Search for Rare Higgs Boson Production and Beyond the Standard Model Physics with Bottom Quarks and Tau Leptons at the CMS Experiment

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Pascal Raffael Bärtschi

von

Rüegsau BE

Promotionskommission

Prof. Dr. Florencia Canelli (Vorsitz)

Prof. Dr. Cristina Botta

Prof. Dr. Massimiliano Grazzini

Zürich, 2024

Abstract

The fundamental constituents of matter, called fermions, as described by the Standard Model of particle physics, are grouped in three generation with increasing mass. The heaviest fermions are particularly interesting, as they have the strongest coupling to the Higgs boson and are often most sensitive to theories going beyond the Standard Model.
The heaviest lepton is the tau lepton and the heaviest quarks are the top and the bottom quarks. In this thesis, two searches are described with tau leptons and bottom quarks measured in the CMS detector from proton-proton collisions taken in the years 2016 to 2018. The CMS experiment is one of the main experiments at the Large Hadron Collider and measures the properties of particles produced in proton-proton or heavy ion collisions. One search is for a rare and so far not observed process predicted by the Standard Model, the Higgs boson production in association with bottom quarks. The Higgs boson is identified through its decay to a tau lepton pair or a W boson pair. This is the first search for the single Higgs boson production in association with bottom quarks, resulting in an upper limit on the signal cross section. The observed (expected) upper limit at the 95% confidence level on the signal strength is 3.7 (6.1) times the value expected in the Standard Model.
The second search is for vector-like leptons, predicted by the 4321 model, a beyond the Standard Model theory that introduces additional gauge symmetries. The model also predicts leptoquarks, which couple leptons to quarks with the strongest coupling to third generation fermions. The flavor-dependent coupling strengths can address flavor hierarchies and potential lepton flavor violations. The vector-like leptons couple through leptoquarks to Standard Model fermions and are expected to have mass of smaller than 2 TeV. A search for vector-like leptons in a fully hadronic final state, performed by the CMS Collaboration, found an excess of data events with a deviation from the background-only hypothesis. The excess is checked by the described search in a final state with one electron or muon, one hadronic tau lepton decay and at least three bottom quarks.
In the future, the High Luminosity Large Hadron Collider will increase the proton collision rate to deliver large amounts of data that enables the CMS and other experiments to search for processes even more rare. The change in conditions requires an upgrade of the CMS detector, including the innermost part, the silicon pixel detector. The best pixel sensors for the upgrade are evaluated in a research and development campaign, which includes testing pixel sensors in testbeams. My contribution to the testbeam campaign is described in this thesis.

Acknowledgements

I want to thank the CMS group at the University of Zürich for their support throughout the creation of this thesis, especially Adinda de Wit and Kyle Cormier for sharing their knowledge with me and always being ready for a chat when I needed help, and my supervisor, Florencia Canelli, for offering me the opportunity to join her group and her guidance.
I would also like to thank Anna and Stefanos for their supervision of the test beam data analysis.
Special thanks to the UZH group, including Anson, Arash, Armin, Eslam, Fabian, Fanqiang, Filip, Marc, Valeria and Weijie for all the interesting discussions we had at lunch and with some beers at the Irchel Bar, and to the CERN Group with Giovanni, Jona and Raffaella, for making each trip to CERN so much nicer.
I am thankful to the former PhD candidates and Postdocs Eduardo, Kunal, Matias, Sascha, Sergio, Umberto and Vini, and Izaak for his enthusiasm in creating Feynman diagrams and templates and providing them to all PhD students.
I sincerely thank Alexei, Andrea and Maryam from DESY for their valuable contributions to the bbH search.
My genuine thanks go to my friends at ASVZ for helping me maintain a healthy work-sport balance.
Finally, I am deeply grateful to my family, friends, and Veronika who always supported me in my endeavors and allowed me to focus on the work required for this thesis.

Contents

1 Introduction

Since the beginning of mankind certain individuals were curious how nature works and constructed models based on their observations. However, the composition of matter and how matter interacts remained a mystery for a very long time. Over the last century great progress was made in understanding the building blocks of matter, the elementary particles, and how they interact with each other. In the 1970s a theoretical framework of particle physics was constructed, the Standard Model (SM). In the following decades many new elementary particles have been discovered, but one missing piece of the SM remained, the Higgs boson. In 2012, around 40 years after its prediction, the "Compact Muon Solenoid" (CMS) and "A Toroidal LHC Apparatus" (ATLAS) Collaborations at CERN discovered the Higgs boson. It plays a special role in the SM by giving mass to elementary particles through the Higgs mechanism. Studying the Higgs boson's coupling to other particles through its production modes and decay rates helps validate the SM and can reveal deviations that may indicate new physics. The production modes of the single Higgs boson occur at vastly different rates. The most likely one is the gluon fusion, followed by the vector boson fusion, the production in association with a vector boson and the production in association with top quarks. Even rarer is the production in association with bottom quarks, which is the only production mode among those with the largest cross sections that has not yet been observed. The CMS experiment recorded data amounting to 138 fb^{-1} from proton-proton collision at a center-of-mass energy of 13 TeV during the Large Hadron Collider (LHC) Run 2, which spanned from 2016 to 2018. This data is offering the opportunity to search for the rare Higgs boson production in associated with bottom quarks. The Higgs boson predominantly decays into a bottom quark antiquark pair, with less frequent decays into a W boson pair, gluon-gluon pair, a tau lepton pair, and even rarer decays. In this search, the Higgs boson is identified through its decay to a tau lepton pair. The tau lepton is short lived and decays to an electron, muon or hadrons, which are measured in the detector. The decay to a tau lepton pair was chosen because of its relatively high decay fraction and distinct decay signatures.

A similar final state including a tau lepton and several bottom quark jets is used to search for vector-like leptons, the second search described in this thesis. Vector-like leptons are predicted by the 4321 model, a beyond the SM theory that introduces additional gauge symmetries. This model predicts new couplings between quarks and leptons via a massive leptoquark, with the strongest coupling to third generation fermions. These flavor-dependent coupling strengths can address flavor hierarchies and potential lepton flavor violations. In order for the mixing between generations to be present, vector-like leptons are introduced, which couple to SM fermions via leptoquarks. The search focuses on vector-like leptons coupling to third-generation fermions in the mass range from 500 to 1000 GeV.

In the future CMS and the other experiments at CERN will search for even rarer processes. Starting in 2029, the High Luminosity Large Hadron Collider (HL-LHC), an upgrade of the Large Hadron Collider with increased proton collision rate, will begin operation. CMS and the other experiments will collect large amount of data, that will improve the sensitivity to rare

Standard Model processes and new physics predicted by beyond the Standard Model theories. The increased proton collision rate results in more radiation close to the beamline, more proton collisions at the same time, and requires a faster trigger system. Therefore, the CMS detector must be upgraded to handle these new conditions. In this thesis the upgrade work for a new silicon pixel detector for the CMS experiment is described with emphasis on the evaluation of pixel sensors using the testbeam data.
This thesis is structured as follows:
In chapter 2.1 the Standard Model of particle physics is explained, focussing on Higgs physics and beyond the Standard Model theories including vector-like leptons as those are relevant for the searches described in this thesis. In chapter 3 the Large Hadron Collider and the CMS detector are described. Chapter 4 focuses on the reconstruction of the final states particles used in the searches. In chapter 5 one of the two searches of this thesis is described, the search for a Standard Model Higgs boson produced in association with bottom quarks in final states with leptons. In the following chapter 6 the second search is discussed, the search for vector-like leptons with electrons or muons in the final state. Chapter 7 describes my contribution towards the CMS pixel detector upgrade. A conclusion is given in Chapter 8.

2 Theoretical framework

The searches described in this thesis require a theoretical framework for predictions of expected and known processes and potential signal events. The fundamental framework is given by the Standard Model of particle physics, explained in section 2.1. The search for bottom quark associated production of the Higgs boson requires knowledge of the different production modes and decay processes of the Higgs boson, which are described in section 2.2. The search for vector-like leptons is based on beyond the Standard Model theories, explained in section 2.3. Both searches in this thesis include bottom quarks and tau leptons, whose properties are described in section 2.4.

2.1 Standard Model

The Standard Model (SM) of particle physics is a theoretical framework that includes all known elementary particles and the electromagnetic, weak, and strong interactions. The elementary particles are divided in bosons, which mediate interactions, and fermions, consisting of three generations of quarks and leptons. The SM is a quantum field theory, describing the particles as excited states of quantum fields. The mathematical function to describe the SM is given through the Lagrangian:

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{EW}} + \mathcal{L}_{\mathrm{QCD}} + \mathcal{L}_{\mathrm{Higgs}} + \mathcal{L}_{\mathrm{Yukawa}}, \tag{2.1}$$

where $\mathcal{L}_{\mathrm{EW}}$ is the term related to the electroweak interaction, $\mathcal{L}_{\mathrm{QCD}}$ to the strong interaction, $\mathcal{L}_{\mathrm{Higgs}}$ the Higgs mechanism and $\mathcal{L}_{\mathrm{Yukawa}}$ the Yukawa interaction, describing the interaction between the Higgs field and fermion fields.

The SM is gauge invariant under local gauge transformations of the symmetry group $SU(3)_c \times SU(2)_L \times U(1)_Y$, where $SU(3)_C$ represents the strong interaction, $SU(2)_L$ the weak interaction and $U(1)_Y$ the hypercharge, which is related to the electromagnetic interaction. Each symmetry group has generators, that describe the infinitesimal transformations within the group. Associated to the generators are gauge bosons, which mediate the interaction. The SM is renormalizable, which means that all infinities arising from quantum corrections can be absorbed into a finite number of parameters of the theory.

The electroweak theory, described in section 2.1.1, unifies the electromagnetic and the weak interaction. The $SU(2)_L \times U(1)_Y$ gauge symmetry can be spontaneously broken by the Higgs mechanism, described in section 2.1.3, resulting to the massive gauge bosons W and Z, that mediate the weak interaction. The Higgs boson with spin 0 is the excitation of the Higgs field, which gives all the other particles mass. The electromagnetic interaction is mediated by photons. The leptons consist of the electron, muon and tau lepton, and their corresponding neutrinos. The charged leptons interact via the electromagnetic and weak force, the neutral and almost massless neutrinos only via the weak force. The strong interaction is mediated by the massless gluons and described in section 2.1.2. The quarks are the only particles with color charge and interact via the strong force aside from also interacting via the electromagnetic and weak force. There are quarks with charge +2/3, up, charm and top, and quarks with charge -1/3, down, strange and bottom. The Standard Model is shown in Fig. 2.1. For each particle the mass, charge and spin is listed. The mass is given in units of $\mathrm{eV/c^2}$, where eV represents electrovolt, that is equivalent to the kinetic energy gained by an electron accelerated in vacuum through an electric potential of one volt, and c the speed of light. In particle physics the speed of light is set to one in order to simplify calculations. The same is done in this thesis. The prefixes to the mass unit of eV include the kilo (k) that corresponds to 10^3, Mega (M) with 10^6, Giga (G) with 10^9 and Tera (T) with 10^{12}. The charge is given in unit of the electron charge e. The spin is the intrinsic angular momentum of each particle and given in units of $\hbar$, the Planck constant h divided by 2π. In particle physics the $\hbar$ is set to one. The description of the Standard Model is mainly based on Refs. [1, 2]

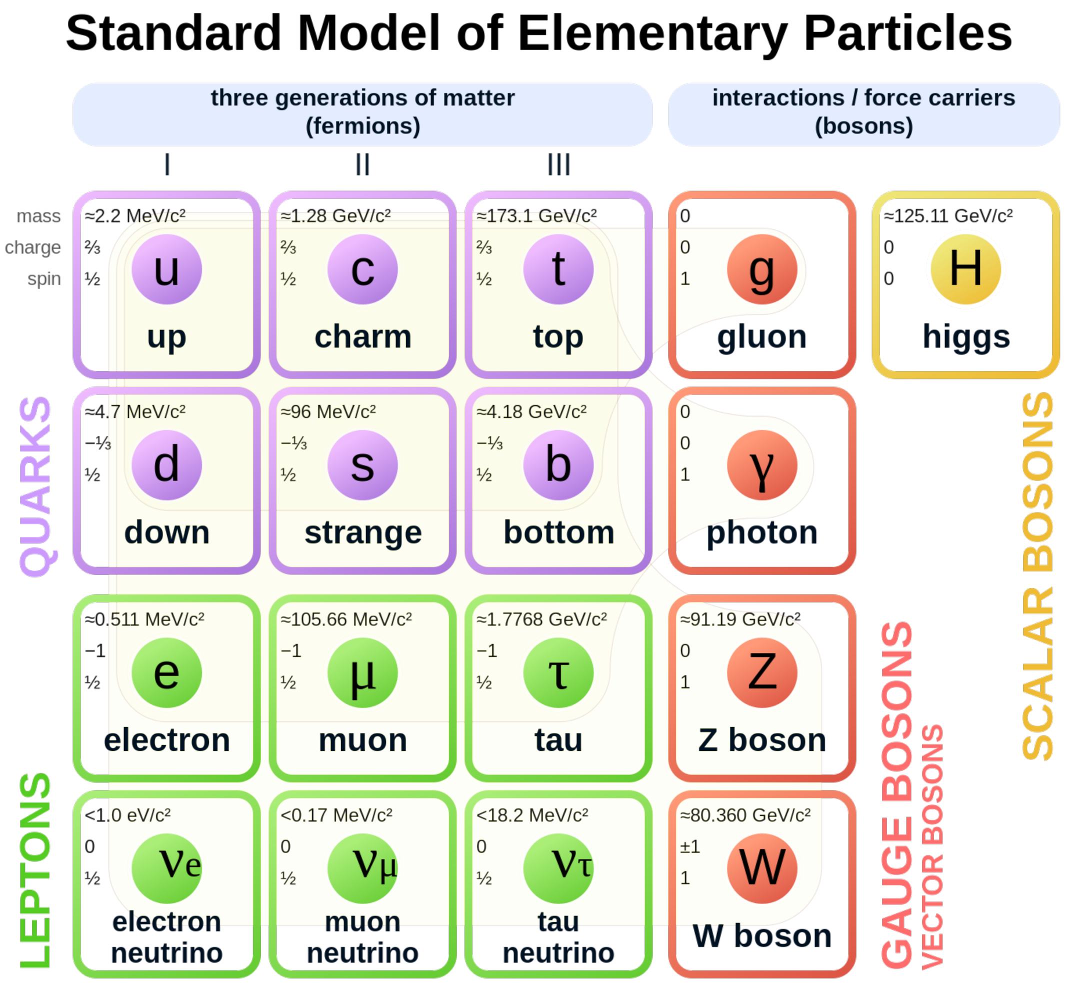


Figure 2.1: The Standard Model of particle physics. Figure from [3].

2.1.1 Electroweak interactions

The electroweak theory [4–6] unifies the electromagnetic and the weak interaction and is local gauge invariant under the $SU(2)_L \times U(1)_Y$ gauge group. The generators of $SU(2)$ are the three components of the weak isospin operator $T_i = \frac{\sigma_i}{2}$, with σ_i being the 2×2 Pauli matrices. The generator of $U(1)$ is the weak hypercharge Y. The $SU(2)_L$ gauge group corresponds to the weak interaction with three gauge bosons $W^1_{T_3=+1}$, $W^2_{T_3=-1}$ and $W^3_{T_3=0}$, where T_3 is the 3rd component of the weak isospin. The electromagnetic charge is a combination of the weak isospin and weak hypercharge:

$$Q = T_3 + \frac{1}{2}Y \tag{2.2}$$

The weak interaction only couples to left-handed particles and right-handed antiparticles. Left- and right-handedness is based on chirality, a fundamental property of particles. The left-handed and right-handed projections of the Dirac spinor ψ are defined by:

$$\psi_L = \frac{1}{2}(1-\gamma^5)\psi; \qquad \psi_R = \frac{1}{2}(1+\gamma^5)\psi, \tag{2.3}$$

where γ^5 is defined in terms of the gamma matrices as $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$.
Left-handed fermions are doublets under $SU(2)_L$ and right-handed fermions are singlets under $SU(2)_L$:

$$Q_L = \begin{pmatrix} u_\mathrm{L} \\ d_\mathrm{L} \end{pmatrix}, \quad u_\mathrm{R}, \quad d_\mathrm{R}, \quad L_L = \begin{pmatrix} \nu_\mathrm{L} \\ e_\mathrm{L} \end{pmatrix}, \quad e_\mathrm{R}, \tag{2.4}$$

where u (d) refers to all the up (down) type quarks, e the leptons and ν the neutrinos. Right-handed neutrinos are not included in the SM. All left-handed fermions have a weak isospin of $\pm\frac{1}{2}$ and all right-handed fermions a weak isospin of 0.
The weak interaction is unique in its ability to change fermion flavor via charged current interactions mediated by W bosons. In these interactions, Charge Parity (CP) symmetry can be violated. Charge conjugation (C) changes a particle into its antiparticle and Parity is a spatial inversion that changes the chirality of particles. The neural current interactions mediated by Z bosons do not change the flavor of fermions.
The electroweak term of the SM Lagrangian is written as:

$$\mathcal{L}_\mathrm{EW} = \overline{\psi}_f i\gamma^\mu D_\mu \psi_f - \frac{1}{4}W_a^{\mu\nu}W^a_{\mu\nu} - \frac{1}{4}B^{\mu\nu}B_{\mu\nu} \tag{2.5}$$

where ψ_f is a left-handed doublet or a right-handed singlet with flavor f, $W^a_{\mu\nu}$ (a=1,2,3) and $B_{\mu\nu}$ are the field strength tensors associated to the $SU(2)_L$ and the $U(1)$ group, respectively. The covariant derivative D_μ is defined as:

$$D_\mu = \partial_\mu + ig'\frac{1}{2}YB_\mu + igT_jW^j_\mu, \tag{2.6}$$

where B_μ is the $U(1)$ gauge field, W_μ the $SU(2)_L$ gauge field, the $U(1)$ coupling constant g' and the $SU(2)$ coupling constant g.

In section 2.1.3 it will be discussed how the $SU(2)$ symmetry can be spontaneously broken by the Higgs mechanism. The mass eigenstates of the spontaneously broken $SU(2)$ symmetry are two charged W bosons and a neutral Z boson. The photon and the Z boson arise from linear combinations of W^3 and B:

$$\begin{pmatrix} \gamma \\ Z^0 \end{pmatrix} = \begin{pmatrix} \cos\theta_\mathrm{W} & \sin\theta_\mathrm{W} \\ -\sin\theta_\mathrm{W} & \cos\theta_\mathrm{W} \end{pmatrix} \begin{pmatrix} B \\ W^3 \end{pmatrix}, \tag{2.7}$$

where θ_W is the weak mixing angle [7]. The mass eigenstates of the charged W boson are given by

$$W^{\pm} = \frac{1}{\sqrt{2}}(W^1 \mp iW^2) \tag{2.8}$$

2.1.2 Quantum Chromodynamics

Quantum Chromodynamics (QCD) describes the strong interaction between quarks and gluons in the SM. QCD is local gauge invariant under the $SU(3)_C$ gauge group with the generators being the eight components of the *color charge* operator $\Lambda_i = \frac{\lambda_i}{2}$ with λ_i being the 3×3 Gell-Mann matrices. The mediators of the strong interaction are 8 massless gluons.
The QCD term of the SM Lagrangian is given by:

$$\mathcal{L}_{\text{QCD}} = \overline{\psi}_f i\gamma^\mu D_\mu \psi_f - \frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a, \tag{2.9}$$

where ψ_f is the fermion field (described in eq. 2.4) and $G^a_{\mu\nu}$ is the gluon field strength tensor. The covariant derivative D_μ is defined as:

$$D_\mu = \partial_\mu + ig_s \Lambda_a G^a_\mu, \tag{2.10}$$

where G^a_μ is the eight component gluon field.

There are three different types of color charge associated with the colors red, blue and green and their respective anticolors. Each quark carries a color and each antiquark an anticolor. The gluons carry a combination of a color and an anticolor, which enables them to interact with each other and to change the color of quarks. The self coupling of gluons affects how the coupling of the strong interaction α_s changes with the energy scale of the interaction Q. At leading order in perturbation theory the function is given by:

$$\frac{d\alpha_s}{dQ^2} = -\frac{\alpha_s^2}{12\pi Q^2}(33 - 2n_f), \tag{2.11}$$

where n_f is the number of flavors accessible and the strong coupling is described by $\alpha_s = \frac{g_s^2}{4\pi}$ with g_s being the coupling constant. This means that α_s decreases with increasing Q, the interaction is stronger at long distances than at short distances. This effect is known as asymptotic freedom. The quarks cannot exist independently and form composite particles, a process called hadronization. Three quarks form a baryon and a quark-antiquark pair a meson. The composite particles are colorless and do not interact with the strong force. If quarks are separated, the energy increases until two new quark-antiquark pairs are formed in order to preserve color confinement. A consequence of this is the creation of jets, which is a spray of hadrons.

2.1.3 The Higgs mechanism

The Higgs mechanism [8–10] was proposed as a way how particles gain mass by spontaneously breaking the $SU(2)_L$ symmetry. The Higgs Lagrangian can be written as

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi) \tag{2.12}$$

with two complex scalar fields:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{2.13}$$

and with the Higgs potential:

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2, \tag{2.14}$$

with the free parameters μ and λ. For $\mu^2 > 0$ the minimum of the Higgs potential occurs if both scalar fields are zero. But for $\mu^2 < 0$ the Higgs potential has infinite minima at

$$\phi^\dagger \phi = -\frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2} \tag{2.15}$$

with v the vacuum expectation value, whose value is measured as $v = 246$ GeV [11]. With a local gauge transformation called unitary gauge three components become zero:

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h \end{pmatrix}, \tag{2.16}$$

where h is the Higgs field.
At low energies the $SU(2)$ symmetry is spontaneously broken, leading to massive electroweak gauge bosons. The W boson gets a mass of $m_W = \frac{1}{2}gv$ and the Z boson $m_Z = \frac{1}{2}v\sqrt{g^2 + g'^2}$. The Higgs boson is the excited state of the Higgs field with a mass of $m_H = \sqrt{2\lambda}v$. It was discovered in 2012 by the CMS and ATLAS Collaborations with a mass of 125 GeV [12–14].

2.1.4 Yukawa interactions

The fermions gain mass through their Yukawa coupling with the Higgs boson. After spontaneous symmetry breaking the Yukawa Lagrangian can be described as:

$$\mathcal{L}_{\text{Yukawa}} = -\frac{Y_d^{ij}}{\sqrt{2}} \overline{d}_L^i (v + h) d_R^j - \frac{Y_u^{ij}}{\sqrt{2}} \overline{u}_L^i (v + h) u_R^j - \frac{Y_e^{ij}}{\sqrt{2}} \overline{e}_L^i (v + h) e_R^j \tag{2.17}$$

with Y^{ij} the Yukawa matrices and i and j the three flavors. If the Yukawa matrices are diagonalized the mass term for fermions is found to be $m^i = \frac{vY^{ii}}{\sqrt{2}}$. The Yukawa matrices are not diagonal matrices, the weak interaction therefore not only occurs within the same generation. If the up type quark basis is fixed, then the corresponding down type quarks, with which they interact, are a linear combination of the down type quark mass eigenstates. The combinations are given by the *Cabbibo-Kobayashi-Maskawa* (CKM) matrix [15]:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \tag{2.18}$$

2.2 Higgs boson

The Higgs boson is the excitation of the Higgs field. At the Large Hadron Collider (LHC) the Higgs boson is predominantly produced through gluon fusion, where gluons from the colliding protons interact with each other and form a virtual quark loop. The virtual quarks are most likely top or bottom quarks because of their strong interaction with the Higgs boson. With an order of magnitude lower cross section, the Higgs boson can be produced through vector boson fusion. Colliding quarks exchange heavy vector bosons (W or Z boson), which fuse to produce a Higgs boson. The quarks only get deflected slightly and will be visible as jets in the detector. *Cross section* corresponds to the probability that a process will occur in the collision. It is defined as an effective area and given in units of barn, $10^{-24}\mathrm{cm}^2$. The third production mode is called Higgs-strahlung and involves the collision of a quark and antiquark, producing a virtual W or Z boson that emits a Higgs boson. The fourth production mode is called top quark associated production and characterizes the collision of gluons, each decaying to a heavy quark-antiquark pair. The fusion of a quark and an antiquark of each pair produces the Higgs boson. The Feynman diagrams of these processes are shown in Fig. 2.2. Production modes with even smaller cross section are the bottom quark associated production (bbH) and the production associated with a single top quark.

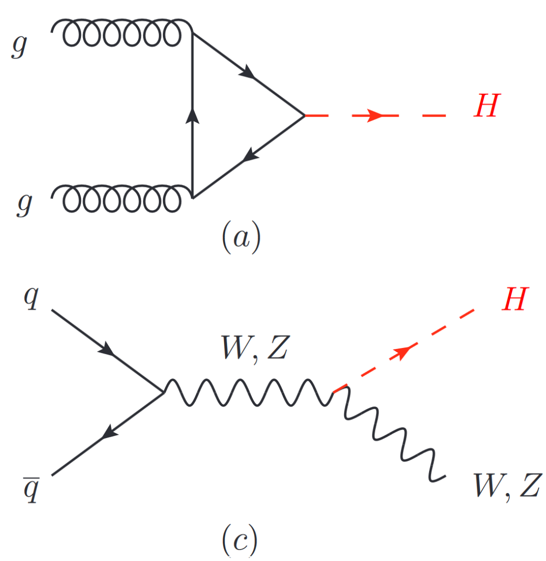


Figure 2.2: Higgs boson production Feynman diagrams for gluon fusion (a), vector boson fusion (b), Higgs-strahlung (c) and top quark associated production (d). Figure from [16].

The described production modes except the bbH production were successfully observed and measured by the CMS and ATLAS experiments, shown in Fig. 2.3. Observation means, that the background only hypothesis can be rejected with a significance of more than five standard deviations. The plots have a logarithmic scale for the y-axis, showing the large differences between the cross sections. The search for the Higgs boson associated production with bottom quarks is the topic in chapter 5.

