (no $M_{\text{cut}}$) | $t\bar{t}c\bar{c}$ (41.5 fb$^{-1}$) ($M_{\text{cut}} = 2$ TeV) |
|---|---|---|---|
| $C_{tq}^{8}$ | $[-6.6, 4.1]$ | $[-0.7, 0.09]$ | $[-5.9, 3.5]$ |
| $C_{tq}^{1}$ | $[-2.6, 2.6]$ | $[-0.3, 0.03]$ | $[-2.0, 1.9]$ |
| $C_{Qd}^{8}$ | $[-9.6, 6.6]$ | $[-1.9, 0.07]$ | $[-10.9, 8.7]$ |
| $C_{Qu}^{8}$ | $[-7.1, 4.6]$ | $[-2.6, 0.1]$ | $[-7.1, 4.9]$ |
| $C_{Qd}^{1}$ | $[-4.0, 4.1]$ | $[-0.9, 0.05]$ | $[-4.0, 3.9]$ |
| $C_{Qu}^{1}$ | $[-2.9, 2.9]$ | $[-0.4, 0.03]$ | $[-2.6, 2.5]$ |
| $C_{td}^{8}$ | $[-9.4, 6.4]$ | $[-1.6, 0.02]$ | $[-12.8, 10.1]$ |
| $C_{tu}^{8}$ | $[-7.4, 5.1]$ | $[-0.9, 0.03]$ | $[-7.4, 5.0]$ |
| $C_{td}^{1}$ | $[-4.0, 4.1]$ | $[-0.6, 0.03]$ | $[-3.5, 3.7]$ |
| $C_{tu}^{1}$ | $[-2.9, 3.0]$ | $[-0.4, 0.03]$ | $[-2.3, 2.4]$ |
| $C_{Qq}^{8,3}$ | $[-5.6, 5.0]$ | $[-0.7, 0.2]$ | $[-6.5, 5.5]$ |
| $C_{Qq}^{8,1}$ | $[-7.0, 4.4]$ | $[-0.6, 0.07]$ | $[-7.4, 4.9]$ |
| $C_{Qq}^{1,3}$ | $[-2.6, 2.6]$ | $[-0.1, 0.09]$ | $[-2.2, 2.2]$ |
| $C_{Qq}^{1,1}$ | $[-2.5, 2.7]$ | $[-0.2, 0.03]$ | $[-2.1, 2.2]$ |

The generator–level jets are required to have $p_T > 20$ GeV and $|\eta| < 2.4$. The generator–level electrons and muons are required to have $p_T > 25$ GeV and $|\eta| < 2.4$. The angular distance in $\Delta$R between any two objects (jets or leptons) is required to be larger than 0.5.

The effects of the CMS detector are simulated using the full detector simulation implemented in the GEANT4 Toolkit [196–198]. The events pass through the standard CMS event reconstruction as outlined in Sec. 5.5 and are only selected if they pass the event selection from the $t\bar{t}$+HF analysis as discussed in Sec. 5.5. The reconstructed phase space in which the $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$LF cross sections have been fitted, serves as a starting point for the EFT analysis. Further selection on dedicated NN discriminators will be added on top of this baseline selection to increase the sensitivity to the SMEFT operators.

### 6.2.2 Validity of the EFT in the $t\bar{t}c\bar{c}$ final–state

As a starting point, a sensitivity study will be conducted to derive limits on the individual Wilson coefficients in the reconstructed phase space. This phase space is defined after the full event reconstruction has taken place and all selections outlined in Sec. 5.5 have been applied. There is no need to unfold the cross section to some fiducial phase space given that also the signal samples including the SMEFT operators can be evaluated directly in the reconstructed phase space. The $t\bar{t}c\bar{c}$ cross section is measured using the template fitting method to the two–dimensional $\Delta_b^c - \Delta_L^c$ distribution as described in Sec. 5.8. Then the dependence of the $t\bar{t}c\bar{c}$ cross section on the value of the Wilson coefficients is derived from simulations in the same phase space. This allows to constrain the values of these Wilson coefficients to a range that is consistent with the measured cross section. This strategy relies on the assumption that the $\Delta_b^c$ and $\Delta_L^c$ distributions do not depend on the value of the Wilson coefficient. Indeed, it can be seen from Fig. 6.19 that within the statistical uncertainties of the MC samples, these distributions have the same differential shape for the pure SM events (red) and for events including both the SM and SMEFT effects from the Wilson coefficient $C_{Qu}^1 = 5$ TeV$^{-2}$ (blue) or 10 TeV$^{-2}$ (green). This is expected given that these NN discriminators are almost exclusively constructed from the flavor information of the additional jets. The presence of EFT couplings is not expected to alter the flavor–tagging discriminators of these jets.

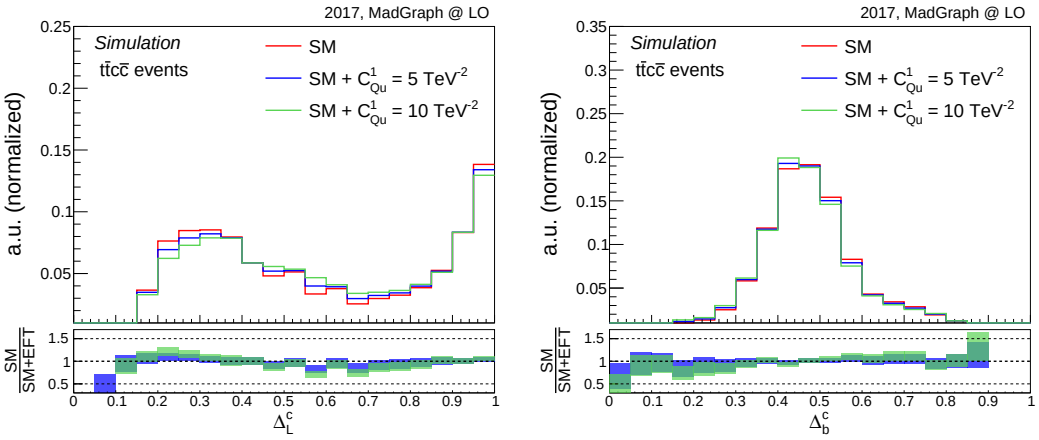


FIGURE 6.19: Normalized distributions of the $\Delta_L^c$ (left) and $\Delta_b^c$ (right) NN discriminators for pure SM events (red) and for events including both the SM and SMEFT effects from the Wilson coefficient $C_{Qu}^1 = 5$ TeV$^{-2}$ (blue) or 10 TeV$^{-2}$ (green). The bottom panels show the ratio $^{\text{SM}}/_{\text{SM+EFT}}$. These distributions are drawn using the LO MADGRAPH simulations for the $t\bar{t}c\bar{c}$ production, including SMEFT operators.

This cross–check justifies to measure $\sigma_{t\bar{t}c\bar{c}}$ with the same strategy as before and interpret this result in terms of potential SMEFT contributions. Remains to be determined is a proper value for $M_{\text{cut}}$ to ensure that the validity requirement in Eq. (6.4) is fulfilled. In what follows, two benchmark operators, $O_{Qu}^1$ and $O_{Qu}^8$, are chosen to demonstrate the different results, whereas the results for other operators are summarized at the end of Sec. 6.2.3. For the two benchmark operators, the dependence of the limits has been investigated for different values of $M_{\text{cut}}$. These limits are displayed by the black lines in Fig. 6.20, with the non–valid region superimposed in light pink. The limits are again insensitive to the value of the Wilson coefficient down to a

value of $M_{\text{cut}} \sim 1.5$ TeV and we safely fix the value of $M_{\text{cut}}$ to 2 TeV throughout the rest of this study. The dark pink shaded regions represents again the more stringent validity requirement chosen such that the upper limits cross the validity boundary at $M_{\text{cut}} = 2$ TeV.

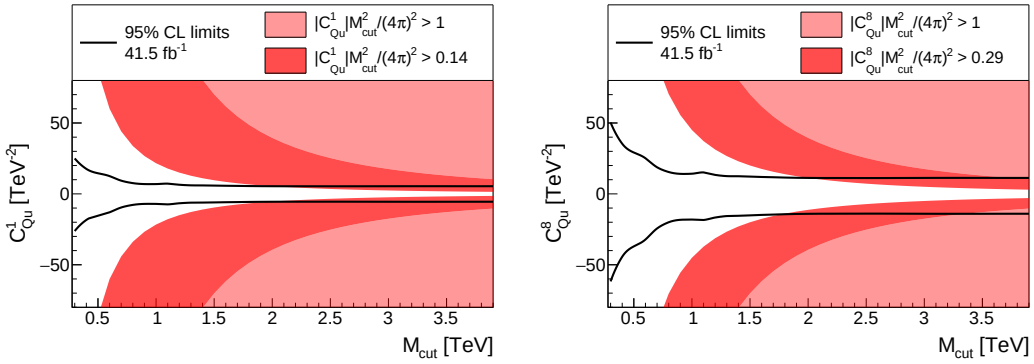


FIGURE 6.20: Limits at 95% CL on $C^1_{Qu}$ (left) and $C^8_{Qu}$ (right) as a function of the mass cut $M_{\text{cut}}$ for an integrated luminosity of 41.5 fb$^{-1}$ (full black line). The non-perturbative regime of the EFT in which $|C_i|M^2_{\text{cut}} > (4\pi)^2$ is indicated with the light pink shaded region. The darker red region represents a more stringent perturbativity requirement for which the upper limit on the Wilson coefficient intersects the perturbativity threshold at $M_{\text{cut}} = 2$ TeV.

Limits on the individual Wilson coefficients have been derived using the $t\bar{t}c\bar{c}$ cross section measurement. In fact the limits are derived using the signal strength modifier $\mu_{t\bar{t}c\bar{c}} = \sigma_{t\bar{t}c\bar{c}}/\sigma^{\text{SM}}_{t\bar{t}c\bar{c}}$, where $\sigma^{\text{SM}}_{t\bar{t}c\bar{c}}$ represents the theoretically predicted $t\bar{t}c\bar{c}$ SM cross section. In contrast to the phenomenological study in Sec. 6.1, no assumptions need to be made on the statistical and systematical uncertainties, as instead the full fit can be conducted with all systematic uncertainties properly taken into account. The result of the measured signal strength modifier (and the corresponding 95% CL interval) is shown by the green band in the upper panels of Fig. 6.21. The cross section dependence on the Wilson coefficients has been determined by fitting the quadratic form $\mu_{t\bar{t}c\bar{c}} = \sigma_{t\bar{t}c\bar{c}}/\sigma^{\text{SM}}_{t\bar{t}c\bar{c}} = 1 + p_1 \cdot C_i + p_2 \cdot C_i^2$ to the simulated sample points ($C_i \in \{-10, -5, -3, -1, 1, 3, 5, 10\}$). This fitted function is shown by the red line for the two benchmark operators $C^1_{Qu}$ (left) and $C^8_{Qu}$ (right). The bottom panels show the resulting $\Delta\chi^2$ and the corresponding limits on the Wilson coefficients. The example color singlet operator is constrained to $\sim [-5, 5]$ (TeV$^{-2}$), whereas the example color octet operator is constrained to $\sim [-14, 11]$ (TeV$^{-2}$). The results for all other operators are shown in red in Fig. 6.29.

### 6.2.3 Learning the operators affecting $t\bar{t}c\bar{c}$ production

In order to improve the sensitivity of the $t\bar{t}c\bar{c}$ final–state to the EFT operators, a NN has been trained with the same inputs as the ones described in Sec. 6.1.4. Its architecture is comprised of two hidden layers with 30 neurons each. This network is trained once again to identify SM events, events from SMEFT operators with a LH top quark current ($t_L$) and those with a RH top quark current ($t_R$). Also here, the training is performed using samples that include only the pure (squared order) EFT contributions and does not take into account the interference effects between the SM and the EFT. The normalized distributions of the input variables for the different output categories are shown in Fig. 6.22 for $t\bar{t}c\bar{c}$ events only. The same

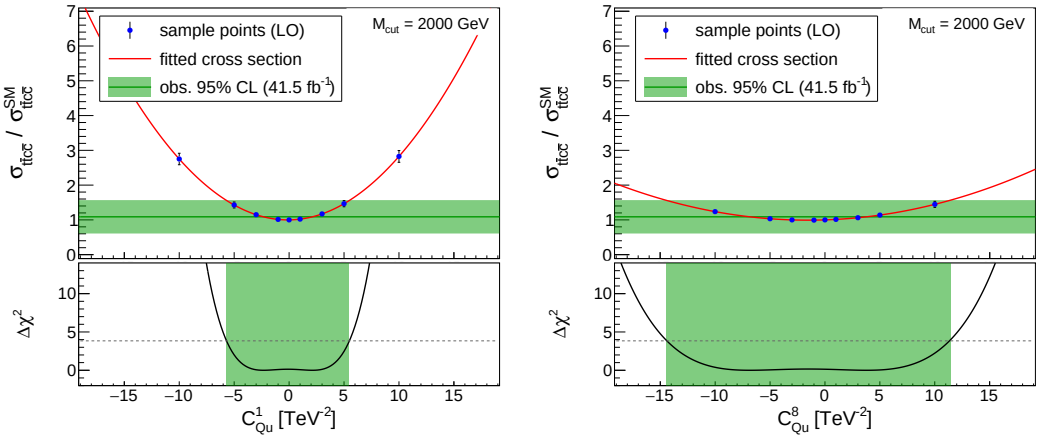


FIGURE 6.21: Limits at 95% CL on $C_{Qu}^{1}$ (left) and $C_{Qu}^{8}$ (right) derived from the measured signal strength modifier $\mu_{t\bar{t}c\bar{c}} = \sigma_{t\bar{t}c\bar{c}}/\sigma_{t\bar{t}c\bar{c}}^{\mathrm{SM}}$ using 41.5 fb$^{-1}$ (green area).

trends appear as for the $t\bar{t}b\bar{b}$ study in the previous section. The high–energy tails of the energy–dependent variable distributions clearly allow to identify the presence of EFT processes whereas the distinction between the $t_L$ and $t_R$ categories lies mostly in the kinematics of the leptons. The training curve showing the value of the loss function and the accuracy over the different epochs is shown in Fig. 6.23.

The three output probabilities are again combined into four final discriminators as defined in Tab. 6.3. Their normalized distributions from the training (and testing) samples are shown on the left in Fig. 6.24, with the corresponding ROC curves on the right. Overall, when comparing the performance of this NN to the one used in the phenomenological $t\bar{t}b\bar{b}$ study (see Fig. 6.12), the network in the $t\bar{t}c\bar{c}$ topology seems to perform slightly better. This is most likely due to the better jet–parton assignment available in the full $t\bar{t}c\bar{c}$ analysis compared to the simplified assignment due to the approximate detector simulation with DELPHES that was used in the $t\bar{t}b\bar{b}$ study.

When focusing on one EFT operator at a time, the two upper discriminators in Tab. 6.3 are used to define an additional event selection that enriches the obtained event sample in EFT contributions, relative to the SM events. The $t\bar{t}c\bar{c}$ cross section is then determined in that selected phase space using the template fitting strategy to the $\Delta_L^c - \Delta_b^c$ distributions. The allowed values of the Wilson coefficients are then determined as explained before in Sec. 6.2.2. Using the two benchmark operators, we first determine the most optimal selection threshold by investigating the dependence of the obtained limits on the lower threshold on the NN discriminator. This is shown in Fig. 6.25 for $C_{Qu}^{1}$ on the left and $C_{Qu}^{8}$ on the right. The most stringent constraints are obtained at a lower threshold of around 0.08 on these operators belonging to the $t_L$ category[11]. Beyond that point, the statistical uncertainty starts dominating the measurement, resulting in worse limits.

By applying the optimal lower threshold on the dedicated NN outputs, limits are obtained on all individual operators. The example color singlet operator is constrained to $\sim [-3, 3]$ (TeV$^{-2}$), whereas the example color octet operator is constrained to $\sim [-7, 5]$ (TeV$^{-2}$), as shown in Fig. 6.26. This is an improvement of a factor of two

[11] For operators from the $t_R$ category, the limits are derived using the dedicated SM vs $t_R$ discriminator and the optimal lower threshold was found to lie around 0.1.

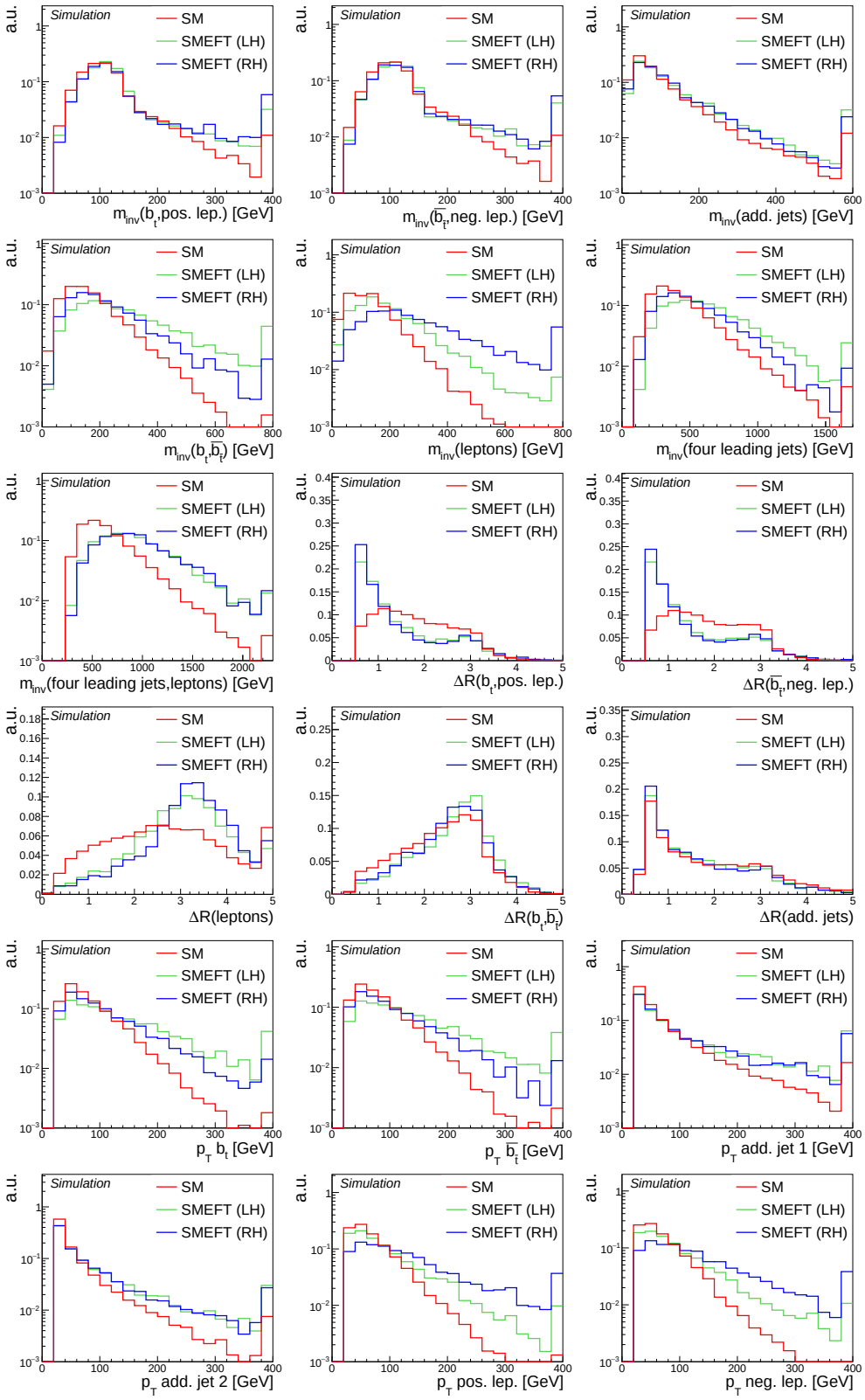


FIGURE 6.22: Normalized distributions of the SMEFT neural network input variables in $t\bar{t}c\bar{c}$ processes for SM events (red), LH two–heavy–two–light operators (green) and RH two–heavy–two–light operators (blue). Overflow is included in the last bin.

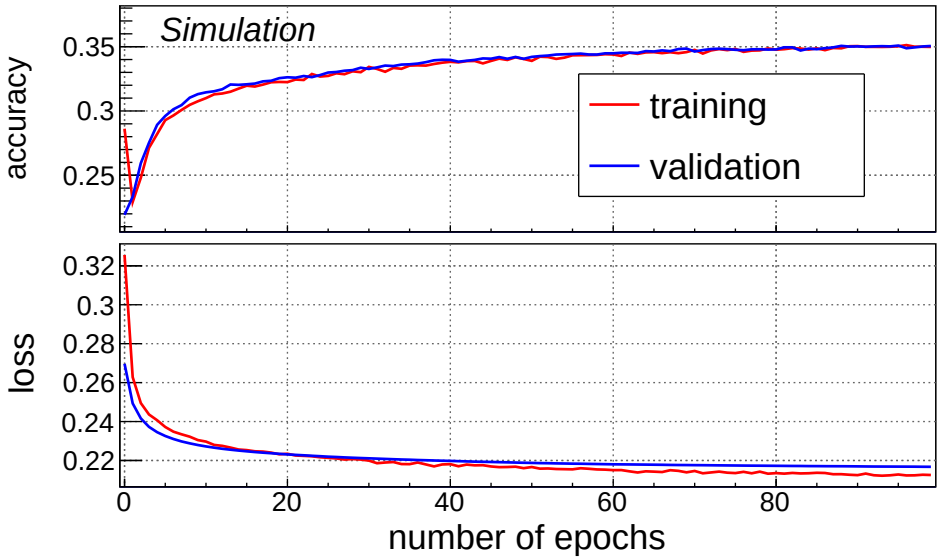


FIGURE 6.23: Training curves of the SMEFT neural network training on $t\bar{t}c\bar{c}$ processes, displaying the evolution of the accuracy (top) and value of the loss function (bottom) for increasing number of epochs. These curves are shown both for the training (red) and for an independent validation data set (blue) and reach a plateau after about 100 epochs.

in the limits compared to those obtained in the original phase space in Sec. 6.2.2. The results for all other operators are shown in blue in Fig. 6.29.

### Template fits to the NN discriminators

Finally, before considering multiple operators at a time, it will be demonstrated how a template fitting procedure to the dedicated NN discriminators can be used to derive constraints on the individual Wilson coefficients. In this demonstration, we will no longer extract the $t\bar{t}c\bar{c}$ cross section using the $\Delta^c_L - \Delta^c_b$ distributions. Instead the templates of the NN discriminators will be used directly to disentangle the EFT effects from the SM ones. These templates are produced from the squared order EFT contributions only (*i.e.* from those samples used for training the NN) in order to make sure the shape of the templates does not depend on the value of the Wilson coefficient. It is important to note that these NN discriminators do not use any information on the jet flavor and are therefore not capable of distinguishing the $t\bar{t}c\bar{c}$ events from the $t\bar{t}b\bar{b}$ or $t\bar{t}$LF events. The resulting constraints are expected to be slightly weaker given that the template–fitting strategy will search for an excess of EFT events over the collection of all $t\bar{t}$+HF events, without first extracting the $t\bar{t}c\bar{c}$ component. Ideally the $\Delta^c_L - \Delta^c_b$ distributions should be fitted simultaneously, or the jet flavor information should be included in the NN that tries to distinguish the EFT effects from the SM ones. Such a full and proper treatment is left for the future when performing a global EFT fit in the $t\bar{t}$+HF final–state (including all operators that potentially affect different flavor categories). The results in this section serve as a proof of concept showing that a template fitting procedure is able to put competitive constraints on the operators compared to a cut–based approach outlined above.