The Higgs boson decays most likely to a bottom quark pair, because its mass of 125 GeV is not large enough to produce a on-shell top quark pair, Z boson pair or W boson pair. On-shell particles satisfy the equation of motion and are considered real particles. Off-shell particles are virtual particles, which can briefly violate the conservation of energy. The probability of

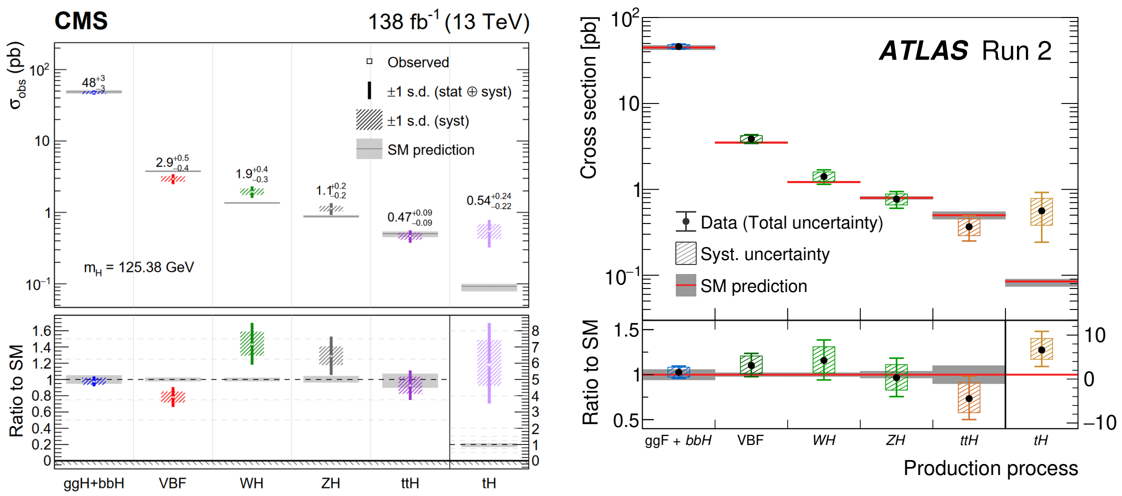


Figure 2.3: Measured Higgs boson production cross sections from CMS and ATLAS Collaboration. Figures from CMS [17] and ATLAS [18].

a decay is given as *branching ratio* or *branching fraction* and corresponds to the fraction of a certain decay of the total number of possible decays. A rarer decay of the Higgs boson is to a W boson pair, where one of the W bosons has to be off-shell. Other decays in order of decreasing branching ratio are the decay to a gluon pair, tau lepton pair, charm quark pair, Z boson pair and photon pair, as shown in Fig. 2.4. The gluon and the photon are massless and do not couple to the Higgs boson. The decay to a gluon pair is induced by virtual top and bottom quark loops and the decay to a photon pair is mediated by virtual W boson and top quark loops [19].
Even though the Higgs boson decay to a bottom quark pair is dominant, it is a challenging final state to search for the Higgs boson, because in proton-proton collisions a large number of quarks and gluons are produced. It is hard to find bottom quark pairs originating from the Higgs boson in a large bottom quark background and to distinguish jets induced from bottom quarks decay from other jets, which can be induced by other quarks or by gluons. Final states with leptons leave a clear signature in the detector and are easier to distinguish from background. Final state refers to the set of particles measured in the detector after a particle interaction or decay process. The decay of the Higgs boson to tau leptons and to a WW pair are the final state searched for in the analysis described in chapter 5.

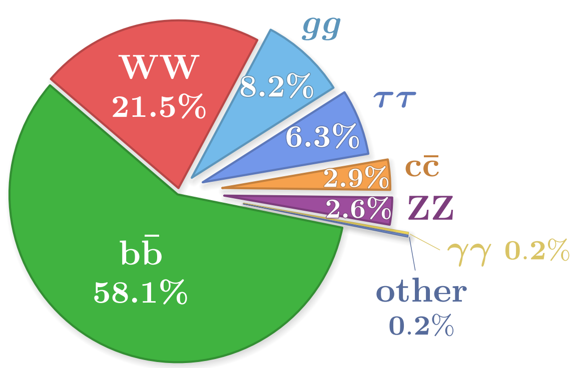


Figure 2.4: Higgs boson decay branching ratios. Chart from [20].

2.3 Beyond the Standard Model

Even though the Standard Model is the best theoretical framework for particle physics to date, it has some shortcomings. The most evident one is the missing of Gravity in the SM. Because gravity is so much weaker than the weak interaction, the effects from gravity are negligible at the energy range probed by the LHC. Only at much higher energies than the LHC can deliver, at the order of 10^{19} GeV, the quantum effects of gravity are expected to be relevant. The SM also does not have an explanation for the hierarchy problem, referring to the fact that the weak force is 10^{24} times stronger than gravity. This is a violation of naturalness, which assumes that all forces should be of the same order of magnitude. The hierarchy problem is also apparent taking the observed Higgs boson mass of 125 GeV into account. The Higgs boson mass gets higher order corrections from quantum loops, dominated by top quark loops. The mass predicted by the quantum loop corrections is much larger than the measured mass of 125 GeV. Only by adding a very large fine tuning cancellation the masses can be brought in agreement.
In the SM the coupling of electroweak gauge bosons to the three generations of leptons (electron, muon and tau) is assumed to be the same, with only small variations due to the different lepton masses. This prediction is known as lepton flavor universality (LFU). Several measurements of the LFU in the W boson decay show good agreement with the SM expected value in the first and second generation with high precision [21–23], and with less constraint for the third generation [21, 24]. Several experiments found hints for lepton flavour violation in the decay of B mesons [25], although these excesses are still disputed in the science community. If lepton flavor is violated, new particles with varying coupling strengths for different lepton flavors can exist. One potential solution are leptoquarks, hypothetical particles that couple to leptons and quarks. They have flavor-dependent coupling strengths, allowing a transition between different lepton flavors. Leptoquarks are predicted by many models, including the 4321 model [26], that extends the SM sector to $SU(4) \times SU(3)' \times SU(2) \times U(1)'$. The fact, that the 4321 model is UV complete and flavor non-universality is required, introduces vector-like leptons (VLL), which couple to SM fermions through a leptoquark [27]. In particular, the mass of the VLL needs to be below 2 TeV to be compatible with measurements of the $B_s - \overline{B}_s$ mixing [28], making it possible to search for them with data collected by the CMS experiment.
A search performed by the CMS collaboration in final states with third generation fermions, bottom quarks and tau leptons, found an excess for mass points ranging from 500 to 1000 GeV, with a significance of 2.8 standard deviations at a representative mass point of 600 GeV [29]. The observed upper limit on the production cross section of VLL is shown in Fig. 2.5. The analysis focused on an all hadronic final state. The inclusion of light leptons (electrons or muons) in the final state with bottom quarks and hadronic tau leptons gives a similar sensitivity to the all hadronic analysis and can be used as an independent test to confirm the excess. This is the second search described in this thesis and topic of chapter 6.

2.4 Bottom quarks and tau leptons

The search for bbH production and the search for VLL both include third generation fermions, bottom quarks and tau leptons, in the final state of the analyses. The coupling of the Higgs boson to particles is dependent on the particle mass, third generation fermions have a stronger coupling than the other generations, as is shown in Figure 2.6.

The largest coupling of leptoquarks is expected to third generation fermions and one order of

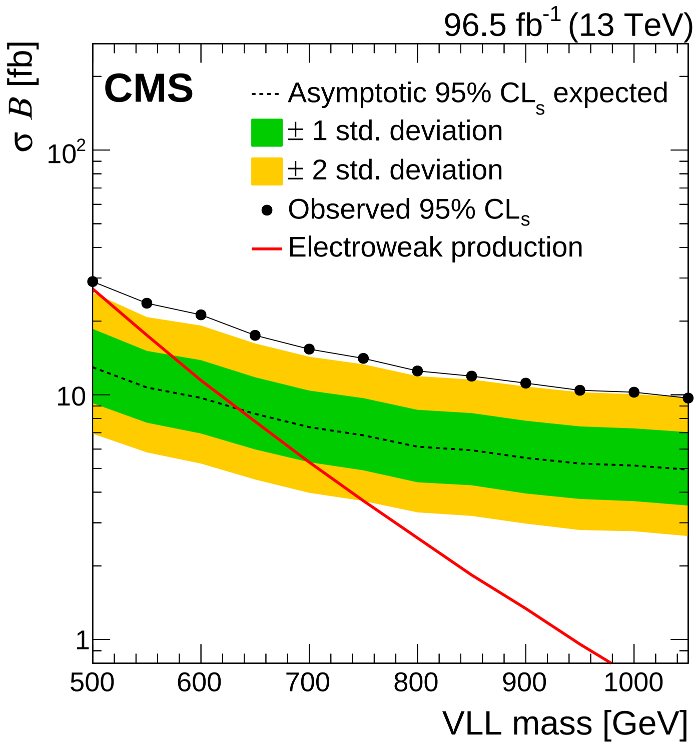


Figure 2.5: Observed and expected 95% confidence level upper limit on the electroweak vector-like lepton production cross section. Figure from the CMS Collaboration search in the all hadronic final state [29].

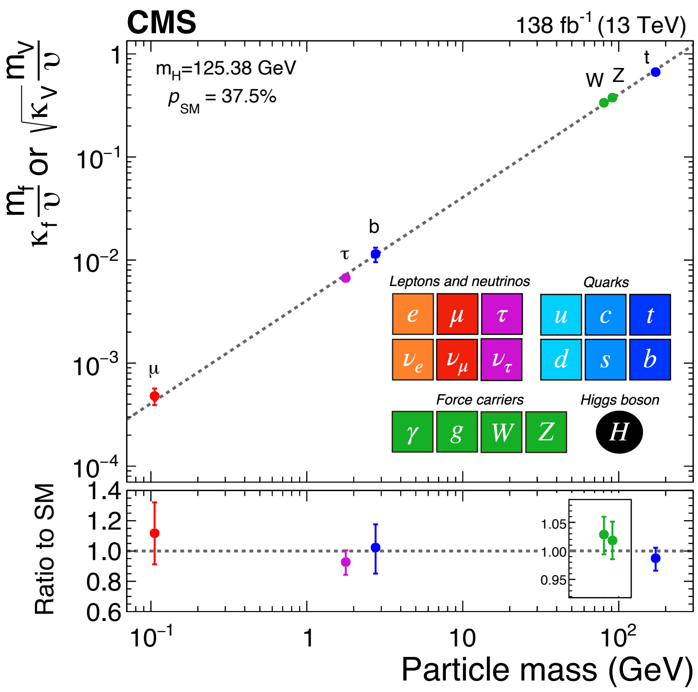


Figure 2.6: Higgs boson couplings measured by the CMS experiment versus the particle mass. The SM expected values are indicated by the dashed line. Figure from [30].

magnitude or more weaker to second and first generation fermions [31]. Therefore the dominant decay channel of VLL is to third generation fermions.

The bottom quark has a mass of 4.18 GeV and confined in a B hadron has a relatively long lifetime of 10^{-12} s. The B hadron travels a certain distance away from the primary vertex (PV), which marks the collision point of the protons, before producing a jet with a secondary vertex (SV), as shown in Figure 2.7. This signature can be used to distinguish jets originating from bottom quarks and jets induced by gluons and other quarks. Most of the bottom quark pairs originate from gluon splitting, quark-antiquark annihilation and from the decay of Z bosons and top quark pairs. The top quark decays after only 5×10^{-25} s almost exclusively to a W boson and a bottom quark.

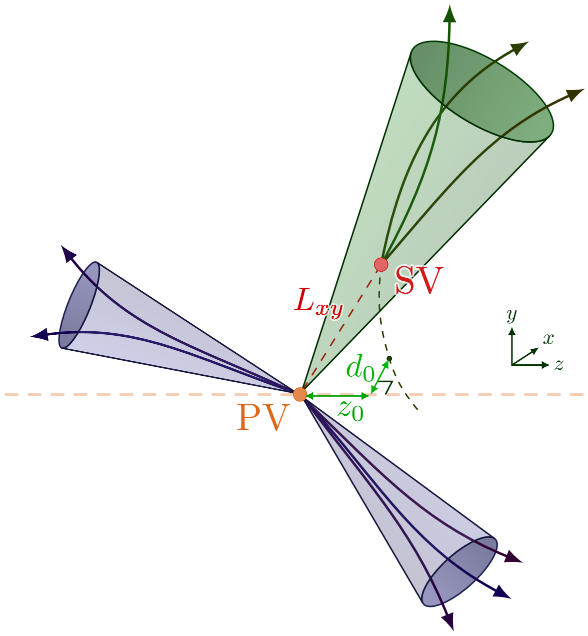


Figure 2.7: Bottom quark tagging of jets with the displaced secondary vertex (SV) from the primary vertex (PV). Taken from [32].

.

The tau lepton has a very short lifetime of 2.9×10^{-13} s, making it impossible to measure the tau leptons directly in the detector. Instead they are identified by their decay particles. The tau lepton has a large mass of 1.776 GeV, allowing it to decay to hadrons through weak interaction. The Feynman diagram of the tau lepton decay is shown in Fig. 2.8.

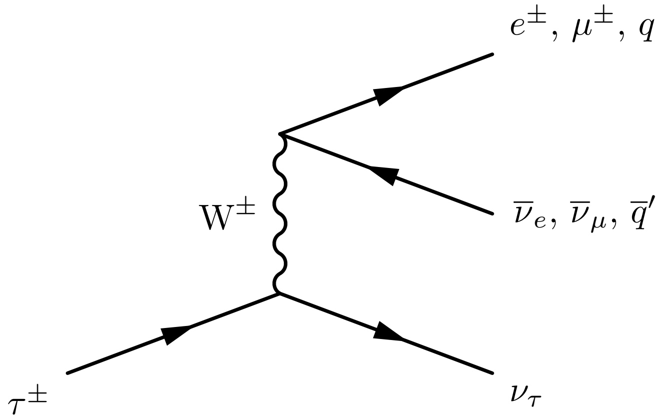


Figure 2.8: Feynman diagram of the tau lepton decay to a light lepton and a neutrino or to hadrons.

With a branching ratio of 64.8% the tau lepton decays most of the time into a number of charged and neutral hadrons (denoted with τ_h), mostly pions. The decay to an electron occurs in 17.8% and into a muon in 17.4% of the time. Hadronic decays of the tau lepton can be identified by looking for collimated hadrons with one (1-prong), two (2-prong) or three charged hadrons (3-prong) and multiple neutral hadrons. In Figure 2.9 the difference in the substructure of jets from hadronic tau decays and a jet induced by a gluon or quark is shown.

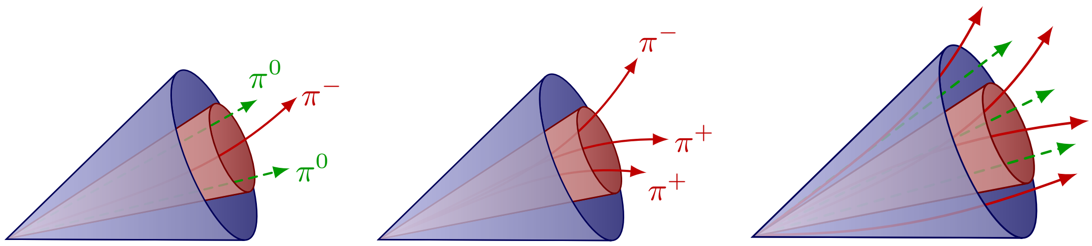


Figure 2.9: Jet from 1-prong (left) and 3-prong (center) hadronic tau decay and gluon/quark initiated jet (right). Figure from [33].

Considering the tau lepton pair, the fully hadronic decay channel ($\tau_h\tau_h$) with both tau leptons decaying into hadrons occurs 42% of the time, the semileptonic decay channels to an electron or a muon and a hadronically decaying tau lepton ($\mathrm{e}\tau_\mathrm{h}$, $\mu\tau_\mathrm{h}$) each occur 23% of the time and the remaining 12% are decays to two light leptons, called fully leptonic decay channel (eμ, μμ, ee) [11]. The branching ratios are visualized in the pie charts in figure 2.10. Most tau lepton pairs measured by the detector are decay particles from Z bosons or virtual photons produced in quark-antiquark annihilation.

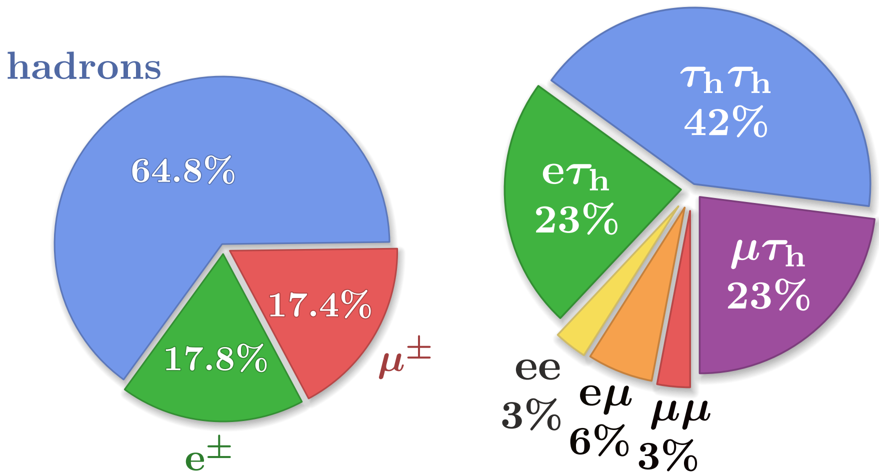


Figure 2.10: Tau lepton decay branching ratios (left) and tau lepton pair decays (right). Charts from [20].

3 Large Hadron Collider and Compact Muon Solenoid Experiment

A powerful particle collider and a large-scale detector are required to probe rare Standard Model processes or to search for beyond the Standard Model theories. The Large Hadron Collider (LHC) is the world's most powerful particle accelerator, located in a circular tunnel near Geneva. The tunnel was formerly occupied by the Large Electron-Positron (LEP) collider, which was in operation from 1989 to 2000. The LHC started its operation in 2008 and delivered a record amount of proton-proton and heavy-ion collision to be used by the experiments. One of the main four experiments is the Compact Muon Solenoid (CMS) experiment. It encloses the interaction point and measures all the particles produced in collisions. The data collected from 2016 to 2018 by the CMS experiment is used for the searches described in this thesis.

3.1 Large Hadron Collider

The Large Hadron Collider (LHC) [34] is a particle accelerator near Geneva at the French-Swiss border, operated by CERN, the European Organization for Nuclear Research. The LHC is an underground circular accelerator with a circumference of 27 km. Two particle beams, with protons or heavy ions, are accelerated in opposite directions close to the speed of light and brought into collision at four interaction points. The particle beams are bent with dipole magnets, which provide a magnetic field of 8.3 Tesla. LHC started operation in 2009 with a center-of-mass-energy of 7 TeV. In 2015, the center-of-mass energy was increased to 13 TeV and in 2022 to 13.8 TeV. In order to achieve energies at the TeV scale, several acceleration stages are needed. The source of the proton beams is hydrogen gas from a hydrogen bottle. In the linear accelerator (LINAC 4) the negative hydrogen ions get accelerated and reach energies of 160 MeV. The ions are stripped of their electrons during injection to the Proton Synchroton Booster, leaving only protons. In the Booster they are accelerated to energies of 2 GeV, before being injected to the Proton Synchrotron (PS), accelerating them to energies of 26 GeV, and afterwards the Super Proton Synchrotron (SPS), where they reach energies of 450 GeV. Then they are injected to the LHC and accelerated to center-of-mass energy of several TeV. The protons are combined into bunches of up to 110 billion protons [34, 35], which travel close to speed of light in two beams going in opposite directions. The maximal number of bunches per beam was 2220 (2556) in 2016 (2017,2018) [35]. Every 25 nanoseconds (with a frequency of 40 MHz), the bunches cross at the four interaction points, where the four main experiments are located: CMS, ATLAS, LHCb, and ALICE. Figure 3.1 shows the CERN accelerator complex with the main experiments.

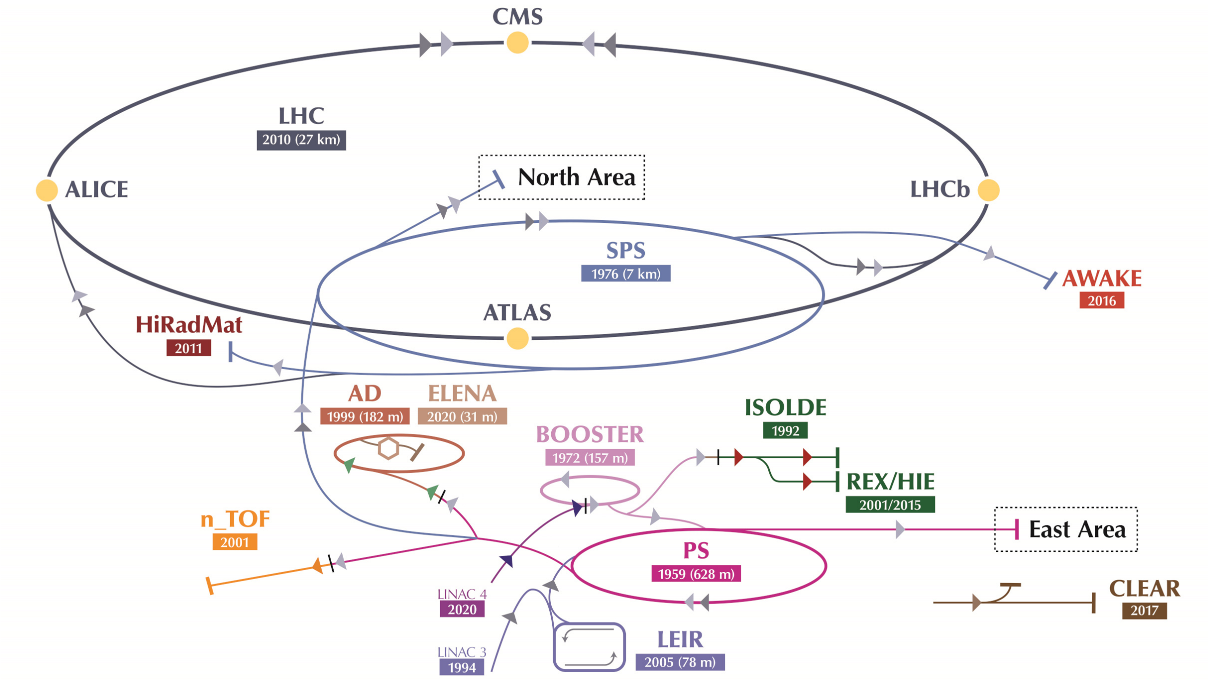


Figure 3.1: CERN accelerator complex. The four main experiments, CMS, ATLAS, LHCb and ALICE are marked with yellow circles. Taken from [36].

3.1.1 Luminosity

The number of proton-proton collisions per unit time and cross section are measured with the *instantaneous luminosity* [11]:

$$\mathcal{L} = \frac{N_1 N_2 f}{A}, \tag{3.1}$$

where f is the frequency of the bunch crossing, N_1 and N_2 are the number of protons in each bunch, and A is the cross sectional area of the bunches. In order to increase the instantaneous luminosity, the cross sectional area should be as small as possible, around 10-20 μm^2. The unit of *barn* (b) is used for the area, which is equivalent to $10^{-24}cm^2$. The amount of data collected over a period of time is measured with the *integrated luminosity*, later referred to as *luminosity*:

$$L = \int \mathcal{L} dt \tag{3.2}$$

Figure 3.2 shows the integrated luminosity for each data taking year since 2015. In the searches presented in this thesis, the data taking years 2016, 2017, and 2018 are used, amounting to collected data of 138 fb^{-1} integrated luminosity.
At each bunch crossing multiple proton collisions happen at the same time, most of them are soft scattering and not interesting for physics analysis. The mean number of proton-proton interactions per crossing is called "pileup" and is shown in Fig. 3.2 for each year. Higher pileup means a larger event rate for the process we are searching for, but also larger noise from additional (not interesting for physics analysis) proton interactions.

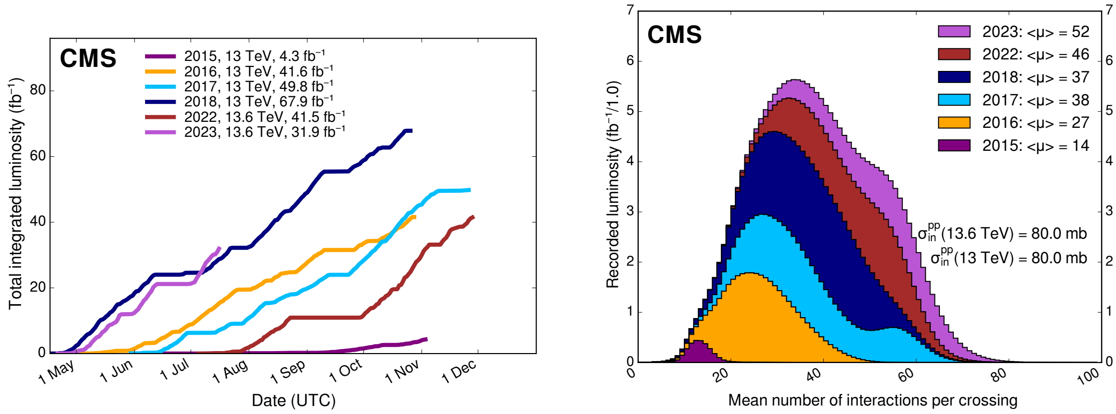


Figure 3.2: Left: Integrated luminosity of data collected by the CMS experiment from 2015 to 2023. Right: Distribution of number of proton-proton interactions per crossing, assuming an inelastic cross section of 80 mb. Figures taken from [37].

3.2 Compact Muon Solenoid detector

The Compact Muon Solenoid (CMS) is located at one of the four LHC beam crossings near the French town of Cessy, 100m underground. It has a length of 21.5 meters and a diameter of 14.6 meters. The detector is built around a large solenoid magnet, which delivers a magnetic field of 3.8 Tesla. The detector consists of several layers of subdetectors, each with a barrel section and two endcaps at either end of the barrel, enclosing the interaction point. The innermost detector is the silicon tracker, which measures the trajectory of charged particles. The next layer is the electromagnetic calorimeter, for measuring the energy of electrons and photons, followed by the hadronic calorimeter, which measures the energy of hadrons. Outside of the solenoid, embedded in the iron return yoke, is the muon detector, designed to measure the kinematics of muons. An illustration of the detector and its subdetectors is shown in Figure 3.3. A more detailed description of the detector can be found in Ref. [38].

3.2.1 Coordinate system

The origin of the coordinate system is at the nominal interaction point, at the center of the detector. The x-axis is pointing towards the center of the LHC, the y-axis points upwards and z-axis along the counterclockwise beam direction. The azimuthal angle ϕ lies in the x-y plane with $\phi = 0$ along the positive x axis. The polar angle θ lies in the z-y plane with $\theta = 0$ along the positive z axis. Because particles often move in forward/backward region close to the beam line, instead of θ, the *pseudorapidity* η is used, which is defined as:

$$\eta = -\ln[\tan\left(\frac{\theta}{2}\right)] \tag{3.3}$$

$\eta = 0$ is along the positive y-axis, $\eta = +\inf$ is along the positive z-axis, and $\eta = -\inf$ is along the negative z-axis.
The radial distance from the beam line in the x-y plane is denoted as $r = \sqrt{x^2 + y^2}$.

3.2.2 Tracker

The silicon tracker is the innermost detector and measures the tracks of charged particles. The spatial resolution of the tracker needs to be smaller the closer it is to the beam line in order to

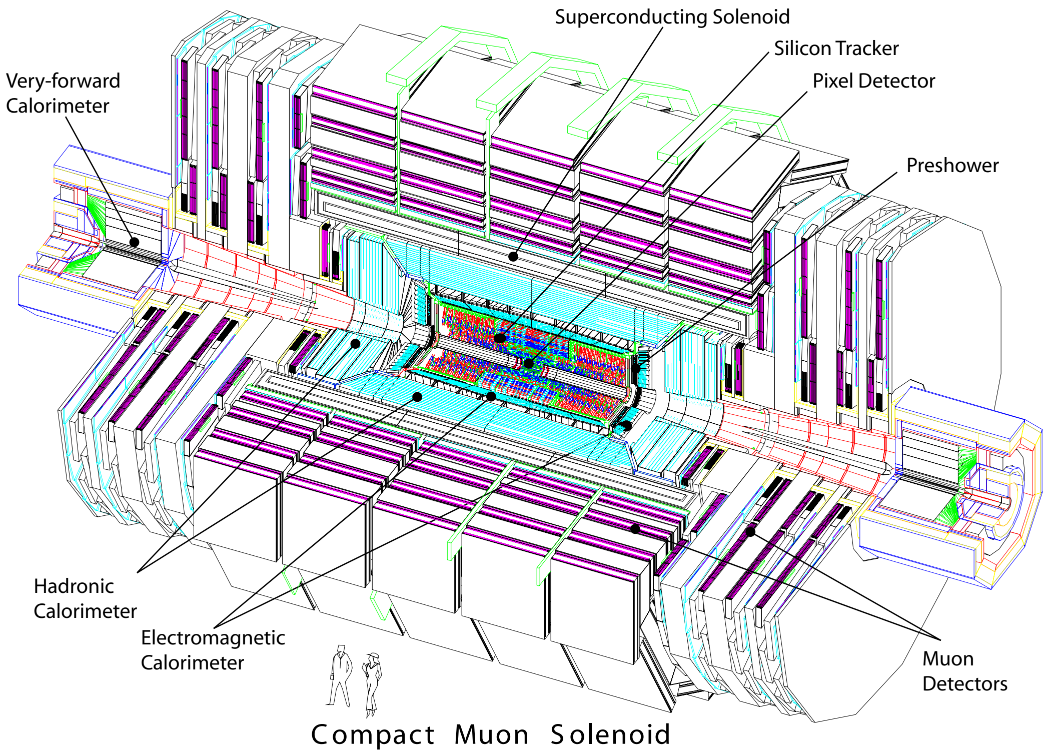


Figure 3.3: The CMS detector, highlighting the multiple layers of subdetectors surrounding the interaction point. Taken from [38].

link the particles to the correct bunch crossing. For that reason, the tracker is split into two subdetectors, the pixel detector, equipped with small pixel cells, and the strip detector, built from pixel strips.

3.2.2.1 Silicon pixel tracker

The silicon pixel detector is built from small silicon pixel cells with a size of $100 \times 150\,\mu\text{m}^2$. If a charged particle passes through, it deposits ionizing charges in one or more pixels, which are collected by a readout system bonded to the pixels. Neighboring pixels with deposited charge are grouped together as a cluster, and the hit position is reconstructed by taking the average of the charge. The pixels are grouped together as modules, consisting of a silicon sensor and a readout chip. The modules are mounted on four cylindrical barrel layers (BPIX) and on three endcap disks in the forward region (FPIX) on either side. The layers of the barrel region are placed 2.9 cm up to 16 cm from the beam line. The endcap disks, positioned between 30 cm and 46.5 cm in the z direction from the interaction point, cover a pseudorapidity range up to $|\eta| < 2.5$. If multiple hits are registered on different layers, consistent with the trajectory of a charged particle deflected by the magnetic field, they are grouped together as a track.
The pixel detector was upgraded at the end of 2016 due to radiation damage of the pixels and to increase the readout bandwidth, to ensure good tracking efficiency with the higher instantaneous luminosity of Run 2. The original design consisted of three barrel layers and two disks and the innermost layer was placed further away from the interaction point, as shown in Figure 3.5.
The spatial resolution of the pixels is 10 µm in the $r-\phi$ direction and 20-30 µm in the z direction. The high precision is required to reconstruct the collision vertices of the proton-proton collisions by extrapolating all the tracks back to their origin in the beam line. It is also important to

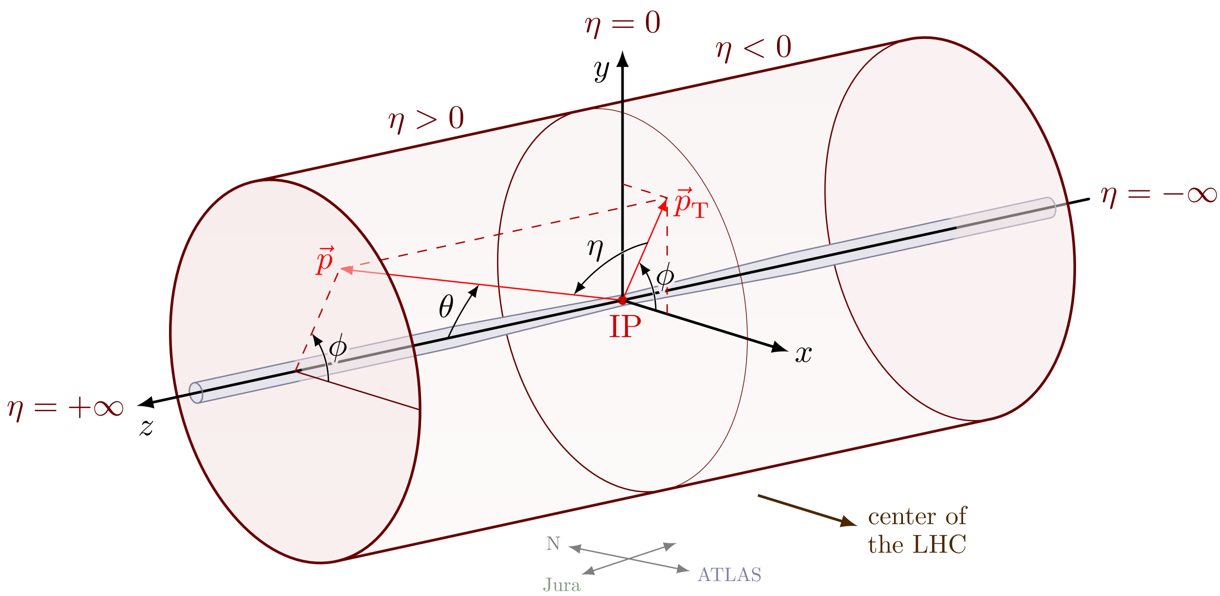


Figure 3.4: Coordinate system of CMS. The interaction point (IP) is set as origin, with the x-axis pointing towards the center of the LHC, the y-axis directed upwards, and the z-axis aligned along the counterclockwise beam direction. Taken from [39].

identify secondary or displaced vertices, created by short-lived particles such as B hadrons or tau leptons, that travel a few millimeters away from the beam line before decaying. A more detailed description of the pixel tracker can be found in Refs. [40, 41].

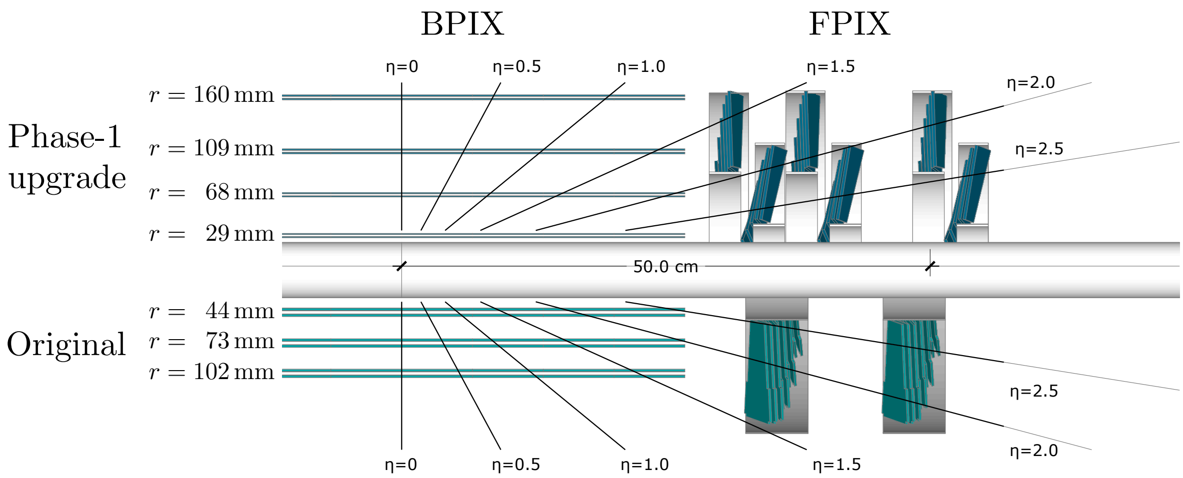


Figure 3.5: Layout of the CMS Phase-1 pixel detector used since the end of 2016 (top) compared to the original layout (bottom). Taken from [40].

3.2.2.2 Silicon strip tracker

The silicon strip detector encloses the pixel detector and lies in the range of 20 cm to 116 cm around the beam line. Because of the lower occupancy compared to the pixel detector, less radiation hardness is required, and a larger spatial resolution is sufficient, making silicon strips the more cost-effective solution for this detector. Radiation hardness refers to the ability of a material to withstand a large dose of radiation without or with only minor defects. The detector is divided into four inner barrel layers (TIB) and two inner endcaps (TID), each consisting of three disks. The six layers of the outer barrel (TOB) enclose the TIB and TID and are closed off on either side by the two endcaps (TEC). The covered pseudorapidity range is $|\eta| < 2.5$. The layout of the tracker system is shown in Figure 3.6. The strip detector is described in more detail in Ref. [41].

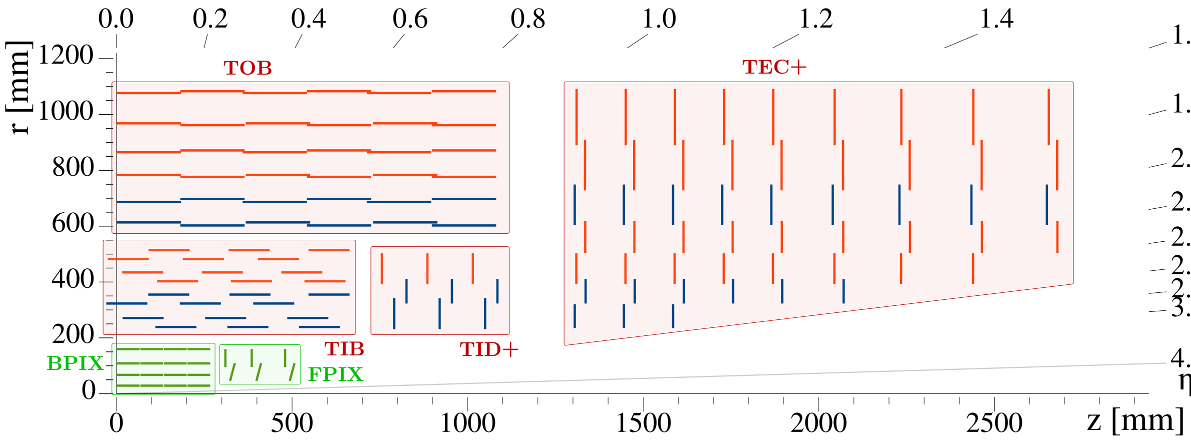


Figure 3.6: One quadrant of the Tracker system. The Phase-1 pixel detector is shown in green and the strip detector in red. Adapted from [42].

3.2.3 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) measures the energy of electrons and photons that induce electromagnetic showers, which are cascades of electrons and photons. The ECAL is built from lead tungstate ($PbWO_4$) crystals, radially placed in a barrel and two endcaps on either side, covering a pseudorapidity range of up to $|\eta| \leq 3$. Lead tungstate is a fast scintillator with high density, short radiation length, and is radiation hard. Radiation length refers to the mean distance over which an electron loses all but $1/e$ of its energy due to bremsstrahlung, or the distance over which a photon reduces its energy by the same fraction due to pair production. Radiation hardness is required due to the short distance to the interaction point, and a short radiation length is important to measure all of the energy of electrons and photons in a small amount of space. In the barrel region, the crystals have a length of 23 cm and in the endcaps 22 cm.
Scintillation light is emitted in the crystal if a particle passes through, and is measured at the top of each crystal by photodetectors. In the barrel region avalanche photodiodes are used and in the endcaps the more radiation hard vacuum phototriodes. The amount of scintillation photons is proportional to the energy deposited in the crystal. In front of the endcap a preshower detector is placed. It is built from two lead plates followed by two orthogonal silicon strip layers that allow the measurement of showers induced by photons hitting the lead. The preshower detector enables the discrimination between single high-energy photons and collimated low-energy photon pairs. A schematic view of the ECAL is shown in Figure 3.7. A more detailed description of the ECAL can be found in Ref. [43].

3.2.4 Hadronic calorimeter

The hadronic calorimeter (HCAL) measures the energy of hadrons that induce hadronic showers through nuclear interactions. The HCAL is a sampling calorimeter, consisting of layers of brass absorbers and plastic scintillators. If a hadronic particle hits an absorber, it produces secondary

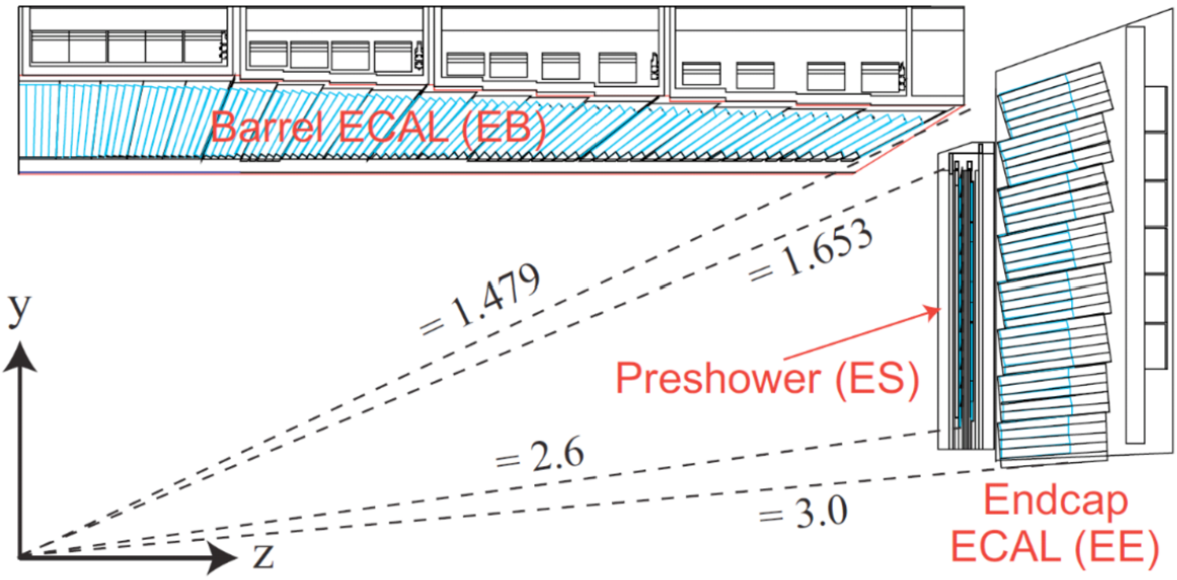


Figure 3.7: Schematic view of the electromagnetic calorimeter. Taken from [44].

particles, which themselves produce cascades of electromagnetic or hadronic showers. Particles passing through the scintillator material emit blue-violet light, which is absorbed by wavelength-shifting fibers, shifting the light to the green spectrum. The light is then collected by hybrid photodiodes through fiber cables. The space inside the solenoid is very restricted and does not allow for sufficient interaction lengths to capture all the energy of the hadrons. Therefore, a small part of the HCAL is placed outside of the solenoid.
The barrel and endcap of the HCAL cover a range of up to $|\eta| \leq 3$. In order to extend the pseudorapidity range up to $|\eta| \leq 5$, a forward calorimeter is placed 11.15 m from the interaction point. Because of the much higher particle flux in the forward region, the forward calorimeter is made of steel plates and quartz fibers. The different components of the HCAL are shown in Figure 3.8. The HCAL is described in more detail in Ref. [45].

3.2.5 Solenoid magnet

The solenoid magnet is made of coils of wire which produce a uniform magnetic field of 3.8 Tesla when an electric current flows through them. The magnet is cooled to 4.65 Kelvin (corresponding to -268.5 degrees Celsius), making the solenoid superconducting. The magnet has an inner diameter of 6 meters and is 13 meters long. The magnetic field is parallel to the beam line and bends the trajectories of charged particles in the transverse plane, orthogonal to the beam line. The momentum of particles in the transverse plane, called transverse momentum (p_{T}), can be measured by the bending radius (r), the charge of the particle (q) and the magnetic field (B):

$$p_{\mathrm{T}} = |q| \cdot r \cdot B \tag{3.4}$$

Outside of the solenoid, 5 steel barrel rings and 3 steel disks at the endcaps act as flux-return yoke, making the field more homogeneous. The barrel rings and disks act as main support system of the detector and weigh more than 10000 tons.

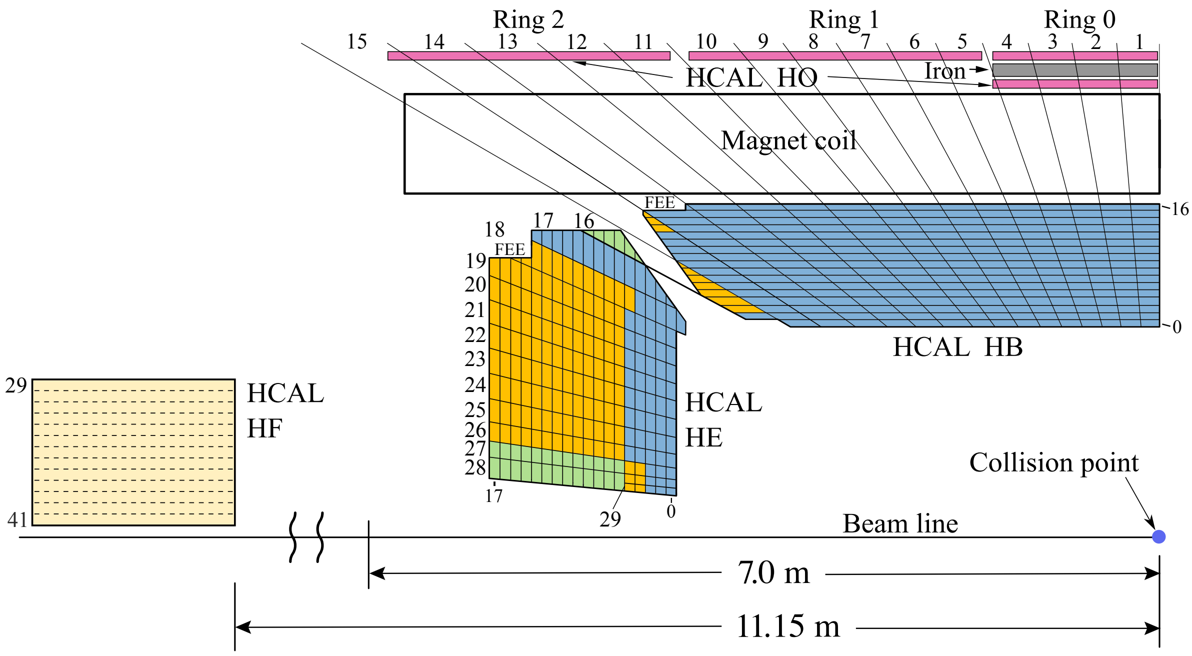


Figure 3.8: Schematic view of the hadron calorimeter with the barrel region (HB), endcap region (HE), forward calorimeter (HF) and the hadron calorimeter outside of the solenoid (HO). Taken from [46].

3.2.6 Muon chambers

The muons are not stopped by the calorimeters and are therefore measured by gas-ionization detectors, interlaid with the iron flux-return yoke plates. A schematic view of one quadrant of the detector is shown in Figure 3.9. Different types of gas ionization detectors are used for the barrel and endcap region because of different conditions. The gas ionization detectors work all in a similar way. There is a positive-charged anode and a negative-charged cathode inside a gas volume. If a muon passes through, it ionizes the gas, the electrons drift towards the anode and the positive charged ions towards the cathode. The electrons hit other atoms and create an avalanche which is measured once it hits the anode. Based on the drift time the position of the muon can be measured.

The barrel region consists of five wheels, each with four rings of drift tubes (DTs) interlayed with the iron flux-return yoke plates. The DTs cover a region of up to $|\eta| \leq 1.2$ and are well suited for the uniform magnetic field and lower particle flux compared to the forward region. The DTs are built from drift cells, which contain a wire within a gas volume. In the forward region, the magnetic field is inhomogeneous and the rate of muons and neutron induced background is much larger. Four disks of cathode strip chambers (CSCs), interlayed with iron plates cover a pseudorapidity range of $0.9 \leq |\eta| \leq 2.4$. The CSCs are built as a grid of positively-charged anode wires and negatively-charged copper cathode strips within a gas volume. The CSCs provide a good spatial and time resolution. In order to get a fast readout for the trigger system, resistive plate chambers (RPCs) are added in the barrel and endcap region. The RPCs are built from two parallel plates of an anode and a cathode, separated by a thin gas volume. The RPCs provide very good time resolution of just one nanosecond, which is important to link the muons to the correct bunch crossing. The muon chambers together with the information from the tracker provide a muon identification efficiency of 96% for muons with p_{T} larger than a few GeV and a muon momentum resolution of 1 to 3 % for muons with $p_{\mathrm{T}} < 100$ GeV [47]. A more detailed description of the muon detector can be found in Ref. [48].

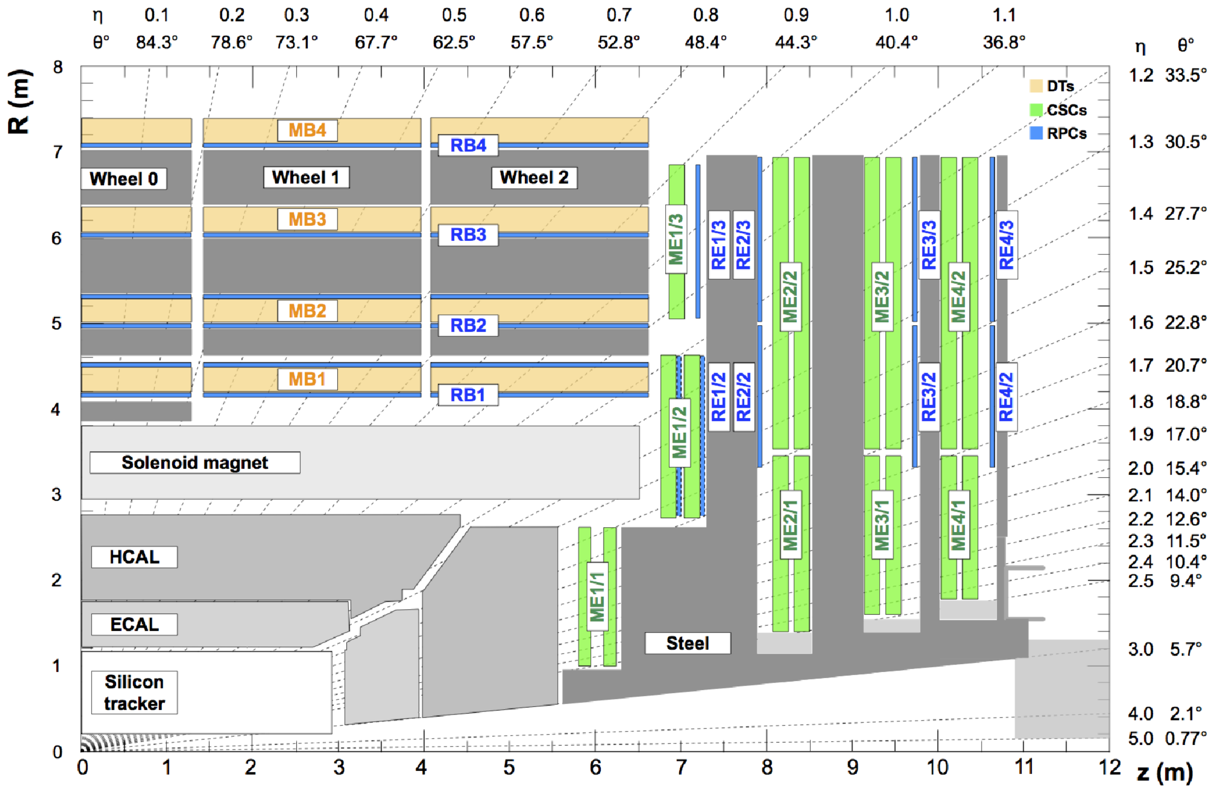


Figure 3.9: Schematic view of the muon chambers, interleaved with the flux-return yoke. Taken from [47].

3.2.7 Trigger system

The collision rate of 40 MHz makes it impossible to store all the data for processing, and most of the events are soft scattering collisions that are not interesting for physics analysis. The decision which events to keep is taken by a trigger system with two levels. The *Level-1* (L1) trigger is implemented in hardware and uses the information from the calorimeter and muon detector to make a decision within 4 μs after the collision and decrease the rate of events down to 100 kHz. The next level, for events which pass the L1 trigger, is the *high-level trigger* (HLT). Software optimized for fast processing, running on a processor farm, reconstructs the events and checks for signatures interesting for physics analysis. If an event passes the HLT, it is stored on disk and available for offline reconstruction. The average rate of events stored on disk is 1 kHz [49].

4 Object reconstruction

This chapter describes how particles are reconstructed from the tracks and energy deposits measured in the CMS detector. All the information from the subdetectors are collected to identify the particles, as explained in section 4.1. In the following sections, the algorithm used to connect all the information, called "Particle-flow", and the reconstruction of the objects used in the searches are described.
The Higgs boson and vector-like leptons cannot be directly identified, instead they are reconstructed from their decay particles. The Higgs boson is reconstructed by its decay to a tau lepton pair, with one tau lepton decaying to an electron or muon and the other decaying to jets. In the tau lepton decay, multiple neutrinos are also produced, which can only be reconstructed by the missing transverse energy. Additionally, at least one jet identified as coming from a bottom quark is required. In the vector-like lepton search one electron or muon, one tau lepton decaying to jets and three or four bottom quark jets are required.
The data is compared to simulation, for which the Monte Carlo method, described in section 4.9, is used.

4.1 Particle identification

The particle identification in CMS is based on the trace the particles leave in the different subdetectors. If particles have a short lifetime, they cannot be measured directly, but need to be reconstructed from their decay particles. Charged particles ionize the silicon of the tracker detector and leave a visible track. Neutral particles, like photons or neutral hadrons, can only be detected by their energy deposition in the calorimeters. Photons and electrons deposit their energy in the electromagnetic calorimeter by inducing an electromagnetic shower. Hadrons decay to a spray of particles, called a jet. The jets can be induced in the electromagnetic calorimeter and are stopped in the hadronic calorimeter. Muons are not stopped by the calorimeters because of their low interaction cross section with matter. They are stopped by the iron flux-return yoke and measured in the muon chambers. Tau leptons have a very short lifetime of 290 fs and can only be reconstructed by their decay to leptons and hadrons [11]. The neutrinos only interact very weakly with matter and leave the detector undetected. They are estimated as a momentum imbalance in the transverse plane by summing up all the transverse momenta of the particles from one primary vertex. The traces particles leave in the subdetectors are shown in Figure 4.1.
The objects are identified with specific identification algorithms. Those algorithms have different working points (WP), ranging from loose to tight. The tighter the WP, the less likely a misidentification is, but at the cost of a smaller sensitivity for the identification. Sensitivity refers to the ability of the detector to measure and identify particles.