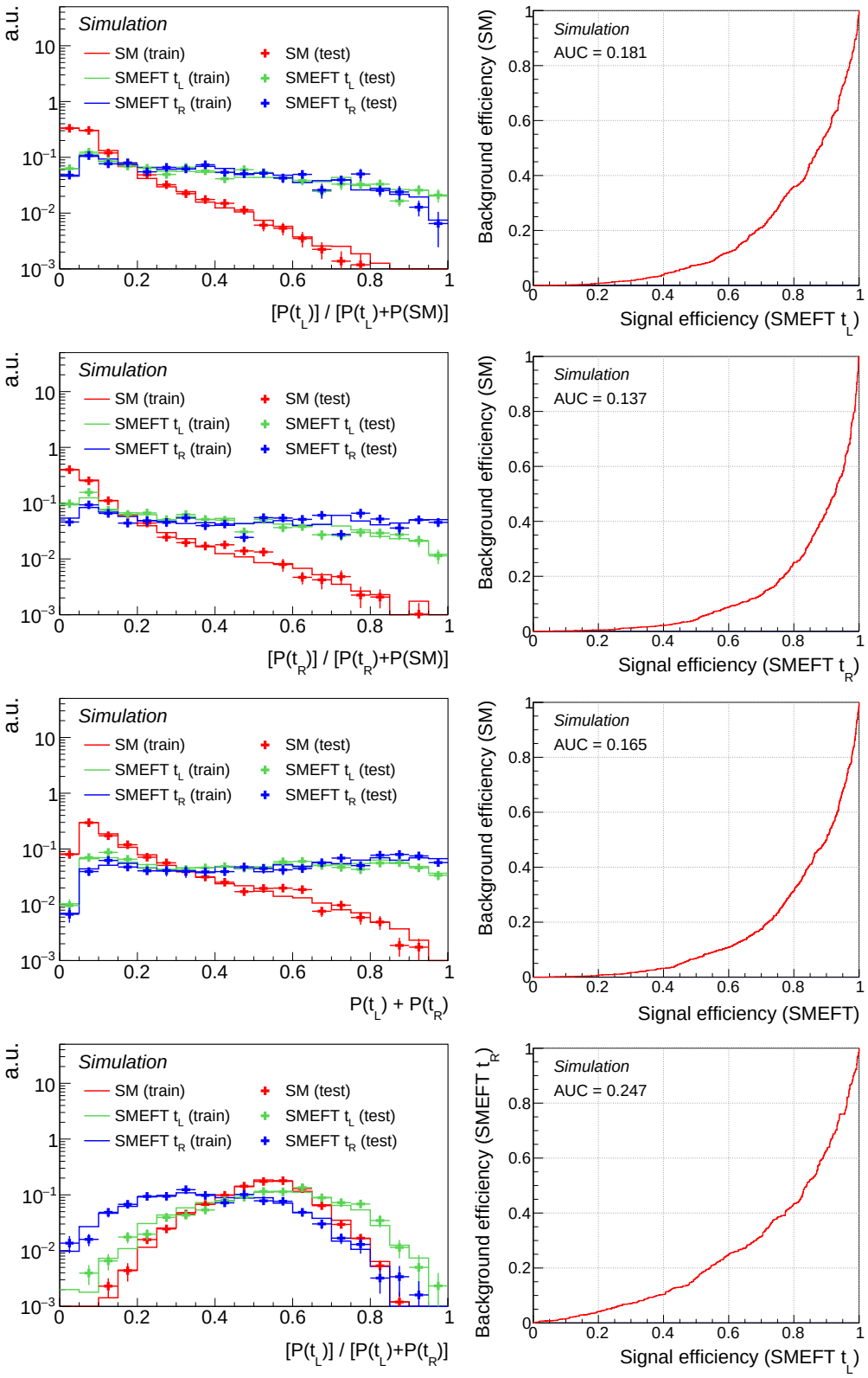


FIGURE 6.24: (left) Discriminator distributions of the combined outputs of the neural network as summarized in Tab. 6.3, used in the $t\bar{t}c\bar{c}$ interpretation. (right) The ROC curves corresponding to each of these distributions. The area under the ROC curve (AUC) is also displayed in the panel.

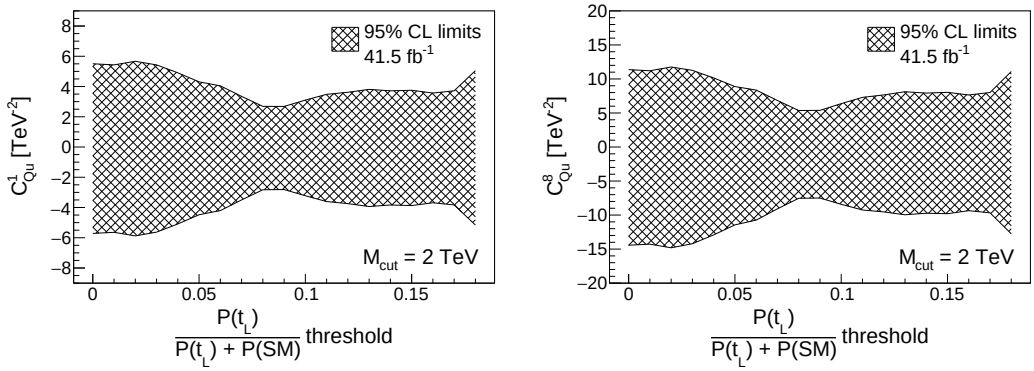


FIGURE 6.25: Limits at 95% CL on $C_{Qu}^{1}$ (left) and $C_{Qu}^{8}$ (right) as a function of the threshold on the network output using the measurement of the $t\bar{t}c\bar{c}$ cross section.

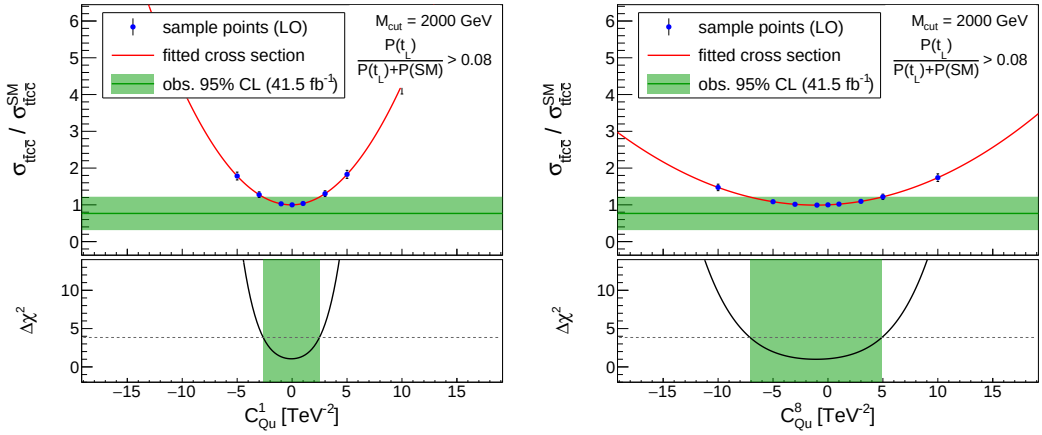


FIGURE 6.26: Limits at 95% CL on $C_{Qu}^{1}$ (left) and $C_{Qu}^{8}$ (right) after requiring the network output to be above 0.08 and using the measurement of the $t\bar{t}c\bar{c}$ cross section in that selected phase space.

The templates (which are shown in Fig. 6.24) are constructed for SM events, for $t_L$ events and for $t_R$ events. These are then fitted to Eq. (6.14). The fit extracts the normalization of the SM events ($N_{SM}$) as well as the value of the Wilson coefficient $C_i$ under consideration. The SMEFT template ($H_{SMEFT}^{norm}$) is chosen to be the one from the SM vs $t_L$ discriminator or the SM vs $t_R$ discriminator depending on the class of the operator that is considered. It was already mentioned that these discriminators can not separate $t\bar{t}c\bar{c}$ events from non–$t\bar{t}c\bar{c}$ SM events. Nevertheless these two templates are separately included given that the SMEFT contributions can only affect the $t\bar{t}c\bar{c}$ yield. The relative fraction of $t\bar{t}c\bar{c}$ events to the inclusive $t\bar{t}jj$ yield is kept constant by the prediction in the simulation ($N_{t\bar{t}c\bar{c}}^{MC}/N_{All,SM}^{MC}$). The SMEFT yield is then allowed to alter the $t\bar{t}c\bar{c}$ yield with a factor $[p_1 \cdot C_i + p_2 \cdot C_i^2]$ corresponding to Eq. (6.8). The values of $p_{1,2}$ are derived from the samples that contain both the SM and the EFT contributions.

$$f\left(\mathrm{N_{SM}}, C_i\right) = \mathrm{N_{SM}} \cdot \left( \frac{N^{\mathrm{MC}}_{t\bar{t}c\bar{c}}}{N^{\mathrm{MC}}_{\mathrm{All,SM}}} \cdot \mathrm{H}^{\mathrm{norm}}_{t\bar{t}c\bar{c}} + \frac{N^{\mathrm{MC}}_{\text{non-}t\bar{t}c\bar{c}}}{N^{\mathrm{MC}}_{\mathrm{All,SM}}} \cdot \mathrm{H}^{\mathrm{norm}}_{\text{non-}t\bar{t}c\bar{c}} \right) + \mathrm{N_{SM}} \cdot \left( \frac{N^{\mathrm{MC}}_{t\bar{t}c\bar{c}}}{N^{\mathrm{MC}}_{\mathrm{All,SM}}} \right) \cdot \left[ p_1 \cdot C_i + p_2 \cdot C_i^2 \right] \cdot \mathrm{H}^{\mathrm{norm}}_{\mathrm{SMEFT}} \tag{6.14}$$

The templates are then fitted to the data using 41.5 $\mathrm{fb}^{-1}$ of integrated luminosity. An example of the postfit distribution of data and simulated events, using the $t_L$ template in the fit, is shown in Fig. 6.27. The data show a good agreement with the SM expectations and limits can thus be derived on the Wilson coefficients. By scanning the negative logarithm of the likelihood ($-\Delta\log(\mathrm{L})$), the constraints on the Wilson coefficients are obtained. Examples are once again shown for the two benchmark operators in Fig. 6.28, whereas results for other operators are summarized with the green lines in Fig. 6.29. The obtained sensitivity from the template fitting method is very similar to the one from the selection on the NN discriminators. Nevertheless the template fitting strategy yields consistently slightly worse limits over all operators. This however shows that such a strategy allows to obtain competitive bounds even without the use of jet–flavor information to extract the $t\bar{t}c\bar{c}$ component over the large $t\bar{t}$LF background. Improvements are thus expected when this information is added to this method.

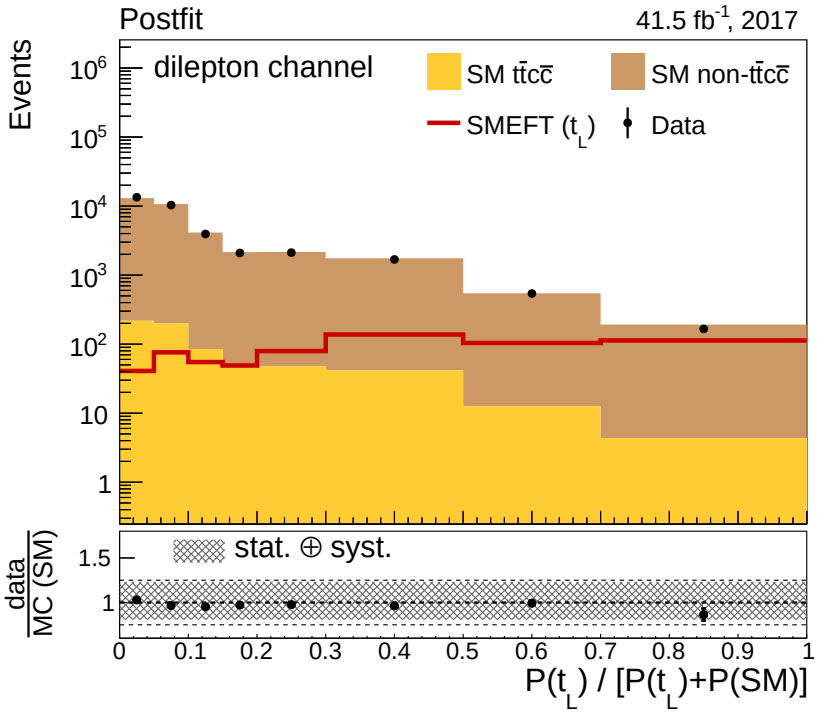


FIGURE 6.27: Data–to–simulation agreement after the one–dimensional template fit of the data to the SM and the $t_L$ templates. The $t_L$ template is superimposed (red) with an arbitrary normalization that results in the same yield as the $t\bar{t}c\bar{c}$ component. The data show a good agreement with the SM expectations and limits can thus be derived on the Wilson coefficients.

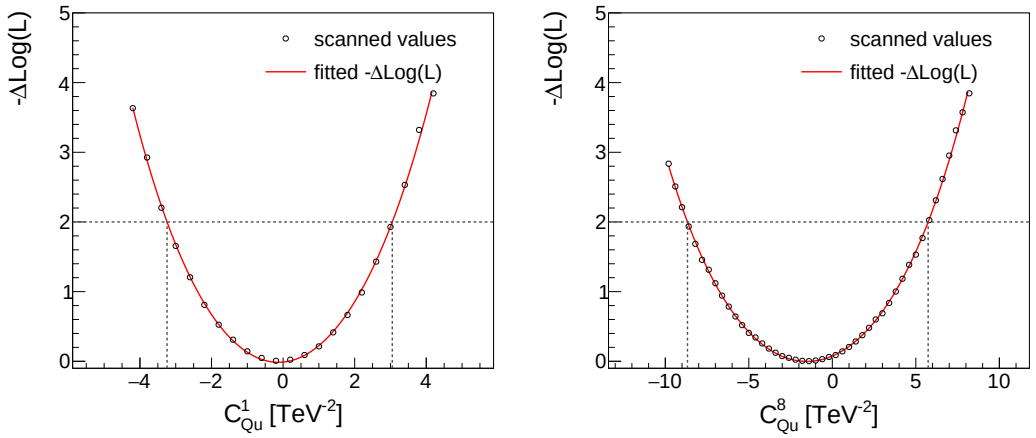


FIGURE 6.28: Scan of the negative logarithm of the likelihood from the template fit as a function of the value of the Wilson coefficient $C^1_{Qu}$ (left) and $C^8_{Qu}$ (right). The intersections with $-\Delta\log(\mathrm{L}) = 2$ are drawn using the dotted lines and denote the 95% CL limits on the Wilson coefficients.

### 6.2.4 Multiple two–heavy–two–light operators

In analogy to what has been discussed in Sec. 6.1.5, we will now demonstrate the improved sensitivity in the $t\bar{t}$+HF analysis when two operators are allowed to have non–zero Wilson coefficients simultaneously. The choice has been made to use the $C^1_{Qu}$ operator from the $t_L$ category and the $C^1_{tu}$ operator from the $t_R$ category. Again the two discriminators in the bottom rows of Tab. 6.3 will be exploited to disentangle the contributions from SM, $t_L$ and $t_R$ events and to identify phase space regions that are enriched in SMEFT contributions with either LH or RH top–quark currents. The normalized two–dimensional distributions of these NN discriminators are drawn from simulations in Fig. 6.30 for SM $t\bar{t}c\bar{c}$ events (top left), SM non–$t\bar{t}c\bar{c}$ events (bottom left), SM+EFT $t\bar{t}c\bar{c}$ events with $C^1_{Qu} = 10$ TeV$^{-2}$ (top right) and SM+EFT $t\bar{t}c\bar{c}$ events with $C^1_{tu} = 10$ TeV$^{-2}$ (bottom right). Indeed it can be seen that the SM events are located to the left in this two–dimensional phase–space, whereas $t_L$ and $t_R$ SMEFT contributions are distributed more towards the top and bottom right corners respectively.

Two signal regions are defined that select the $t_L$ and $t_R$ enriched phase–space regions, denoted SR 1 and SR 2 respectively. The specific values of the selections have been determined by scanning over a wide range of phase–space points, resulting in the following definitions of the two signal regions

$$\underline{\texttt{SR 1}}: \quad \mathrm{P}(t_L) + \mathrm{P}(t_R) > 0.16, \quad \frac{\mathrm{P}(t_L)}{\mathrm{P}(t_L) + \mathrm{P}(t_R)} > 0.5, \tag{6.15}$$

$$\underline{\texttt{SR 2}}: \quad \mathrm{P}(t_L) + \mathrm{P}(t_R) > 0.16, \quad \frac{\mathrm{P}(t_L)}{\mathrm{P}(t_L) + \mathrm{P}(t_R)} < 0.5. \tag{6.16}$$

Additionally, the sensitivity is derived with a one–dimensional selection on the SM versus EFT discriminator, *i.e.* $\mathrm{P}(t_L) + \mathrm{P}(t_R) > 0.16$ as a reference. In each of these signal regions, the $t\bar{t}c\bar{c}$ cross section is extracted with the traditional template–fitting strategy to the $\Delta^c_L - \Delta^c_b$ distributions and the parametric dependence of the signal strength modifier $\mu_{t\bar{t}c\bar{c}}$ on the value of the two Wilson coefficients is fitted to the

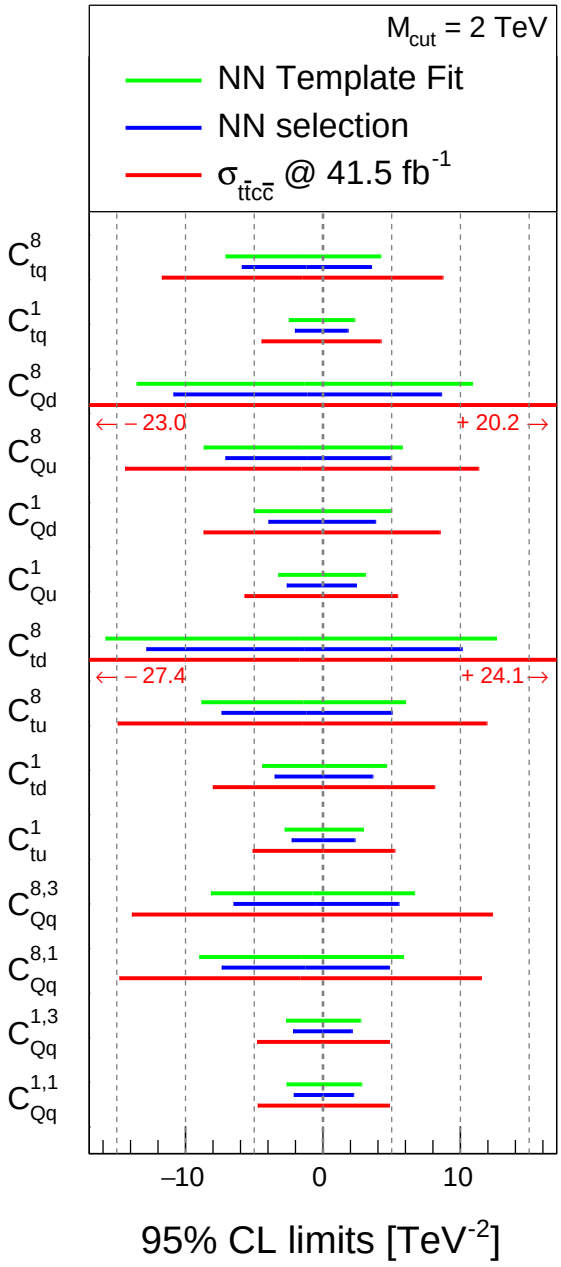


FIGURE 6.29: Summary of the individual limits at 95% CL on all the Wilson coefficients of the two–heavy–two–light operators, using the inclusive $t\bar{t}c\bar{c}$ cross section measurement of CMS with 41.5 fb$^{-1}$ (red), by making an additional selection on the NN discriminator output (blue) and by applying template fitting techniques to the network outputs (green). An upper cut on every energy scale of the process of $M_{\text{cut}} = 2$ TeV has been applied throughout.

simulated yields according to

$$\mu_{t\bar{t}c\bar{c}} = \frac{\sigma_{t\bar{t}c\bar{c}}}{\sigma^{\text{SM}}_{t\bar{t}c\bar{c}}} = \{1 + p_A \cdot C^1_{Qu} + p_B \cdot C^1_{tu} + p_{AA} \cdot \left(C^1_{Qu}\right)^2 + p_{BB} \cdot \left(C^1_{tu}\right)^2 + p_{AB} \cdot C^1_{Qu} C^1_{tu}\}. \tag{6.17}$$

The (two–dimensional) $\Delta\chi^2$ is constructed according to Eq. (6.9) and the resulting

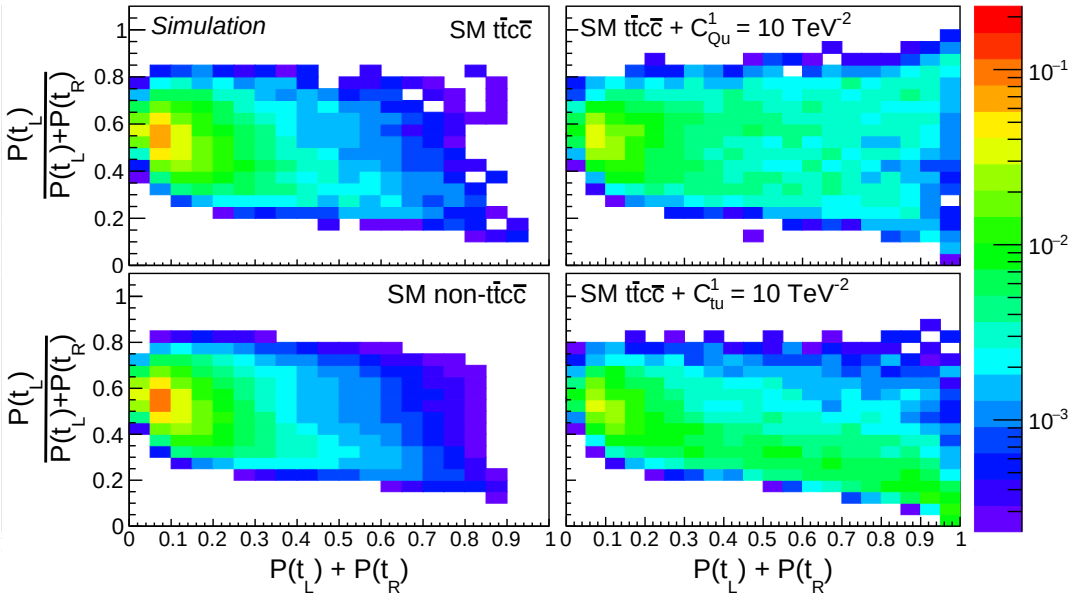


FIGURE 6.30: Normalized two–dimensional distributions of $\frac{P(t_L)}{P(t_L)+P(t_R)}$ on the y–axis and $P(t_L) + P(t_R)$ on the x–axis for SM $t\bar{t}c\bar{c}$ events (top left), SM non–$t\bar{t}c\bar{c}$ events (bottom left), SM+EFT $t\bar{t}c\bar{c}$ events with $C^1_{Qu} = 10$ TeV$^{-2}$ (top right) and SM+EFT $t\bar{t}c\bar{c}$ events with $C^1_{tu} = 10$ TeV$^{-2}$ (bottom right).