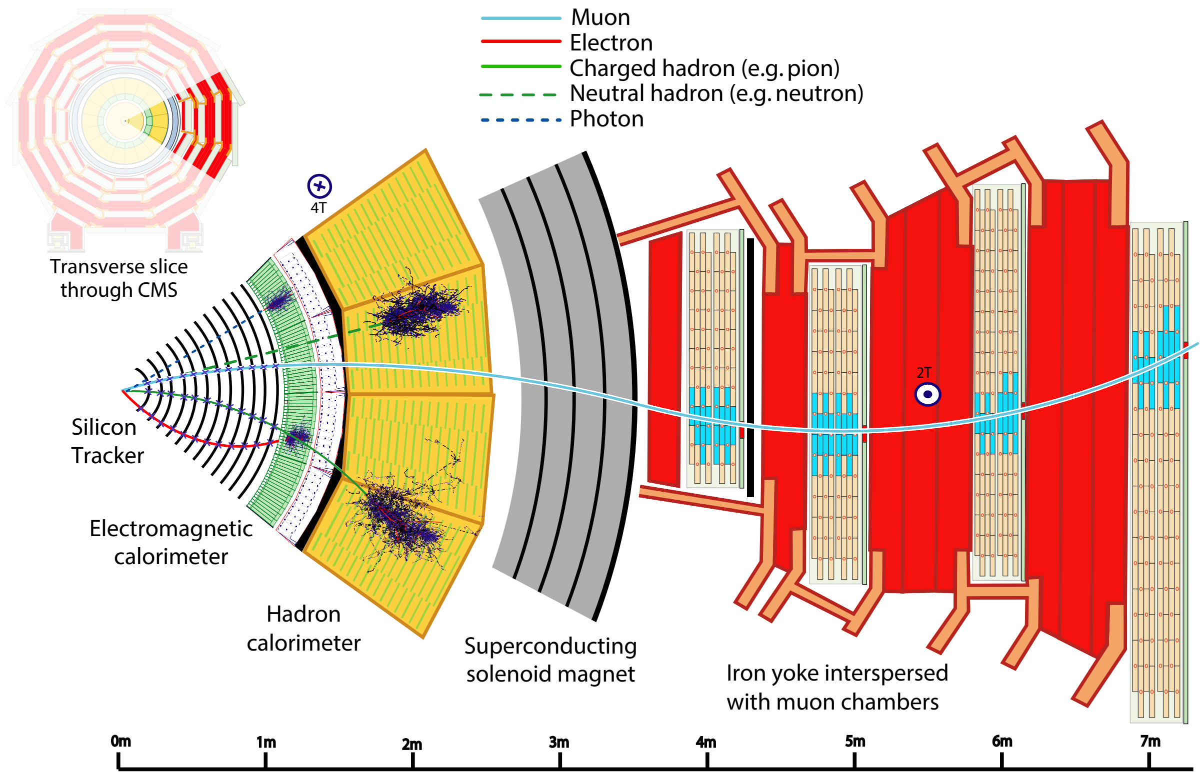


Figure 4.1: Particle identification in the CMS subdetectors. Adapted from [50].

4.2 Particle flow

Particle-flow [51] is an algorithm to reconstruct and identify particles by using the measured tracks from the tracker and muon system and the clusters from the calorimeters.

The tracks are reconstructed based on Kalman Filtering [52, 53] in three stages. The initial seed is generated if there are several hits compatible with a charged particle trajectory. The trajectory is extrapolated to other tracker layers, and hits are added if they match the trajectory. Finally, a fit is made to determine the origin, transverse momentum, and the direction of the particle.

The clusters in the calorimeters are identified by a clustering algorithm, except for the forward hadronic calorimeter, where no clustering is performed. A cluster seed is identified as cells with deposited energy above a certain threshold and larger than in the neighboring cells. The clusters are expanded by including neighboring cells with energy larger than twice the noise level.

The tracks are extrapolated to the calorimeters, and clusters are associated with the tracks if the extrapolated position is within the cluster boundaries.

Muons are identified by linking tracks with hits in the muon chambers. For the electrons, the frequent photon emissions due to Bremsstrahlung are accounted for by a dedicated reconstruction algorithm. All the tracks not associated with muons or electrons are identified as charged hadrons. Photons and neutral hadrons are identified as clusters in the ECAL and HCAL without corresponding track in the inner tracker. The identified particles, including muons, electrons, photons as well as neutral and charged hadrons, are then used to reconstruct jets, identify tau leptons, and determine the missing transverse energy.

4.3 Primary vertex

The measured tracks are extrapolated to the beam line to find the origin of the proton collisions, the vertices. Most of the collisions are soft-scattering and not interesting for physics analysis. One vertex gets identified as primary vertex (PV) if the vertex is close to the nominal interaction point and has the highest quadratic track momenta sum. Particles are associated with the PV if they have a small impact parameter, describing the perpendicular distance between the closest approach of the track to the PV.
All reconstructed tracks not associated to the primary vertex are considered pile up and removed by particle-flow in a process called charged hadron subtraction (CHS). It is not possible to link neutral particles from pile-up vertices to the primary vertex, requiring other correction methods to account for them in the reconstruction of jets, tau leptons, and missing transverse energy.

4.4 Electrons

The electrons are reconstructed by matching tracks with clusters in the ECAL if they are in the fiducial pseudorapidity range of the ECAL, up to $|\eta| < 2.5$. The electrons lose energy even before they reach the ECAL because of Bremsstrahlung radiation. This leads to changes in the curvature and the electron energy is spread over several ECAL crystals, especially in ϕ direction. Therefore, the tracks are refitted using Gaussian sum filter [54] and the clusters are combined to superclusters (SC). The transition region from the ECAL barrel to the endcaps ($1.444 < |\eta| < 1.566$) is excluded because of worse reconstruction performance in this region. The impact parameters in the x-y plane and in z direction have to be small to ensure the electron is from the PV.
In both searches in this thesis, the electrons are identified based on a multivariate analysis (MVA). A boosted decision tree was trained with simulated $Z \to ee$ samples with prompt electrons (electrons from hard-scattering in the PV) as signal and non-prompt electrons (electrons from decays and misidentified electrons) as background. The inputs include track-cluster matching, shower shapes, and tracking measurements. In the bbH search the electrons are required to pass the 90% efficiency WP, with no isolation components as inputs, and in the VLL search the 80% efficiency WP with isolation components as inputs. The identification is discussed in more detail in Ref. [55].
The isolation criteria for electrons sums the p_{T} of all particles inside a cone of $\Delta R = 0.3$ around the electron track, excluding the electron p_{T}, and estimates the contribution from pile-up with the $\rho - area$ method [56]. The isolation is defined as:

$$I^{e}_{\mathrm{rel}} = \frac{\sum_{\mathrm{charged}} p_{\mathrm{T}} + \max(0, \sum_{\mathrm{neutral}} p_{\mathrm{T}} - \rho A_{eff})}{p^{e}_{\mathrm{T}}}, \tag{4.1}$$

where $\sum_{\mathrm{charged}} p_{\mathrm{T}}$ is the sum of the transverse momenta of all charged particles, $\sum_{\mathrm{neutral}} p_{\mathrm{T}}$ is the sum of the transverse momenta of all neutral hadrons and photons. Only charged hadrons associated with the PV are included. The contribution from neutral hadrons not associated with the PV is estimated with ρA_{eff}, where ρ is the median energy density and A_{eff} is the electron effective area.
In the bbH search, the isolation requirement is set as $I^{e}_{\mathrm{rel}} < 0.1$. In the VLL search, the isolation is added as input for the MVA-based identification and is therefore not additionally applied.

4.5 Muons

Muons are reconstructed within the tracker detector acceptance up to $|\eta| < 2.5$ using the information from the tracker and the muon system. The muons are reconstructed as *standalone muons*, *tracker muons* or *global muons*. The *standalone muons* are reconstructed using only the hits in the muon system, and *tracker muons* using the hits from the silicon tracker detector and at least one hit in the muon chambers. *Global muons* are standalone muons that can be linked to tracks in the inner tracker. The impact parameters in the x-y plane and in z direction have to be small to ensure the muon is from the PV.

In the bbH search, the *medium* working WP is required for the muons. This includes the identification as a tracker or global muon and requirements on the quality of the track, with an efficiency of 98%.

In the VLL search, the *tight* WP is required for the muons, defined as a global muon with additional requirements on the quality of the muon, with an efficiency of 97%. Muons with a *loose* WP (99% efficiency) are used as veto in the electron channel. They have to be identified either as tracker or as global muon.

A more detailed description of the identification can be found in Ref. [47].

The isolation criteria for the muon takes all the particles inside a cone of $\Delta R = 0.4$ around the muon track into account, and is calculated as follows:

$$I^{\mu}_{\text{rel}} = \frac{\sum_{\text{charged}} p_{\text{T}} + \max(0, \sum_{\text{neutral}} p_{\text{T}} - \Delta\beta \sum_{\text{charged}} p_{\text{T}}^{PU})}{p_{\text{T}}^{\mu}}, \tag{4.2}$$

where $\sum_{\text{charged}} p_{\text{T}}$ is the sum of the transverse momenta of all charged particles, excluding the p_{T} of the muon itself, and $\sum_{\text{neutral}} p_{\text{T}}$ is the sum of the transverse momenta of all neutral hadrons and photons. Only charged hadrons associated with the PV are included. The contribution from neutral hadrons not associated with the PV is estimated with $\Delta\beta \sum_{\text{charged}} p_{\text{T}}^{PU}$, where $\sum_{\text{charged}} p_{\text{T}}^{PU}$ is the sum of the transverse momenta of all charged particles within the cone from pile-up vertices, and $\Delta\beta$ is the ratio of neutral to charged particles, estimated as 0.5.

In both searches the requirement on the muon isolation is $I^{\mu}_{\text{rel}} < 0.15$.

4.6 Taus

The hadronic decays of tau leptons are reconstructed with an algorithm called Hadron plus Strips (HPS) [57]. HPS skims all the jets and only keeps τ_{h} candidates if certain criteria are fulfilled: The charge of the τ_{h} has to be ± 1, the invariant mass matches the mass of intermediate resonances, and the charged and neutral hadrons must be in a small signal cone around the τ_{h} momentum. In the bbH and VLL search the considered decay modes are:

- DM0: one-prong + 0 π^0
- DM1: one-prong + 1 π^0
- DM10: three-prong + 0 π^0
- DM11: three-prong + 1 π^0

and additionally in the VLL search:

- DM2: one-prong + 2 π^0

- DM7: two-prong + 2 π^0

with $DM = 5\times(N_{\text{charged}}-1)+N_{\text{neutral}}$. Other decay modes are excluded because of negligible sensitivity.
The impact parameter in the z direction has to be small to ensure the τ_h is originating from the PV.
The isolation of hadronic tau leptons (τ_h) against electrons, muons and jets wrongly identified as hadronic tau candidates is performed with a convolutional deep neural network called DEEP-TAU [58]. It has four outputs with a relative probability that an object is an electron (y_e), a muon (y_μ), a jet (y_{jet}) or a genuine tau lepton (y_τ).
Three discriminators against electrons, muons, and jets are reconstructed:

$$\text{DeepTauVSe} = \frac{y_e}{y_\tau + y_e} \tag{4.3}$$

$$\text{DeepTauVSmu} = \frac{y_\mu}{y_\tau + y_\mu} \tag{4.4}$$

$$\text{DeepTauVSjet} = \frac{y_{jet}}{y_\tau + y_{jet}} \tag{4.5}$$

In the bbH search, the *medium* WP for the jet discriminator is used, which has an efficiency of 70% for genuine τ_h and 0.1% misidentification rate for quark and gluon jets [58]. In the $\mu\tau_\text{h}$ channel, the *tight* WP for the muon discriminator (99.5% efficiency) and the *very very very loose* WP for the electron discriminator (99.5% efficiency) are required. For the $\text{e}\tau_\text{h}$ channel, the *very loose* WP for the muon discriminator (99.95% efficiency) and the *tight* WP for the electron discriminator (80% efficiency) are used.
In the VLL search, the *tight* WP for the jet discriminator with an efficiency of 60% and a misidentification rate of 0.15% is required. For the muon and electron discriminator the *very loose* (99.95% efficiency) and *very very loose* WPs (99.5% efficiency) are sufficient [58].

4.7 Jets and b-tagging

Clustering algorithms combine all measured hadrons to jets. The algorithms must be infrared and collinear safe, which means that the radiation of a gluon does not change the jet properties and that the algorithm is not sensitive to a collinear splitting of a particle. The algorithm used in this thesis is ANTI-k_T [59]. The distance between particles i and j is defined as:

$$d_{ij} = \min(p_{T,i}^2, p_{T,j}^2)\frac{\Delta_{ij}^2}{\Delta R^2}, \tag{4.6}$$

where $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, $p_{T,i}$, y_i, and ϕ_i describe the transverse momentum, the rapidity and azimuthal angle of particle i. The rapidity is defined with the energy and momentum component along the beam axis as:

$$y = \frac{1}{2}\ln\left(\frac{E+p_z}{E-p_z}\right) \tag{4.7}$$

ΔR is the cone size of the jet. The distance between object i and the beam is defined as:

$$d_{iB} = p_{T,i}^2 \tag{4.8}$$

The algorithm iteratively pairs particles together with minimal d_{ij}. After each step, the pairs are replaced with an object with summed four momentum of particles i and j. If $d_{ij} > d_{iB}$, the object is considered a jet and removed from the list of pairs. The jets are clustered with a cone size of $\Delta R = 0.4$ using the $FASTJET$ library [60]. Charged hadrons linked to pile up vertices are removed using CHS, as mentioned already in section 4.3. In order to account for neutral hadrons from pile up vertices and detector noise, the energy of jets needs to be corrected in simulation, including *jet energy corrections* (JEC) and *jet energy resolution* (JER) [61].
The jets are required to pass identifications to distinguish genuine jets from pile-up jets [62] and to reject jets originating from HCAL detector noise [63].
In the bbH search, the jets have to be at least $\Delta R > 0.4$ away from any isolated electrons or muons.
In the VLL search, the jets are required to be at least $\Delta R > 0.4$ away from any hadronic tau leptons with a loose selection.
The bottom quarks, or b quarks for short, are identified by the existence of a secondary vertex, as discussed in section 4.9. B-tagging algorithms are based on artificial neural networks and identify jets originating from b quarks. They use track parameters as inputs and the output is a discriminator value from 0 to 1, stating the probability that a jet is from a b quark and not from other quarks or gluons.
In both searches the DeepJet b-tagging algorithm with a *medium* WP is used, corresponding to a b-tagging efficiency of 75%-82% depending on the jet p_{T} and a mis-tagging efficiency of light-flavor jets of 1% [64].

4.8 Missing transverse Energy

Particles not measured by the detector, neutrinos or other so far unknown particles, can be estimated with the missing transverse energy (MET). The transverse momenta of the partons inside the protons are assumed to be negligible in the hard scattering, therefore the MET can be estimated from the total transverse momentum imbalance summing up all reconstructed particles, after removing the contribution from pile-up.

$$\vec{p}_T^{\,\mathrm{miss}} = -\sum_i \vec{p}_T^{\,i} \tag{4.9}$$

The MET relies on a good calibration of all the reconstructed particles. Therefore, the jet energy corrections need to be propagated to the MET [65].

4.9 Event simulation

The data is compared to simulated events generated with the *Monte Carlo* method (MC). It is based on a large number of random generators that mimic the probabilistic nature of all the processes happening inside the detector. Each proton-proton collision is simulated as an event, from the collision to the resulting signals that would be measured in the detector. In a first step, the hard scattering, that happens when protons collide, is derived with matrix element (ME) calculations. The protons are composite particles made from two up and one down quark, called valence quarks, gluons and quark-antiquark pairs of sea quarks, which are created by quantum fluctuations of the gluon field inside the proton. In high energy proton collisions at the LHC, to a good approximation, all of the momentum is exchanged between one parton of each proton. The probability to find a parton with a momentum fraction x in a collision at a certain momentum scale Q is given by the parton distribution function $f(x, Q)$ (PDF). Most of the momentum is exchanged between gluons and valence quarks, less by the sea quarks. The cross section of a hard scattering process $pp \to X$ can be calculated with:

$$\sigma(pp \to X) = \sum_{i,j} \int_0^1 \int_0^1 f_i(x_1, Q^2) f_j(x_2, Q^2) \hat{\sigma}_{ij \to X}(x_1, x_2, Q^2) dx_1 dx_2, \tag{4.10}$$

where $\hat{\sigma}_{ij \to X}(x_1, x_2, Q^2)$ is the cross section at parton level. It is summed over all the partons of both protons, i and j.

In the calculation of the PDFs two scales play an important role, the renormalization scale (μ_R), defining the scale at which the strong coupling constant α_s is renormalized, and the factorization scale (μ_F), which is the scale that separates long-distance (non-perturbative) effects from short-distance (perturbative) effects.

Different flavor schemes are used for generating events, the five-flavor scheme (5FS) where the bottom quarks are considered as partons and the four-flavor scheme (4FS) where they are not.

The next step is the simulation of parton showers (PS). Initial- and final-state radiation (ISR, FSR), describing the emission of particles before and after the hard scattering process, are simulated. In the same step, hadronization is also simulated, referring to the forming of composite particles from partons.

The ME calculation has to be merged with the PS simulation and emissions in an overlapping phase space have to be removed. There are two main methods used, either the subtraction of the ME calculation emissions with aMC@NLO [66] or the subtraction of emissions from the PS with POWHEG [67–69]. Both of these generators calculate the process at a next-to-leading (NLO) order accuracy, which includes additional quark loops. For some samples in the VLL search also the base MadGraph [70] generator is used, where the processes are calculated at leading-order (LO). The generator for the simulation of the parton shower and hadronization used in the bbH and VLL search is PYTHIA [71].

In the last step, the CMS detector response is modeled, where every simulated event is treated in the same way as data. The detector response is simulated with the Geant4 simulation [72].

5 Search for the bottom quark associated production of the Higgs boson

5.1 Motivation

This chapter describes the search for a Higgs boson produced in association with two bottom quarks (bbH) in the final states with $H \to \tau\tau$ and $H \to$ WW. The coupling of the Higgs boson to third generation fermions has been measured by the CMS and ATLAS Collaborations in the Higgs boson decays to $\tau\tau$ [73, 74] and bb [75, 76], and in the production associated with top quarks [77, 78]. As mentioned in Section 2.2, out of the production modes that have not been observed yet, the bbH production mode has the largest production cross section [79, 80]. The bbH production mode is difficult to observe, because of its small cross section and because the largest contribution to the final state of the Higgs boson decay products plus two bottom quarks is from Higgs production through gluon fusion with a gluon splitting into a bottom quark pair [81, 82]. In this process, the Higgs boson is produced through an intermediate quark loop which mainly involves top quarks, therefore the production cross section is proportional to the top quark Yukawa coupling (y_t^2). The production mode with the second largest cross section is the bottom quark fusion, with a production cross section proportional to the bottom quark Yukawa coupling (y_b^2). The production cross sections are listed in Tab. 5.1, together with a destructive interference term ($y_b y_t$) which also has to be considered. The total production cross section including all the considered contributions is 1.489 pb. Another process contributing to the final state is the Higgs-strahlung with a Z boson decaying into a bottom quark pair. The different contributions to the topology are shown in Fig. 5.1. The contributions from y_b and y_t are both considered as signal, because of very similar variable distributions, that makes it very hard to separate them. This also has the effect that only simultaneous limits can be set on the Higgs boson coupling to bottom quarks and to top quarks. In the search, only processes involving the Higgs boson coupling to top and bottom quarks are considered, therefore ZH is treated as background. Measuring the bottom quark associated production of the Higgs boson is crucial to separate this process from the double Higgs boson production signal, which is the focus of future searches at the CMS experiment.

Process	σ (pb)
bbH(y_t^2)	1.040 (+0.468, -0.489)
bbH(y_b^2)	0.482 (+0.048, -0.070)
bbH($y_b y_t$)	-0.033 (+0.007, -0.008)
bbH	1.489

Table 5.1: Signal contributions with their cross sections. The interference term is destructive and thus has a negative cross section. Values taken from [81, 84].

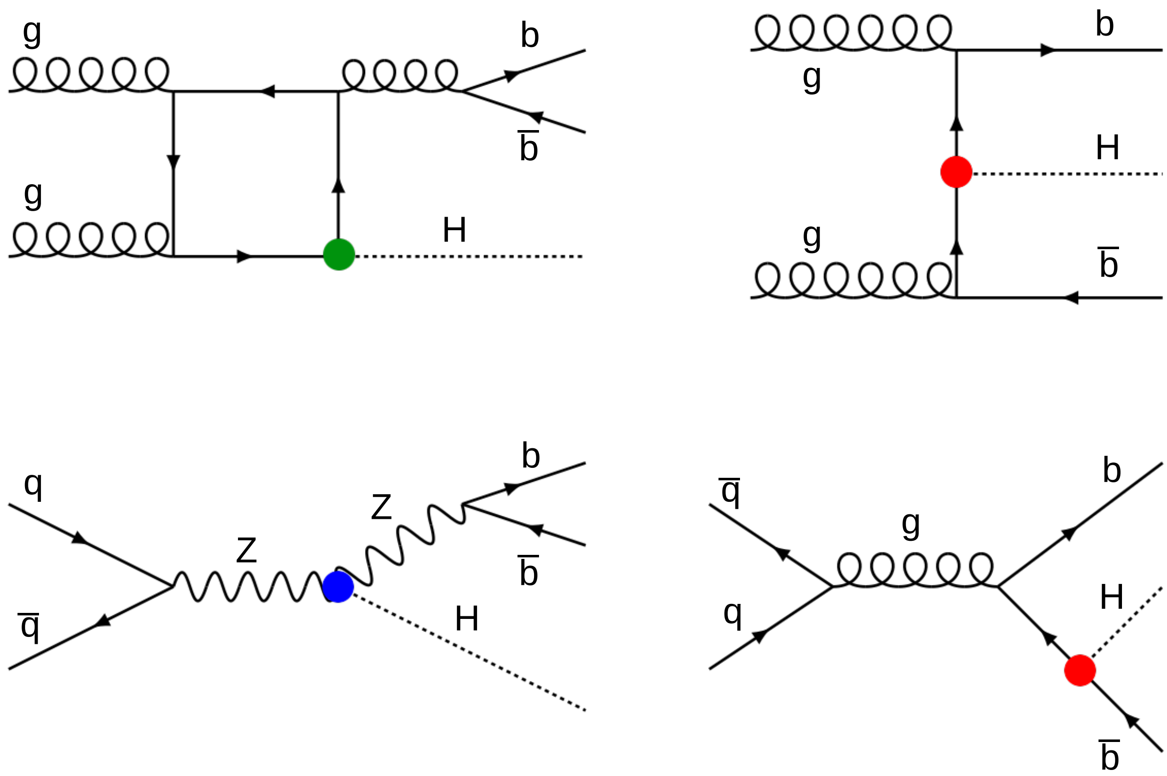


Figure 5.1: Feynman diagrams contributing to the Higgs boson production in association with bottom quarks: Gluon fusion with gluon splitting (top left), bottom quark fusion (top right), radiation of Higgs boson of a bottom quark (bottom right) and the Z boson associated Higgs boson production with a Z boson decaying to bottom quark pair (bottom left). The Higgs boson coupling to bottom quarks is marked with a red dot, the coupling to top quarks with a green dot and the coupling to vector bosons with a blue dot. Figure taken from [83].

5.2 Overview

The search for the bottom quark associated production of the Higgs boson is conducted in five final states: $H \to \tau\tau$ with the final states $\tau_h\tau_h$, $\mu\tau_h$, $e\tau_h$ and $e\mu$, and $H \to WW$ for the $e\mu$ final state. The largest branching fraction for a tau lepton pair decay is into two hadronically decaying tau leptons. The semileptonic tau lepton pair decays have a similar branching ratio and describe the decay of one tau lepton into a light lepton, referring to a muon or electron, and the other tau lepton decaying to hadrons. The fully leptonic decay with one tau lepton decaying to a muon and the other to an electron has the smallest branching ratio. In the $e\mu$ final states also the contribution from $H \to WW$ is considered, which has a similar branching ratio in this final state. The final states and their branching ratios are shown in Tab. 5.2.

final state	$\mathcal{B}(H \to \tau\tau/WW)$	$\mathcal{B}(\tau\tau/WW \to$ final state$)$	$\mathcal{B}(H \to$ final state$)$
$H \to \tau\tau \to \tau_h\tau_h$	6.3%	42%	2.65%
$H \to \tau\tau \to \mu\tau_h$		22%	1.39%
$H \to \tau\tau \to e\tau_h$		23%	1.45%
$H \to \tau\tau \to e\mu$		6%	0.38%
$H \to WW \to e\mu$	22%	2.3%	0.51%

Table 5.2: Final states in which the search is conducted along with their corresponding branching ratios ($\mathcal{B}$). Values from [11, 84].

The analysis is divided into four channels based on the final states, because each final state has different background processes that require different techniques. I led the analysis in the two semileptonic channels ($\mu\tau_h$, $e\tau_h$). The following sections will describe the work done in those two channels. A brief summary of the fully hadronic channel and the fully leptonic channel will be given in section 5.10. The results in section 5.11 will include all final states and their combination. The region where we expect to observe the signal is referred to as signal region. There are other physics processes with the same final state that can contribute to the signal region, which we refer to as *background* processes. The main background arises from $t\bar{t}$ events, where the top quark decays into a bottom quark and a W boson. The W boson decays into a lepton or hadrons, which produce jets that can be misidentified as hadronic tau leptons, resulting in the same final state as the signal process. Other major backgrounds are Drell-Yan and W boson plus additional jets, and a fake τ_h background. Drell-Yan refers to a virtual photon or Z boson produced through quark-antiquark annihilation and decaying to leptons. The fake τ_h background comes from gluon or quark initiated jets misidentified as a hadronic tau decay. The fake τ_h contribution is estimated using a data-driven method called fake-factor method, described in more detail in section 5.6. Minor backgrounds are diboson, single top quarks and other Higgs boson production modes including vector-boson fusion and Higgs boson production associated with top quarks or vector bosons. The background contributions, except the fake τ_h, are estimated from simulated Monte Carlo samples, described in section 5.3.3. Because differences between data and simulation occur, corrections are applied to the simulation. The corrections are described in detail in section 5.4.

In the $\mu\tau_h$ channel, events are selected by a single muon trigger, and in the $e\tau_h$ channel by a single electron trigger. In the signal region one muon or electron, one hadronic tau lepton and at least one bottom quark jet are required. The charge of the light lepton and the hadronic tau lepton have to be opposite and the separation defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ between the light lepton and the hadronic tau lepton has to be larger than 0.5. The transverse mass, defined in eq. 5.5, is required to be lower than 60 GeV to reduce the $t\bar{t}$ and W+jets background. In the $\mu\tau_h$ channel a veto on electrons and on di-muon pairs close to each other and with opposite charge are set. And similar for the $e\tau_h$ channel, but there with a veto on muons and di-electron pairs. The selections used for the $\mu\tau_h$ and $e\tau_h$ channel are shown in Tab. 5.3.

$\mu\tau_h$ channel	$e\tau_h$ channel
1 μ, 1τ_h	1 e, 1τ_h
$\geq$ 1 b-jet	$\geq$ 1 b-jet
$m_T < 60$ GeV	$m_T < 60$ GeV
$q_\mu q_{\tau_h} = -1$	$q_e q_{\tau_h} = -1$
$\Delta R(\mu, \tau_h) > 0.5$	$\Delta R(e, \tau_h) > 0.5$
veto on e and 2μ	veto on μ and 2e

Table 5.3: Signal region selection requirements for the $\mu\tau_h$ and $e\tau_h$ channel.