95% confidence level contours are summarized in Fig. 6.31. The full red contour is the result from the one–dimensional selection on the SM versus EFT discriminator alone, whereas the sensitivity in SR 1 and SR 2 is shown by the green and blue contours respectively. The limits in SR 1 and SR 2 show an improved sensitivity to $C^1_{Qu}$ and $C^1_{tu}$ respectively. By combining the obtained limits from each of the signal regions, an improved sensitivity is observed as shown by the red dotted line. These results are consistent with the observations in the phenomenological $t\bar{t}b\bar{b}$ study in Fig. 6.17 on the left.

### Two–dimensional template fits to the NN discriminators

Finally, a two–dimensional template–fitting strategy is adopted to derive limits on both operators simultaneously using the templates shown in Fig. 6.30. These templates are binned using the binning[12]

$$\mathrm{P}(t_L) + \mathrm{P}(t_R) \otimes \frac{\mathrm{P}(t_L)}{\mathrm{P}(t_L) + \mathrm{P}(t_R)} : [0, 0.25, 0.4, 0.65, 1] \otimes [0, 0.2, 0.4, 0.6, 0.8, 1],$$

and are then unrolled onto a one–dimensional histogram that is fitted according to Eq. (6.18). This function takes the same form as Eq. (6.14), with the addition of an extra SMEFT template to accommodate simultaneously the $t_L$ and $t_R$ categories in the fit. These two templates are scaled according to their linear and quadratic contributions to the $t\bar{t}c\bar{c}$ yield. The interference between $C^1_{Qu}$ and $C^1_{tu}$ (corresponding to the coefficient $p_{AB}$) is not included, but was found to be a factor of 20 smaller

[12] Bins with too few events are excluded from the fit to avoid unexpected behavior of the templates from the variations of the systematic uncertainties. This resulted in only 17 out of 20 bins being used in the final fit.

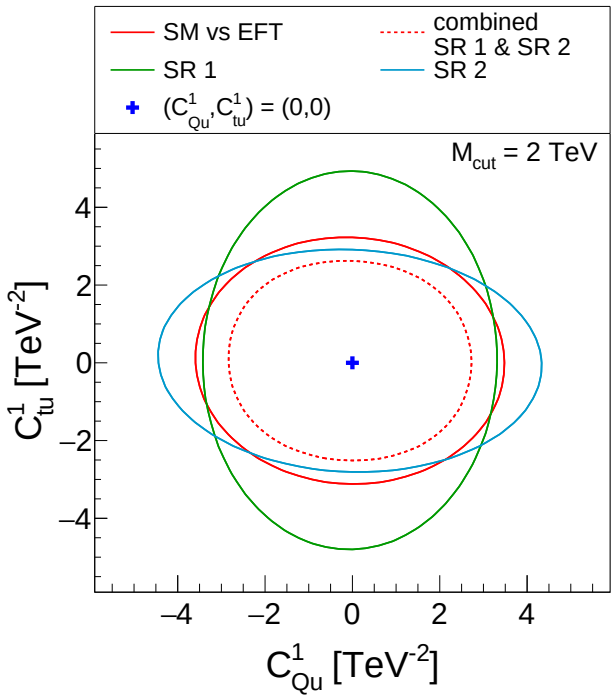


FIGURE 6.31: Two-dimensional limits at 95% CL using 41.5 fb$^{-1}$ of data in the $t\bar{t}$+HF analysis and allowing two couplings, $C^1_{Qu}$ and $C^1_{tu}$ to vary simultaneously: (red) one dimensional cut on $P(t_L) + P(t_R)$; (green) SR 1; (blue) SR 2; (red dashed) combination of SR 1 and SR 2; (black) two–dimensional template fit.

than the squared order contributions of each individual Wilson coefficient.

$$
\begin{aligned}
f\left(\mathrm{N_{SM}}, C^1_{Qu}, C^1_{tu}\right) = \mathrm{N_{SM}} \cdot & \left( \frac{N^{\mathrm{MC}}_{t\bar{t}c\bar{c}}}{N^{\mathrm{MC}}_{\mathrm{All,SM}}} \cdot \mathrm{H}^{\mathrm{norm}}_{t\bar{t}c\bar{c}} + \frac{N^{\mathrm{MC}}_{\text{non-}t\bar{t}c\bar{c}}}{N^{\mathrm{MC}}_{\mathrm{All,SM}}} \cdot \mathrm{H}^{\mathrm{norm}}_{\text{non-}t\bar{t}c\bar{c}} \right) \\
& + \mathrm{N_{SM}} \cdot \left( \frac{N^{\mathrm{MC}}_{t\bar{t}c\bar{c}}}{N^{\mathrm{MC}}_{\mathrm{All,SM}}} \right) \cdot \left[ p_{\mathrm{A}} \cdot C^1_{Qu} + p_{\mathrm{AA}} \cdot \left( C^1_{Qu} \right)^2 \right] \cdot \mathrm{H}^{\mathrm{norm}}_{t_L} \\
& + \mathrm{N_{SM}} \cdot \left( \frac{N^{\mathrm{MC}}_{t\bar{t}c\bar{c}}}{N^{\mathrm{MC}}_{\mathrm{All,SM}}} \right) \cdot \left[ p_{\mathrm{B}} \cdot C^1_{tu} + p_{\mathrm{BB}} \cdot \left( C^1_{tu} \right)^2 \right] \cdot \mathrm{H}^{\mathrm{norm}}_{t_R}
\end{aligned}
\tag{6.18}
$$

The postfit distributions of the unrolled histograms for data and simulation are shown in Fig. 6.32 on the left. The $t_L$ and $t_R$ templates are superimposed in red and blue respectively, each with an arbitrary normalization that results in the same yield as the $t\bar{t}c\bar{c}$ component. The data once again show a good agreement with the SM expectations and limits are derived on the Wilson coefficients. The right panel in Fig. 6.32 shows a scan of the two–dimensional negative logarithm of the likelihood and the 95% CL contour is superimposed with the red line. This contour is repeated in black Fig. 6.33 and is compared to the red contour from a one–dimensional selection on the SM versus EFT discriminator. The template fit does not use any information on the jet flavor to extract the $t\bar{t}c\bar{c}$ component. From the previous results summarized in Fig. 6.29 for individual operators, it was observed that this missing flavor information resulted in a slightly worse sensitivity of the template fit (green) compared to the one–dimensional neural network selection (blue). Analogously, one could expect also to see a worse sensitivity with the template–fitting strategy in the case when two operators are allowed to vary simultaneously. Nevertheless, the template

fit reaches a better sensitivity than the one–dimensional selection. Even though the nature of the two methods is perhaps too different for a direct comparison, this observation demonstrates the benefit of a two–dimensional template–fitting method over a selection–based approach that uses the cross section in a sensitive phase space. In any case, it is encouraging to realize that further improvements can be expected from the template fit when flavor information of the jets is added to the neural network training. Finally it is observed that this strategy overall yields a better sensitivity to the $t_R$ category because its template is more distinct from the SM one when compared to the $t_L$ category (see Figs. 6.24 and 6.30).

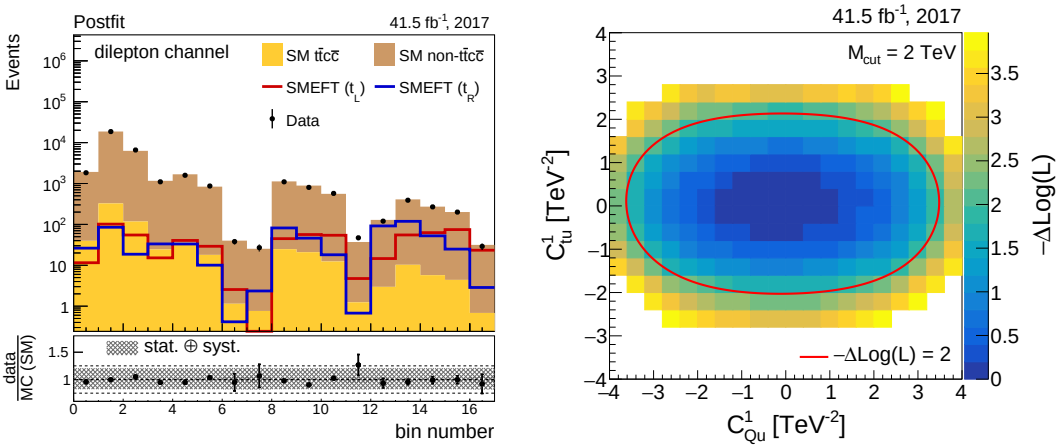


FIGURE 6.32: (left) Data–to–simulation agreement after the template fit of the data to the unrolled (one–dimensional) SM, $t_L$ and $t_R$ templates. The $t_L$ and $t_R$ templates are superimposed in red and blue respectively, each with an arbitrary normalization that results in the same yield as the $t\bar{t}c\bar{c}$ component.
(right) Two–dimensional scan of the negative logarithm of the likelihood from the template fit as a function of the value of the Wilson coefficients $C^1_{Qu}$ and $C^1_{tu}$. The intersections with $-\Delta\log(\mathrm{L}) = 2$ are drawn by the red contour and denote the 95% CL interval on the Wilson coefficients.

### Discussion on the SMEFT interpretation of the t$\bar{\text{t}}$+HF final–state

A novel method has been presented to increase the sensitivity of a given final–state to the presence of EFT operators using Machine Learning algorithms. We have demonstrated this improved strategy by using multi–class neural networks which are trained on a set of reconstructed observables that are sensitive to the presence of EFT operators and therefore learn to identify sensitive phase–space regions. By deriving limits in the most sensitive phase–space, or by performing template fits to the NN discriminators, the sensitivity to these new operators can be significantly increased.

The method has first been introduced using a phenomenological study of the sensitivity of the $t\bar{t}b\bar{b}$ final–state to a set of four–heavy–quark operators. With projections for 300 fb$^{-1}$ after Run–3 of the LHC, it has been shown that limits on the individual Wilson coefficients can be improved by a factor of more than two when using the neural network discriminators. Also when multiple operators are considered simultaneously, the network allows to disentangle effects from different types of operators, as was shown for operators with either a LH or a RH top–quark current. This significantly improves the limits on these operators,

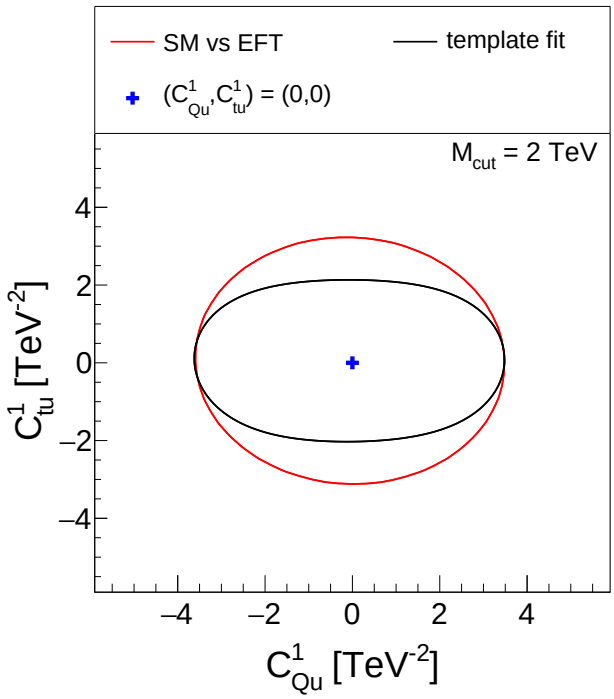


FIGURE 6.33: Two-dimensional limits at 95% CL using 41.5 fb$^{-1}$ of data in the $t\bar{t}$+HF analysis and allowing two couplings, $C^1_{Qu}$ and $C^1_{tu}$ to vary simultaneously: (red) one dimensional cut on $P(t_L) + P(t_R)$; (black) two–dimensional template fit.

and allows to pinpoint more accurately the origin of a hypothetical excess in the data.

The methodology was verified using the first measurement of the $t\bar{t}c\bar{c}$ cross section presented in this manuscript. By using real collision data corresponding to an integrated luminosity of 41.5 fb$^{-1}$, the sensitivity of the $t\bar{t}c\bar{c}$ process to a set of 14 two–heavy–two–light operators has been studied. The findings in the phenomenological $t\bar{t}b\bar{b}$ study have been verified, demonstrating that these methods can successfully be applied in an actual analysis. The obtained limits on many of these operators supersede those derived from a projected limit on the four–top–quark signal strength using 300 fb$^{-1}$ [108], *i.e.* with 7 to 8 times more integrated luminosity (see Tab. 6.4). This clearly motivates the use of these methods in future SMEFT interpretations. It is however important to note that the presented study still only serves as a proof–of–principle and a proper treatment of all four–quark operators, including their effect on both the $t\bar{t}b\bar{b}$ and $t\bar{t}c\bar{c}$ (and perhaps even the $t\bar{t}$LF) final–states is needed to make consistent claims about the allowed ranges of the corresponding Wilson coefficients.

# Chapter 7 Conclusions and prospects

## 7.1 Conclusion

The Large Hadron Collider has fulfilled its primary objective by revealing for the first time the existence of the Brout–Englert–Higgs boson back in 2012, thus confirming the Standard Model of Particle Physics as we know it today. With this discovery several years behind us, we have entered a precision era in measuring the properties of this scalar boson to confirm its consistency with the predictions from the SM, or to reveal new physics phenomena. In measuring the coupling of the Higgs boson to the heaviest quarks (the top and bottom quarks), the need for precise background estimations has triggered the effort to study the production of a top quark pair with additional bottom ($t\bar{t}b\bar{b}$) and charm ($t\bar{t}c\bar{c}$) jets. Whereas the $t\bar{t}b\bar{b}$ process has been measured by the CMS and ATLAS collaborations at different center–of–mass energies, this thesis presents the first measurement of the $t\bar{t}c\bar{c}$ cross section in proton–proton collisions at a center–of–mass energy of 13 TeV using 41.5 fb$^{-1}$ of integrated luminosity collected with the CMS experiment. This measurement is performed in the dileptonic decay channel of the top quark pairs and relies on the use of recently developed charm–jet identification algorithms. These state–of–the–art heavy–flavor tagging algorithms are outlined in detail in the context of this thesis and their performance is discussed. The analysis is based on a template–fitting method using templates of a neural network classifier that is trained to identify the different flavor categories defined by the additional jets. This allows to simultaneously extract the $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$LF cross sections, as well as the ratios $\mathrm{R}_c = {}^{\sigma_{t\bar{t}c\bar{c}}}/_{\sigma_{t\bar{t}jj}}$ and $\mathrm{R}_b = {}^{\sigma_{t\bar{t}b\bar{b}}}/_{\sigma_{t\bar{t}jj}}$. To this end, a novel calibration of the full shape of the $c$–tagging discriminator distributions has been employed for the first time such that this information can be used in the construction of the multivariate discriminator.

The $t\bar{t}c\bar{c}$ cross section is measured to be $\sigma^{\mathrm{vis}}_{t\bar{t}c\bar{c}} = 0.278 \pm 0.028\,(\mathrm{stat.}) \pm 0.049\,(\mathrm{syst.})$ pb in the visible phase space and $\sigma^{\mathrm{full}}_{t\bar{t}c\bar{c}} = 5.86 \pm 0.48\,(\mathrm{stat.}) \pm 1.03\,(\mathrm{syst.})$ pb in the full phase space. The obtained precision of around 20% on the $t\bar{t}c\bar{c}$ cross section will result in a much more accurate estimation of this important background in the measurement of the $t\bar{t}$H (H$\rightarrow b\bar{b}$) process. The ratio of the $t\bar{t}c\bar{c}$ cross section to the inclusive $t\bar{t}jj$ cross section is found to be $\mathrm{R}^{\mathrm{vis}}_c = 1.42 \pm 0.14\,(\mathrm{stat.}) \pm 0.21\,(\mathrm{syst.})$ % in the visible phase space and $\mathrm{R}^{\mathrm{full}}_c = 3.15 \pm 0.29\,(\mathrm{stat.}) \pm 0.48\,(\mathrm{syst.})$ % in the full phase space. The dominant uncertainties in these measurements come from the $b$– and $c$–tagging calibrations, as well as theoretical uncertainties on the choice of the renormalization and factorization scales. The measured values for $\sigma_{t\bar{t}c\bar{c}}$

and $R_c$ coincide well with the theoretical expectations from the POWHEG and MG5_aMC@NLO simulations, as summarized in Figs. 5.26 and 5.27 for the visible and full phase space respectively. The measured $t\bar{t}b\bar{b}$ cross section and its ratio $R_b$ to the inclusive $t\bar{t}jj$ production are found to be higher than the predictions from the simulations to the level of up to three standard deviations. These findings are consistent with results from previous dedicated analyses that focus specifically on measuring the $t\bar{t}b\bar{b}$ cross section [11, 12]. It will thus be interesting to await new results from these dedicated analyses, including measurements in the semileptonic and fully hadronic decay channels of the top quark pairs, that could either confirm or counter this observed discrepancy.

Finally, an interpretation of these results in the model–independent framework of the Standard Model Effective Field Theory would provide valuable information on the allowed parameter space for new physics interactions. A new method has been presented that uses multi–class neural network classifiers to increase the sensitivity of a given process of interest to a set of SMEFT operators. By exploiting all the available kinematical information of the final–state particles, the network is trained to identify events with an insertion of an EFT vertex, but also to disentangle effects from different types of operators. The method is introduced via a phenomenological study of the $t\bar{t}b\bar{b}$ final–state and its sensitivity to a set of four–heavy–quark operators. Using the information comprised in the network discriminators, either through a selection or via a template–fitting method, it was shown that more stringent constraints can be obtained on these SMEFT operators. The network was able to learn the difference between operators with a left–handed or right–handed top quark chirality, resulting in the best sensitivity by constructing dedicated discriminators that identify events from either of these classes. It was also demonstrated that in case of a hypothetical excess in the data, this method allows to pinpoint most accurately the responsible operators. Especially when multiple operators are allowed to contribute simultaneously, this allows to lift possible ambiguities in the interpretation and to point towards the operators that caused the hypothetical excess. The production of the $t\bar{t}b\bar{b}$ final–state is an indispensable component in constraining the total set of four–heavy–quark operators. The feasibility of applying this new method in an analysis was demonstrated by considering the effect of a set of 14 two–heavy–two–light operators on the $t\bar{t}c\bar{c}$ production cross section and consequently interpreting the obtained results from the measurement conducted in this thesis. Indeed, the best sensitivity to individual operators was obtained by measuring the $t\bar{t}c\bar{c}$ cross section after making a selection on a dedicated neural network output that was trained to separate SM events from events involving SMEFT operators with either a LH or a RH top–quark current. A similar sensitivity was obtained by making use of a template–fitting method of the simulated templates for SM and EFT events to the data. Even though the template–fitting strategy did not make use of any jet–flavor information to extract the $t\bar{t}c\bar{c}$ process from the overwhelming $t\bar{t}$LF background, such a strategy is already able to constrain these operators with a comparable strength. Therefore an even more enhanced sensitivity is expected when the jet–flavor information is included in the training of this network. This study leads to the most stringent individual limits that are currently available on some of these operators, under the assumption that the operators affect only the $t\bar{t}c\bar{c}$ process in the full $t\bar{t}$+HF production. A full interpretation of the $t\bar{t}$+HF production in terms of all four–fermion SMEFT operators simultaneously is left for future studies, exploiting the full Run–2 dataset of the LHC. This thesis presents the novel groundwork for this to emerge.

## 7.2 Prospects for the future

This thesis contains multiple novel techniques, such as the $c$ jet identification algorithms with the corresponding $c$-tagging shape calibration, the first measurement of the $t\bar{t}c\bar{c}$ cross section and the ML–based SMEFT interpretation. Even though large efforts have been made to present the most performant versions of these new techniques, there remains room to further optimize and improve upon the presented results. In this final section of the thesis the potential improvements, some of which are actively being worked on, are summarized.

### Prospects on the $t\bar{t}$+HF measurement

Although the analysis presented in Chapter 5 makes use of the DeepCSV tagger, it was clearly demonstrated that a more powerful jet–flavor identification can be achieved with the most recent DeepFlavor tagger. To get an idea of this expected improvement, Fig. 7.1 compares the ROC curves that show how well the $t\bar{t}c\bar{c}$ events can be distinguished from the $t\bar{t}b\bar{b}$ events (top) and from $t\bar{t}$LF events (bottom) solely based on the CvsL and CvsB $c$–tagging discriminants of the first (full lines) and the second (dotted lines) additional jet. The top panel clearly demonstrates an improved performance of the DeepFlavor CvsB discriminator (orange) compared to the DeepCSV algorithm (blue), whereas the bottom panel conveys the same message for the DeepFlavor CvsL discriminant (red) compared to the DeepCSV CvsL discriminant (green). It is therefore expected that by replacing the DeepCSV with the DeepFlavor tagger outputs in the $\Delta_b^c$ and $\Delta_L^c$ discriminators, the templates for the different event categories will be more clearly separated, resulting in a lower statistical uncertainty on the template–fitting method. However, given that the measurement of the $t\bar{t}c\bar{c}$ cross section is dominated by systematical uncertainties, the impact of this improved discrimination on the final sensitivity may be limited. Nevertheless, a full shape calibration of the DeepFlavor $b$–tagging discriminants was recently provided by the BTV group of the CMS Collaboration. All ingredients would thus be in place to derive also the DeepFlavor $c$–tagging shape calibration in analogy to Sec. 5.7. The expected improvement thus depends largely on the systematical uncertainty that will reside from the calibration of the DeepFlavor tagger.

The $c$–tagger shape calibration itself has room for further improvements. The current method is based on an enrichment of $b$, $c$ and light jets in a topology of semileptonic top quark pair events. By considering multiple topologies at once, this enrichment can be enhanced even further to obtain control regions which are very pure in either of the jet flavors. The semileptonic top quark pairs are still an ideal candidate to provide a $b$–enriched control region, and may benefit even further from the requirement of having a soft lepton inside the $b$ jets. By considering events with two oppositely–charged, same–flavor leptons with an invariant mass around the Z boson mass, one can extract a sample of Z+jets events which is highly enriched in light–flavor jets. Finally, the W+charm topology has proven to be very pure in $c$ jets in the derivation of the working point scale factors and could therefore also be used in the shape calibration. Especially this topology would provide a clear benefit to the method since the charm enrichment in the semileptonic $t\bar{t}$ events was rather poor (only around 20%). By applying the methodology outlined in Sec. 5.7.2 to these three control regions (instead of considering the jets with the second, third and fourth highest $b$–tagging discriminant in the semileptonic $t\bar{t}$ events), the

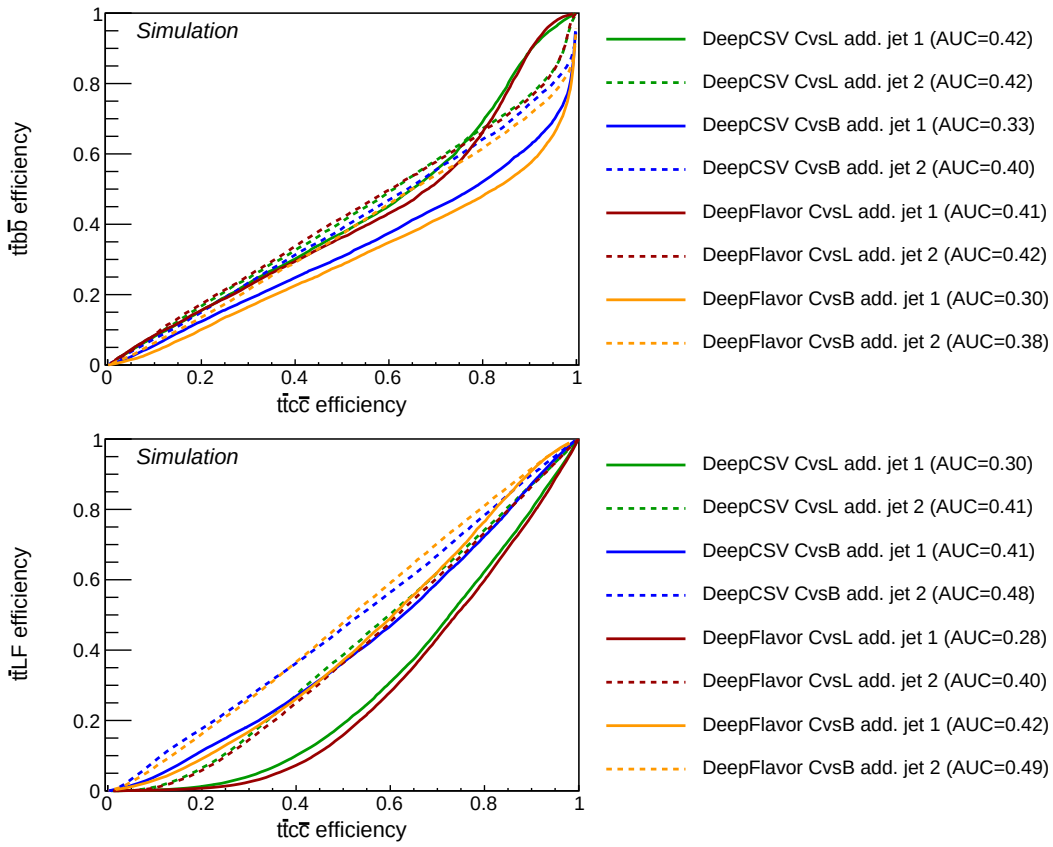


FIGURE 7.1: ROC curves comparing the performance of the DeepCSV to the DeepFlavor CvsL and CvsB $c$–tagging discriminators of the first (full lines) and second (dotted lines) additional jets. The top panel shows how well the $t\bar{t}c\bar{c}$ events can be distinguished from the $t\bar{t}b\bar{b}$ events, whereas the bottom panel shows the discriminating power between $t\bar{t}c\bar{c}$ and $t\bar{t}$LF events.

uncertainties related to this calibration can be further reduced, resulting in a more accurate determination of the $t\bar{t}c\bar{c}$ cross section. Ideally, one could not apply any a priori $b$–tagging shape calibration and instead perform an inclusive heavy–flavor shape calibration for all jet flavors with this method. This could eliminate the large uncertainties related to the $b$–tagging calibration, especially the ad hoc uncertainties on the $c$ jets would no longer be needed.

In order to reduce the theoretical uncertainties, dedicated studies of the simulation of the $t\bar{t}c\bar{c}$ and $t\bar{t}b\bar{b}$ final–states are needed, such as those presented in Refs. [172, 290]. These studies may motivate the use of dedicated simulations for $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$LF production with appropriate choices of the renormalization and factorization scales, instead of considering a single, inclusive $t\bar{t}$ simulation. The left panel in Fig. 7.2 shows the simulated differential cross section at NLO as a function of the number of additional charm jets (not from the top quark decays) in a dedicated $t\bar{t}c\bar{c}$ simulation (purple) and an inclusive $t\bar{t}$ simulation (black) for events with at least one $c$ jet. In the dedicated $t\bar{t}c\bar{c}$ simulation, the $t\bar{t}c\bar{c}$ final–state is generated at the level of the matrix element at NLO, whereas in the inclusive $t\bar{t}$ simulation the additional charm jets reside from the parton shower. Uncertainties related to the choice of $\mu_R$ and $\mu_F$ are shown by the purple and hatched regions for the $t\bar{t}c\bar{c}$ and inclusive $t\bar{t}$ sample respectively. The right panel in Fig. 7.2 shows the simulated

differential cross section as a function of the angular separation in ΔR between the additional charm jets, comparing again several dedicated $t\bar{t}c\bar{c}$ simulations to inclusive $t\bar{t}$ simulations. This is one of the observables used in the neural network that is trained to separate $t\bar{t}c\bar{c}$ from $t\bar{t}b\bar{b}$ and $t\bar{t}$LF events in Sec. 5.8. Both of these figures clearly demonstrate that large differences of up to 50%, both in shape and in normalization, can be observed along different simulations. The uncertainties due to the choice of renormalization and factorization scale also have a large impact on the measured spectra. For more details, the reader is referred to Refs. [172, 290], though it should become clear that more detailed studies are needed to decide on how to improve the simulation of the $t\bar{t}$+HF final–state.

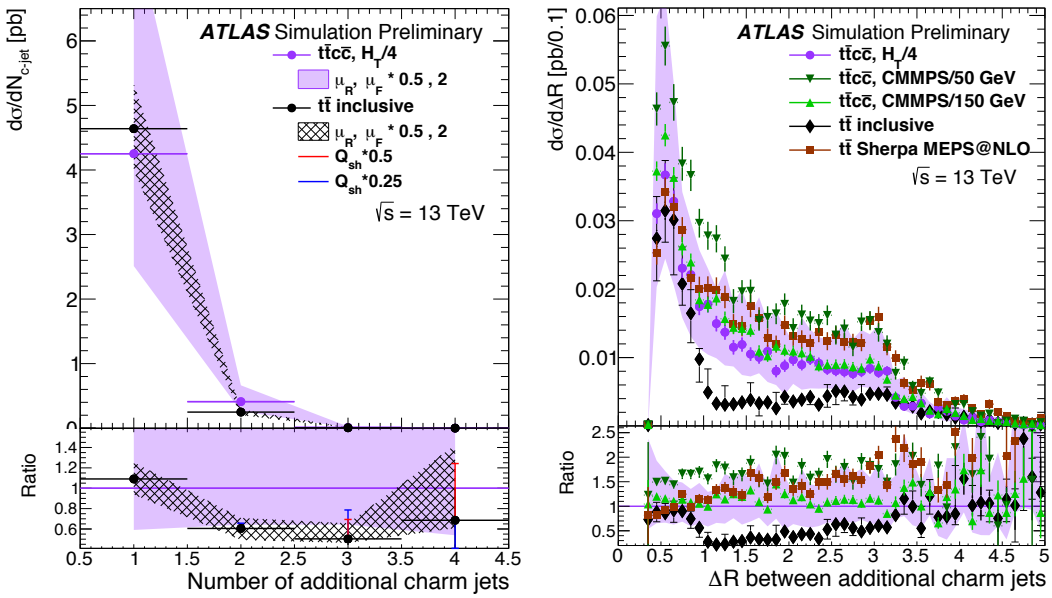


FIGURE 7.2: (left) Differential cross section as a function of the number of additional charm jets (not from the top quark decays) in a dedicated $t\bar{t}c\bar{c}$ simulation (purple) and an inclusive $t\bar{t}$ simulation (black) for events with at least one $c$ jet. Uncertainties related to the choice of $\mu_R$ and $\mu_F$ are shown by the purple and hatched regions for the $t\bar{t}c\bar{c}$ and inclusive $t\bar{t}$ sample respectively. (right) Simulated differential cross section as a function of the angular separation in ΔR between the additional charm jets using simulated $t\bar{t}c\bar{c}$ samples with different scale choices as well as different inclusive $t\bar{t}$ simulations. Figures taken from Ref. [290], in which more information can be found on the notation.

The measured values of $\sigma_{t\bar{t}b\bar{b}}$ and $R_b$ were found to be larger than the predictions from simulations to the level of up to three standard deviations (see Figs. 5.26 and 5.27). This is in line with previous observations in CMS at 8 and 13 TeV from dedicated $t\bar{t}b\bar{b}$ analyses [11, 12]. Also the most recent measurement of the $t\bar{t}b\bar{b}$ process from the ATLAS Collaboration at 13 TeV [15] consistently observes an underestimation of $\sigma_{t\bar{t}b\bar{b}}$ in the simulations, both in the dileptonic and semileptonic decay channels of the top quark pairs. These results are summarized in Fig. 7.3, showing differences up to a factor of two, corresponding to a disagreement of up to two standard deviations. It will thus be interesting to await new results from these dedicated $t\bar{t}b\bar{b}$ analyses using larger datasets and to see if the observed differences are indeed further confirmed or not. If so, dedicated efforts will be needed to better understand the modeling of this final–state, before searching for alternative

explanations due to new physics phenomena.

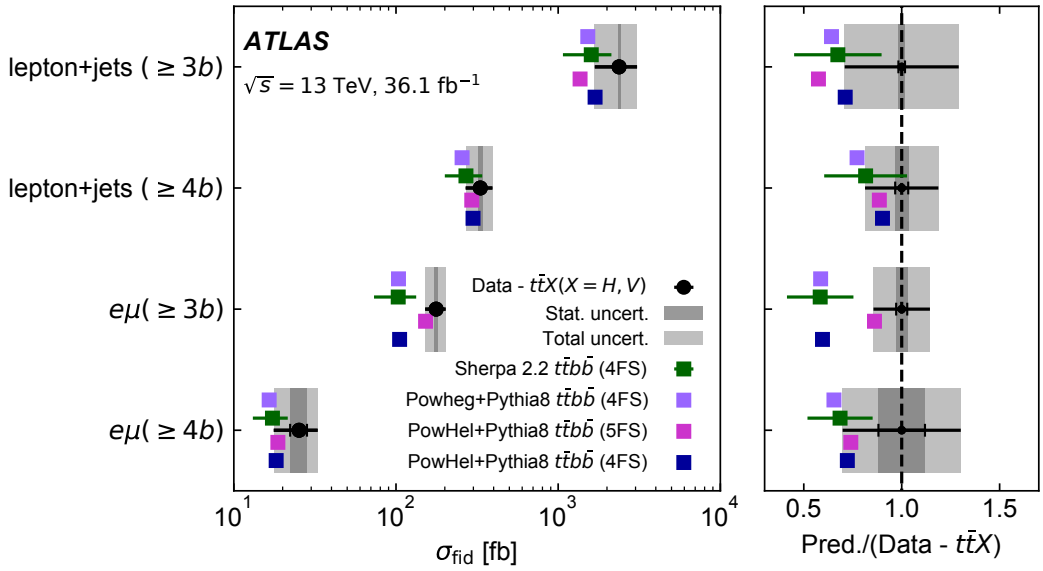


FIGURE 7.3: The measured $t\bar{t}b\bar{b}$ cross section by ATLAS at 13 TeV using 36.1 fb$^{-1}$ in different fiducial phase space definitions, compared to several theoretical predictions from different simulations. Figure taken from Ref. [15], in which more information can be found on the notation.

The currently most sensitive measurement of the $t\bar{t}$H process in the H $\to b\bar{b}$ decay channel by CMS at 13 TeV results in a best fitted signal strength $\mu = 0.72 \pm 0.24\,(\text{stat.}) \pm 0.38\,(\text{syst.})$ [80]. The dominant uncertainty in this measurement results from a conservative 50% rate uncertainty that is applied on the total $t\bar{t}$+HF background yield. The impact on the fitted signal strength due to the combination of this $t\bar{t}$+HF normalization and the parton shower uncertainties is found to be $\Delta\mu(\sigma_{t\bar{t}+\text{HF}} + \text{PS}) = {}^{+0.24}_{-0.28}$ in absolute value. In terms of relative uncertainty, this accounts for ${}^{\Delta\mu}/_{\mu}(\sigma_{t\bar{t}+\text{HF}} + \text{PS}) = {}^{+33}_{-39}$ %. The measurement of the $t\bar{t}c\bar{c}$ and $t\bar{t}b\bar{b}$ cross section presented in Chapter 5 resulted in a relative uncertainty of around 20%. Even when being conservative, the rate uncertainty on the $t\bar{t}$+HF background in the $t\bar{t}$H can be reduced by a factor of two compared to the current measurement. Even though it is hard to assess the dedicated impact on the final result of the $t\bar{t}$H (H $\to b\bar{b}$) signal strength, it can be expected to reduce by a factor 1/3 the theoretical uncertainty on this measurement from the improved $t\bar{t}$+HF background estimation. A comparable improvement in the statistical uncertainty would require a significant amount of additional integrated luminosity. An accurate measurement of the $t\bar{t}$+HF process thus deserves to be recognized as a vital component in the future observation of the $t\bar{t}$H process in the H $\to b\bar{b}$ decay channel.

### Prospects on the SMEFT interpretations

The results presented in Chapter 6 have clearly demonstrated the potential of the new Machine Learning based method for SMEFT interpretations, together with its feasibility in the $t\bar{t}$+HF analysis. There is however still a lot of room for extending and improving this method.

- The effect of more advanced Machine Learning methods could result in a significantly larger sensitivity of the method. Parametrized learning approaches

could potentially also be applied to learn about the interference effects between the SM and the SMEFT operators. The use of recurrent layers could also prove their power by learning about sequential ordering in the final–state objects, in analogy to what is used for the DeepFlavor algorithm as explained in Sec. 4.4. Given that one expects at least two particles from the EFT vertex with on average a larger energy, it could be useful to rank the observed particles according to their $p_T$, and feed that ordered list into a recurrent layer. Other studies have also shown the benefits from optimized loss functions that encapsulate information from the event generators themselves into the training [236].

- The power of a template–fitting method to the (multi–dimensional) neural network discriminators was already shown in Chapter 6. Further significant improvements are expected by including also jet–flavor information into these Machine Learning classifiers such that they are also able to extract the different flavor categories together with the different EFT contributions.

- Currently the implementation of the four–quark operators in the UFO model exists only at leading order. Once these calculations become available also at NLO accuracy, more precise predictions can be made, which will result in more accurate constrains on the SMEFT operators.

- Finally, a complete interpretation of the $t\bar{t}$+HF analysis in terms of all possible SMEFT operators that can affect this topology would be needed. This should include effects of all four–quark operators (and perhaps even the top and bottom quark chromomagnetic dipole operators from Eqs. (1.30) and (1.31)) on the production of $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ and $t\bar{t}$LF events. Ideally this measurement could be included in a global fit of top–quark operators in the SMEFT to a variety of top–quark related measurements, such as performed in Refs. [105, 106].

- These methods could also be applied in other analyses, not necessarily including top quarks.

In conclusion, the techniques developed in this thesis allowed for a first measurement of the $t\bar{t}c\bar{c}$ cross section, and to derive stringent constraints on a set of four–quark EFT operators. The use of advanced flavor–tagging algorithms and other Machine Learning based methods has played a vital role in this achievement. The presented analyses use a dataset of 13 TeV proton–proton collisions with an integrated luminosity of 41.5 fb$^{-1}$ collected with the CMS detector. By now the LHC has already delivered a high–quality dataset of around 150 fb$^{-1}$ that is ready to be analyzed, of which over 100 fb$^{-1}$ resides from 13 TeV collisions. Applying the developed analysis techniques on this full dataset would not only result in a reduction of the statistical uncertainty, but would allow for a more accurate assessment of several experimental systematical uncertainties as well. I eagerly look forward to exploring this unprecedented dataset and hope it will take us one step closer towards unravelling the mysteries of the universe. Who knows, it might even open up an entirely new box of questions!

# Contributions and achievements

This section aims to emphasize the research topics in this thesis to which I have directly contributed, and to list my additional contributions to the CMS Collaboration that are not directly discussed in this manuscript.

**A. First measurement of the $t\bar{t}c\bar{c}$ cross section**
I have developed the full analysis presented in Chapter 5. Starting from the event selections which have been optimized in previous $t\bar{t}b\bar{b}$ analyses, several innovative methods have been developed to measure for the first time the $t\bar{t}c\bar{c}$ cross section. These results have been endorsed by the CMS TOP Physics Analysis Group.

- An improved neural network based jet–parton assignment has been employed with a increased efficiency compared to the assignment used in existing $t\bar{t}b\bar{b}$ analyses (see Sec. 5.6).
- I have developed a novel method to calibrate the differential shape of the c–tagger distributions based on an iterative fitting approach using semileptonic top quark pair events (see Sec. 5.7). This calibration strategy has been picked up by other analyses within the CMS Collaboration and will most likely become a standard calibration supported by the $b$–tagging and vertexing (BTV) group in CMS.
- The analysis presents for the first time a measurement of the $t\bar{t}c\bar{c}$ cross section with a precision of around 20%, using 41.5 fb$^{-1}$ of proton–proton collision data at 13 TeV center–of–mass energy, collected by the CMS Collaboration during the 2017 data–taking period.

**B. Machine Learning for SMEFT interpretations**
I have introduced a novel method to increase the sensitivity to SMEFT operators using multi–class neural networks, as demonstrated in Chapter 6.

- The phenomenological study presented in Sec. 6.1 has been published in Ref. [109], of which I am the main contact author.
- I presented this work in the form of a poster during the 11$^{\text{th}}$ International Workshop on Top Quark Physics, 19$^{\text{th}}$ September 2018, Bad Neuenahr, Germany, for which I was awarded a best poster prize.
- The application of this method as explained in Sec. 6.2 is a first demonstration that these methods can indeed be applied within the experimental collaborations to provide an improved sensitivity of a given final–state to a set of SMEFT operators.

**C. Top quark flavor–changing interactions with Dark Matter**
I have contributed to a phenomenological study exploring possible flavor–changing interactions between top– and charm quarks and dark matter ($\chi$).

These results are published in Ref. [84], in which we explore the collider phenomenology as well as the relic dark matter abundance using an EFT description and a simplified model with a heavy Z' mediator to parametrize the $tc\chi\chi$ interactions.

**D. CMS b–tagging and vertexing (BTV) activities**

For the past five years, I have been an active member of the *b*–tagging and vertexing physics object group (BTV POG) in the CMS Collaboration.

- I have developed the first charm–tagging algorithm available for the CMS Collaboration [252]. I was selected by the CMS BTV group to present these results during the 8$^{\text{th}}$ International Workshop on Charm Physics (CHARM2016), September 2016, Bologna, Italy, from which the proceedings can be found in Ref. [291].
- I have contributed to the maintenance and retraining of the CSVv2 tagger when it was still in use in 2016.
- I was involved in the development of the first heavy–flavor tagging algorithms based on deep neural networks (DeepCSV and DeepFlavor), as explained in Sec. 3.3.3 and 4.4.
- I have performed commissioning studies on properties of heavy–flavor jets in data collected during the start of the 2017 data–taking period of CMS, when the new pixel detector had just been installed. These findings are summarized in a detector performance summary note [292].
- I have developed an offline monitoring system that allows to monitor the quality of simulated tracks resulting from *b* and *c* hadron decays whenever new software releases become available.
- I have contributed significantly to the reference paper on heavy–flavor tagging for Run–2 in CMS [225]. More precisely, I have created all the figures shown in Sec. 3.3.2 of this thesis, and I have provided input to Chapter 5.2 (on charm jet identification) in Ref. [225].
- I have been level–3 convener of the performance and calibration subgroup of the BTV POG from September 2017 to September 2018, followed by my current level–3 convenership of the software and algorithm subgroup of the BTV POG from September 2018 to September 2019. During these mandates I was/am responsible for the results provided by a group of around 20 to 30 fellow CMS members. This includes commissioning and calibration studies as well as software developments.

**E. CMS central shifts for the detector control system**

I have performed service shifts by monitoring the central detector control system (DCS) for the CMS experiment during the start of the 2017 data–taking period of the LHC. During this period I have worked closely together with the different sub–detector experts to make sure the detector works properly and is able to provide high–quality data.

# Summary

The Standard Model (SM) of Particle Physics describes the elementary building blocks of the universe and the forces that are responsible for the interactions amongst them. The discovery of the Brout–Englert–Higgs boson at the Large Hadron Collider (LHC) in 2012 marked the final milestone in completing this model as we know it today. Nevertheless, the Standard Model is not a complete theory, as it does not explain phenomena such as dark matter, neutrino masses, and it does not include gravity. An era of precision measurements is ahead to further confirm the predictions of the SM or perhaps reveal the origin of its shortcomings.

Motivated by the need for precise background estimates in the top–Higgs sector, the CMS and ATLAS Collaborations at the LHC have developed an ambitious program in measuring precisely the production of a top quark pair with additional bottom quark jets ($t\bar{t}b\bar{b}$). Such measurements rely on the use of dedicated bottom–jet identification algorithms. It is however of equal importance to measure also the production of a top–quark pair with additional charm quark jets ($t\bar{t}c\bar{c}$). Recently, the development of dedicated algorithms for charm–jet identification ($c$–taggers) have opened up the possibility to disentangle and measure also the $t\bar{t}c\bar{c}$ cross section for the first time. Such a measurement would provide an improved background estimation in the Higgs sector and would complement the $t\bar{t}b\bar{b}$ measurement to obtain a global picture of the SM production of top–quark pairs with additional heavy–flavor jets ($t\bar{t}$+HF).

This thesis presents the first measurement of the $t\bar{t}c\bar{c}$ process with the CMS detector at the LHC using 13 TeV proton-proton collision data collected in 2017, corresponding to an integrated luminosity of 41.5 fb$^{-1}$. The analysis strategy relies heavily on the information that resides in the charm–jet identification discriminants, as well as on the use of Machine–Learning based techniques to classify the events according to the additional jet flavor. To this end, a novel calibration of the differential $c$–tagger discriminant distributions has been developed. The $t\bar{t}c\bar{c}$ cross section is measured to be $\sigma^{\text{vis}}_{t\bar{t}c\bar{c}} = 0.278 \pm 0.028\,(\text{stat.}) \pm 0.049\,(\text{syst.})$ pb in the visible phase space and $\sigma^{\text{full}}_{t\bar{t}c\bar{c}} = 5.86 \pm 0.48\,(\text{stat.}) \pm 1.03\,(\text{syst.})$ pb in the full phase space. The obtained precision of around 20% on the $t\bar{t}c\bar{c}$ cross section will result in a much more accurate estimation of this important background in the measurement of the $t\bar{t}$H (H$\rightarrow b\bar{b}$) process. The ratio of the $t\bar{t}c\bar{c}$ cross section to the inclusive $t\bar{t}jj$ cross section is also measured and is found to be $\text{R}_c^{\text{vis}} = 1.42 \pm 0.14\,(\text{stat.}) \pm 0.21\,(\text{syst.})$ % in the visible phase space and $\text{R}_c^{\text{full}} = 3.15 \pm 0.29\,(\text{stat.}) \pm 0.48\,(\text{syst.})$ % in the full phase space. Furthermore, the presented analysis strategy allows for a simultaneous extraction of the $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ and $t\bar{t}$ + light–jet cross sections and their ratios to the inclusive $t\bar{t}jj$ cross section. Overall, the obtained results were found to be consistent with the predictions from the simulations, though the measured value of the $t\bar{t}b\bar{b}$

cross section and its ratio to the inclusive $t\bar{t}jj$ production cross section is larger than the predictions to the level of around three standard deviations. This observation is consistent with previous measurements from dedicated $t\bar{t}b\bar{b}$ analyses, triggering the need for more detailed investigations with more data on the source of this discrepancy.