In order to enhance the signal to background separation, a multiclass Boosed Decision Tree (BDT) is trained for each year and channel with three classes: $t\bar{t}$, Drell-Yan and signal. The input variables to the BDT include low-level and high-level variables. Low-level variables refer to parameters measured in the detector, like the kinematics from jets. High-level variables refer to reconstructed variables using multiple objects, like the Higgs boson mass. The BDT is described in more detail in section 5.7. The output of the BDT is simultaneously fit for all channels to derive upper limits on the signal process cross section. Limits are simultaneously set on the Higgs boson coupling to bottom quarks and to top quarks.

5.3 Data and simulation

This search uses data collected by the CMS experiment in the years from 2016 to 2018. The data is compared with Monte Carlo simulation, described in section 4.9.

5.3.1 Data

The data used in the analysis was collected by the CMS experiment from 2016 to 2018 with proton-proton collisions at a center of mass energy of 13 TeV. The integrated luminosity for 2016 amounts to 36.3 fb^{-1}, 41.5 fb^{-1} for 2017 and 59.97 fb^{-1} for 2018, which gives a combined integrated luminosity of 138 fb^{-1}. 2016 is split in two different eras, pre and postVFP (Vertex Fitting Procedure) because of a substantial change of the conditions in the strip tracker detector during the 2016 data taking. Data events are selected by a set of single muon and single electron triggers shown in Tab. 5.4. Multiple triggers are used in order to achieve optimal efficiency across the p_T and pseudorapidity range, and to ensure redundancy. Depending on the year, different combinations of triggers have to be used. In 2017 and 2018 the triggers with the lower p_T were prescaled, which means that only a certain fraction of events got stored in order to reduce the amount of data. The prescaling became essential due to the higher instantaneous luminosity in those years. For the single isolation muon triggers of 2016, including triggers that require a tracker muon instead of a global muon (as described in section 4.5) was found to provide the best efficiency.

channel	Trigger type	p_T thresholds [GeV]
$\mu\tau_h$	Single isolated muon	22 in 2016 24 or 27 in 2017 & 2018
$e\tau_h$	Single electron	25 or 27 in 2016 32 or 35 in 2017 & 2018

Table 5.4: Trigger requirements in the $\mu\tau_h$ and $e\tau_h$ channel.

5.3.2 Signal simulated samples

As introduced in section 5.1, the signal consists of a bottom quark fusion process ($bbH(y_b^2)$), a gluon fusion process with gluon splitting to bottom quark pair ($bbH(y_t^2)$) and a destructive interference term ($bbH(y_b y_t)$). They are simulated with the MadGraph5_aMC@NLO generator at next-to-leading order (NLO) in QCD in the 4-flavor scheme. Events from the b-fusion and interference sample are selected if they have at least one generator level bottom quark jet, otherwise they are considered as background. The gluon fusion sample $bbH(y_t^2)$ is stitched with an inclusive gluon fusion sample. Events from $bbH(y_t^2)$ are selected if there are two generator level bottom quark jets or if there is one generator level bottom quark jet with at least two generator level B hadrons matched to it with $\Delta R < 1.5$. Events from the inclusive gluon fusion sample are selected if they have one generator level bottom quark jet with one generator level B hadron matched to it. Events with no generator level b-jets are considered as background. The reason for the stitching is to cover the full phase space. The $bbH(y_t^2)$ sample has two B hadrons in the matrix element, but the signal is defined with one bottom quark jet at generator level. The cross section for the two B hadron phase space is different in the $bbH(y_t^2)$ sample compared to the inclusive sample. The choice of matching the B hadrons to b-jets with $\Delta R < 1.5$ is taken to ensure the cross sections are similar enough, while also ensuring that a large fraction of the

Process	Generators	$\sigma \times \mathcal{B}$ [pb]
$bbH(y_t^2) \to \tau\tau$	aMC@NLO, PYTHIA	1.040×0.06208
$bbH(y_b^2) \to \tau\tau$	aMC@NLO, PYTHIA	0.4822×0.06208
$bbH(y_b y_t) \to \tau\tau$	aMC@NLO, PYTHIA	-0.033×0.06208
$ggH \to \tau\tau$	aMC@NLO, PYTHIA	48.52×0.06208

Table 5.5: Simulated samples for signal processes with the production cross section times branching ratio. Production cross section values are from [81, 84], based on calculations from Refs. [85–91].

$bbH(y_t^2)$ sample is classified as signal.
The simulated signal samples and their cross sections are shown in Tab. 5.5.

5.3.3 Background simulated samples

Several processes have the same final state with a tau lepton pair and bottom quark jets as our signal. The largest contribution come from $\mathrm{t\bar{t}}$, Drell-Yan plus additional jets and quarks or gluon initiated jets misidentified as τ_h, mainly from the $\mathrm{t\bar{t}}$, W+jets and QCD multijet background. The simulated background samples with the generators used and their cross sections are shown in Tab.5.6.
$\mathrm{t\bar{t}}$: This process decays predominantly via $\mathrm{t\bar{t}} \to b\bar{b}W^+W^-$, as shown in Fig. 5.2. The final state can contain genuine tau leptons from the leptonic decay of the W boson or jets misidentified as τ_h. The $\mathrm{t\bar{t}}$ background process is generated with the POWHEG generator at NLO in QCD.

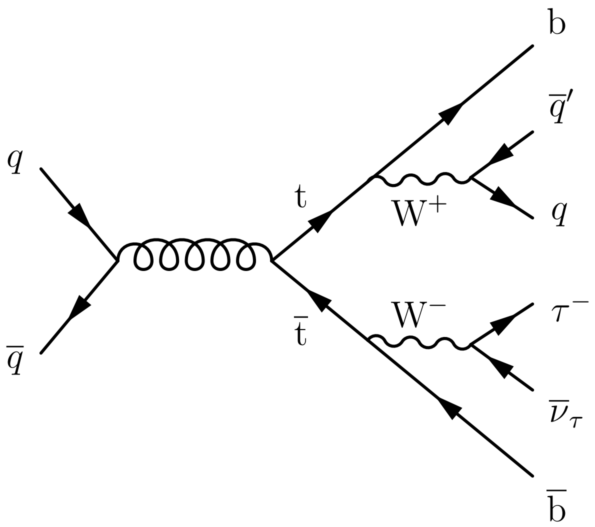


Figure 5.2: Feynman diagram of $\mathrm{t\bar{t}}$ background process, showing one possible final state where one W boson decays into leptons and the other W boson decays into hadrons. The production of $\mathrm{t\bar{t}}$ can also occur through gluon-gluon fusion, which is not shown in this figure.

Drell-Yan+jets: Drell-Yan refers to the $Z/\gamma* \to \ell^+\ell^-$ process, which decays in one third of the time to a tau lepton pair, and can be accompanied by jets from initial or final state radiation. A Feynman diagram of the process is shown in Fig. 5.3. The $\mu\tau_\mathrm{h}$ and $\mathrm{e}\tau_\mathrm{h}$ final states could also be produced by a Drell-Yan process to a light lepton pair with one light lepton misidentified as τ_h, but most of this contribution is suppressed by applying the DeepTau lepton discriminators, as described in section 4.6. Aside from the DeepTau jet discriminator also vetoing additional muons and electrons helps in suppressing this background. The Drell-Yan+jets background is simulated at NLO in QCD with the MadGraph5_aMC@NLO generator. An inclusive sample is stitched with jet binned samples to increase the statistics. The events in the inclusive sample and the jet binned samples are weighted to ensure that the total event yield remains as if only the inclusive sample is used. The weight applied to the inclusive sample depends on the number of partons at NLO (denoted with N_p):

$$w_{\mathrm{incl},N_p} = \frac{\frac{\sigma_{N_p}}{\sigma_{\mathrm{sum}}} \cdot N_{\mathrm{eff}}^{\mathrm{incl}}}{\frac{\sigma_{N_p}}{\sigma_{\mathrm{sum}}} \cdot N_{\mathrm{eff}}^{\mathrm{incl}} + N_{\mathrm{eff}}^{N_p}}, \tag{5.1}$$

where $N_{\mathrm{eff}} = \mathrm{sum}w_{\mathrm{gen}}/w_{\mathrm{gen}}$ with w_{gen} the generator level weight per event, sumw_{gen} the sum of all generator level weights, σ is the cross section of the sample and $\sigma_{\mathrm{sum}} = \sigma_{0\mathrm{J}} + \sigma_{1\mathrm{J}} + \sigma_{2\mathrm{J}}$.
The weight applied to the jet binned samples depends on the number of jets (N_J):

$$w_{N_J} = \frac{N_{\mathrm{eff}}^{N_J}}{\frac{\sigma_{N_J}}{\sigma_{\mathrm{sum}}} \cdot N_{\mathrm{eff}}^{\mathrm{incl}} + N_{\mathrm{eff}}^{N_J}}. \tag{5.2}$$

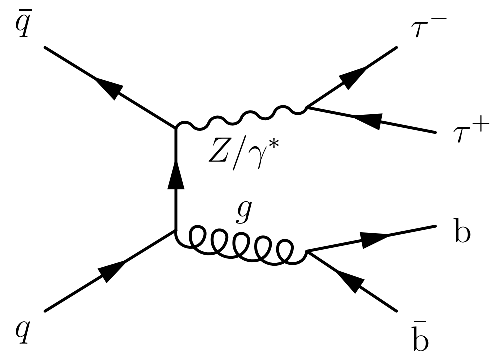


Figure 5.3: Feynman diagram of the Drell-Yan+jets background process with the Z boson decaying to a tau lepton pair and a gluon splitting into a bottom quark-antiquark pair.

Fake τ_h: The background from jets misidentified as hadronic tau decays arises from the $\mathrm{t\bar{t}}$, W+jets and QCD multijet processes. The $\mu\tau_\mathrm{h}$ and $\mathrm{e}\tau_\mathrm{h}$ final states are produced in the $\mathrm{t\bar{t}}$ and W+jets processes by the leptonic decay of the W boson and one jet misidentified as hadronic tau, and in the QCD multijet because of leptons from hadron decays. The fake τ_h background is estimated with a data-driven method, because it is difficult to model the misidentification rate from simulation. The fake-factor method is described in section 5.6.
Minor backgrounds arise from W+jets, single top quarks, two vector bosons and other Higgs boson production modes.
W+jets: The W+jets simulation is mainly used for the estimation of the fake τ_h background. The simulated samples are generated with MadGraph5_aMC@NLO at NLO in QCD, combining an inclusive sample, requiring zero jets at the Matrix Element level, with jet binned samples.
Single top: Samples for the production of a single top quark are simulated with the POWHEG generator at NLO in QCD. Inclusive t-channel samples are produced in the 4-flavor, tW samples in the 5-flavor scheme. For the s-channel samples only the leptonic decay channels are considered.
Diboson: The diboson background includes WW, WZ and ZZ with their semi-leptonic final states. They are simulated at NLO in QCD with the MadGraph5_aMC@NLO generator.
$H \to \tau\tau$: The background from other Higgs boson production modes include vector-boson-fusion, vector boson associated production and top quark associated production. The samples are simulated at NLO in QCD with the POWHEG generator.

Process	Generators	Cross section σ [pb]
$t\bar{t}$ + jets		
Fully leptonic	POWHEG, PYTHIA	88.29 (NNLO)
Semi-leptonic	POWHEG, PYTHIA	365.34 (NNLO)
Fully Hadronic	POWHEG, PYTHIA	377.96 (NNLO)
DY+jets, $Z/\gamma^* \to \ell^+\ell^-$		
+ jets, $m_{\ell\ell} = 50\,\text{GeV}$	aMC@NLO, PYTHIA	6077.22 (NNLO)
+ 0 jets, $m_{\ell\ell} = 50\,\text{GeV}$	aMC@NLO, PYTHIA	5125 (NLO)
+ 1 jets, $m_{\ell\ell} = 50\,\text{GeV}$	aMC@NLO, PYTHIA	951.4 (NLO)
+ 2 jets, $m_{\ell\ell} = 50\,\text{GeV}$	aMC@NLO, PYTHIA	358.6 (NLO)
W + jets, $W \to \ell\nu$		
+ jets	aMC@NLO, PYTHIA	61526.7 (NNLO)
+ 0 jets	aMC@NLO, PYTHIA	53300 (NLO)
+ 1 jets	aMC@NLO, PYTHIA	8949 (NLO)
+ 2 jets	aMC@NLO, PYTHIA	3335 (NLO)
Single top		
$t + W^-$	POWHEG, PYTHIA	35.85 (NNLO)
$\bar{t} + W^+$	POWHEG, PYTHIA	35.85 (NNLO)
Single t, s channel ($t \to b\ell^+\nu_\ell$)	POWHEG, PYTHIA	7.2 (NNLO)
Single $\bar{t}$, s channel ($\bar{t} \to \bar{b}\ell^-\bar{\nu}_\ell$)	POWHEG, PYTHIA	4.16 (NNLO)
Single t, t channel	POWHEG, PYTHIA	136.02 (NNLO)
Single $\bar{t}$, t channel	POWHEG, PYTHIA	80.95 (NNLO)
Diboson		
$WW \to \ell\nu qq$	POWHEG, PYTHIA	45.99 (NLO)
$WW \to \ell^+\ell^-\nu\nu$	POWHEG, PYTHIA	11.08 (NLO)
$WW \to qqqq$	POWHEG, PYTHIA	47.73 (NLO)
$WZ \to \ell^+\ell^-\ell\nu$	aMC@NLO, PYTHIA	5.052 (NLO)
$WZ \to \ell^+\ell^- qq$	aMC@NLO, PYTHIA	6.331 (NLO)
$WZ \to \ell\nu\nu\nu$	aMC@NLO, PYTHIA	3.3 (NLO)
$WZ \to \ell\nu qq$	aMC@NLO, PYTHIA	11.66 (NLO)
$ZZ \to \ell^+\ell^-\ell^+\ell^-$	aMC@NLO, PYTHIA	1.369 (NLO)
$ZZ \to \ell^+\ell^- qq$	aMC@NLO, PYTHIA	3.688 (NLO)
$ZZ \to qq\nu\nu$	aMC@NLO, PYTHIA	4.561 (NLO)
$H \to \tau\tau$		$\sigma \times \mathcal{B}$
VBF $H \to \tau\tau$	POWHEG, PYTHIA	3.771×0.06208 (NLO)
$W^-H \to \tau\tau$	POWHEG, PYTHIA	0.831×0.06208 (NLO)
$W^+H \to \tau\tau$	POWHEG, PYTHIA	0.527×0.06208 (NLO)
$ZH \to \tau\tau$	POWHEG, PYTHIA	0.877×0.06208 (NLO)
$ttH \to \tau\tau$	POWHEG, PYTHIA	0.503×0.06208 (NLO)

Table 5.6: Summary of simulated background processes with the generators and the corresponding production cross sections.

5.4 Data and MC comparison and corrections

The simulated samples can differ from data because of mismodeling of the reconstruction and identification of particles, of the detector noise or because of mismodeled physics processes. Corrections are applied to the simulated samples to improve the agreement with the observed data. The corrections are applied as scale factors or as energy scales. The scale factor is defined as the ratio of the efficiency in data ϵ_{Data} over the efficiency in the simulated Monte Carlo (MC) samples ϵ_{MC}: SF = $\epsilon_{\mathrm{Data}}/\epsilon_{\mathrm{MC}}$. The lepton efficiency is measured using a tag-and-probe technique, which selects lepton pairs based on the tag lepton and measures the efficiency with the probe lepton. This ensures that the probe leptons are not biased by any other selection other than the one for which the efficiency is measured. The efficiency is measured by comparing the number of probe leptons that pass the identification requirements, for which the efficiency is evaluated, to the total number of probe lepton. The numbers are extracted from a fit to the reconstructed lepton pair mass around the Z boson mass peak. This ensures that the probe lepton is indeed the lepton for which we want to measure the efficiency.
The energy scale corrections are applied to the p_{T} and mass of the jets, as discussed in section 4.7, the p_{T} and ϕ of the $p_{\mathrm{T}}^{\mathrm{miss}}$, described in section 4.8, or as a multiplicative factor to the τ_{h} four-momentum, defined as $(p_{\mathrm{T}}, \eta, \phi, m)$. The trigger scale factors and the normalization corrections were derived by me specifically for this analysis. The other corrections are provided by the respective Physics Object Group (POG) of the CMS Collaboration.

5.4.1 Pileup Reweighting

The simulated samples are generated with a pileup distribution that approximates the conditions during data taking. In order to ensure an accurate modeling of the pileup effects, it is important to apply pileup reweighting to the simulated samples. The mean number of interactions per bunch crossing is rebinned in 100 bins and reweighted in each bin to match the distribution in data. The number of interactions per bunch crossing in data is estimated as the product of the instantaneous luminosity and assuming a total inelastic cross section of $\sigma_{pp} = 69.2$ mb.

5.4.2 Trigger corrections

The scale factors for the single muon and the single electron triggers were privately derived for the analysis. The efficiency is measured in data and in an inclusive Drell-Yan plus additional jets simulated sample. The muons and electrons are required to satisfy the same selection required for the signal region, defined in chapter 4 and section 5.5. Two muons or two electrons are required for the measurement of the trigger efficiency. The tag leptons are selected if they activated the single lepton trigger for which the efficiency is measured with the probe leptons. The tag leptons must be matched to the single lepton trigger object within $\Delta R < 0.2$, the probe leptons do not have any additional selection requirements. The efficiency is measured as a function of the lepton p_{T}, with small bins close to the trigger threshold, and the lepton η in the di-lepton mass distribution in the range of the Z boson mass from 75 to 125 GeV. The measurement is conducted in a *pass* and a *fail* region. The *pass* region includes all events where the probe lepton activated the trigger, the *fail* region all the events where the probe lepton did not activate the trigger. A maximum likelihood fit is performed simultaneously on the pass and on the fail region. A superposition of two Voigtian functions with the width of the Lorentzian part set to the Z decay width of 2.495 GeV is used as signal model, while the background is modeled with an exponential. The fit in the pass and the fail region is shown in Fig. 5.4. The measured efficiencies are shown in Figs. 5.5 and 5.6 as an example for the 2018 $\mu\tau_h$ and $e\tau_h$ channel.

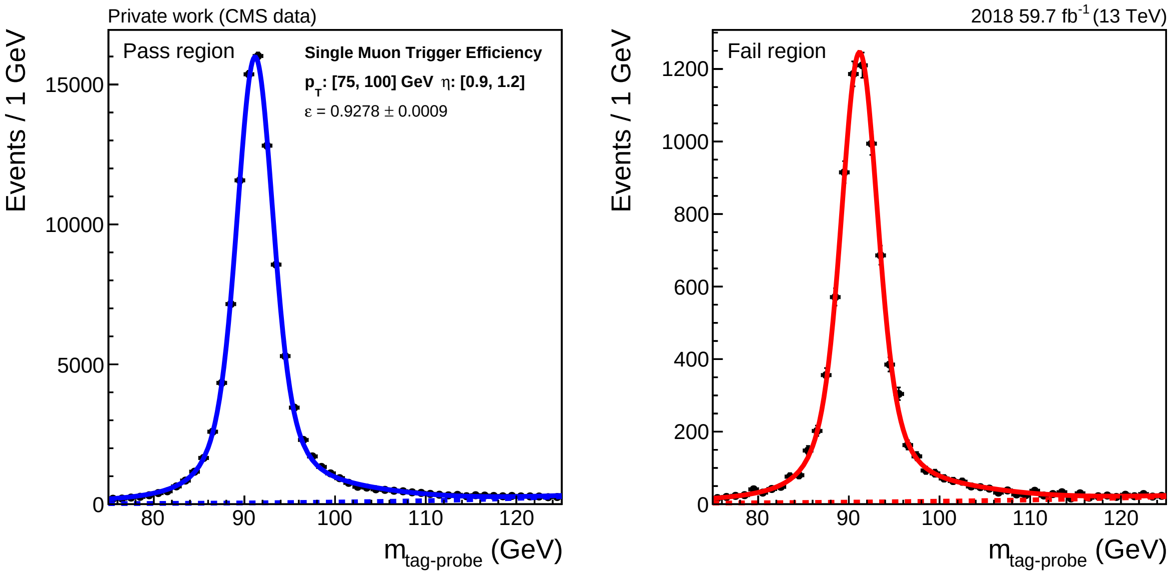


Figure 5.4: Fit to the dilepton mass in the pass (left) and the fail region (right) for a p_T range from 75 to 100 GeV and $0.9 < |\eta| < 2.1$ for the single muon triggers of 2018.

The scale factors are shown as function of the lepton p_T and for each η bin in Fig. 5.7 for 2018, 2017, 2016 post VFP and 2016 pre VFP.

Systematic uncertainties for the scale factors are derived by changing the mass window for the dilepton mass distribution upwards (80 GeV to 130 GeV) or downwards (70 GeV to 120 GeV), by using alternative fit functions and by using the generator level dilepton mass distribution for the simulated samples.

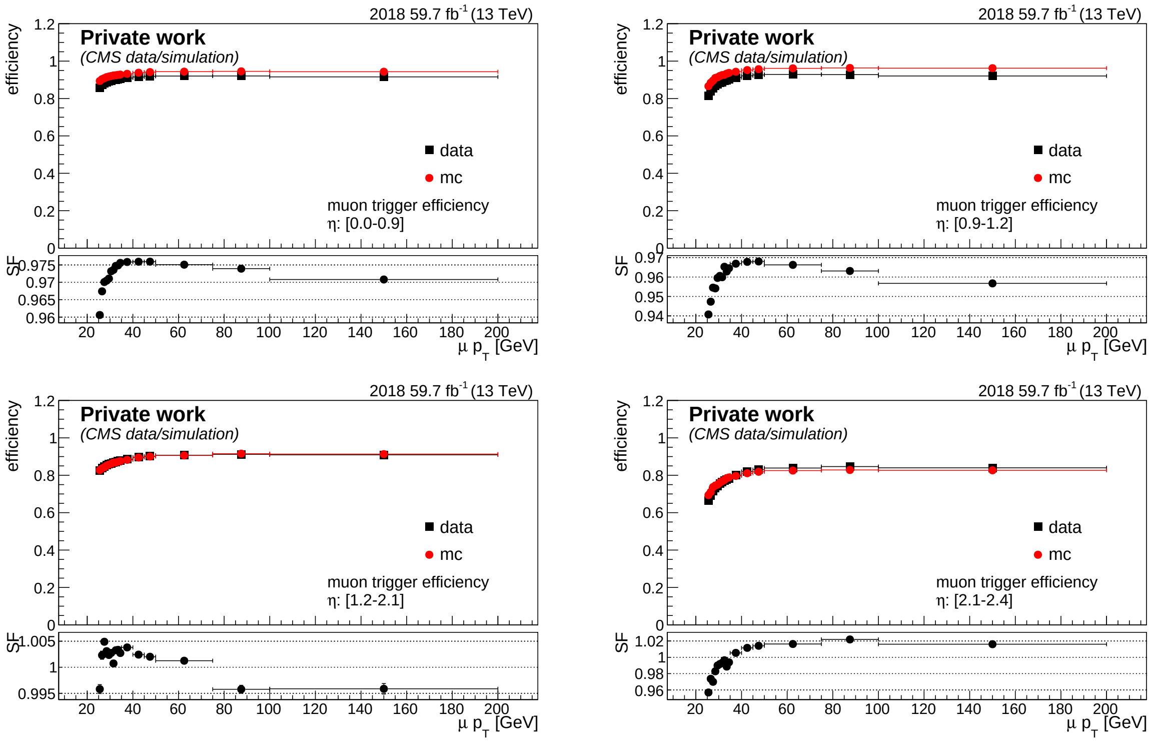


Figure 5.5: Trigger efficiency for 2018 single muon trigger as function of the muon p_T in the range $0.0 \leq |\eta| \leq 0.9$ (top left), $0.9 \leq |\eta| \leq 1.2$ (top right), $1.2 \leq |\eta| \leq 2.1$ (bottom left) and $2.1 \leq |\eta| \leq 2.4$ (bottom right).

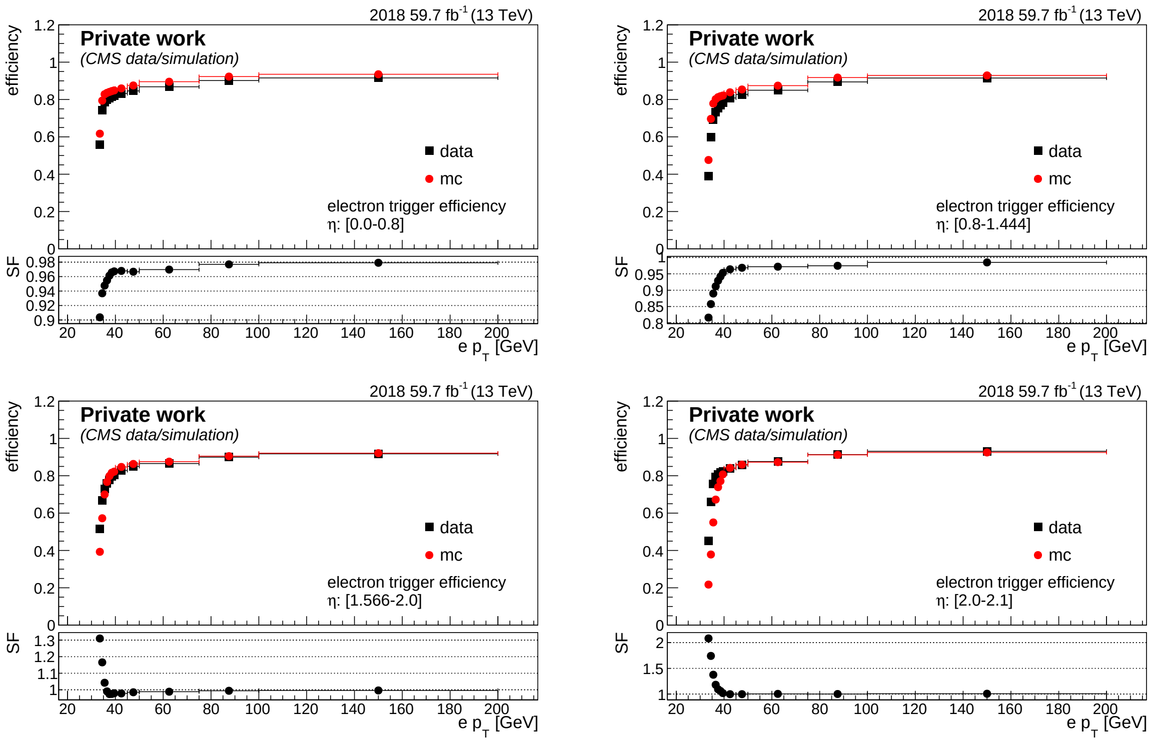


Figure 5.6: Trigger efficiency for 2018 single electron trigger as function of the electron p_T in the range $0.0 \leq |\eta| \leq 0.8$ (top left), $0.8 \leq |\eta| \leq 1.444$ (top right), $1.556 \leq |\eta| \leq 2.0$ (bottom left) and $2.0 \leq |\eta| \leq 2.1$ (bottom right). The $|\eta|$ range from 1.444 to 1.566 is between the barrel and the endcaps of the ECAL and excluded in the analysis.

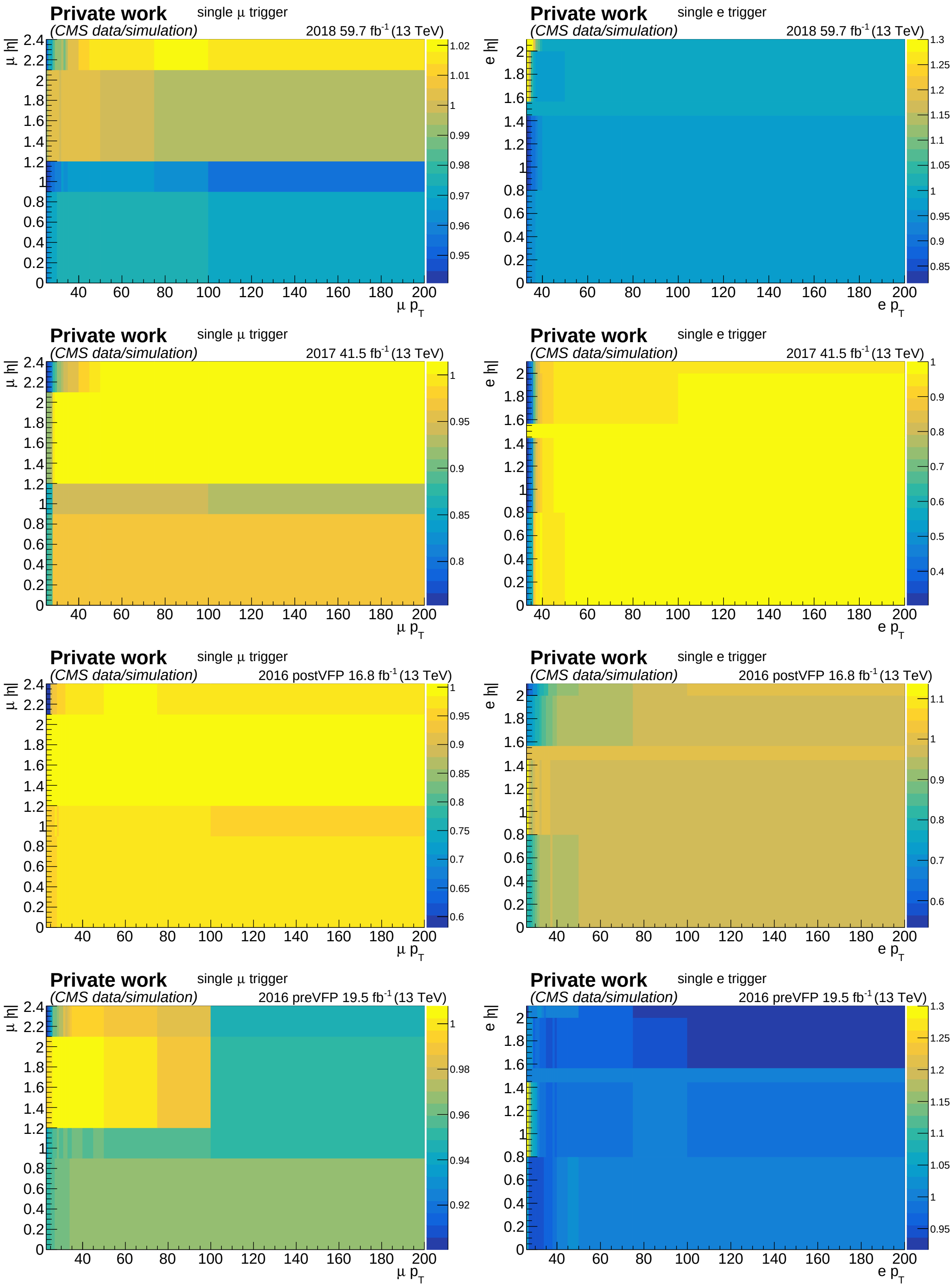


Figure 5.7: Scale factors for the single muon triggers (left) and single electron triggers (right) from top to bottom for 2018, 2017, 2016 post VFP and 2016 pre VFP.

5.4.3 Muon and electron corrections

Lepton identification, isolation and reconstruction efficiencies are measured in data and simulation using a tag-and-probe technique and applied as scale factors for each event to correct the simulation. Efficiencies are measured in $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ events as function of the lepton p_{T} and $|\eta|$. The tag lepton must past the kinematic selection, identification and isolation, while the probe lepton must only pass the kinematic selection. The efficiency is defined as the ratio of number of probe leptons, which pass the identification and isolation, to the total number of probe leptons, and is extracted by a fit to the dilepton mass distribution around the Z boson mass [47, 55]. The lepton identification, isolation and reconstruction scale factors are provided by the respective POG of the CMS Collaboration.

5.4.4 Tau corrections

Scale factors for the DeepTau discriminator against jets are derived in the $Z \rightarrow \tau\tau \rightarrow \mu\tau_{\mathrm{h}}$ process. The muon is tagged as one of the tau leptons decaying leptonically and the probe side that reconstructs a Z boson is assumed to be the τ_{h} with only a loose requirement on the identification. The *pass* region includes all probes which pass the working point of the DeepTau identification against jets, for which the scale factors are derived. Probes that do not pass the working point are included in the *fail* region. The efficiency is extracted in bins of the τ_{h} p_{T} from a fit to the visible invariant mass of the $\mu\tau_{\mathrm{h}}$ pair.

The scale factors for the DeepTau electron and muon discriminators are derived in a $Z \rightarrow ee/Z \rightarrow \mu\mu$ process with one electron/muon misidentified as τ_{h}, considered as probe. If the probe passes the working point of the DeepTau identification against electrons or DeepTau identification against muons it is selected for the *pass* region, otherwise for the *fail* region. The efficiencies are extracted in several bins of the τ_{h} η as fit on the visible invariant mass.

The τ_{h} energy scale correction is derived with a tag-and-probe method in the same $\mu\tau_{\mathrm{h}}$ event sample as described before, but using the mass of the reconstructed hadronic tau lepton and the visible invariant mass of the $\mu\tau_{\mathrm{h}}$ pair as distributions. A maximum likelihood fit is performed for each τ_{h} decay mode to extract the τ_{h} energy scale. The relative difference in the τ_{h} energy scale between data and simulation is applied as a multiplicative factor to the τ_{h} four-momentum in simulation.

The scale factors are provided by the Tau POG of the CMS Collaboration and are described in more detail in Refs. [57, 58].

5.4.5 Bottom quark tagging corrections

The difference in the efficiency of the b-tagging in simulation and data is corrected by applying scale factors. Since the b-tagging discriminator value is used as input to the BDT, the entire discriminator distribution is corrected, not just at a specific working point as is done for other objects. The scale factors are derived with an iterative fit to the discriminator value to match the distribution in simulation with the one in data. A $\mathrm{t\bar{t}}$ sample with a dilepton final state is used as b-jet enriched region and a Drell-Yan+jets sample with a dilepton final state as light-jet enriched region, including the up, down and strange quarks as well as gluons. The fit is performed for every bin of the discriminator value in the b-jet enriched region, assuming the nominal light-jet distribution. Then the discriminator distribution in the light-jet enriched region is fit, taking into account the just derived b-jet scale factors. The fit is iterated several times, at each stage

taking into account the scale factors derived in the previous stage [92]. The b-tagging scale factors are provided by the BTV POG of the CMS Collaboration.
The scale factors are applied to each jet in an event depending on its flavor (light-jet or b-jet), p_T, η and b-tagging discriminator value. The product of all the jet scale factors yields the per event scale factor applied to the simulated samples. The scale factor only changes the shape of the discriminator value, not its normalization. Therefore the normalization needs to be corrected for each sample. A control region is defined with the selection of the signal region, but without a cut on the number of b-tagged jets. The correction is derived as the ratio of the normalization in the leading jet p_T distribution without the b-tag scale factor applied over the normalization of the distribution where the b-tag scale factor is applied, as shown for the $t\bar{t}$ sample in figure 5.8 for 2018 as an example.

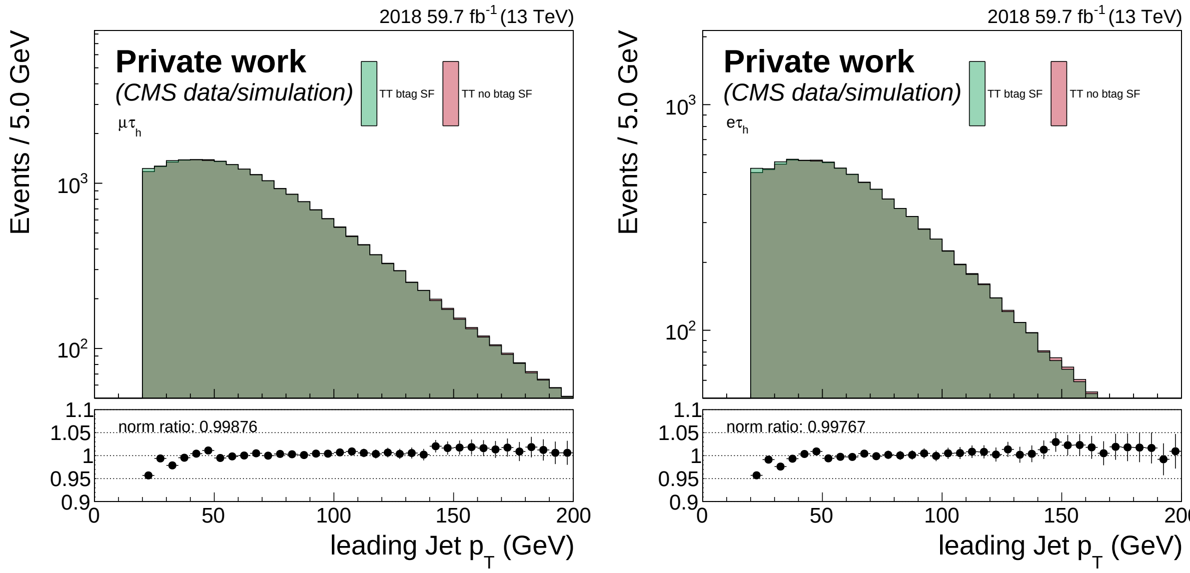


Figure 5.8: $t\bar{t}$ simulated sample with and without b-tag scale factor applied for the leading jet p_T distribution for the 2018 $\mu\tau_h$ (left) and $e\tau_h$ (right) channel.

5.4.6 L1 Prefiring

In 2016 and 2017 the gradual timing shift of the ECAL crystals was not properly propagated to the level-1 trigger primitives. This had the effect, that the trigger primitives in the forward region of $2 < |\eta| < 3$ could be incorrectly assigned to the previous bunch crossing, which can lead to an incorrect self-veto of the level-1 trigger in the correct bunch crossing [93]. This effect was not taken into account in the simulation. A correction is derived with 2017 simulated samples by assigning all the jets and photons in an event a prefiring probability. The probability that no self-veto occurred in an event is derived from taking the product of all the jets and photons not prefiring. This probability of no self-veto is applied as a weight to each event in the 2016 and 2017 simulated samples, resulting in a small correction.

5.4.7 Top-quark pt correction

The p_T distribution of the top quarks in the $t\bar{t}$ events is observed to be softer in data than in simulation in many measurements. Corrections were derived based on measurements of the $t\bar{t}$ cross section [94, 95]. The difference between data and simulation was fitted and a correction weight depending on the generator level top quark p_{T} was derived. For each top quark a weight is calculated:

$$\mathrm{weight}_t = \exp(a + bp_{\mathrm{T}} + cp_{\mathrm{T}}^2), \tag{5.3}$$

where $a = 0.088$, $b = -0.00087$, $c = 9.2e - 7$. The approximation is only valid up to a p_{T} value of 472 GeV. If the p_{T} is larger, the weight for $p_{\mathrm{T}} = 472$ GeV is used.

The weight is then derived as the square root of the product of the weight for the top quark and the top antiquark:

$$\mathrm{weight}_{\mathrm{t\bar{t}}} = \sqrt{w_t \cdot w_{\bar{t}}} \tag{5.4}$$

5.4.8 Normalization corrections

The simulation can differ from data because of uncertainties in the theoretical models and because of detector effects. Therefore the normalization of the DY+jets, W+jets and $\mathrm{t\bar{t}}$ simulated samples are corrected by using appropriate control regions, separately for the years 2018, 2017 and 2016.

For the DY+jets samples a $\mu\mu$ control region is defined. The muon pair must have opposite charge and must be separated by at least $\Delta R > 0.15$. The correction is derived as a function of the muon pair p_{T} and mass in the regions with one and with two bottom quark jets. Figure 5.9 shows the distributions of the muon pair p_{T} and mass as an example for 2018 and in Fig. 5.10 the correction applied to the simulated DY+jets sample for each year. In the two bottom quark region the $\mathrm{t\bar{t}}$ contribution is large, therefore an uncertainty is derived by varying the $\mathrm{t\bar{t}}$ cross section by its uncertainty (6%) and measuring the effect on the DY+jets correction weight.

The normalization correction of the $\mathrm{t\bar{t}}$ simulated sample is derived in an $\mathrm{e}\mu$ control region. The lepton pair must have opposite charge and be separated by at least $\Delta R > 0.15$. In order to be exclusive to the $\mathrm{e}\mu$ signal region, $D_\zeta < -35$ is required (definition of D_ζ in eq. 5.12). The correction is studied as function of the dilepton p_{T}, and due to the flat trend applied as average of the corrections. The dilepton p_{T} distribution is shown as an example for 2018 in Fig. 5.11, together with the corrections for each year.

The normalization of the W+jets simulated sample is corrected separately for the $\mu\tau_{\mathrm{h}}$ and the $\mathrm{e}\tau_{\mathrm{h}}$ channel. The control region is defined with requiring a large transverse mass of $m_T > 80$ GeV and no bottom quark jets in the event. The correction is derived as a linear fit as function of the τ_{h} p_{T}. Figure 5.12 shows the τ_{h} p_{T} distribution as an example for 2018 and in Fig. 5.13 the derived corrections are shown for each year.

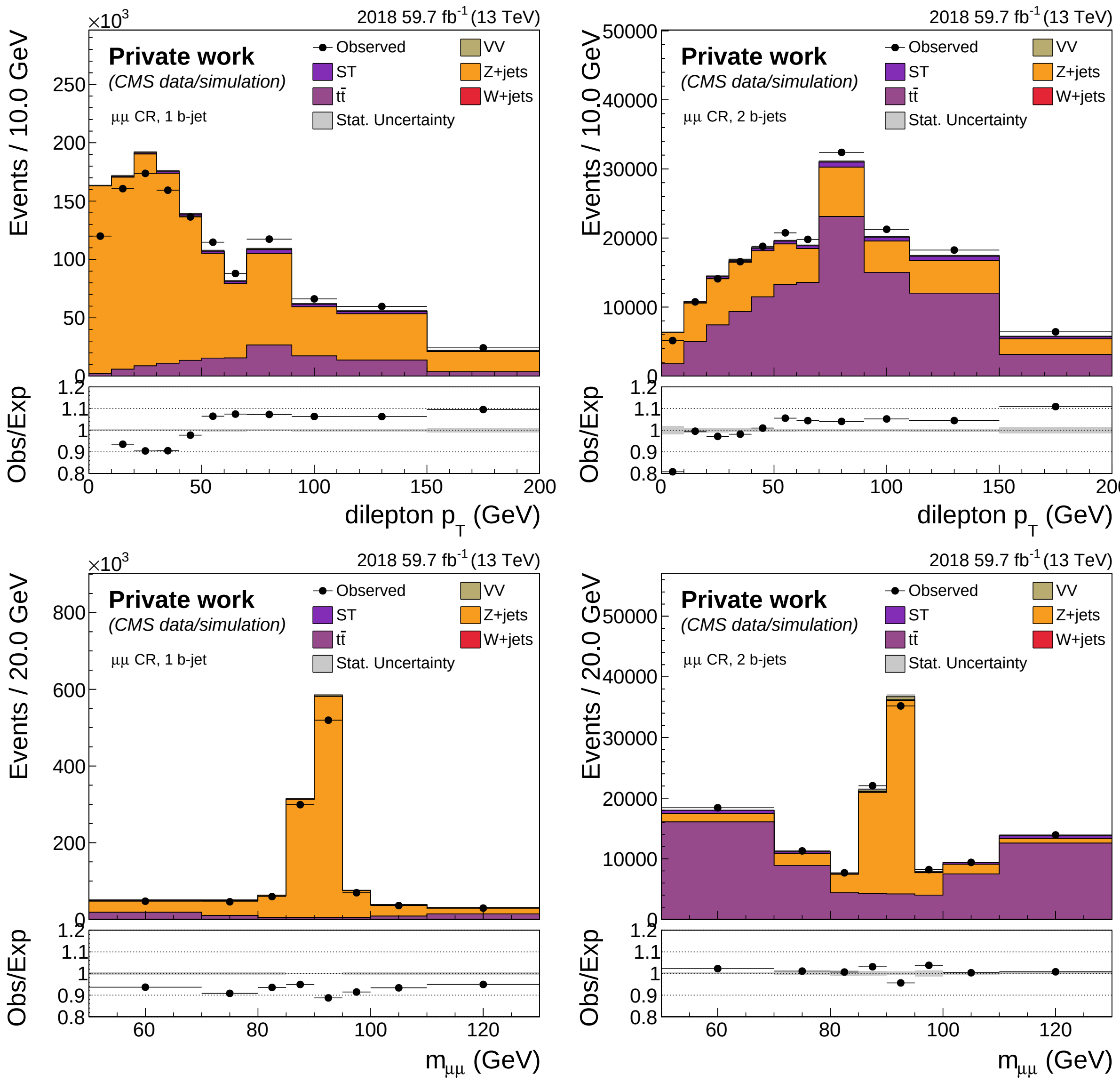


Figure 5.9: Dilepton p_T (top) and mass (bottom) in the 2018 DY+jets $\mu\mu$ control region for one b-tagged jet (left) and two b-tagged jets (right).

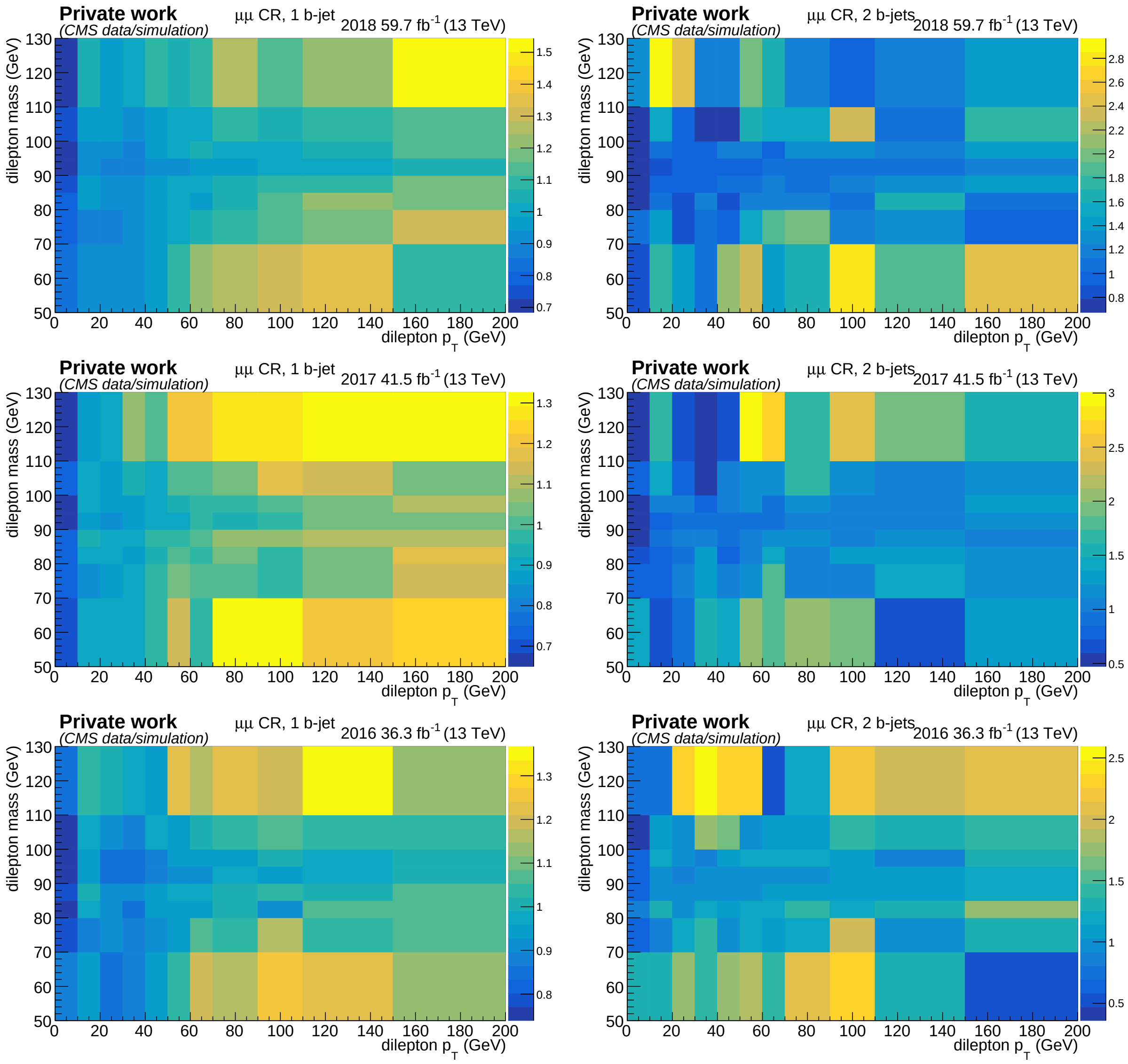


Figure 5.10: Drell-Yan+jets normalization correction as function of dilepton p_T and mass for one b-tagged jet (left) and two b-tagged jets (right) for 2018 (top), 2017 (center) and 2016 (bottom).

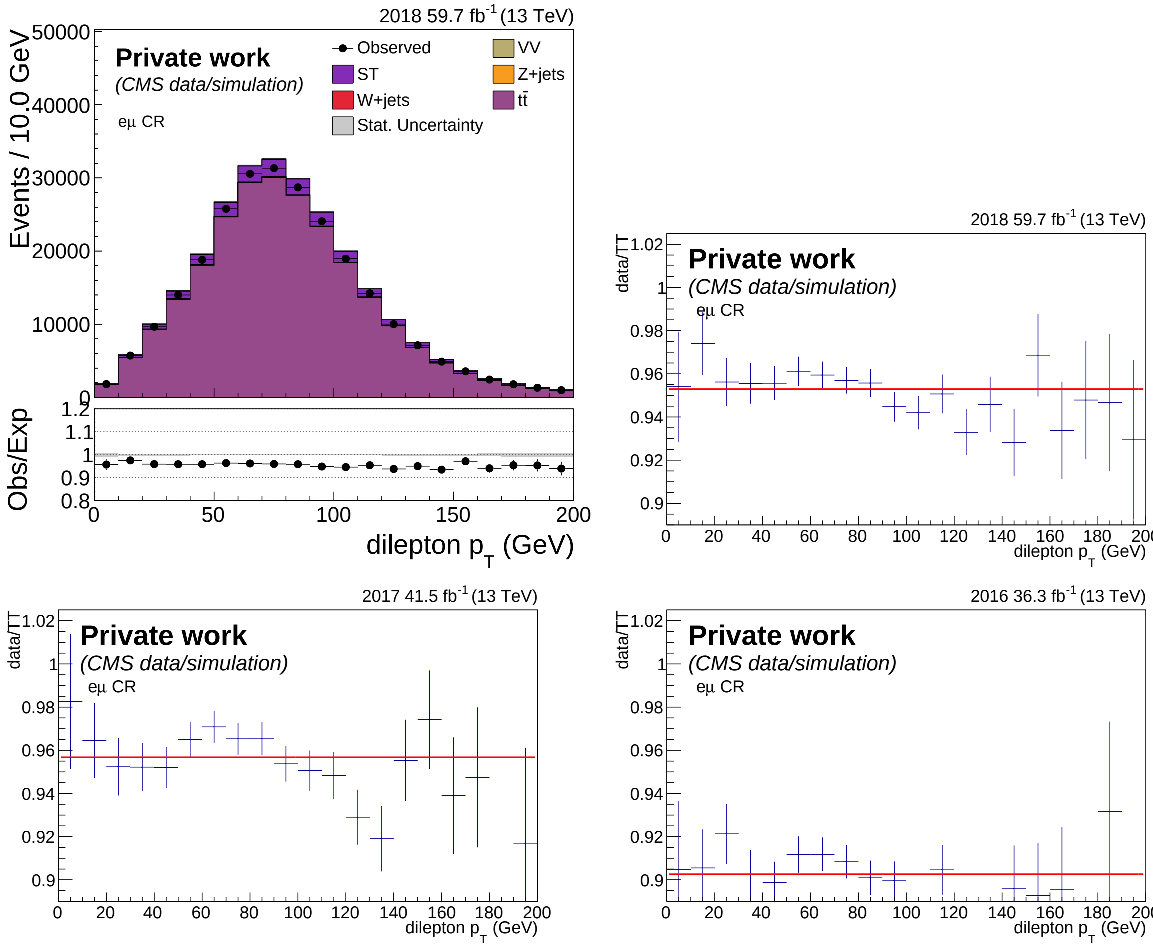


Figure 5.11: Dilepton p_{T} distribution in the 2018 $\mathrm{t\bar{t}}$ $1e1\mu$ control region (top left) and the ratio of the data over simulation normalization as function of the dilepton p_{T}, with their average for 2018 (top right), for 2017 (bottom left) and 2016 (bottom right). The normalization correction to the $\mathrm{t\bar{t}}$ samples is applied as a multiplicative scalar.

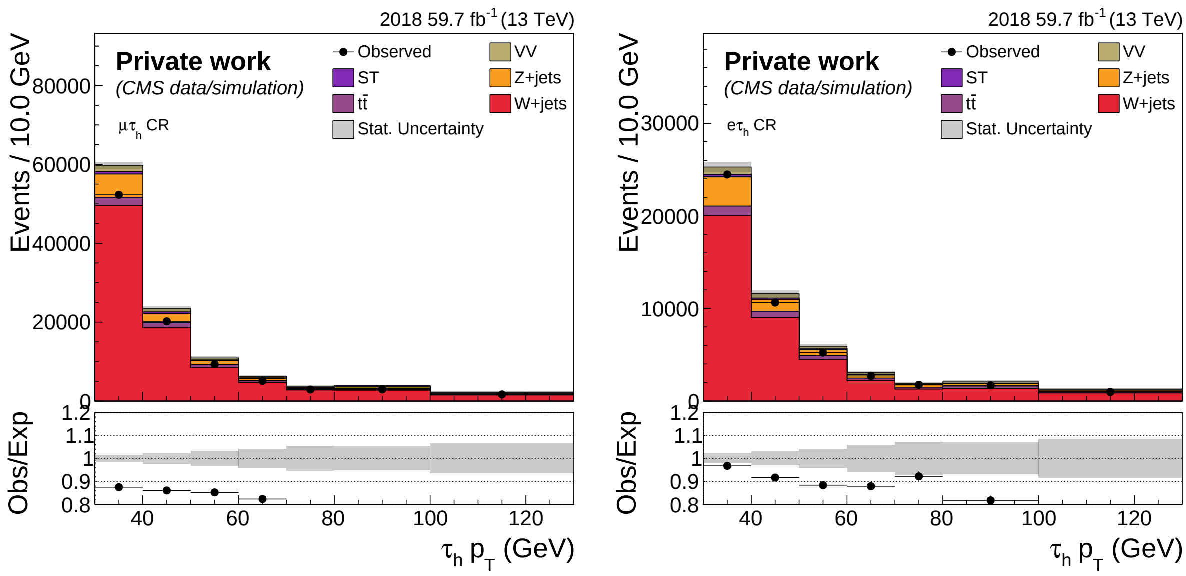


Figure 5.12: τ_{h} p_{T} distribution in the W+Jets control regions for the 2018 $\mu\tau_{\mathrm{h}}$ channel (left) and $\mathrm{e}\tau_{\mathrm{h}}$ channel (right).