Finally, the Standard Model Effective Field Theory framework (SMEFT) provides a model–independent way to interpret these measurements in terms of new physics interactions. A novel methodology is presented in which Machine Learning methods are used to provide more stringent constraints on the Wilson coefficients that parametrize the new physics interactions in the SMEFT. A multi–class shallow neural network was trained to distinguish SM from EFT effects, as well as to differentiate between the effects from EFT operators with different top–quark chirality, based on the final–state kinematics. The method has been introduced via a phenomenological study of the $t\bar{t}b\bar{b}$ process and its sensitivity to a set of four–heavy–quark operators of dimension six. It was shown that this allows for a competitive sensitivity of the $t\bar{t}b\bar{b}$ final–state compared to the four–top–quark process, with the additional advantage that the $t\bar{t}b\bar{b}$ process is sensitive to some previously unconstrained operators. This proposed method does not only result in more stringent constraints on the Wilson coefficients, but additionally allows to pinpoint more accurately the origin of a hypothetical excess observed in the data. The feasibility of applying this method in a real analysis was demonstrated using the results obtained from the CMS measurement of the $t\bar{t}c\bar{c}$ cross section presented in this thesis. Assuming that the production of the $t\bar{t}c\bar{c}$ final–state is possibly affected by a set of two–heavy–two–light–quark operators of dimension six, the most stringent individual constraints were obtained by making an additional selection on the neural network discriminators. When multiple operators are allowed to vary simultaneously, the use of a two–dimensional template–fitting method to the neural network discriminators has proven its power by obtaining some of the most stringent constraints in parts of the parameter space. An even larger sensitivity is expected from this method when additional information on the jet flavor is used in the neural network training to extract the different flavor categories in the $t\bar{t}$+HF final–state.

# Samenvatting

*"Eerste meting van de werkzame doorsnede van een top quark paar met extra charm jets, gebruikmakend van het CMS experiment"*

Het Standaard Model (SM) van de deeltjesfysica beschrijft de elementaire bouwstenen van het universum en de krachten die verantwoordelijk zijn voor de interacties tussen deze deeltjes. De ontdekking van het Brout–Englert–Higgs boson aan de Large Hadron Collider (LHC) in 2012 was de laatste mijlpaal om dit model te vervolledigen. Nochtans is het SM niet volledig, aangezien het er niet in slaagt om fenomenen zoals donkere materie, de massa van neutrino's of zelfs zwaartekracht te beschrijven. Een tijdperk van precisiemetingen staat voor de deur om de voorspellingen van het SM verder te bevestigen, of de oorsprong van zijn tekortkomingen te onthullen.

De nood aan precieze voorspellingen van de achtergrond processen in de top–Higgs sector heeft ertoe geleid dat de CMS en ATLAS Collaboraties nauwkeurig de productie van een top quark paar met extra bottom quark jets ($t\bar{t}b\bar{b}$) zijn gaan meten. Deze metingen zijn gebaseerd op gespecialiseerde algoritmes die ontworpen zijn om bottom quark jets te identificeren. Het is nochtans even belangrijk om ook de productie van een top quark paar met extra charm quark jets ($t\bar{t}c\bar{c}$) te meten. De recente ontwikkeling van algoritmes om charm quark jets te identificeren ($c$–taggers) opent de mogelijkheid om voor de eerste keer ook de $t\bar{t}c\bar{c}$ component af te zonderen en zijn werkzame doorsnede te meten. Deze meting zou een nauwkeurigere beschrijving kunnen geven van de belangrijkste achtergrond processen in de top–Higgs sector, alsook bijdragen tot een globaal beeld van de SM top–quark paar productie met extra heavy–flavor jets ($t\bar{t}$+HF).

Deze thesis presenteert de eerste meting van het $t\bar{t}c\bar{c}$ proces met de CMS detector aan de LHC, door de data van de 13 TeV proton–proton botsingen te gebruiken die in 2017 werden verzameld, overeenkomend met een geïntegreerde luminositeit van 41.5 fb$^{-1}$. De analyse berust voor een groot deel op de informatie die omvat is in de charm–jet identificatie algoritmes, alsook op het gebruik van Machine Learning om de botsingen te classificeren naargelang de smaak van de extra jets. Daarvoor is een nieuwe methode ontwikkeld om de verdeling van de $c$-tagger discriminanten te kalibreren. De bekomen waarde van de $t\bar{t}c\bar{c}$ werkzame doorsnede is $\sigma^{\text{vis}}_{t\bar{t}c\bar{c}} = 0.278 \pm 0.028\,(\text{stat.}) \pm 0.049\,(\text{syst.})$ pb in de zichtbare faseruimte en $\sigma^{\text{full}}_{t\bar{t}c\bar{c}} = 5.86 \pm 0.48\,(\text{stat.}) \pm 1.03\,(\text{syst.})$ pb in de volledige faseruimte. De precisie van rond de 20% op dit resultaat zal leiden tot een nauwkeurigere schatting van deze belangrijke achtergrond in de meting van het $t\bar{t}$H (H$\rightarrow b\bar{b}$) proces. De verhouding van de $t\bar{t}c\bar{c}$ werkzame doorsnede tot de inclusieve $t\bar{t}jj$ werkzame doorsnede werd

ook bepaald, waarvoor werd gevonden dat $\mathrm{R}_c^{\mathrm{vis}} = 1.42 \pm 0.14\,(\mathrm{stat.}) \pm 0.21\,(\mathrm{syst.})$ % in de zichtbare faseruimte en $\mathrm{R}_c^{\mathrm{full}} = 3.15 \pm 0.29\,(\mathrm{stat.}) \pm 0.48\,(\mathrm{syst.})$ % in de volledige faseruimte. Bovendien laat de strategie die wordt gebruikt in deze analyse toe om tegelijkertijd de $t\bar{t}c\bar{c}$, $t\bar{t}b\bar{b}$ en $t\bar{t}$ + light jet werkzame doorsnede, alsook hun verhouding tot de inclusieve $t\bar{t}jj$ werkzame doorsnede te meten. Over het algemeen komen de gemeten waarden goed overeen met de theoretische voorspellingen met behulp van de simulaties. Echter komen de gemeten waarden van de $t\bar{t}b\bar{b}$ werkzame doorsnede en zijn verhouding tot de $t\bar{t}jj$ werkzame doorsnede hoger uit dan wordt voorspeld, tot op het niveau van ongeveer drie standaardafwijkingen. Deze observatie is consistent met vorige metingen die specifiek toegespitst zijn op het meten van het $t\bar{t}b\bar{b}$ process, wat aanleiding geeft om dit process in meer detail te gaan bestuderen met meer botsingsdata.

Tot slot voorziet het theoretische kader van de "Standard Model Effective Field Theory" (SMEFT) een model–onafhankelijke manier om deze metingen te interpreteren aan de hand van nog onbekende interacties. Een nieuwe methode wordt gepresenteerd, waarbij Machine Learning algoritmes gebruikt worden om sterkere limieten af te leiden op de Wilson coëfficiënten die de nieuwe SMEFT interacties parametriseren. Een neuraal netwerk bestaande uit meerdere klassen, werd getraind op basis van de kinematische eigenschappen van de deeltjes in de eindtoestand, om een onderscheid te maken tussen de effecten ten gevolge van SM en EFT processen, alsook om te differentiëren tussen EFT interacties die top quarks met verschillende chiraliteit bevatten. De methode werd geïntroduceerd aan de hand van een fenomenologische studie van het $t\bar{t}b\bar{b}$ proces en zijn gevoeligheid aan een reeks vier–zware–quark operatoren van dimensie zes. De methode resulteert in een vergelijkbare gevoeligheid van het $t\bar{t}b\bar{b}$ proces in vergelijking met het vier–top–quark proces, met als bijkomstig voordeel dat het $t\bar{t}b\bar{b}$ proces gevoelig is aan een verzameling operatoren, dewelke hun toegestane bereik nog niet eerder werd beperkt aan de hand van metingen. De voorgestelde methode resulteert niet enkel in de sterkste limieten op de operatoren, maar laat ook toe om de oorsprong van een hypothetisch signaal in de data met een grotere nauwkeurigheid te bepalen. De toepasbaarheid van deze methode in een volledige analyse werd gedemonstreerd aan de hand van de meting van de $t\bar{t}c\bar{c}$ werkzame doorsnede uit deze thesis. In de veronderstelling dat het $t\bar{t}c\bar{c}$ proces kan worden beïnvloed door een reeks twee–zware–twee–lichte operatoren, werden de sterkste individuële limieten bekomen door een extra selectie toe te passen op de discriminatoren van het neurale netwerk. Indien meerdere operatoren tegelijkertijd mogen variëren, werd de kracht van template–fitting methodes duidelijk gemaakt, door de sterkste limieten te bekomen in specifieke delen van de faseruimte. Er wordt verwacht dat de gevoeligheid van deze methode nog zal toenemen door informatie over de smaak van de jets toe te voegen tijdens het trainen van het neurale netwerk om zo de verschillende smaak–gerelateerde categorieën in de $t\bar{t}$+HF productie van elkaar te onderscheiden.

# Acknowledgements

After the realization that I might have to finish my PhD thesis four months earlier than initially expected, some nerves rushed through my veins. Nevertheless, after an energizing honeymoon, I fully committed to this task. This thesis is proof that I succeeded! An achievement that I could never have fulfilled without the help and support of many people. To those people I dedicate this final chapter of the thesis.

First of all I would like to express my gratitude towards my supervisor, Jorgen D'Hondt, who allowed me to perform state–of–the–art research in a stimulating environment in which I had all the freedom I could wish for. Thanks for pointing me in the right direction when needed. Whenever we sat together to discuss results and new ideas, you fully devoted your time and attention to my project and I consider that one of the best gifts a young scientist can receive. With the same gratitude I thank also my co–supervisor, Alberto Mariotti. You have from the start (already during my Masters thesis) included me in the phenomenological research topics that we have collaborated on and you have introduced me to many fantastic people. To me, this has opened up a wide landscape beyond the experimental world, which has greatly contributed to my current knowledge. Thanks to the both of you for the great support!
I would in this context also like to thank specifically Kirill Skovpen, who was always available to help and share his valuable expertise on a variety of analysis topics. I truly appreciate this and I can honestly say that I owe the quality of many parts of this thesis to your help.
I additionally would like to thank all members of the jury for their availability and interest as well as for the interesting discussions. Your comments, questions and suggestions have without a doubt improved further the quality of this thesis.

For the past five years I had the privilege to work at the Inter–University Institute for High Energies (IIHE) at the Vrije Universiteit Brussel. The people at this institute have given me a warm welcome and helped creating a perfect environment in which to perform my research. There was always enough time for cheerful activities, ranging from the cakes and other sweets during the coffee breaks, the Christmas parties, annual meetings, St. Nicholas breakfasts, barbecues, sport activities, and many more. I would like to thank Marleen, for all of her much–appreciated help concerning organization and administration of any kind. But above all for your cheerfulness and for always putting a smile on everyone's face! Also I would like to thank the IT team for their fast responses and for providing us with such a great computing infrastructure. Furthermore, I consider myself lucky to have worked in a large office with multiple colleagues on which I could always rely for questions and discussions (and laughter!). Thanks Isis, Kevin, Lieselotte, Isabelle, Shimaa, Annik, Quentin, Nadir, Gerrit, Jarne, Kirill, Doug, Dom, Simon,

Denys, Ivan, Petra, Leonidas, Giannis, Emil, A.R. and Dimitrios! Even though we had to battle the sunlight blinding our screens for many years, using an oversized indoor parasol, we are now finally rewarded with proper sun blinds. Nevertheless the parasol remains in the office as a memorial to all who had to suffer in the past!

Already five years have passed since I became a member of the BTV community within the CMS Collaboration. I am proud to be a part of this team of fantastic and motivated people and I look very much forward to staying a part of that in the future. I would like to specifically express my gratitude towards Mauro Verzetti, who has taught me a great deal of computing skills towards the start of my PhD and pushed me towards the use of Python, for which I will be forever grateful! Also thanks to Petra Van Mulders, specifically for nominating me as L3 subgroup convener (this was an incredibly important step in my career!) and for making me responsible for the uniform style of all the plots in the BTV paper (not!).

As an Aspirant – Fonds Wetenschappelijk Onderzoek Vlaanderen, I would like to thank the FWO for supporting and funding my research over the past four years.

Finally, and above all, I want to thank my family for their support and for always being there for me. I am for sure one of the luckiest people with such a warm and loving family by my side, which is in the end all that really matters! Special thanks goes out to my parents, who have always believed in me and on whom I could always rely for anything that my heart could ever desire! I will never forget how you used to drive me to Brussels on the craziest hours, provided with a bag full of food and drinks so that I would never fall short! I think this book is the proof that justifies my choice to study science instead of Latin (still sorry for the late notice!) when I was only 12 years old!
Perhaps the biggest gratitude, the largest golden medal, the Nobel price for love and support (if it would exist) goes out to my wife, Laura. You have been my lucky charm for the past 12 years (and many more to come) and you have supported all of my decisions and choices. You make me the happiest husband and soon also the happiest father alive! I love you unconditionally.

To my baby–girl who is on her way to join our happy family, I want to say that I can not wait to meet you! Finding out about you has given me the final boost needed to finish this thesis. I am already incredibly proud and I hope that one day I can explain to you (at least part of) what is written in this book. But for now, no need to worry about quarks and heavy flavors, as there are many other flavors for you to discover first! I love you already!

# Appendix A
# Impact of the nuisance parameters

This appendix lists the impact of the nuisance parameters on each of the parameters of interest which are determined in Sec. 5.10.

## A.1 Impacts in the visible phase space

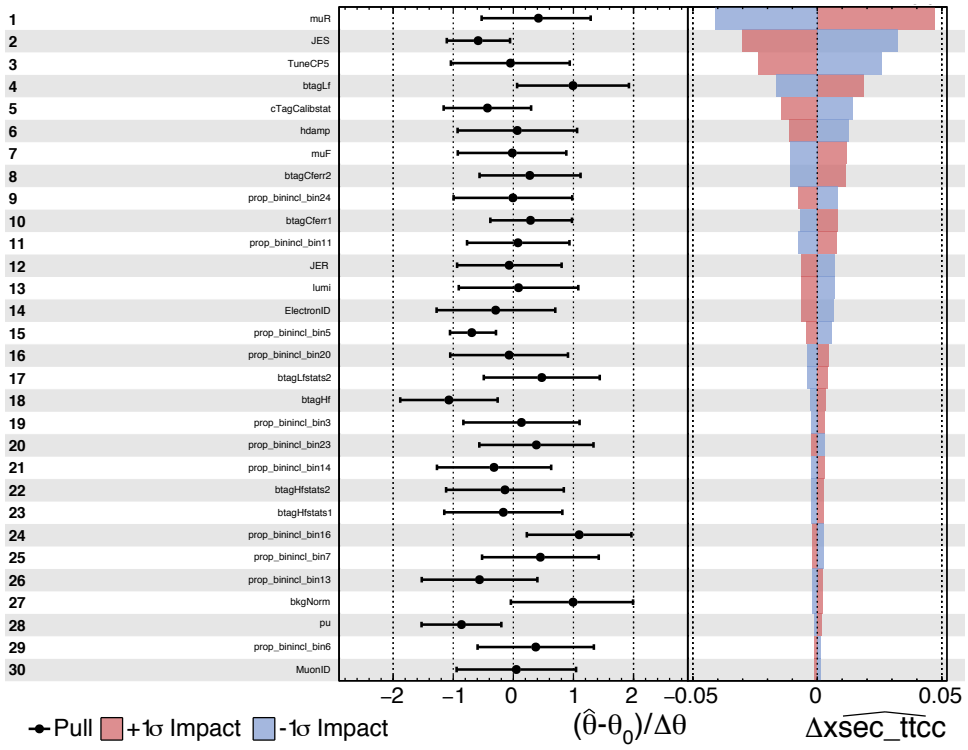


Figure A.1: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $\sigma_{t\bar{t}c\bar{c}}$ in the visible phase space.

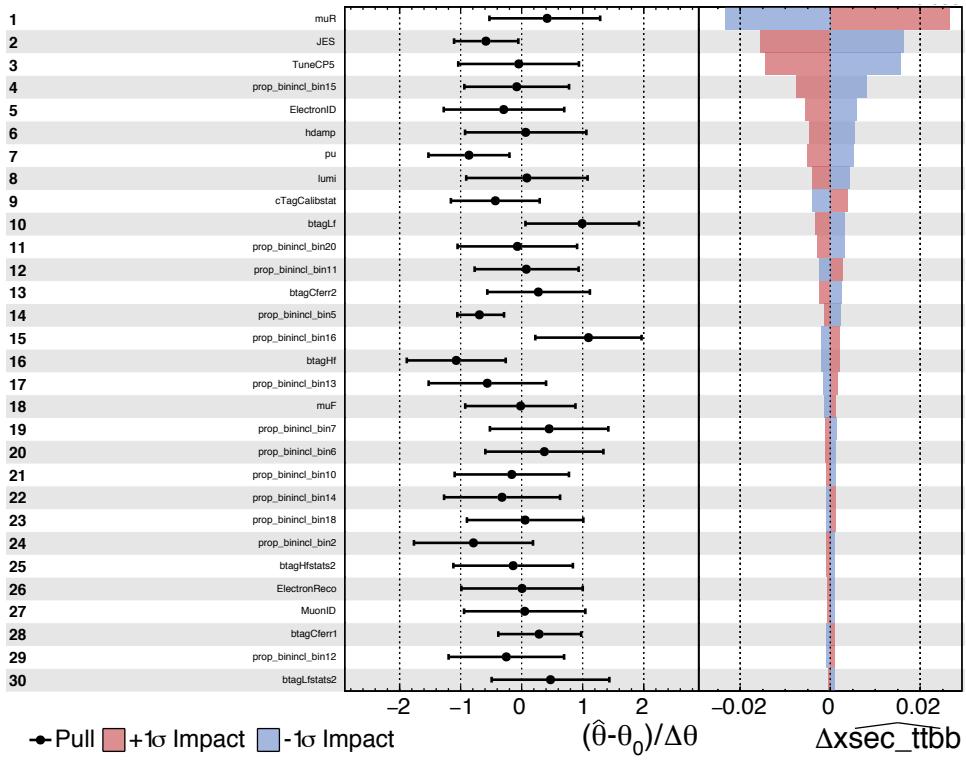


FIGURE A.2: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $\sigma_{t\bar{t}b\bar{b}}$ in the visible phase space.

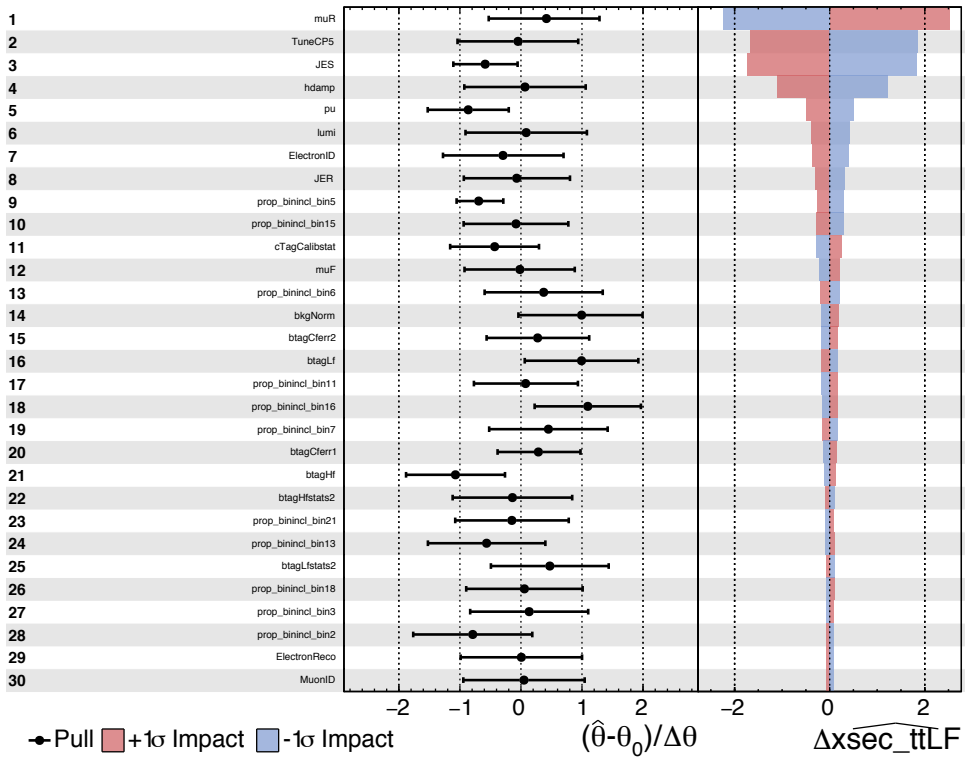


FIGURE A.3: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $\sigma_{t\bar{t}\mathrm{LF}}$ in the visible phase space.

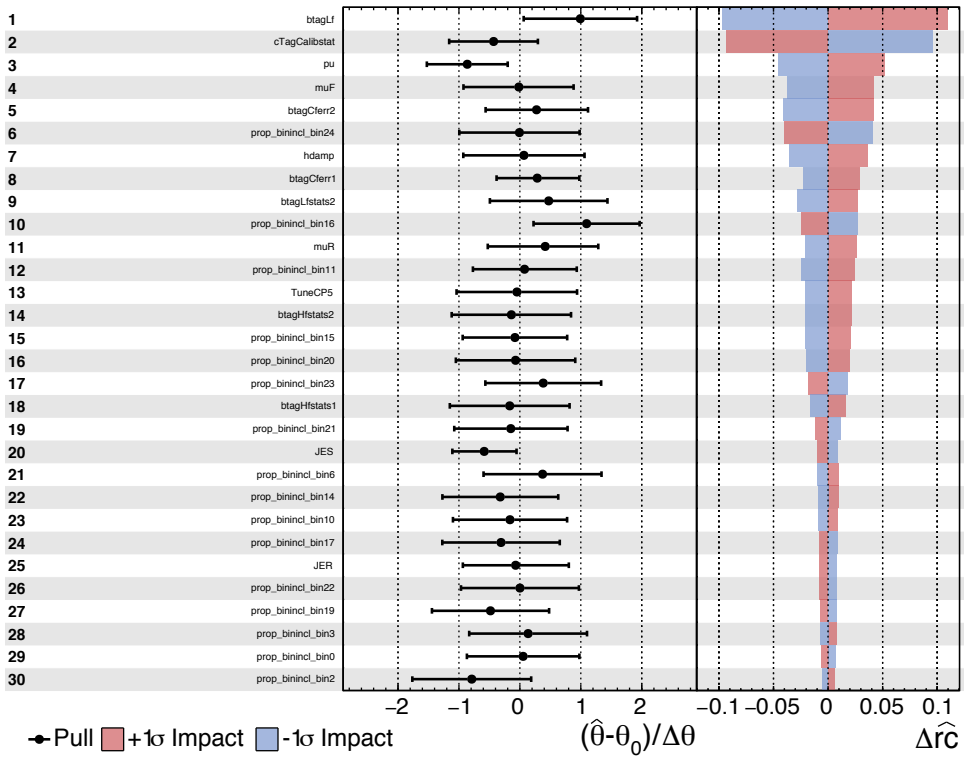


FIGURE A.4: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $R_c$ in the visible phase space.

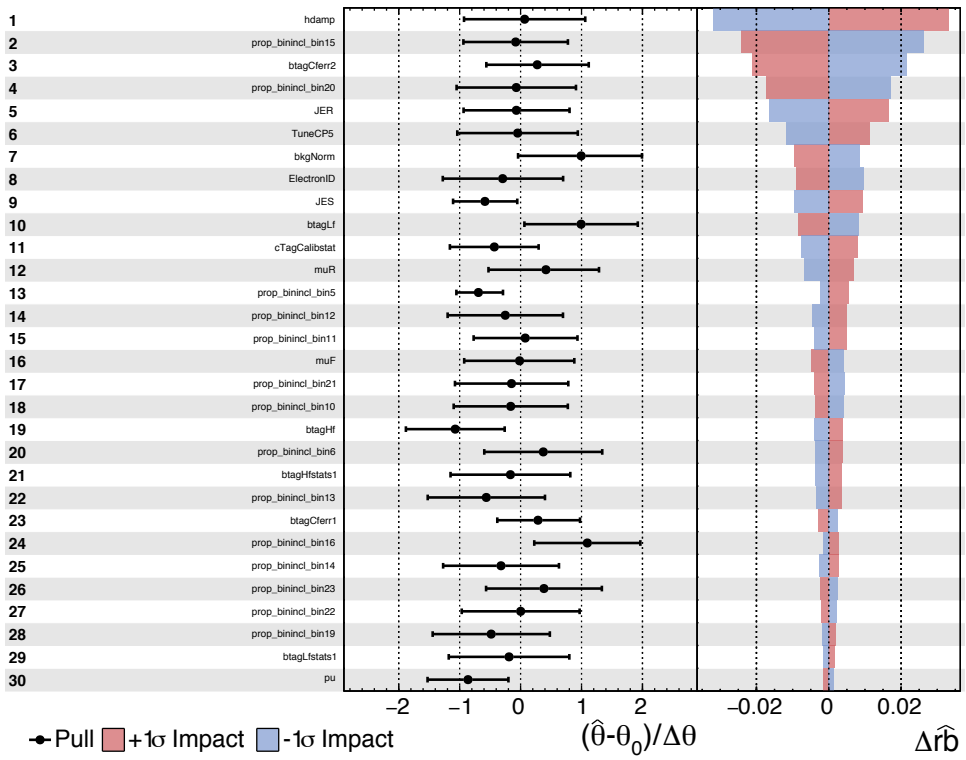


FIGURE A.5: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $R_b$ in the visible phase space.

## A.2 Impacts in the full phase space

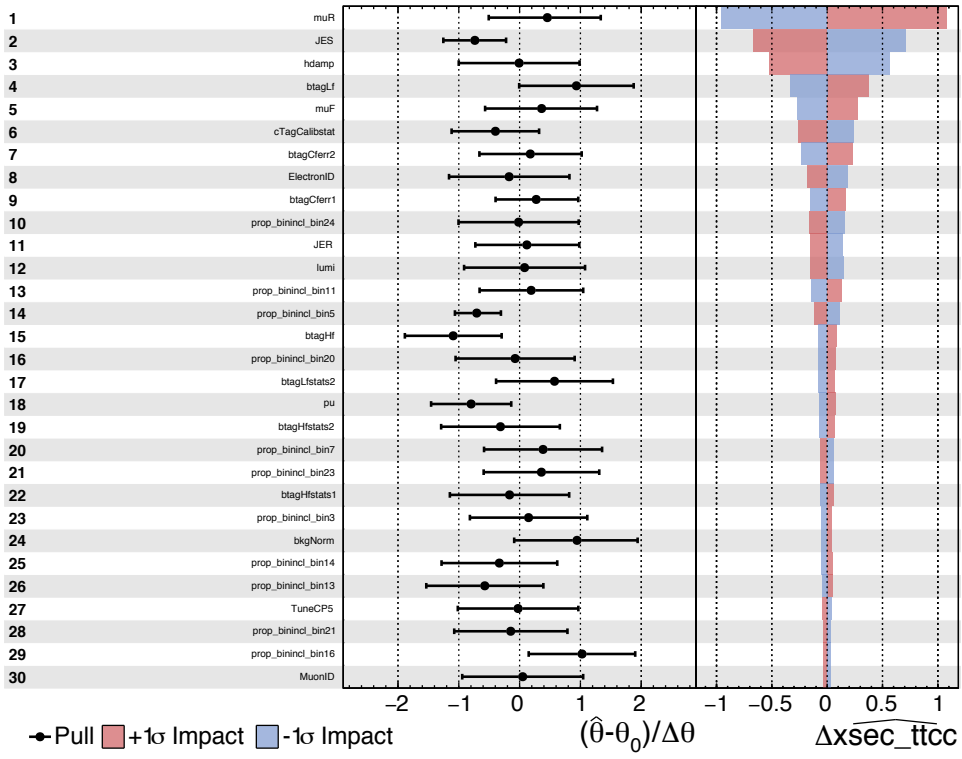


FIGURE A.6: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $\sigma_{t\bar{t}c\bar{c}}$ in the full phase space.

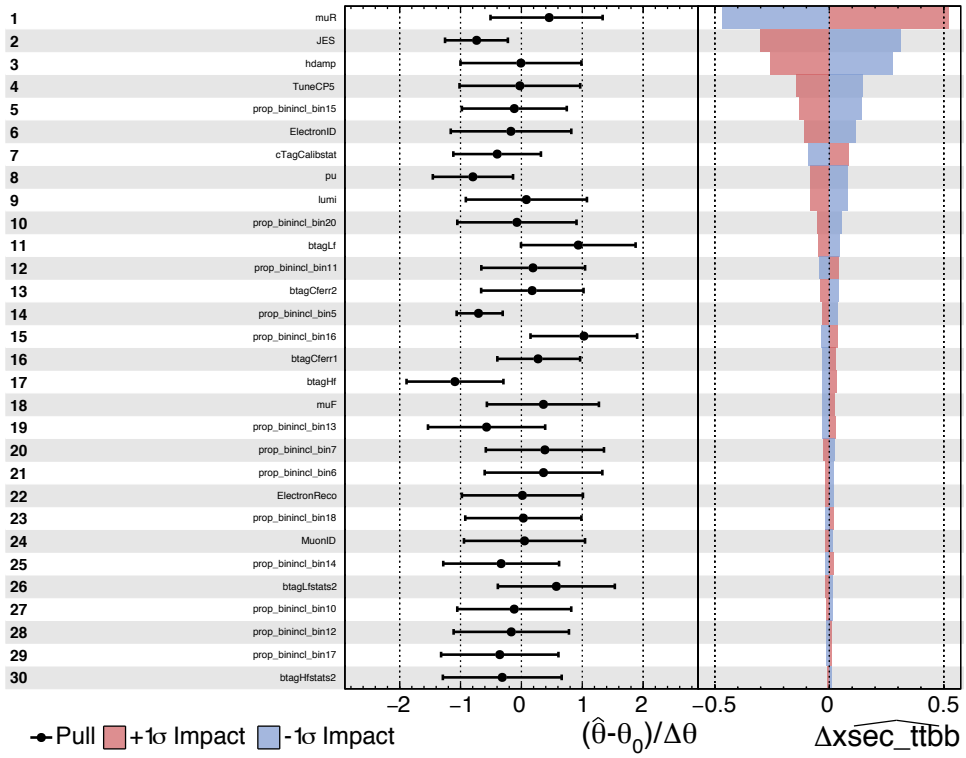


FIGURE A.7: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $\sigma_{t\bar{t}b\bar{b}}$ in the full phase space.

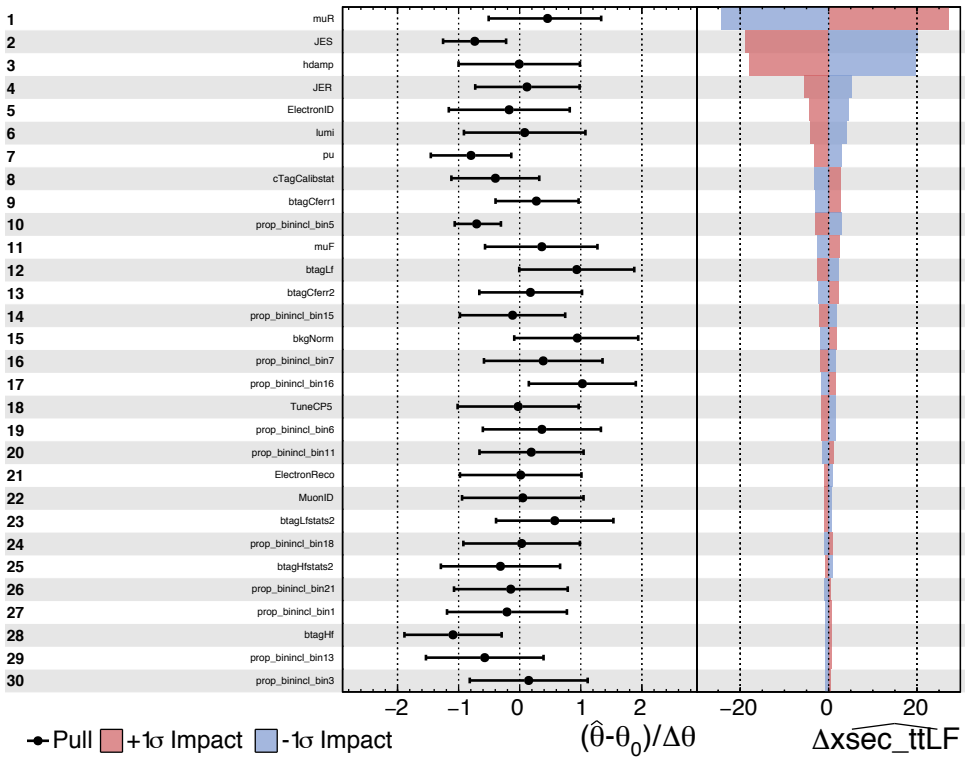


FIGURE A.8: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $\sigma_{t\bar{t}\mathrm{LF}}$ in the full phase space.

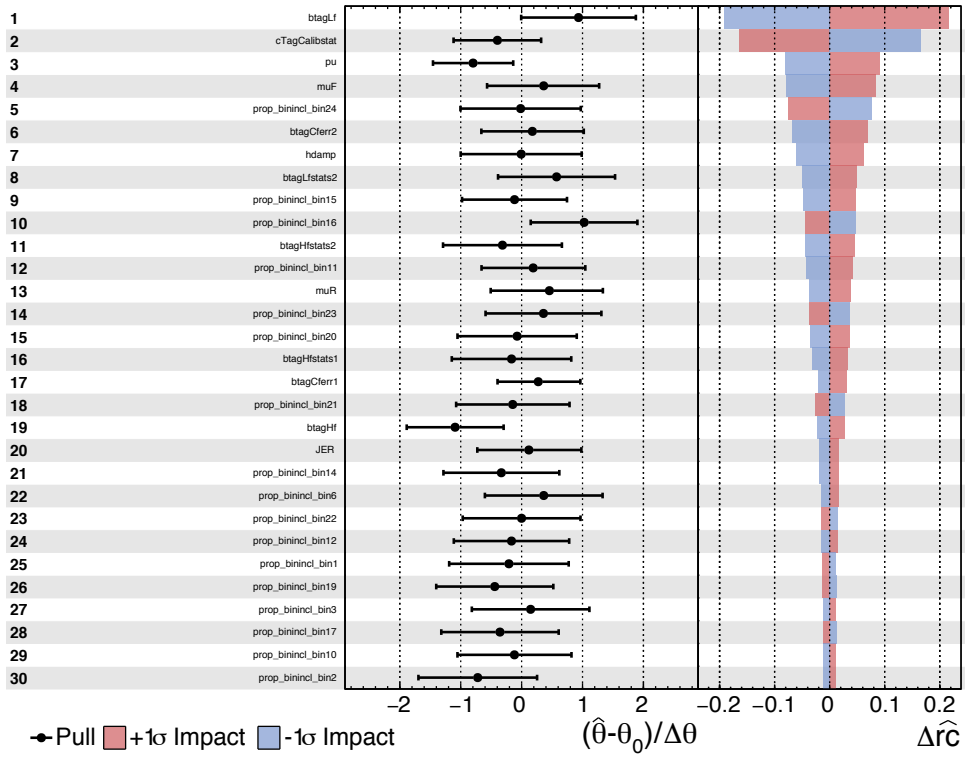


FIGURE A.9: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $R_c$ in the full phase space.

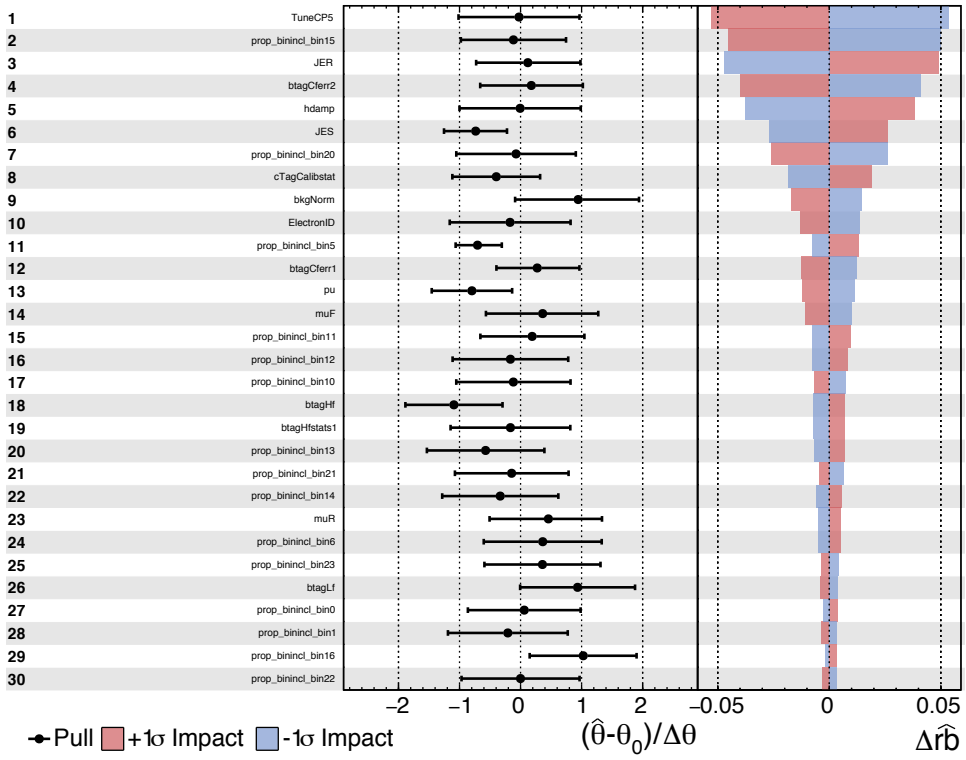


FIGURE A.10: Summary of the pull, constraint and impact of the 30 most important nuisance parameters on the value of $R_b$ in the full phase space.

# Bibliography


[1] CMS Collaboration, **Event displays from the b–tagging and vertexing (BTV) group of CMS, using 2016 data**, (Accessed: January 27th, 2019). https://twiki.cern.ch/twiki/bin/viewauth/CMS/BTVEventDisplaysInternal.

[2] K. R. Popper, **The logic of scientific discovery**. Routledge, 1959.

[3] LIGO Scientific, Virgo Collaboration, B. P. Abbott et al., **Observation of Gravitational Waves from a Binary Black Hole Merger**, *Phys. Rev. Lett.* **116** (2016), no. 6 061102, [arXiv:1602.03837].

[4] CMS Collaboration, S. Chatrchyan et al., **Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC**, *Phys. Lett.* **B716** (2012) 30–61, [arXiv:1207.7235].

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

[6] ATLAS Collaboration, G. Aad et al., **Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC**, *Phys. Lett.* **B716** (2012) 1–29, [arXiv:1207.7214].

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

[8] P. W. Higgs, **Broken Symmetries and the Masses of Gauge Bosons**, *Phys. Rev. Lett.* **13** (1964) 508–509.

[9] CMS Collaboration, A. M. Sirunyan et al., **Observation of $t\bar{t}H$ production**, *Phys. Rev. Lett.* **120** (2018), no. 23 231801, [arXiv:1804.02610].

[10] ATLAS Collaboration, M. Aaboud et al., **Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector**, *Phys. Lett.* **B784** (2018) 173–191, [arXiv:1806.00425].

[11] CMS Collaboration, V. Khachatryan et al., **Measurement of the cross section ratio $\sigma_{t\bar{t}b\bar{b}}/\sigma_{t\bar{t}jj}$ in pp collisions at $\sqrt{s} = 8$ TeV**, *Phys. Lett.* **B746** (2015) 132–153, [arXiv:1411.5621].

[12] CMS Collaboration, A. M. Sirunyan et al., **Measurements of $t\bar{t}$ cross sections in association with b jets and inclusive jets and their ratio using dilepton final states in pp collisions at $\sqrt{s}$ = 13 TeV**, *Phys. Lett.* **B776** (2018) 355–378, [arXiv:1705.10141].

[13] ATLAS Collaboration, G. Aad et al., **Measurements of fiducial cross-sections for $t\bar{t}$ production with one or two additional b-jets in pp collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector**, *Eur. Phys. J.* **C76** (2016), no. 1 11, [arXiv:1508.06868].

[14] ATLAS Collaboration, G. Aad et al., **Study of heavy-flavor quarks produced in association with top-quark pairs at $\sqrt{s}$ = 7 TeV using the ATLAS detector**, *Phys. Rev.* **D89** (2014), no. 7 072012, [arXiv:1304.6386].

[15] ATLAS Collaboration, M. Aaboud et al., **Measurements of fiducial and differential cross-sections of $t\bar{t}$ production with additional heavy-flavour jets in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector**, [arXiv:1811.12113].

[16] C. Bailey, **The greek atomists and epicurus: A study**. Russell & Russell, 1964.

[17] D. Mendelejew, **Über die Beziehungen der Eigenschaften zu den Atomgewichten der Elemente**, *Zeitschrift für Chemie* **12** (1869) 405–406.

[18] E. Rutherford, **The scattering of alpha and beta particles by matter and the structure of the atom**, *Phil. Mag. Ser.6* **21** (1911) 669–688.

[19] J. J. Thomson, **Cathode rays**, *Phil. Mag. Ser.5* **44** (1897) 293–316.

[20] E. D. Bloom et al., **High-Energy Inelastic e p Scattering at 6-Degrees and 10-Degrees**, *Phys. Rev. Lett.* **23** (1969) 930–934.

[21] M. Breidenbach et al., **Observed Behavior of Highly Inelastic Electron-Proton Scattering**, *Phys. Rev. Lett.* **23** (1969) 935–939.

[22] M. Thomson, **Modern particle physics**. Cambridge University Press, New York, 2013.

[23] M. E. Peskin and D. V. Schroeder, **An introduction to quantum field theory**. Addison-Wesley, Reading, USA, 1995.

[24] F. Mandl and G. Shaw, **Quantum field theory**. A Wiley-Interscience publication. J. Wiley, 1993.

[25] M. Krämer, **The Standard Model of Particle Physics**, (Accessed: January 4th, 2019). Lectures given at the 2015 BND summer school: https://indico.cern.ch/event/357886/contributions/849361/attachments/1150733/1651725/mkraemer_BND_2015_SM1.pdf.

[26] F. Reines and C. L. Cowan, **The neutrino**, *Nature* **178** (1956) 446–449.

[27] S. H. Neddermeyer and C. D. Anderson, **Note on the Nature of Cosmic Ray Particles**, *Phys. Rev.* **51** (1937) 884–886.

[28] G. Danby et al., **Observation of High-Energy Neutrino Reactions and the Existence of Two Kinds of Neutrinos**, *Phys. Rev. Lett.* **9** (1962) 36–44.

[29] M. L. Perl et al., **Evidence for Anomalous Lepton Production in $e^+$-$e^-$ Annihilation**, *Phys. Rev. Lett.* **35** (1975) 1489–1492.

[30] DONUT Collaboration, K. Kodama et al., **Observation of tau neutrino interactions**, *Phys. Lett.* **B504** (2001) 218–224, [hep-ex/0012035].

[31] ALEPH Collaboration, D. Décamp et al., **Determination of the number of light neutrino species**, *Phys. Lett. B* **231** (Oct, 1989) 519–529. 20 p.

[32] S. L. Glashow, J. Iliopoulos, and L. Maiani, **Weak Interactions with Lepton-Hadron Symmetry**, *Phys. Rev.* **D2** (1970) 1285–1292.

[33] SLAC-SP-017 Collaboration, J. E. Augustin et al., **Discovery of a Narrow Resonance in $e^+e^-$ Annihilation**, *Phys. Rev. Lett.* **33** (1974) 1406–1408. [Adv. Exp. Phys.5,141(1976)].

[34] J. J. Aubert et al., **Experimental Observation of a Heavy Particle J**, *Phys. Rev. Lett.* **33** (1974) 1404–1406.

[35] S. W. Herb et al., **Observation of a Dimuon Resonance at 9.5 GeV in 400 GeV Proton-Nucleus Collisions**, *Phys. Rev. Lett.* **39** (1977) 252–255.

[36] CDF Collaboration, F. Abe et al., **Observation of top quark production in $\bar{p}p$ collisions**, *Phys. Rev. Lett.* **74** (1995) 2626–2631, [hep-ex/9503002].

[37] D0 Collaboration, S. Abachi et al., **Observation of the top quark**, *Phys. Rev. Lett.* **74** (1995) 2632–2637, [hep-ex/9503003].

[38] CMS Collaboration, A. M. Sirunyan et al., **Search for vector-like T and B quark pairs in final states with leptons at $\sqrt{s} = 13$ TeV**, *JHEP* **08** (2018) 177, [arXiv:1805.04758].

[39] ATLAS Collaboration, M. Aaboud et al., **Combination of the searches for pair-produced vector-like partners of the third-generation quarks at $\sqrt{s} = 13$ TeV with the ATLAS detector**, *Phys. Rev. Lett.* **121** (2018), no. 21 211801, [arXiv:1808.02343].

[40] Particle Data Group Collaboration, M. Tanabashi et al., **Review of particle physics**, *Phys. Rev. D* **98** (Aug, 2018) 030001.

[41] C. D. Anderson, **The Positive Electron**, *Phys. Rev.* **43** (1933) 491–494.

[42] **CERN website: Antimatter**, Accessed: January 9th, 2019. https://home.cern/science/physics/antimatter.

[43] S. M. Carroll, **Spacetime and geometry: An introduction to general relativity**. San Francisco, USA: Addison-Wesley (2004) 513 p, 2004.

[44] E. Noether, **Invariante variationsprobleme**, *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* **1918** (1918) 235–257.