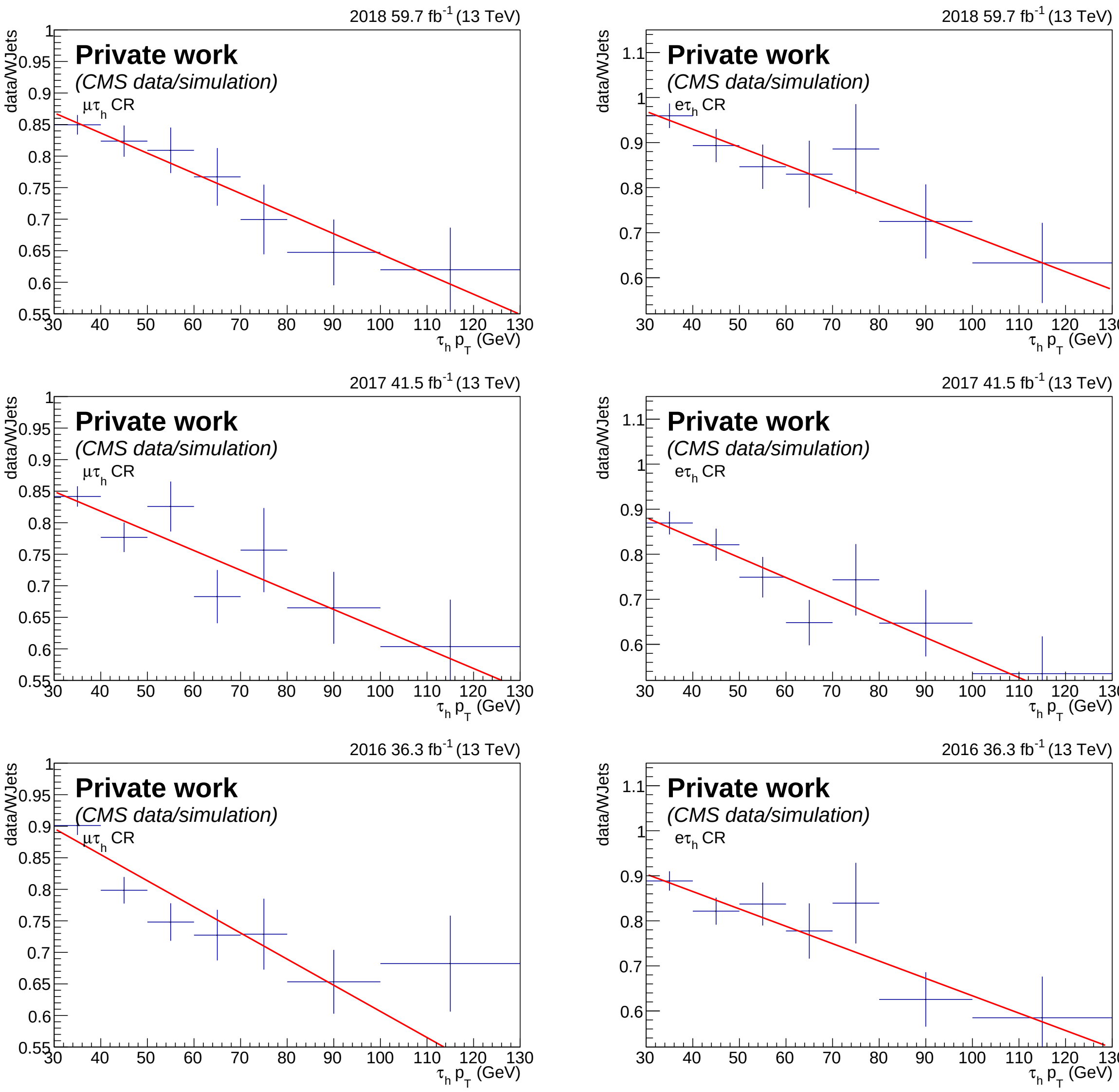


Figure 5.13: W+jets control region ratio of the data over simulation normalization as function of the τ_h p_T, fitted with a linear function, for the $\mu\tau_\mathrm{h}$ channel (left) and $e\tau_\mathrm{h}$ channel (right) for 2018 (top), 2017 (center) and 2016 (bottom). The normalization correction to the W+jets samples is applied as function of the τ_h p_T.

5.5 Selection of events

Events have to pass a certain number of selection criteria before being considered signal candidates, identically in both data and simulation. The selections are reported as follows for each object.

Muon and electron: The muon (for the $\mu\tau_h$ channel) or the electron (for the $e\tau_h$ channel) that activated the trigger is required to have a transverse momentum of at least the trigger threshold plus 1 GeV. This is to avoid the trigger turn-on, referring to lower efficiency close to the trigger threshold. The turn-on is visible in the derivation of the trigger scale factors, discussed in section 5.4. The muon is required to have $|\eta| < 2.4$, identified with the *medium* WP identification and isolated with $I^{\mu}_{\text{rel}} < 0.15$ as discussed in section 4.5. The electron is required to have $|\eta| < 2.1$, identified by the MVA based identification at the 90% efficiency WP and $I^{e}_{\text{rel}} < 0.1$, described in more detail in section 4.4.

Hadronic tau lepton: In both channels at least one hadronic tau with p_T larger than 30 GeV, $|\eta| < 2.3$ and identified with the *medium* WP of the DEEPTAU discriminator against jets is required. In the $\mu\tau_h$ channel the *loose* WP of the DEEPTAU discriminator against electrons and the *tight* WP of the DEEPTAU discriminator against muons is required, in the $e\tau_h$ channel the *tight* WP of the DEEPTAU discriminator against electrons and the *loose* WP of the DEEPTAU discriminator against muons. More detail can be found in section 4.6.

Pair selection: The $\mu\tau_h$ and $e\tau_h$ pairs are selected as particles of opposite charge and a separation of at least $\Delta R > 0.5$. If multiple pairs in the event pass this requirement, the pair is selected by:

- Select the pair with the most isolated light lepton.
- If the isolation of the light lepton is the same in both pairs (one lepton can be in multiple pairs in combination with different τ_h), prefer the pair with the highest light lepton p_T.
- If the p_T of the light lepton is the same in both pairs, prefer the pair with the most isolated τ_h.
- If the isolation of the τ_h is the same in both pairs, prefer the pair with the highest τ_h p_T.

Veto on addition leptons: In the $\mu\tau_h$ channel no additional muons or electrons are allowed. Only one *medium* WP muon with $p_T > 10$ GeV and $I^{\mu}_{\text{rel}} < 0.3$ is allowed. If there is a muon pair with *loose* WP and each $p_T > 15$ GeV, $I^{\mu}_{\text{rel}} < 0.3$, opposite charge and $\Delta R(\mu_i, \mu_j) > 0.15$ the event is discarded. If there are any electrons with $p_T > 10$ GeV and $I^{e}_{\text{rel}} < 0.3$, the events are discarded.
In the $e\tau_h$ channel no additional electrons or muons are allowed. Only one electron with $p_T > 10$ GeV, $I^{e}_{\text{rel}} < 0.3$ is allowed in an event. If there is an electron pair with each $p_T > 15$GeV, $I^{e}_{\text{rel}} < 0.3$, opposite charge and $\Delta R(e_i, e_j) > 0.15$ the event is not used in the analysis. If there are any muons with $p_T > 10$ GeV, $I^{\mu}_{\text{rel}} < 0.3$ and identified with the *medium* WP, the events are discarded.

Jets: Jets are included in the events if they have at least a p_T of 20 GeV, $|\eta| < 2.4$, identified by a *tight* WP jet identification and a separation of $\Delta R > 0.4$ to any isolated muons or electrons, as discussed in section 4.7. At least one jet needs to be identified as bottom quark jet by the *medium* WP of the DEEPJET b-tagging algorithm. The jets are ordered according to their b-tagging discrimination score. The leading jet refers to the jet with the highest b-tag score.

Transverse mass: The contributions from the $t\bar{t}$ and W+jets background are reduced by requiring the transverse mass to be smaller than 60 GeV. The transverse mass is the invariant mass between $p_{\mathrm{T}}^{\mathrm{miss}}$ and the lepton p_{T} and is derived as:

$$m_T = \sqrt{2p_{\mathrm{T}}^{\ell}p_{\mathrm{T}}^{\mathrm{miss}}(1 - cos(\Delta\phi(\vec{\ell}, \vec{p}_T^{miss})} \tag{5.5}$$

The neutrino and the lepton in the leptonic decay of the W boson are typically emitted in opposite directions, leading to a large $\Delta\phi(\vec{\ell}, \vec{p}_T^{miss})$ close to π, that maximizes the m_T variable. In Fig. 5.14 the transverse mass distribution is shown before the cut as an example for the 2018 $\mu\tau_{\mathrm{h}}$ and $\mathrm{e}\tau_{\mathrm{h}}$ channel.

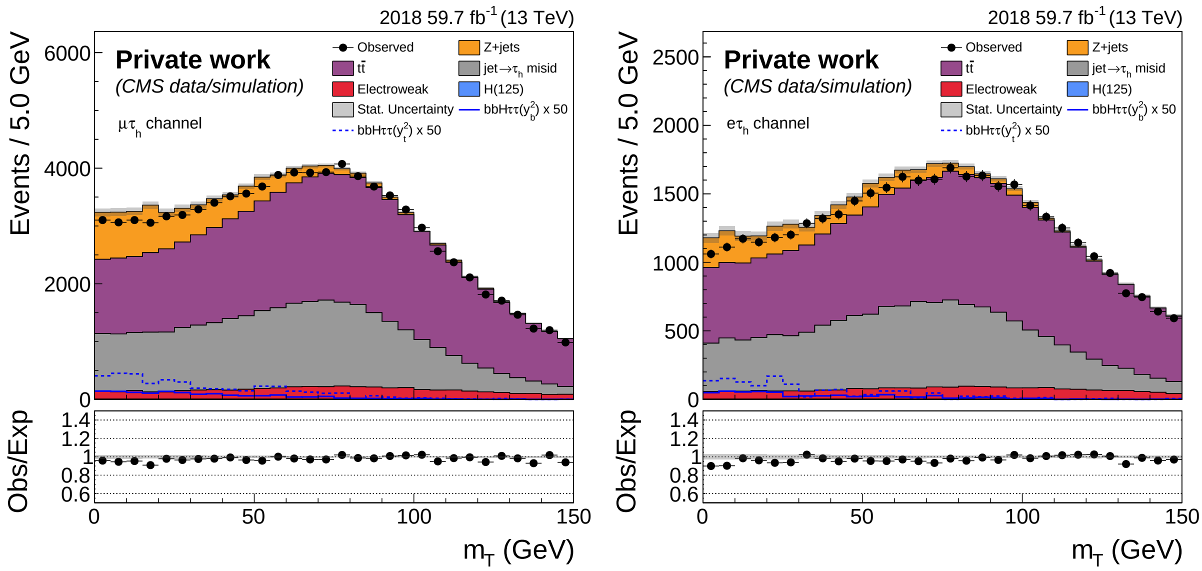


Figure 5.14: Transverse mass distribution in the 2018 $\mu\tau_{\mathrm{h}}$ channel (left) and $\mathrm{e}\tau_{\mathrm{h}}$ channel (right) before the selection of $m_T < 60$ GeV is applied to reduce the $\mathrm{t\bar{t}}$ and W+jets background, which is included in the jet$\rightarrow$ τ_{h} fakes contribution.

Matching to generator level tau leptons: The contribution from jets misidentified as τ_{h} is estimated with a data-driven method, discussed in section 5.6. In order to avoid double counting of such events in the simulated samples, the τ_{h} needs to be matched to a generator level tau lepton with $\Delta R < 0.2$ to remove any fake τ_{h}.

A summary of the selections is given in Tab. 5.7.

	$\mu\tau_h$	$e\tau_h$
μ/e p_{T}	μ/e p_{T} >trigger threshold+1 GeV	
μ/e η	μ $\|\eta\| < 2.4$	e $\|\eta\| < 2.1$
μ/e Id	μ *medium* WP	e *90%* WP
μ/e Iso	$I^{\mu}_{\mathrm{rel}} < 0.15$	$I^{e}_{\mathrm{rel}} < 0.1$
τ_{h} p_{T}	τ_h $p_{\mathrm{T}} > 30\,\mathrm{GeV}$	
τ_{h} η	τ_h $\|\eta\| < 2.3$	
DeepTauVSjet τ_{h} Id	τ_h *medium* WP DeepTauVSjet	
	vvvloose DeepTauVSe	*tight* DeepTauVSe
	tight DeepTauVSmu	*vloose* DeepTauVSmu
Pair selection	$q_\mu \cdot q_{\tau_{\mathrm{h}}} = -1$ $\Delta R(\mu, \tau_{\mathrm{h}}) > 0.5$	$q_e \cdot q_{\tau_{\mathrm{h}}} = -1$ $\Delta R(e, \tau_{\mathrm{h}}) > 0.5$
Jet p_{T}	Jet $p_{\mathrm{T}} > 20$ GeV	
Jet η	Jet $\|\eta\| < 2.4$	
Jet Id	*tight* jet ID	
Jet separation	$\Delta R(\mu, j) > 0.4$	$\Delta R(e, j) > 0.4$
Jet b-tagging	*medium* WP DeepJet algorithm	
Veto on μ	only 1 μ_{medium}, 0 $\mu\mu_{\mathrm{loose}}$	0 μ_{medium}
Veto on e	0 e_{loose}	only 1 e_{medium}, 0 ee_{loose}
Transverse mass	$m_T < 60$ GeV	

Table 5.7: Summary of the selection requirements for the $\mu\tau_{\mathrm{h}}$ and $e\tau_{\mathrm{h}}$ channel.

5.6 Data-driven background estimation

Jets initiated by quarks or gluons and misidentified as hadronically decaying tau leptons are estimated with a data-driven method called fake factor method. For this method two exclusive regions are defined, the *isolation* region that has the same DeepTau discriminator against jets requirement as in the signal region and the *anti-isolation* region, where the discriminator is reverted, only selecting τ_h if the have a DeepTau identification against jets working point that is looser than the medium working point. The main contribution to the fake τ_h background comes from the $t\bar{t}$ process. Smaller contributions also come from the W+jets process, and in the low τ_h p_T region, the QCD multijet process contributes as well. The contributions differ depending on the τ_h decay mode. The anti-isolation region, exclusive of the signal region, is called application region (AR). It is used to estimate the contributions to the fake τ_h background and to apply the derived fake factors. All the contributions except QCD multijet are estimated with simulation, QCD multijet as data minus the other simulated contributions. In figure 5.15 the AR is shown for the $\mu\tau_h$ and $e\tau_h$ channel of 2018, split into 1-prong and 3-prong τ_h decay modes, and in Fig. 5.16 for 2017 and 2016. The "true τ" contribution refers to genuine τ_h decays, which do not contribute to the fake τ_h background.

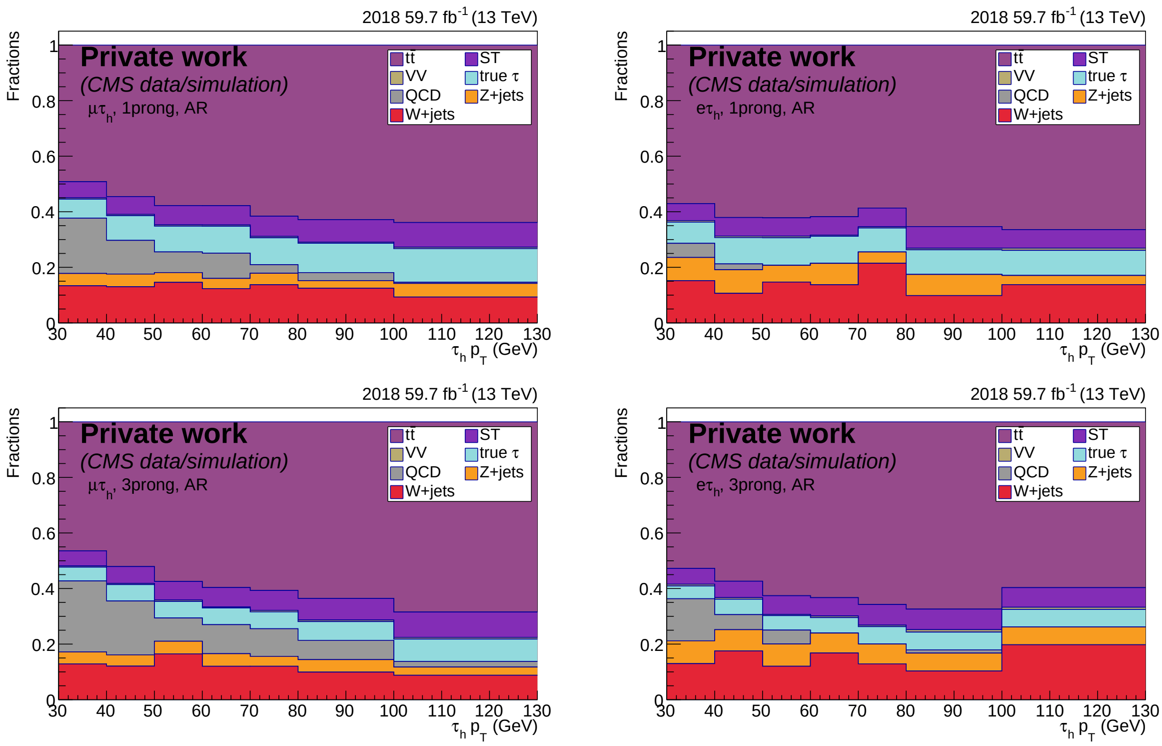


Figure 5.15: Fraction of background processes in the application region as function of the τ_h p_T for the 2018 $\mu\tau_h$ channel (left) and $e\tau_h$ channel (right) for the 1-prong (top) and 3-prong (bottom) τ_h decay mode.

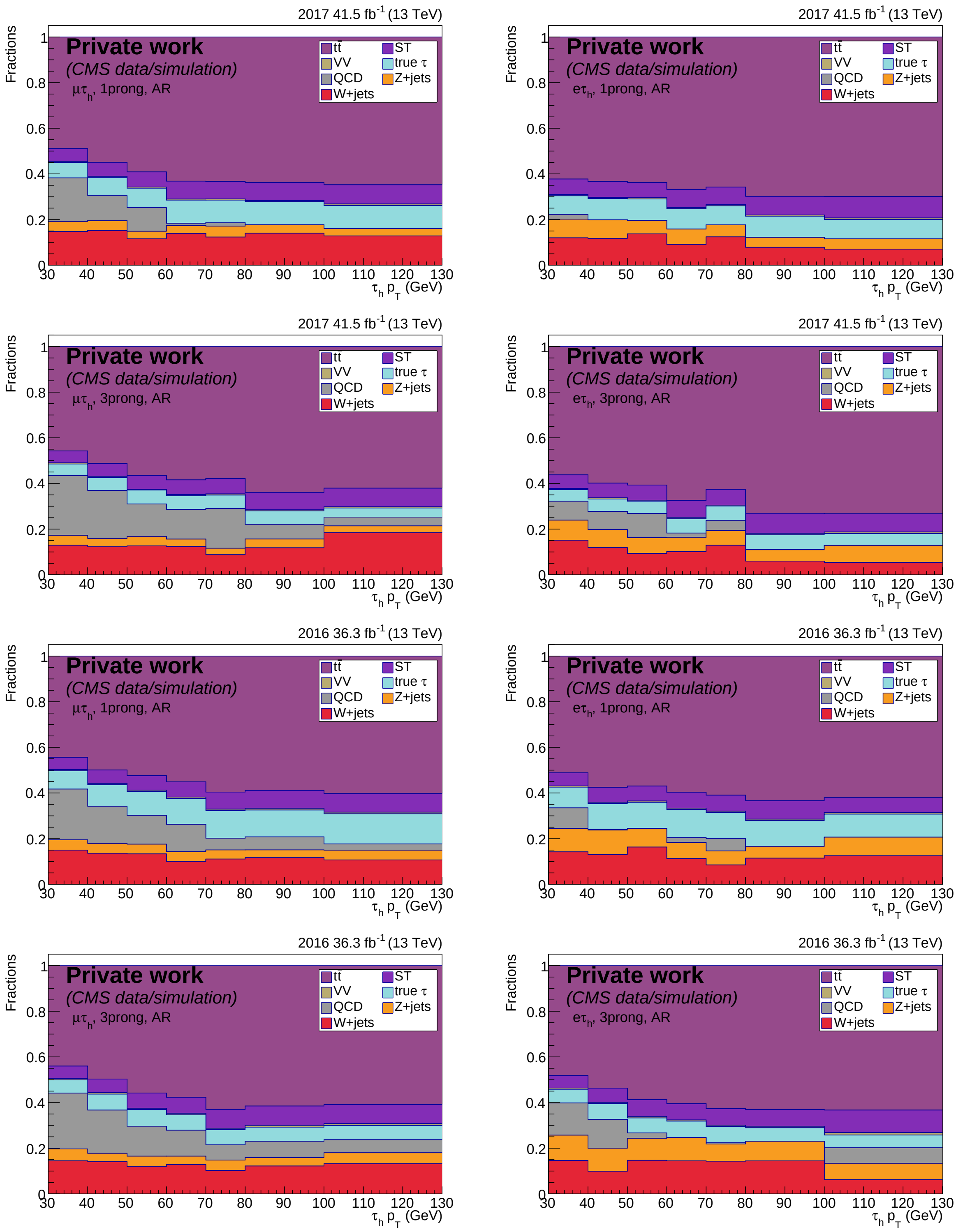


Figure 5.16: Fraction of background processes in the application region as function of the τ_h p_T for the $\mu\tau_h$ channel (left) and $e\tau_h$ channel (right) for the 2017 (top) and 2016 (bottom) data taking year, each for 1-prong and 3-prong τ_h decay mode.

The fake factors are derived in three determination regions, each enriched with a particular type of background: $t\bar{t}$, W+jets and QCD multijet. The fake factor is calculated as the ratio of events in data in the isolation over the anti-isolation region as shown in formula 5.6. The impurity of other backgrounds than the one examined in the determination region is estimated with simulated samples and subtracted from the number of events.

$$FF_i = \frac{n_{iso} - n_{iso}^{rest}}{n_{anti-iso} - n_{anti-iso}^{rest}} \tag{5.6}$$

The measured fake factors in the three determination regions are applied as a weighted average as function of the tau p_T, taking the contribution of each background in the application region into account, as shown in Fig. 5.15.
The estimated fake tau contribution is derived by applying the fake factors to the number of data events in the application region, subtracting the true tau contribution estimated with simulated samples:

$$n_{j \to \tau_{had}} = (n_{AR} - n_{true}) \cdot \sum_i f_i FF_i \tag{5.7}$$

The fake factor is derived separately for 2018, 2017 and 2016, $\mu\tau_h$ and $e\tau_h$ channel and split in 1-prong and 3-prong τ_h decay modes.

5.6.1 TT determination region

The estimation of the contribution of the $t\bar{t}$ process is done by selecting events with transverse mass larger than 70 GeV and smaller than 90 GeV. The isolation and anti-isolation region of the $t\bar{t}$ determination region are shown in Fig. 5.17 as an example for the 2018 $\mu\tau_h$ channel.
The genuine τ_h contribution needs to be subtracted from data for the derivation of the fake factors. Because the genuine τ_h contribution is large in the isolation region, any mismodeling in the genuine τ_h contribution would have a large effect on the fake factors. To avoid this dependence, the fake factors are directly derived from the simulated $t\bar{t}$ samples after subtracting the genuine τ_h contribution and fitted with a first degree polynomial. The fake factors are shown for each year in Fig. 5.18.
The agreement of the fake factors derived with data to the fake factors derived with simulated samples is tested in a looser control region, where the genuine τ_h contribution is smaller. The difference between the fake factors derived with data or only simulated samples is found to be small, therefore no corrections are applied.

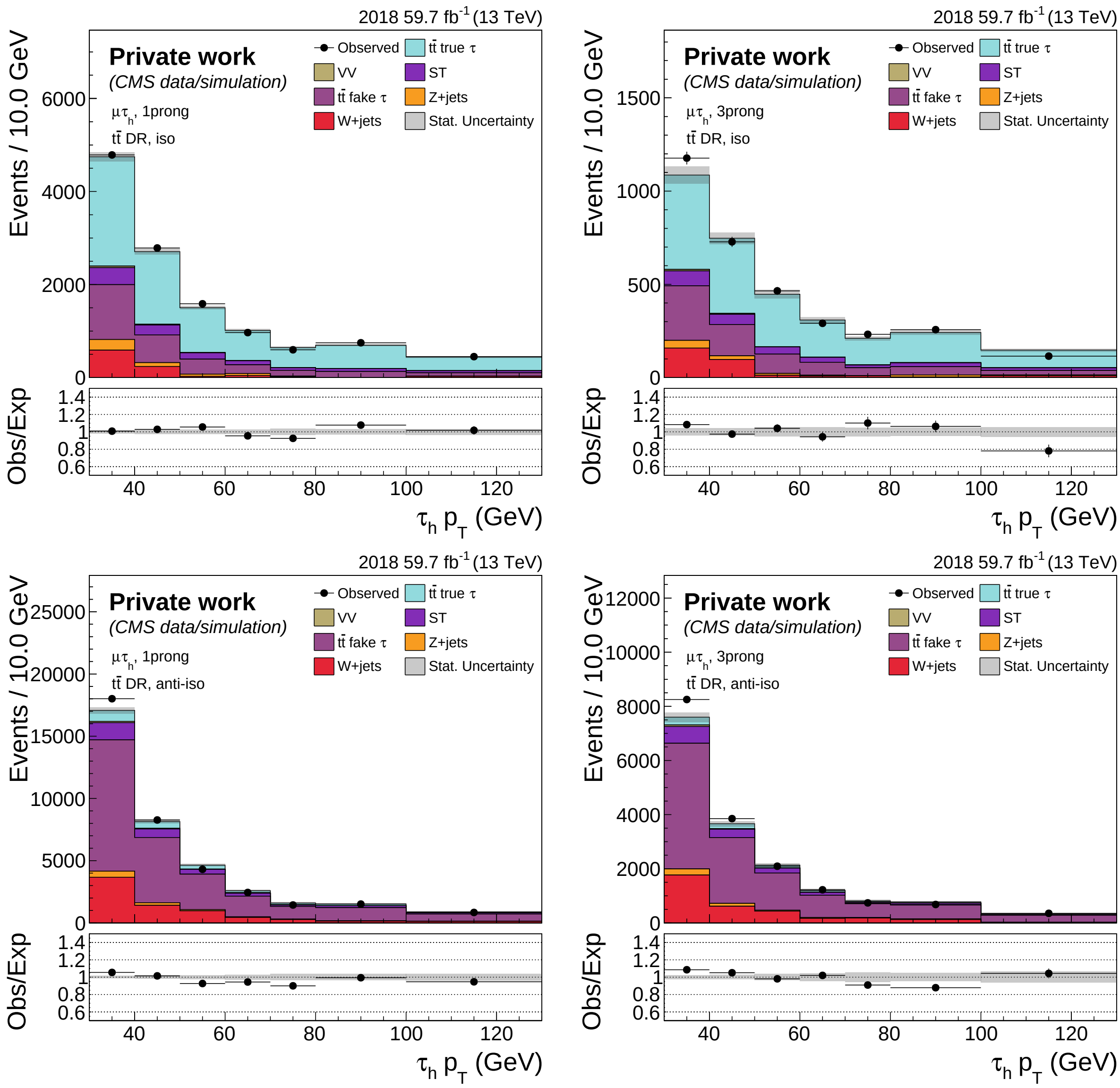


Figure 5.17: Fake-factor $t\bar{t}$ determination region for the isolation (top) and anti-isolation region (bottom) for the 1-prong (left) and 3-prong (right) τ_h decay mode for the 2018 $\mu\tau_h$ channel.