[45] J. A. Aguilar-Saavedra, **Top flavor-changing neutral interactions: Theoretical expectations and experimental detection**, *Acta Phys. Polon.* **B35** (2004) 2695–2710, [hep-ph/0409342].

[46] CMS Collaboration, V. Khachatryan et al., **Search for top quark decays via Higgs-boson-mediated flavor-changing neutral currents in pp collisions at $\sqrt{s} = 8$ TeV**, *JHEP* **02** (2017) 079, [arXiv:1610.04857].

[47] CMS Collaboration, A. M. Sirunyan et al., **Search for the flavor-changing neutral current interactions of the top quark and the Higgs boson which decays into a pair of b quarks at $\sqrt{s} = 13$ TeV**, *JHEP* **06** (2018) 102, [arXiv:1712.02399].

[48] CMS Collaboration, V. Khachatryan et al., **Search for Anomalous Single Top Quark Production in Association with a Photon in pp Collisions at $\sqrt{s} = 8$ TeV**, *JHEP* **04** (2016) 035, [arXiv:1511.03951].

[49] CMS Collaboration, V. Khachatryan et al., **Search for anomalous Wtb couplings and flavour-changing neutral currents in t-channel single top quark production in pp collisions at $\sqrt{s} = 7$ and 8 TeV**, *JHEP* **02** (2017) 028, [arXiv:1610.03545].

[50] CMS Collaboration, A. M. Sirunyan et al., **Search for associated production of a Z boson with a single top quark and for tZ flavour-changing interactions in pp collisions at $\sqrt{s} = 8$ TeV**, *JHEP* **07** (2017) 003, [arXiv:1702.01404].

[51] CMS Collaboration, **Search for flavour changing neutral currents in top quark production and decays with three-lepton final state using the data collected at $\sqrt{s} = 13$ TeV**, Tech. Rep. CMS-PAS-TOP-17-017, CERN, Geneva, 2017.

[52] ATLAS Collaboration, M. Aaboud et al., **Search for top-quark decays $t \to Hq$ with 36 fb$^{-1}$ of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector**, *Submitted to: JHEP* (2018) [arXiv:1812.11568].

[53] ATLAS Collaboration, M. Aaboud et al., **Search for flavour-changing neutral current top-quark decays $t \to qZ$ in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector**, *JHEP* **07** (2018) 176, [arXiv:1803.09923].

[54] ATLAS Collaboration, G. Aad et al., **A search for flavour changing neutral currents in top-quark decays in pp collision data collected with the ATLAS detector at $\sqrt{s} = 7$ TeV**, *JHEP* **09** (2012) 139, [arXiv:1206.0257].

[55] ATLAS Collaboration, G. Aad et al., **Search for flavour-changing neutral current top quark decays $t \to Hq$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector**, *JHEP* **12** (2015) 061, [arXiv:1509.06047].

[56] ATLAS Collaboration, G. Aad et al., **Search for single top-quark production via flavour-changing neutral currents at 8 TeV with the ATLAS detector**, *Eur. Phys. J.* **C76** (2016), no. 2 55, [arXiv:1509.00294].

[57] ATLAS, CMS Collaboration, M. Aaboud et al., **Combinations of single-top-quark production cross-section measurements and $|f_{\mathrm{LV}}V_{tb}|$ determinations at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS experiments**, [arXiv:1902.07158].

[58] Z. Maki, M. Nakagawa, and S. Sakata, **Remarks on the unified model of elementary particles**, *Prog. Theor. Phys.* **28** (1962) 870–880.

[59] CMS Collaboration, V. Khachatryan et al., **Measurement of the ratio $\mathcal{B}(\mathrm{t} \to \mathrm{Wb})/\mathcal{B}(\mathrm{t} \to \mathrm{Wq})$ in pp collisions at $\sqrt{s} = 8$ TeV**, *Phys. Lett.* **B736** (2014) 33–57, [arXiv:1404.2292].

[60] J. Haller, A. Hoecker, R. Kogler, K. Mönig, T. Peiffer, and J. Stelzer, **Update of the global electroweak fit and constraints on two-Higgs-doublet models**, *Eur. Phys. J.* **C78** (2018), no. 8 675, [arXiv:1803.01853].

[61] CMS Collaboration, V. Khachatryan et al., **Measurement of the top quark mass using proton-proton data at $\sqrt{s} = 7$ and 8 TeV**, *Phys. Rev.* **D93** (2016), no. 7 072004, [arXiv:1509.04044].

[62] ATLAS Collaboration, M. Aaboud et al., **Measurement of the top quark mass in the $\mathrm{t\bar{t}} \to$ lepton+jets channel from $\sqrt{s} = 8$ TeV ATLAS data and combination with previous results**, *Submitted to: Eur. Phys. J.* (2018) [arXiv:1810.01772].

[63] A. H. Hoang, **The Top Mass: Interpretation and Theoretical Uncertainties**, in *Proceedings, 7th International Workshop on Top Quark Physics (TOP2014): Cannes, France, September 28-October 3, 2014*, 2014. [arXiv:1412.3649].

[64] LHC Top Working Group, **LHCTopWG Summary Plots**, (Accessed: January 18th, 2019). https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCTopWGSummaryPlots.

[65] M. Cacciari et al., **Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation**, *Phys. Lett.* **B710** (2012) 612–622, [arXiv:1111.5869].

[66] LHC Top Working Group, **ATLAS-CMS recommended predictions for single-top cross sections using the Hathor v2.1 program**, (Accessed: January 19th, 2019). https://twiki.cern.ch/twiki/bin/view/LHCPhysics/SingleTopRefXsec.

[67] P. Kant et al., **HatHor for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions**, *Comput. Phys. Commun.* **191** (2015) 74–89, [arXiv:1406.4403].

[68] M. Aliev et al., **HATHOR: HAdronic Top and Heavy quarks crOss section calculatoR**, *Comput. Phys. Commun.* **182** (2011) 1034–1046, [arXiv:1007.1327].

[69] N. Kidonakis, **Two-loop soft anomalous dimensions for single top quark associated production with a W- or H-**, *Phys. Rev.* **D82** (2010) 054018, [arXiv:1005.4451].

[70] N. Kidonakis, **Top Quark Production**, in *Proceedings, Helmholtz International Summer School on Physics of Heavy Quarks and Hadrons (HQ 2013): JINR, Dubna, Russia, July 15-28, 2013*, pp. 139–168, 2014. [arXiv:1311.0283].

[71] G. Bevilacqua and M. Worek, **Constraining BSM Physics at the LHC: Four top final states with NLO accuracy in perturbative QCD**, *JHEP* **07** (2012) 111, [arXiv:1206.3064].

[72] CMS Collaboration, A. M. Sirunyan et al., **Search for standard model production of four top quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV**, *Phys. Lett.* **B772** (2017) 336–358, [arXiv:1702.06164].

[73] CMS Collaboration, A. M. Sirunyan et al., **Search for standard model production of four top quarks with same-sign and multilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV**, *Eur. Phys. J.* **C78** (2018), no. 2 140, [arXiv:1710.10614].

[74] ATLAS Collaboration, M. Aaboud et al., **Search for four-top-quark production in the single-lepton and opposite-sign dilepton final states in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector**, *Submitted to: Phys. Rev.* (2018) [arXiv:1811.02305].

[75] CMS Collaboration, A. M. Sirunyan et al., **Combined measurements of Higgs boson couplings in proton-proton collisions at $\sqrt{s} = 13$ TeV**, *Submitted to: Eur. Phys. J.* (2018) [arXiv:1809.10733].

[76] ATLAS Collaboration, G. Aad et al., **Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment**, *Eur. Phys. J.* **C76** (2016), no. 1 6, [arXiv:1507.04548].

[77] ATLAS Collaboration, **Combined measurements of Higgs boson production and decay using up to 80 fb$^{-1}$ of proton–proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment**, Tech. Rep. ATLAS-CONF-2018-031, CERN, Geneva, Jul, 2018.

[78] CMS Collaboration, A. M. Sirunyan et al., **Observation of Higgs boson decay to bottom quarks**, *Phys. Rev. Lett.* **121** (2018), no. 12 121801, [arXiv:1808.08242].

[79] ATLAS Collaboration, M. Aaboud et al., **Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector**, *Phys. Lett.* **B786** (2018) 59–86, [arXiv:1808.08238].

[80] CMS Collaboration, A. M. Sirunyan et al., **Search for $t\bar{t}H$ production in the $H \to b\bar{b}$ decay channel with leptonic $t\bar{t}$ decays in proton-proton collisions at $\sqrt{s} = 13$ TeV**, *JHEP* **03** (2019) 026, [arXiv:1804.03682].

[81] ATLAS Collaboration, M. Aaboud et al., **Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector**, *Phys. Rev.* **D97** (2018), no. 7 072016, [arXiv:1712.08895].

[82] Planck Collaboration, N. Aghanim et al., **Planck 2018 results. VI. Cosmological parameters**, [arXiv:1807.06209].

[83] C. Grupen, **Astroparticle physics**. Springer, 2005.

[84] J. D'Hondt, A. Mariotti, K. Mawatari, S. Moortgat, P. Tziveloglou, and G. Van Onsem, **Signatures of top flavour-changing dark matter**, *JHEP* **03** (2016) 060, [arXiv:1511.07463].

[85] A. D. Sakharov, **Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe**, *Pisma Zh. Eksp. Teor. Fiz.* **5** (1967) 32–35.

[86] J. H. Christenson, J. W. Cronin, V. L. Fitch, and R. Turlay, **Evidence for the $2\pi$ Decay of the $K_2^0$ Meson**, *Phys. Rev. Lett.* **13** (1964) 138–140.

[87] BaBar Collaboration, B. Aubert et al., **Measurement of CP violating asymmetries in $B^0$ decays to CP eigenstates**, *Phys. Rev. Lett.* **86** (2001) 2515–2522, [hep-ex/0102030].

[88] Belle Collaboration, K. Abe et al., **Observation of large CP violation in the neutral B meson system**, *Phys. Rev. Lett.* **87** (2001) 091802, [hep-ex/0107061].

[89] LHCb Collaboration, **Observation of CP violation in charm decays**, Tech. Rep. CERN-EP-2019-042. LHCB-PAPER-2019-006, CERN, Geneva, Mar, 2019.

[90] F. Capozzi et al., **Global constraints on absolute neutrino masses and their ordering**, *Phys. Rev.* **D95** (2017), no. 9 096014, [arXiv:1703.04471].

[91] R. Davis, Jr., D. S. Harmer, and K. C. Hoffman, **Search for neutrinos from the sun**, *Phys. Rev. Lett.* **20** (1968) 1205–1209.

[92] SNO Collaboration, Q. R. Ahmad et al., **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, [nucl-ex/0106015].

[93] W. Rodejohann, **Neutrino-less Double Beta Decay and Particle Physics**, *Int. J. Mod. Phys.* **E20** (2011) 1833–1930, [arXiv:1106.1334].

[94] R. M. Wald, **Quantum Field Theory in Curved Space-Time and Black Hole Thermodynamics**. Chicago Lectures in Physics. University of Chicago Press, Chicago, IL, 1995.

[95] P. Langacker, **The Standard Model and beyond; 2nd edition**. High energy physics, cosmology and gravitation. CRC Press, Jul, 2017.

[96] ATLAS, CMS Collaboration, G. Aad et al., **Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments**, *Phys. Rev. Lett.* **114** (2015) 191803, [arXiv:1503.07589].

[97] E. Fermi, **Tentativo di una teoria dell'emissione dei raggi beta**, *Ric. Sci.* **4** (1933) 491–495.

[98] F. L. Wilson, **Fermi's Theory of Beta Decay**, *Am. J. Phys.* **36** (1968), no. 12 1150–1160.

[99] C. N. Leung, S. T. Love, and S. Rao, **Low-Energy Manifestations of a New Interaction Scale: Operator Analysis**, *Z. Phys.* **C31** (1986) 433.

[100] W. Buchmuller and D. Wyler, **Effective Lagrangian Analysis of New Interactions and Flavor Conservation**, *Nucl. Phys.* **B268** (1986) 621–653.

[101] B. Grzadkowski et al., **Dimension-Six Terms in the Standard Model Lagrangian**, *JHEP* **10** (2010) 085, [arXiv:1008.4884].

[102] S. Weinberg, **Baryon and Lepton Nonconserving Processes**, *Phys. Rev. Lett.* **43** (1979) 1566–1570.

[103] D. Barducci et al., **Interpreting top-quark LHC measurements in the standard-model effective field theory**, [arXiv:1802.07237].

[104] G. D' Ambrosio et al., **Minimal flavor violation: An effective field theory approach**, *Nucl. Phys.* **B645** (2002) 155–187, [hep-ph/0207036].

[105] N. P. Hartland et al., **A Monte Carlo global analysis of the Standard Model Effective Field Theory: the top quark sector**, [arXiv:1901.05965].

[106] A. Buckley et al., **Constraining top quark effective theory in the LHC Run II era**, *JHEP* **04** (2016) 015, [arXiv:1512.03360].

[107] CMS Collaboration, **Search for standard model production of four top quarks with the single-lepton and opposite-sign dilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV**, Tech. Rep. CMS-PAS-TOP-17-019, CERN, Geneva, 2019.

[108] C. Zhang, **Constraining qqtt operators from four-top production: a case for enhanced EFT sensitivity**, *Chin. Phys.* **C42** (2018), no. 2 023104, [arXiv:1708.05928].

[109] J. D'Hondt, A. Mariotti, K. Mimasu, S. Moortgat, and C. Zhang, **Learning to pinpoint effective operators at the LHC: a study of the $t\bar{t}b\bar{b}$ signature**, *JHEP* **11** (2018) 131, [arXiv:1807.02130].

[110] **LEP design report**. CERN, Geneva, 1984. Copies shelved as reports in LEP, PS and SPS libraries.

[111] S. Myers, **The LEP Collider, from design to approval and commissioning**. John Adams' Lecture. CERN, Geneva, 1991. Delivered at CERN, 26 Nov 1990.

[112] LEP Working Group for Higgs boson searches, ALEPH, DELPHI, L3, OPAL Collaboration, R. Barate et al., **Search for the Standard Model Higgs boson at LEP**, *Phys. Lett.* **B565** (2003) 61–75, [hep-ex/0306033].

[113] **Design Report Tevatron 1 project**, Tech. Rep. FERMILAB-DESIGN-1984-01, 1984.

[114] CDF, D0 Collaboration, The T. E. V. N. P. H. Working Group, **Combined CDF and D0 Upper Limits on Standard Model Higgs-Boson Production with up to 6.7 fb$^{-1}$ of Data**, in *Proceedings, 35th International Conference on High energy physics (ICHEP 2010): Paris, France, July 22-28, 2010*, 2010. [arXiv:1007.4587].

[115] E. Mobs, **The CERN accelerator complex. Complexe des accéérateurs du CERN**, Jul, 2016.

[116] LHCb Collaboration, A. A. Alves, Jr. et al., **The LHCb Detector at the LHC**, *JINST* **3** (2008) S08005.

[117] ALICE Collaboration, K. Aamodt et al., **The ALICE experiment at the CERN LHC**, *JINST* **3** (2008) S08002.

[118] ATLAS Collaboration, G. Aad et al., **The ATLAS Experiment at the CERN Large Hadron Collider**, *JINST* **3** (2008) S08003.

[119] CMS Collaboration, S. Chatrchyan et al., **The CMS Experiment at the CERN LHC**, *JINST* **3** (2008) S08004.

[120] TOTEM Collaboration, G. Anelli et al., **The TOTEM experiment at the CERN Large Hadron Collider**, *JINST* **3** (2008) S08007.

[121] LHCf Collaboration, O. Adriani et al., **The LHCf detector at the CERN Large Hadron Collider**, *JINST* **3** (2008) S08006.

[122] MoEDAL Collaboration, B. Acharya et al., **The Physics Programme Of The MoEDAL Experiment At The LHC**, *Int. J. Mod. Phys.* **A29** (2014) 1430050, [arXiv:1405.7662].

[123] LHC Study Group, T. S. Pettersson, and P. Lefèvre, **The Large Hadron Collider: conceptual design**, Tech. Rep. CERN-AC-95-05-LHC, Oct, 1995.

[124] O. S. Brüning et al., **LHC Design Report**. CERN Yellow Reports: Monographs. CERN, Geneva, 2004.

[125] L. Evans and P. Bryant, **LHC Machine**, *JINST* **3** (2008) S08001.

[126] J.-L. Caron, **Cross section of LHC dipole. Dipole LHC: coupe transversale.**, AC Collection. Legacy of AC. Pictures from 1992 to 2002., May, 1998.

[127] G. Apollinari et al., **High-Luminosity Large Hadron Collider (HL-LHC): Preliminary Design Report**. CERN Yellow Reports: Monographs. CERN, Geneva, 2015.

[128] CMS Collaboration, **The Phase-2 Upgrade of the CMS Tracker**, Tech. Rep. CERN-LHCC-2017-009. CMS-TDR-014, CERN, Geneva, Jun, 2017.

[129] CMS Collaboration, **CMS luminosity results 2017 proton–proton collisions at 13 TeV**, (Accessed: January 24th, 2019). https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[130] CMS Collaboration, **CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV**, Tech. Rep. CMS-PAS-LUM-17-004, CERN, Geneva, 2018.

[131] CMS Collaboration, G. L. Bayatian et al., **CMS Physics: Technical Design Report Volume 1: Detector Performance and Software**. Technical Design Report CMS. CERN, Geneva, 2006.

[132] CMS Collaboration, S. Chatrchyan et al., **Description and performance of track and primary-vertex reconstruction with the CMS tracker**, *JINST* **9** (2014), no. 10 P10009, [arXiv:1405.6569].

[133] A. Dominguez et al., **CMS Technical Design Report for the Pixel Detector Upgrade**, Tech. Rep. CERN-LHCC-2012-016. CMS-TDR-11, Sep, 2012.

[134] CMS Collaboration, **The CMS electromagnetic calorimeter project: Technical Design Report**. Technical Design Report CMS. CERN, Geneva, 1997.

[135] CMS Collaboration, V. Khachatryan et al., **Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV**, *JINST* **10** (2015) P06005, [arXiv:1502.02701].

[136] CMS Collaboration, V. Khachatryan et al., **Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV**, *JINST* **10** (2015) P08010, [arXiv:1502.02702].

[137] CMS Collaboration, **The CMS hadron calorimeter project: Technical Design Report**. Technical Design Report CMS. CERN, Geneva, 1997.

[138] CMS Collaboration, S. Chatrchyan et al., **Performance of CMS Hadron Calorimeter Timing and Synchronization using Test Beam, Cosmic Ray, and LHC Beam Data**, *JINST* **5** (2010) T03013, [arXiv:0911.4877].

[139] CMS Collaboration, **The CMS muon project: Technical Design Report**. Technical Design Report CMS. CERN, Geneva, 1997.

[140] C. Grupen and B. Schwartz, **Particle detectors**, vol. 26. Cambridge Univ. Pr., Cambridge, UK, 2008.

[141] CMS Collaboration, S. Chatrchyan et al., **Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV**, *JINST* **7** (2012) P10002, [arXiv:1206.4071].

[142] CMS Collaboration, S. Cittolin, A. Rácz, and P. Sphicas, **CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. CMS trigger and data-acquisition project**. Technical Design Report CMS. CERN, Geneva, 2002.

[143] CMS Collaboration, V. Khachatryan et al., **The CMS trigger system**, *JINST* **12** (2017), no. 01 P01020, [arXiv:1609.02366].

[144] CMS Collaboration, **The Phase-2 Upgrade of the CMS L1 Trigger Interim Technical Design Report**, Tech. Rep. CERN-LHCC-2017-013. CMS-TDR-017, CERN, Geneva, Sep, 2017.

[145] I. Bird et al., **Update of the Computing Models of the WLCG and the LHC Experiments**, Tech. Rep. CERN-LHCC-2014-014. LCG-TDR-002, Apr, 2014.

[146] CMS Collaboration, A. M. Sirunyan et al., **Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV**, *JHEP* **07** (2018) 161, [arXiv:1802.02613].

[147] E. Gladysz-Dziadus et al., **CASTOR: Centauro and strange object research in nucleus-nucleus collisions at LHC**, in *21st International Workshop on Nuclear Theory (IWNT 2002) Rila Mountains, Bulgaria, June 10-15, 2002*, 2002. [hep-ex/0209008].

[148] W. K. Hastings, **Monte Carlo Sampling Methods Using Markov Chains and Their Applications**, *Biometrika* **57** (1970) 97–109.

[149] B. R. Webber, **Monte Carlo Simulation of Hard Hadronic Processes**, *Ann. Rev. Nucl. Part. Sci.* **36** (1986) 253–286.

[150] A. Buckley et al., **General-purpose event generators for LHC physics**, *Phys. Rept.* **504** (2011) 145–233, [arXiv:1101.2599].

[151] M. A. Dobbs et al., **Les Houches guidebook to Monte Carlo generators for hadron collider physics**, in *Physics at TeV colliders. Proceedings, Workshop, Les Houches, France, May 26-June 3, 2003*, pp. 411–459, 2004. [hep-ph/0403045].

[152] K. Melnikov, **Lectures on QCD for hadron colliders**, *CERN Yellow Reports: School Proceedings* **3** (2018), no. 0 p. 37.

[153] S. Höche, **Introduction to parton-shower event generators**, in *Proceedings, Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders (TASI 2014): Boulder, Colorado, June 2-27, 2014*, pp. 235–295, 2015. [arXiv:1411.4085].

[154] NNPDF Collaboration, R. D. Ball et al., **Parton distributions from high-precision collider data**, *Eur. Phys. J.* **C77** (2017), no. 10 663, [arXiv:1706.00428].

[155] A. D. Martin et al., **Parton distributions for the LHC**, *Eur. Phys. J.* **C63** (2009) 189–285, [arXiv:0901.0002].

[156] S. Dulat et al., **New parton distribution functions from a global analysis of quantum chromodynamics**, *Phys. Rev.* **D93** (2016), no. 3 033006, [arXiv:1506.07443].

[157] V. N. Gribov and L. N. Lipatov, **Deep inelastic e p scattering in perturbation theory**, *Sov. J. Nucl. Phys.* **15** (1972) 438–450. [Yad. Fiz.15,781(1972)].

[158] G. Altarelli and G. Parisi, **Asymptotic Freedom in Parton Language**, *Nucl. Phys.* **B126** (1977) 298–318.

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

[160] V. Bertone, S. Carrazza, and J. Rojo, **APFEL: A PDF Evolution Library with QED corrections**, *Comput. Phys. Commun.* **185** (2014) 1647–1668, [arXiv:1310.1394].

[161] S. Carrazza et al., **APFEL Web**, *J. Phys.* **G42** (2015), no. 5 057001, [arXiv:1410.5456].

[162] J. C. Collins, D. E. Soper, and G. F. Sterman, **Factorization of Hard Processes in QCD**, *Adv. Ser. Direct. High Energy Phys.* **5** (1989) 1–91, [hep-ph/0409313].

[163] S. Frixione, P. Nason, and C. Oleari, **Matching NLO QCD computations with Parton Shower simulations: the POWHEG method**, *JHEP* **11** (2007) 070, [arXiv:0709.2092].

[164] S. Alioli et al., **A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX**, *JHEP* **06** (2010) 043, [arXiv:1002.2581].

[165] J. M. Campbell et al., **Top-Pair Production and Decay at NLO Matched with Parton Showers**, *JHEP* **04** (2015) 114, [arXiv:1412.1828].

[166] J. Alwall et al., **The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations**, *JHEP* **07** (2014) 079, [arXiv:1405.0301].

[167] T. Sjostrand, S. Mrenna, and P. Z. Skands, **A Brief Introduction to PYTHIA 8.1**, *Comput. Phys. Commun.* **178** (2008) 852–867, [arXiv:0710.3820].

[168] A. Alloul et al., **FeynRules 2.0 - A complete toolbox for tree-level phenomenology**, *Comput. Phys. Commun.* **185** (2014) 2250–2300, [arXiv:1310.1921].

[169] C. Degrande et al., **UFO - The Universal FeynRules Output**, *Comput. Phys. Commun.* **183** (2012) 1201–1214, [arXiv:1108.2040].

[170] A. Bredenstein et al., **NLO QCD corrections to $t\bar{t}b\bar{b}$ production at the LHC: 1. Quark-antiquark annihilation**, *JHEP* **08** (2008) 108, [arXiv:0807.1248].

[171] A. Bredenstein et al., **NLO QCD corrections to pp $\to t\bar{t}b\bar{b} + X$ at the LHC**, *Phys. Rev. Lett.* **103** (2009) 012002, [arXiv:0905.0110].

[172] T. Ježo et al., **New NLOPS predictions for $t\bar{t} + b$ jet production at the LHC**, *Eur. Phys. J.* **C78** (2018), no. 6 502, [arXiv:1802.00426].

[173] G. Bevilacqua and M. Worek, **On the ratio of $t\bar{t}b\bar{b}$ and $t\bar{t}jj$ cross sections at the CERN Large Hadron Collider**, *JHEP* **07** (2014) 135, [arXiv:1403.2046].

[174] J. L. Kneur and A. Neveu, **Lambda$^{\mathbf{QCD}}_{\mathbf{MS}}$ from Renormalization Group Optimized Perturbation**, *Phys. Rev.* **D85** (2012) 014005, [arXiv:1108.3501].

[175] M. Bahr et al., **Herwig++ Physics and Manual**, *Eur. Phys. J.* **C58** (2008) 639–707, [arXiv:0803.0883].

[176] S. Hoeche et al., **Matching parton showers and matrix elements**, in *HERA and the LHC: A Workshop on the implications of HERA for LHC physics: Proceedings Part A*, pp. 288–289, 2005. [hep-ph/0602031].

[177] R. Frederix and S. Frixione, **Merging meets matching in MC@NLO**, *JHEP* **12** (2012) 061, [arXiv:1209.6215].

[178] X. Artru and G. Mennessier, **String model and multiproduction**, *Nucl. Phys.* **B70** (1974) 93–115.

[179] B. Andersson et al., **Parton Fragmentation and String Dynamics**, *Phys. Rept.* **97** (1983) 31–145.

[180] B. Andersson, **The lund model**. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 1998.

[181] R. D. Field and S. Wolfram, **A QCD Model for e+ e- Annihilation**, *Nucl. Phys.* **B213** (1983) 65–84.

[182] D. Amati and G. Veneziano, **Preconfinement as a Property of Perturbative QCD**, *Phys. Lett.* **83B** (1979) 87–92.

[183] D. J. Lange, **The EvtGen particle decay simulation package**, *Nucl. Instrum. Meth.* **A462** (2001) 152–155.

[184] R. K. Ellis, W. J. Stirling, and B. R. Webber, **QCD and collider physics**, *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.* **8** (1996) 1–435.

[185] J. D. Bjorken, **Properties of Hadron Distributions in Reactions Containing Very Heavy Quarks**, *Phys. Rev.* **D17** (1978) 171–173.

[186] C. Peterson et al., **Scaling Violations in Inclusive e+ e- Annihilation Spectra**, *Phys. Rev.* **D27** (1983) 105.

[187] V. G. Kartvelishvili, A. K. Likhoded, and V. A. Petrov, **On the Fragmentation Functions of Heavy Quarks Into Hadrons**, *Phys. Lett.* **78B** (1978) 615–617.

[188] P. D. B. Collins and T. P. Spiller, **The Fragmentation of Heavy Quarks**, *J. Phys.* **G11** (1985) 1289.

[189] G. Colangelo and P. Nason, **A Theoretical study of the c and b fragmentation function from e+ e- annihilation**, *Phys. Lett.* **B285** (1992) 167–171.

[190] M. G. Bowler, **e+ e- Production of Heavy Quarks in the String Model**, *Z. Phys.* **C11** (1981) 169.

[191] E. Braaten et al., **Perturbative QCD fragmentation functions as a model for heavy quark fragmentation**, *Phys. Rev.* **D51** (1995) 4819–4829, [hep-ph/9409316].

[192] The CMS collaboration, **Charged particle multiplicities in pp interactions at $\sqrt{s} = 0.9$, 2.36, and 7 TeV**, *Journal of High Energy Physics* **2011** (Jan, 2011) 79.

[193] ATLAS Collaboration, G. Aad et al., **Charged-particle multiplicities in pp interactions measured with the ATLAS detector at the LHC**, *New J. Phys.* **13** (2011) 053033, [arXiv:1012.5104].

[194] CMS Collaboration, V. Khachatryan et al., **Event generator tunes obtained from underlying event and multiparton scattering measurements**, *Eur. Phys. J.* **C76** (2016), no. 3 155, [arXiv:1512.00815].

[195] CMS Collaboration, **Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements**, Tech. Rep. CMS-PAS-GEN-17-001, CERN, Geneva.

[196] GEANT4 Collaboration, S. Agostinelli et al., **GEANT4: A Simulation toolkit**, *Nucl. Instrum. Meth.* **A506** (2003) 250–303.

[197] J. Allison et al., **Geant4 developments and applications**, *IEEE Trans. Nucl. Sci.* **53** (2006) 270.

[198] J. Allison et al., **Recent developments in Geant4**, *Nucl. Instrum. Meth.* **A835** (2016) 186–225.

[199] CMS Collaboration, A. M. Sirunyan et al., **Particle-flow reconstruction and global event description with the cms detector**, *JINST* **12** (2017) P10003, [arXiv:1706.04965].

[200] S. Cucciarelli et al., **Track reconstruction, primary vertex finding and seed generation with the Pixel Detector**, Tech. Rep. CMS-NOTE-2006-026, CERN, Geneva, Jan, 2006.

[201] W. Adam et al., **Track Reconstruction in the CMS tracker**, Tech. Rep. CMS-NOTE-2006-041, CERN, Geneva, Dec, 2006.

[202] R. Kalman, **A New Approach To Linear Filtering and Prediction Problems**, *Journal of Basic Engineering (ASME)* **82D** (01, 1960) 35–45.

[203] CMS Collaboration, **CMS Tracking POG Performance Plots For 2017 with Phase I pixel detector**, (Accessed: February 5th, 2019). https://twiki.cern.ch/twiki/bin/view/CMSPublic/TrackingPOGPerformance2017MC.

[204] T. Speer et al., **Vertex Fitting in the CMS Tracker**, Tech. Rep. CMS-NOTE-2006-032, CERN, Geneva, Feb, 2006.

[205] R. Fruhwirth, W. Waltenberger, and P. Vanlaer, **Adaptive vertex fitting**, *J. Phys.* **G34** (2007) N343.

[206] CMS Collaboration, A. M. Sirunyan et al., **Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV**, *JINST* **13** (2018), no. 06 P06015, [arXiv:1804.04528].

[207] W. Adam et al., **Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC**, *Journal of Physics G: Nuclear and Particle Physics* **31** (jul, 2005) N9–N20.

[208] CMS Collaboration, **Electron and Photon performance in CMS with the full 2017 data sample and additional 2016 highlights for the CALOR 2018 Conference**, Tech. Rep. CMS-DP-2018-017, May, 2018.

[209] CMS Collaboration, **Electron Identification Based on Simple Cuts**, (Accessed: February 7th, 2019). https://twiki.cern.ch/twiki/bin/view/CMSPublic/EgammaPublicData#Implementing_the_Simple_Cut_Base.

[210] CMS Collaboration, **Electron Cut Based ID for 94X samples**, (Accessed: February 7th, 2019). https://indico.cern.ch/event/732971/contributions/3022843/attachments/1658685/2656462/eleIdTuning.pdf.

[211] M. Cacciari and G. P. Salam, **Pileup subtraction using jet areas**, *Phys. Lett.* **B659** (2008) 119–126, [arXiv:0707.1378].

[212] M. Cacciari, G. P. Salam, and G. Soyez, **The Catchment Area of Jets**, *JHEP* **04** (2008) 005, [arXiv:0802.1188].

[213] CMS Collaboration, **Electron Cut Based ID for 94X samples**, (Accessed: February 7th, 2019). https://indico.cern.ch/event/697576/contributions/2940576/attachments/1620927/2578913/eleIdTuning.pdf.

[214] CMS Collaboration, **Baseline muon selections for Run-II**, (Accessed: February 7th, 2019). https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideMuonIdRun2#Tight_Muon.

[215] CMS Collaboration, **Muon identification and isolation efficiencies with 2017 and 2018 data**, Tech. Rep. CMS-DP-2018-042, Jul, 2018.

[216] G. P. Salam, **Towards Jetography**, *Eur. Phys. J.* **C67** (2010) 637–686, [arXiv:0906.1833].

[217] CMS Collaboration, **Performance of Jet Algorithms in CMS**, Tech. Rep. CMS-PAS-JME-07-003, CERN, Geneva.

[218] M. Cacciari, G. P. Salam, and G. Soyez, **The anti-$k_t$ jet clustering algorithm**, *JHEP* **04** (2008) 063, [arXiv:0802.1189].

[219] S. D. Ellis and D. E. Soper, **Successive combination jet algorithm for hadron collisions**, *Phys. Rev.* **D48** (1993) 3160–3166, [hep-ph/9305266].

[220] CMS Collaboration, V. Khachatryan et al., **Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV**, *JINST* **12** (2017) P02014, [arXiv:1607.03663].

[221] CMS Collaboration, S. Chatrchyan et al., **Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS**, *JINST* **6** (2011) P11002, [arXiv:1107.4277].

[222] CMS Collaboration, **Jet Energy Resolution Scale Factors Measurement**, (Accessed: February 7th, 2019). https://indico.cern.ch/event/776439/contributions/3228597/attachments/1759162/2854405/JER_MEETING_26_11_2018.pdf.

[223] CMS Collaboration, **Performance of missing transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV using the CMS detector**, CMS Physics Analysis Summary CMS-PAS-JME-17-001, 2018.

[224] CMS Collaboration, **Search for new physics in monojet events in pp collisions at $\sqrt{s} = 8$ TeV**, (Accessed: February 7th, 2019). Event display from CMS PAS EXO-12-048, https://indico.cern.ch/event/776439/contributions/3228597/attachments/1759162/2854405/JER_MEETING_26_11_2018.pdf.

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

[226] ATLAS Collaboration, M. Aaboud et al., **Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s} = 13$ TeV**, *JHEP* **08** (2018) 089, [arXiv:1805.01845].

[227] ATLAS Collaboration, G. Aad et al., **Performance of b-jet Identification in the ATLAS Experiment**, *JINST* **11** (2016), no. 04 P04008, [arXiv:1512.01094].

[228] CMS Collaboration, S. Chatrchyan et al., **Identification of b-quark jets with the CMS experiment**, *JINST* **8** (2013) P04013, [arXiv:1211.4462].

[229] CMS Collaboration, V. Khachatryan et al., **Measurement of $B\bar{B}$ Angular Correlations based on Secondary Vertex Reconstruction at $\sqrt{s} = 7$ TeV**, *JHEP* **03** (2011) 136, [arXiv:1102.3194].

[230] CMS Collaboration, **Performance of b tagging algorithms in proton-proton collisions at 13 TeV with Phase 1 CMS detector**, Tech. Rep. CMS-DP-2018-033, Jun, 2018.

[231] CMS Collaboration, **Performance of Deep Tagging Algorithms for Boosted Double Quark Jet Topology in Proton-Proton Collisions at 13 TeV with the Phase-0 CMS Detector**, Tech. Rep. CMS-DP-2018-046, Jul, 2018.

[232] L. Lyons, D. Gibaut, and P. Clifford, **How to Combine Correlated Estimates of a Single Physical Quantity**, *Nucl. Instrum. Meth.* **A270** (1988) 110.

[233] CMS Collaboration, **Performance of the DeepJet b tagging algorithm using 41.9 fb$^{-1}$ of data from proton-proton collisions at 13 TeV with Phase 1 CMS detector**, Tech. Rep. CMS-DP-2018-058, Nov, 2018.

[234] CMS Collaboration, V. Khachatryan et al., **Search for the associated production of the Higgs boson with a top-quark pair**, *JHEP* **09** (2014) 087, [arXiv:1408.1682].

[235] S. Caron et al., **The BSM-AI project: SUSY-AI generalizing LHC limits on supersymmetry with machine learning**, *Eur. Phys. J.* **C77** (2017), no. 4 257, [arXiv:1605.02797].

[236] J. Brehmer et al., **A Guide to Constraining Effective Field Theories with Machine Learning**, *Phys. Rev.* **D98** (2018), no. 5 052004, [arXiv:1805.00020].

[237] M. Farina, Y. Nakai, and D. Shih, **Searching for New Physics with Deep Autoencoders**, [arXiv:1808.08992].

[238] G. Louppe, M. Kagan, and K. Cranmer, **Learning to Pivot with Adversarial Networks**, [arXiv:1611.01046].

[239] CERN, **TrackML: particle tracking challenge**, (Accessed: February 11th, 2019). https://sites.google.com/site/trackmlparticle/home.

[240] J. Duarte et al., **Fast inference of deep neural networks in FPGAs for particle physics**, *JINST* **13** (2018), no. 07 P07027, [arXiv:1804.06913].

[241] O. Behnke et al., **Data Analysis in High Energy Physics**. Wiley-VCH, Weinheim, Germany, 2013.

[242] I. Narsky and F. C. Porter, **Statistical analysis techniques in particle physics**. Wiley-VCH, Weinheim, Germany, 2014.

[243] F.-F. Li, J. Johnson, and S. Yeung, **Convolutional Neural Networks for Visual Recognition**, (Accessed: February 11th, 2019). Lecures given at Stanford University, Spring 2018, http://cs231n.stanford.edu/syllabus.html.

[244] H. Robbins and S. Monro, **A stochastic approximation method**, *Ann. Math. Statist.* **22** (09, 1951) 400–407.

[245] C. Cortes and V. Vapnik, **Support-vector networks**, *Machine Learning* **20** (Sep, 1995) 273–297.

[246] H. Zhang, **The optimality of naive bayes**, in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)* (V. Barr and Z. Markov, eds.), AAAI Press, 2004.

[247] J. Neyman and E. S. Pearson, **On the Problem of the Most Efficient Tests of Statistical Hypotheses**, pp. 73–108. Springer New York, New York, NY, 1992.

[248] K. Cranmer et al., **Experiments using machine learning to approximate likelihood ratios for mixture models**, *Journal of Physics: Conference Series* **762** (oct, 2016) 012034.

[249] J. L. Bentley, **Multidimensional Binary Search Trees Used for Associative Searching**, *Commun. ACM* **18** (Sept., 1975) 509–517.

[250] L. Breiman, **Random forests**, *Machine Learning* **45** (Oct, 2001) 5–32.

[251] J. H. Friedman, **Stochastic gradient boosting**, *Comput. Stat. Data Anal.* **38** (Feb., 2002) 367–378.

[252] CMS Collaboration, **Identification of c-quark jets at the CMS experiment**, Tech. Rep. CMS-PAS-BTV-16-001, CERN, Geneva, 2016.

[253] M. M. Waldrop, **Complexity: the emerging science at the edge of order and chaos**. Simon & Schuster, New York, 1992.

[254] W. S. McCulloch and W. Pitts, **A logical calculus of the ideas immanent in nervous activity**, *The bulletin of mathematical biophysics* **5** (Dec, 1943) 115–133.

[255] F. Rosenblatt, **The perceptron: A probabilistic model for information storage and organization in the brain**, *Psychological Review* (1958) 65–386.

[256] I. Goodfellow, Y. Bengio, and A. Courville, **Deep learning**. MIT Press, 2016. `http://www.deeplearningbook.org`.

[257] D. Guest et al., **Jet Flavor Classification in High-Energy Physics with Deep Neural Networks**, *Phys. Rev.* **D94** (2016), no. 11 112002, [arXiv:1607.08633].

[258] F. Chollet et al., **Keras**, 2015. `https://github.com/keras-team/keras`.

[259] M. Abadi et al., **TensorFlow: Large-scale machine learning on heterogeneous systems**, 2015. Software available from `https://www.tensorflow.org/`.

[260] CMS Collaboration, **CMS Phase 1 heavy flavour identification performance and developments**, Tech. Rep. CMS-DP-2017-013, May, 2017.

[261] P. Nason, **A New method for combining NLO QCD with shower Monte Carlo algorithms**, *JHEP* **11** (2004) 040, [hep-ph/0409146].

[262] P. Skands, S. Carrazza, and J. Rojo, **Tuning PYTHIA 8.1: the Monash 2013 Tune**, *Eur. Phys. J.* **C74** (2014), no. 8 3024, [arXiv:1404.5630].

[263] NNPDF Collaboration, R. D. Ball et al., **Parton distributions for the LHC Run II**, *JHEP* **04** (2015) 040, [arXiv:1410.8849].

[264] Y. Li and F. Petriello, **Combining QCD and electroweak corrections to dilepton production in FEWZ**, *Phys. Rev.* **D86** (2012) 094034, [arXiv:1208.5967].

[265] S. Frixione et al., **Single-top hadroproduction in association with a W boson**, *JHEP* **07** (2008) 029, [arXiv:0805.3067].

[266] E. Re, **Single-top Wt-channel production matched with parton showers using the POWHEG method**, *Eur. Phys. J.* **C71** (2011) 1547, [arXiv:1009.2450].

[267] N. Kidonakis, **NNLL threshold resummation for top-pair and single-top production**, *Phys. Part. Nucl.* **45** (2014), no. 4 714–722, [arXiv:1210.7813].

[268] J. M. Campbell and R. K. Ellis, **MCFM for the Tevatron and the LHC**, *Nucl. Phys. Proc. Suppl.* **205-206** (2010) 10–15, [arXiv:1007.3492].

[269] T. Gehrmann et al., **$W^+W^-$ Production at Hadron Colliders in Next to Next to Leading Order QCD**, *Phys. Rev. Lett.* **113** (2014), no. 21 212001, [arXiv:1408.5243].

[270] CMS Collaboration, **Performance of the Particle-Flow jet identification criteria using proton-proton collisions at 13 TeV with the 2015 dataset**, Tech. Rep. CMS-AN-15-269, CERN, Geneva, 2015.

[271] CMS Collaboration, **Muon reconstruction, identification, isolation and trigger efficiencies (and data/MC scale factors) for 2017 data**, (Accessed: February 26th, 2019). https://twiki.cern.ch/twiki/bin/view/CMS/MuonReferenceEffs2017.

[272] CMS Collaboration, **Electron Scale Factors Run 2**, (Accessed: February 26th, 2019). https://twiki.cern.ch/twiki/bin/view/CMS/ElectronScaleFactorsRun2.

[273] CMS Collaboration, **Investigations of the impact of the parton shower tuning in Pythia 8 in the modelling of $t\bar{t}$ at $\sqrt{s} = 8$ and 13 TeV**, Tech. Rep. CMS-PAS-TOP-16-021, CERN, Geneva, 2016.

[274] J. S. Conway, **Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra**, in *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN,Geneva, Switzerland 17-20 January 2011*, pp. 115–120, 2011. [arXiv:1103.0354].

[275] CMS Collaboration, A. M. Sirunyan et al., **Searches for W' bosons decaying to a top quark and a bottom quark in proton-proton collisions at 13 TeV**, *JHEP* **08** (2017) 029, [arXiv:1706.04260].

[276] C. Degrande et al., **Non-resonant New Physics in Top Pair Production at Hadron Colliders**, *JHEP* **03** (2011) 125, [arXiv:1010.6304].

[277] J. de Blas, M. Chala, and J. Santiago, **Renormalization Group Constraints on New Top Interactions from Electroweak Precision Data**, *JHEP* **09** (2015) 189, [arXiv:1507.00757].

[278] ATLAS Collaboration, **Search for new phenomena in $t\bar{t}$ final states with additional heavy-flavour jets in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector**, Tech. Rep. ATLAS-CONF-2016-104, CERN, Geneva, Sep, 2016.

[279] L. B. Okun, **Leptons and Quarks**. North-Holland, Amsterdam, Netherlands, 1982. p. 321-325.

[280] E. Alvarez et al., **Four Tops for LHC**, *Nucl. Phys.* **B915** (2017) 19–43, [arXiv:1611.05032].

[281] G. F. Giudice et al., **The Strongly-Interacting Light Higgs**, *JHEP* **06** (2007) 045, [hep-ph/0703164].

[282] DELPHES 3 Collaboration, **DELPHES 3, A modular framework for fast simulation of a generic collider experiment**, *JHEP* **02** (2014) 057, [arXiv:1307.6346].

[283] CMS Collaboration, **Identification of b quark jets at the CMS Experiment in the LHC Run 2**, Tech. Rep. CMS-PAS-BTV-15-001, CERN, Geneva, 2016.

[284] S. Pozzorini, **Theory progress on $t\bar{t}H(b\bar{b})$ background**, 11th International Workshop on Top Quark Physics (TOP2018): Bad Neuenahr, Germany, September 116-21, 2018. https://indico.cern.ch/event/690229/contributions/2979729/attachments/1719226/2774671/pozzorini_top2018.pdf.

[285] G. Heiman, **Basic statistics for the behavioral sciences**. Cengage Learning, 2010.

[286] A. Quadt, **Top quark physics at hadron colliders**, *Eur. Phys. J.* **C48** (2006) 835–1000. p. 924, Fig. 89.

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

[288] R. Brun and F. Rademakers, **ROOT: An object oriented data analysis framework**, *Nucl. Instrum. Meth.* **A389** (1997) 81–86.

[289] O. Mattelaer, **On the maximal use of Monte Carlo samples: re-weighting events at NLO accuracy**, *Eur. Phys. J.* **C76** (2016), no. 12 674, [arXiv:1607.00763].

[290] ATLAS Collaboration, **Studies of tt+cc production with MadGraph5_aMC@NLO and Herwig++ for the ATLAS experiment**, Tech. Rep. ATL-PHYS-PUB-2016-011, CERN, Geneva, May, 2016.

[291] CMS Collaboration, S. Moortgat, **Charm jet identification at the CMS experiment**, in *Proceedings, 8th International Workshop on Charm Physics (Charm 2016): Bologna, Italy, September 5-9, 2016*, vol. CHARM2016, p. 060, 2016.

[292] CMS Collaboration, **Commissioning studies on the Phase I pixel detector of CMS in early 2017 proton-proton collisions at 13 TeV**, Tech. Rep. CMS-DP-2017-037, Aug, 2017.

Printed by
Crazy Copy Center Productions
VUB Pleinlaan 2, 1050 Brussel
Tel: +32 2 629 33 44
crazycopy@vub.ac.be
www.crazycopy.be

ISBN : 9789493079243
NUR CODE : 924, 926