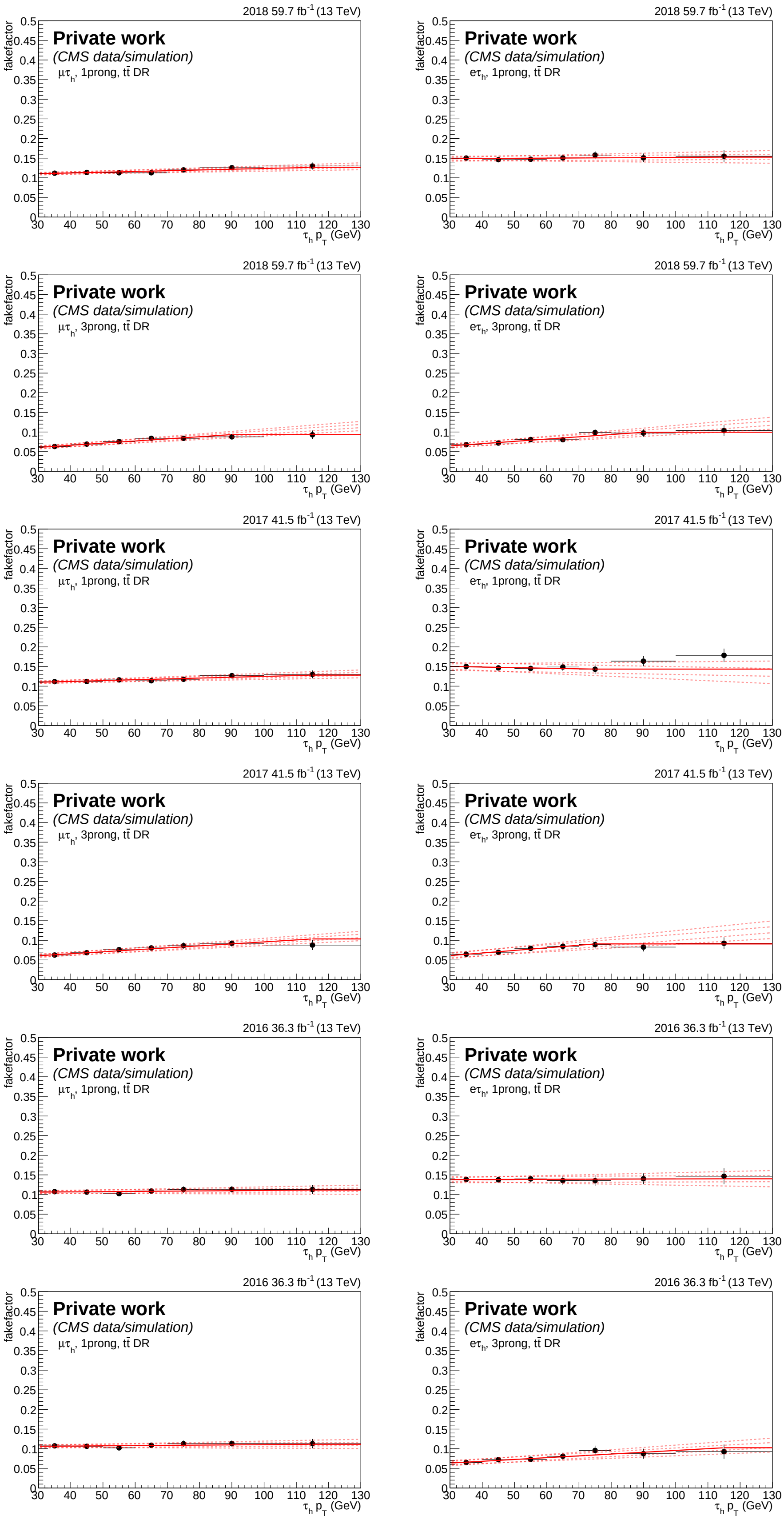


Figure 5.18: Fake factors derived for the $t\bar{t}$ contribution for the $\mu\tau_h$ (left) and $e\tau_h$ (right) channel from top to bottom for 2018, 2017 and 2016, each for 1-prong and 3-prong τ_h decay mode. The fake factors are fitted with a first degree polynomial. The dashed lines have the fit uncertainties applied as an up and down variation for each parameter of the fit function.

5.6.2 W+jets determination region

The estimation of the contribution of the W+jets process is done by selecting events with a transverse mass larger than 80 GeV and requiring zero b-tagged jets. The isolation and anti-isolation region of the W+jets determination region are shown in Fig. 5.19 as an example for the 2018 $\mu\tau_h$ channel. The fake factors are derived as ratio of the data in the isolation over anti-isolation region with subtracting the contribution from DY+jets and $t\bar{t}$ and fitted with a second degree polynomial. The derived fake factors are shown in Fig. 5.20 for each year.

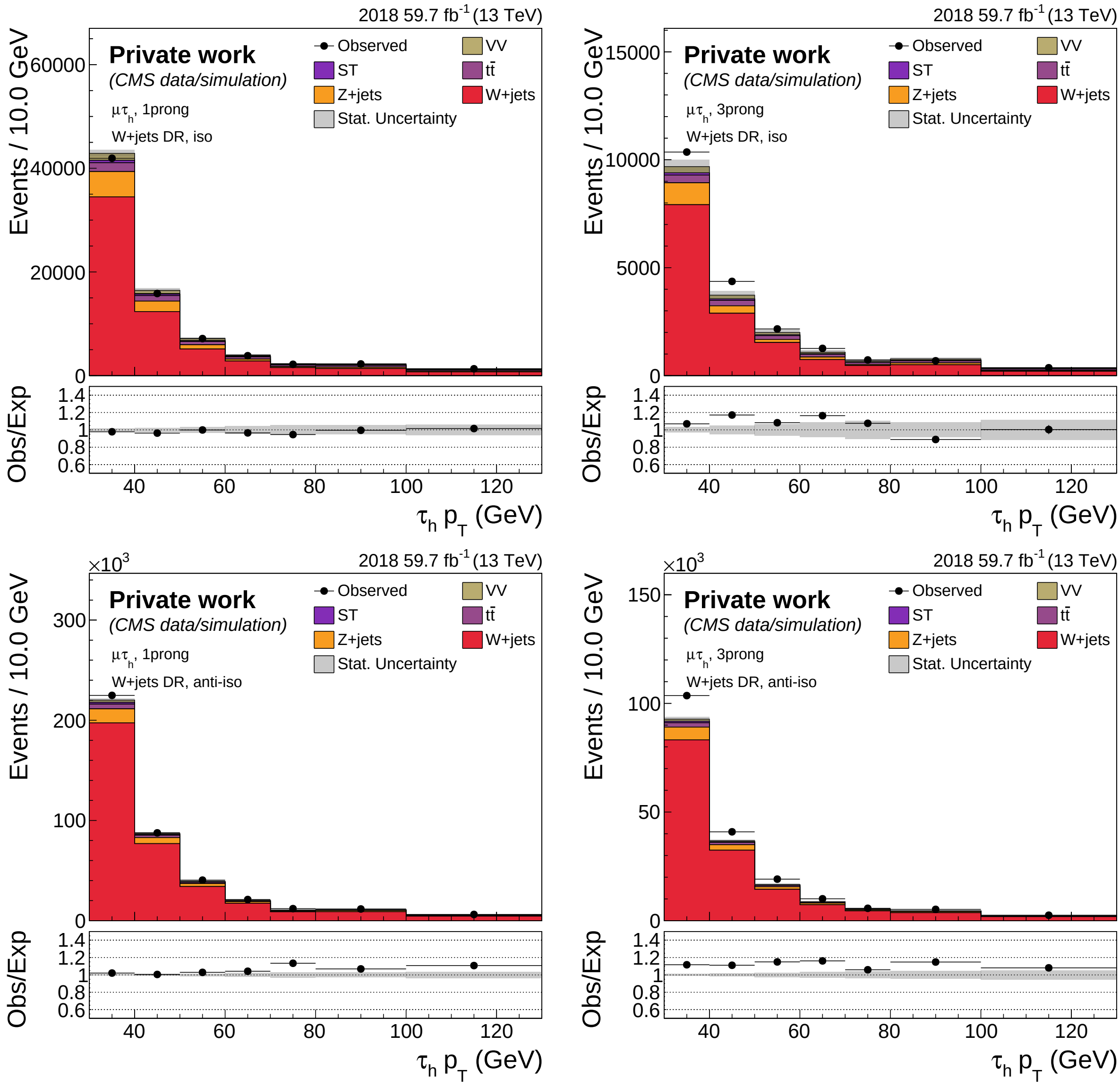


Figure 5.19: Fake-factor W+jets determination region for the isolation (top) and anti-isolation region (bottom) for the 1-prong (left) and 3-prong (right) τ_h decay mode for the 2018 $\mu\tau_h$ channel.

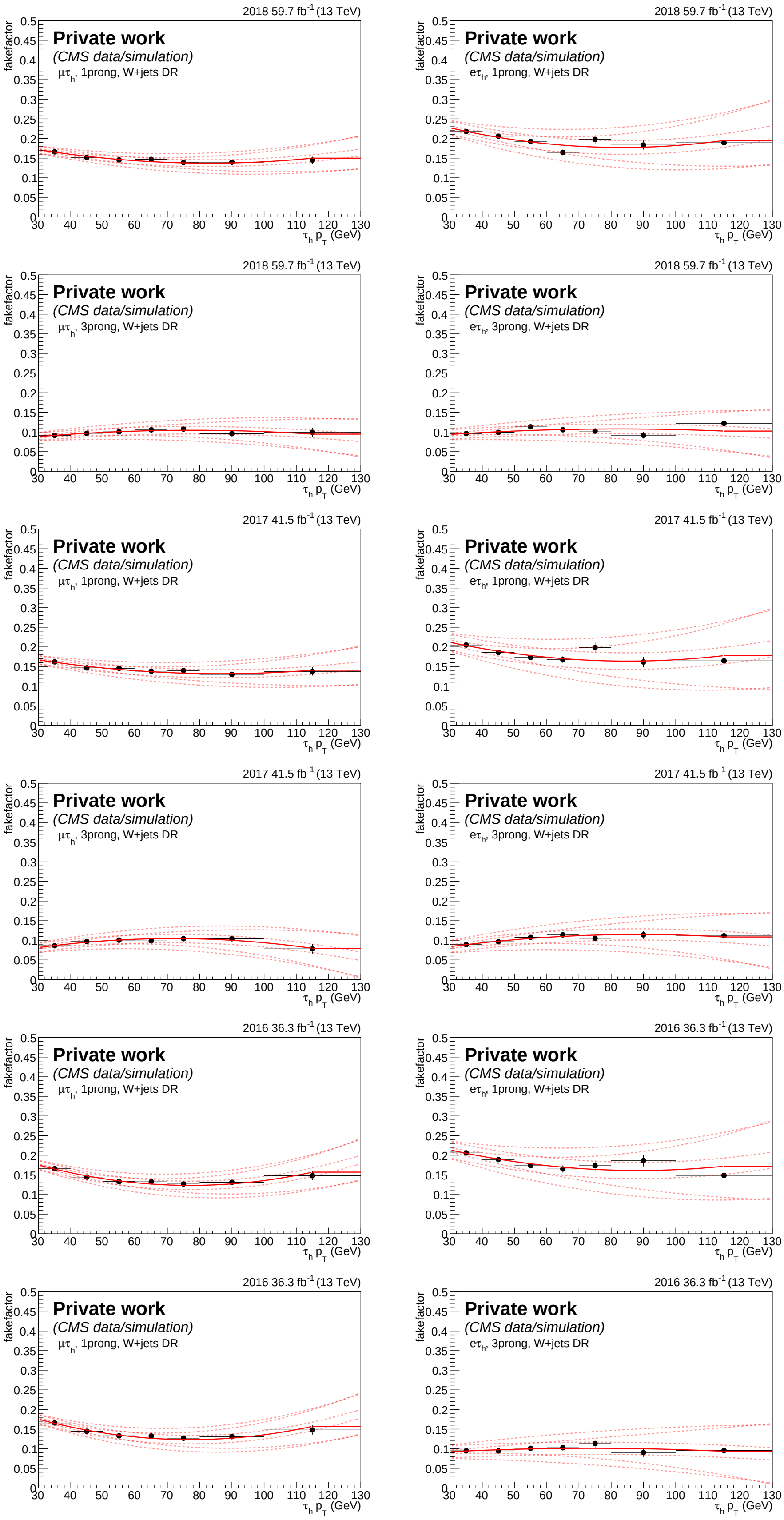


Figure 5.20: Fake factors derived for the W+jets contribution for the $\mu\tau_\mathrm{h}$ channel (left) and $\mathrm{e}\tau_\mathrm{h}$ channel (right) from top to bottom for 2018, 2017 and 2016, each for 1-prong and 3-prong τ_h decay mode. The fake factors are fitted with a second degree polynomial. The dashed lines have the fit uncertainties applied as an up and down variation for each parameter of the fit function.

5.6.3 QCD multijet determination region

The estimation of the contribution of the QCD multijet process is done by selecting events with the same charge of the light lepton and the τ_h. The isolation and anti-isolation region of the QCD multijet determination region are shown in Fig. 5.21 as an example for the 2018 $\mu\tau_h$ channel.

The fake factors are derived as ratio of the data in the isolation over anti-isolation region with subtracting the contributions from DY+jets, W+jets and $t\bar{t}$ and fitted with a first degree polynomial. The fake factors are shown in Fig. 5.22 for each year.

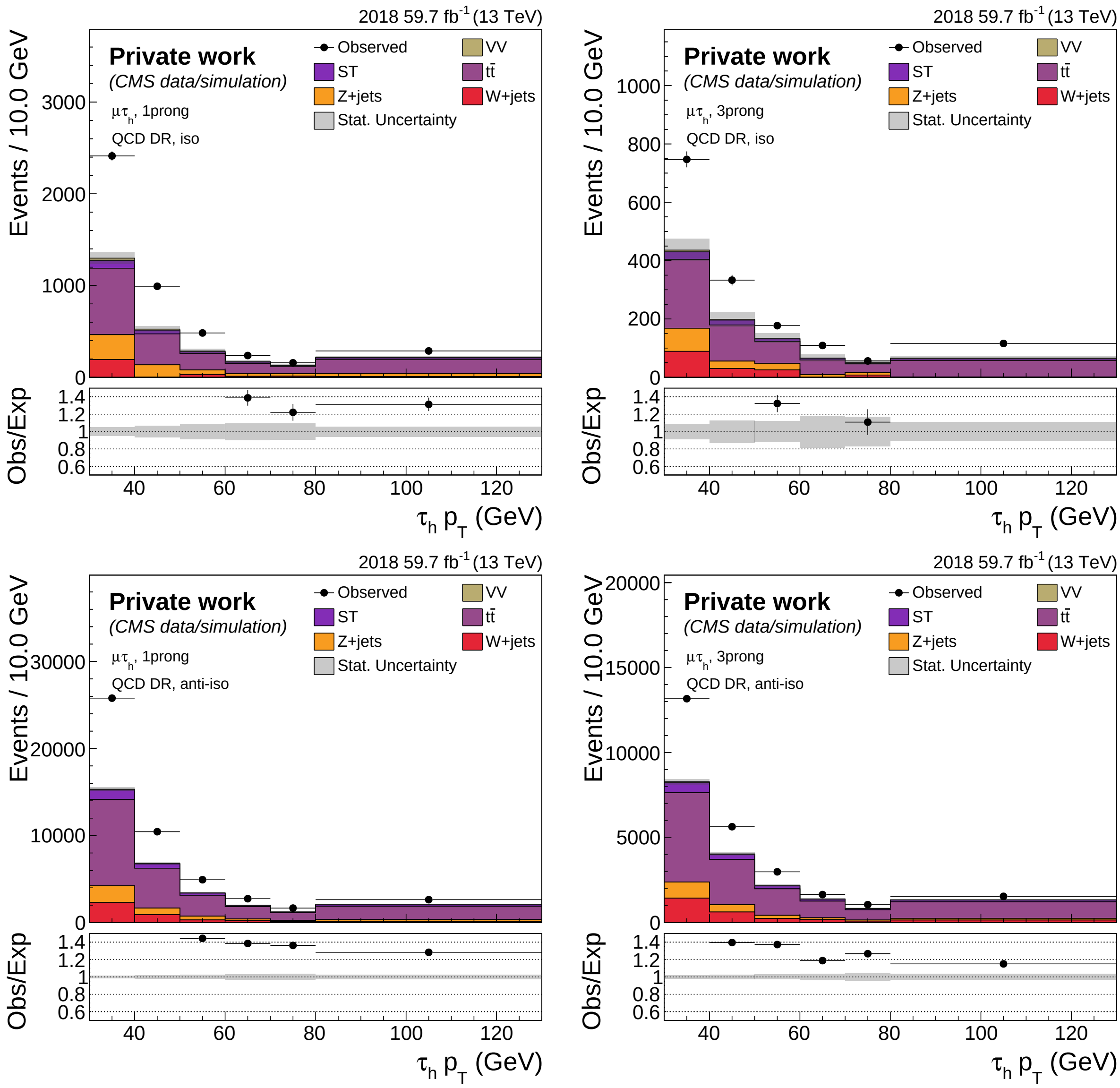


Figure 5.21: Fake-factor QCD multijet determination region for the isolation (top) and anti-isolation region (bottom) for the 1-prong (left) and 3-prong (right) τ_h decay mode for the 2018 $\mu\tau_h$ channel.

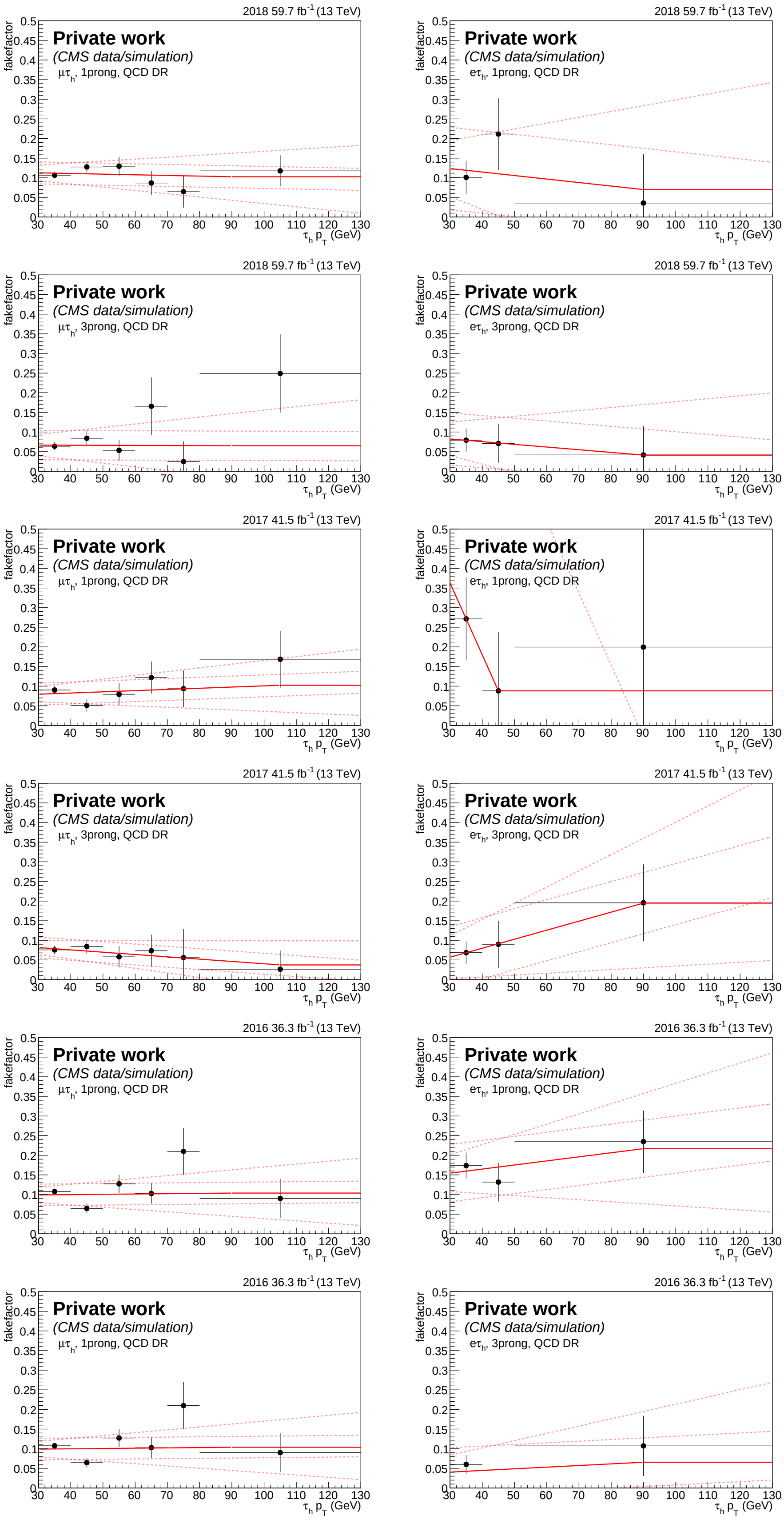


Figure 5.22: Fake factors derived for the QCD multijet contribution for the $\mu\tau_h$ (left) and $e\tau_h$ channel (right) from top to bottom for 2018, 2017 and 2016, each for 1-prong and 3-prong τ_h decay mode. The fake factors are fitted with a first degree polynomial. The dashed lines have the fit uncertainties applied as an up and down variation for each parameter of the fit function.

5.6.4 Corrections

The fake factors are estimated as a function of the τ_h p_T, they therefore need to be corrected for other variables. The closure is estimated in the $t\bar{t}$ and QCD multijet determination region as function of the four variables with the highest importance for the boosted decision tree discrimination, described in section 5.7. These are the reconstructed Higgs boson mass, the leading jet p_T, collinear mass and mass of τ_h plus leading jet. The closure test is performed by applying the fake factor to the anti-isolation region and deriving the ratio from data events in the isolation over anti-isolation region (with subtracting contributions from other simulated backgrounds, same as for the derivation of the fake factor). In Figures 5.23 and 5.24 the closure test for the mentioned variables is shown in the $t\bar{t}$ and QCD multijet determination regions as an example for the 2018 $\mu\tau_h$ 1-prong channel.

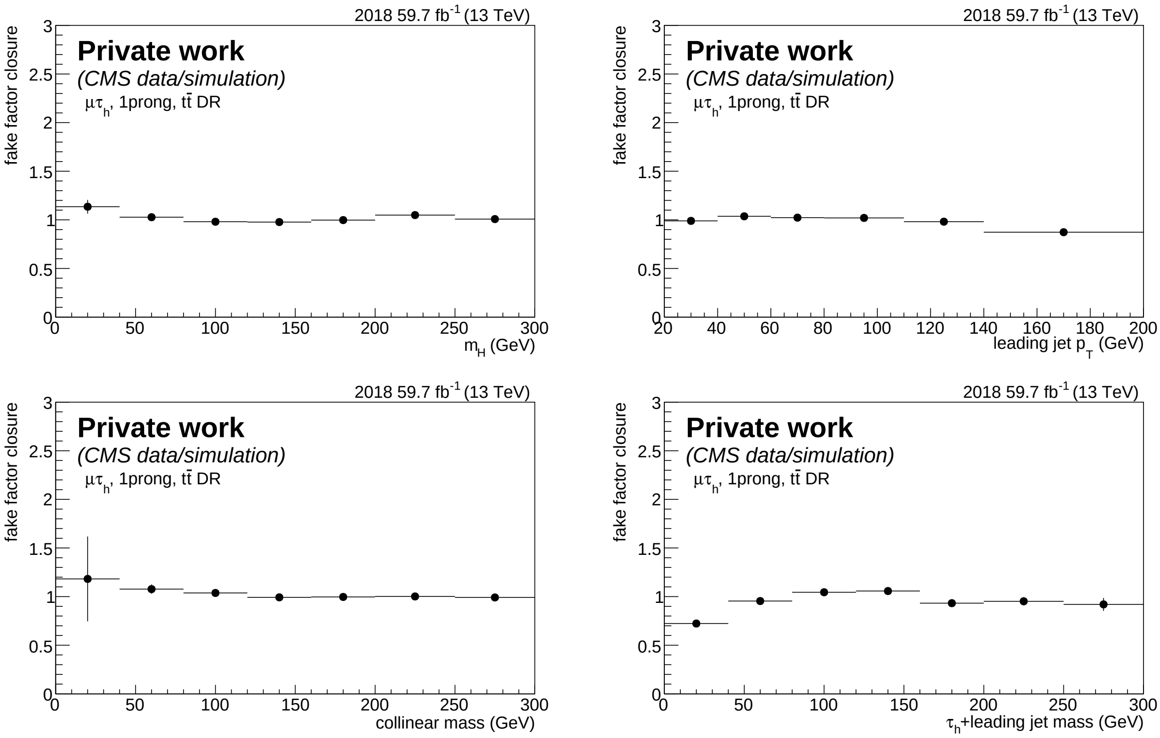


Figure 5.23: Closure of fake factors derived in the $t\bar{t}$ determination region for the 2018 $\mu\tau_h$ 1-prong channel as function of Higgs boson mass (top left), p_T of jet with highest b-tagging discrimination score (top right), collinear mass (bottom left) and mass of τ_h and leading jet (bottom right).

The correction factor is derived as:

$$1 + \frac{x-1}{(1+\sigma)^3}, \tag{5.8}$$

where x is the content of the bin and σ its error. The functional form was chosen in order to apply a large correction if a large deviation is observed in a bin with large statistics, and a smaller correction otherwise. The correction is then applied to the fake factor (as described in formula 5.6) as a multiplicative factor. The correction factors of the four variables are not independent. The closure test for the Higgs boson mass is performed with the fake factor, but for the leading jet p_T the correction factor for the Higgs boson mass is applied to the fake factor

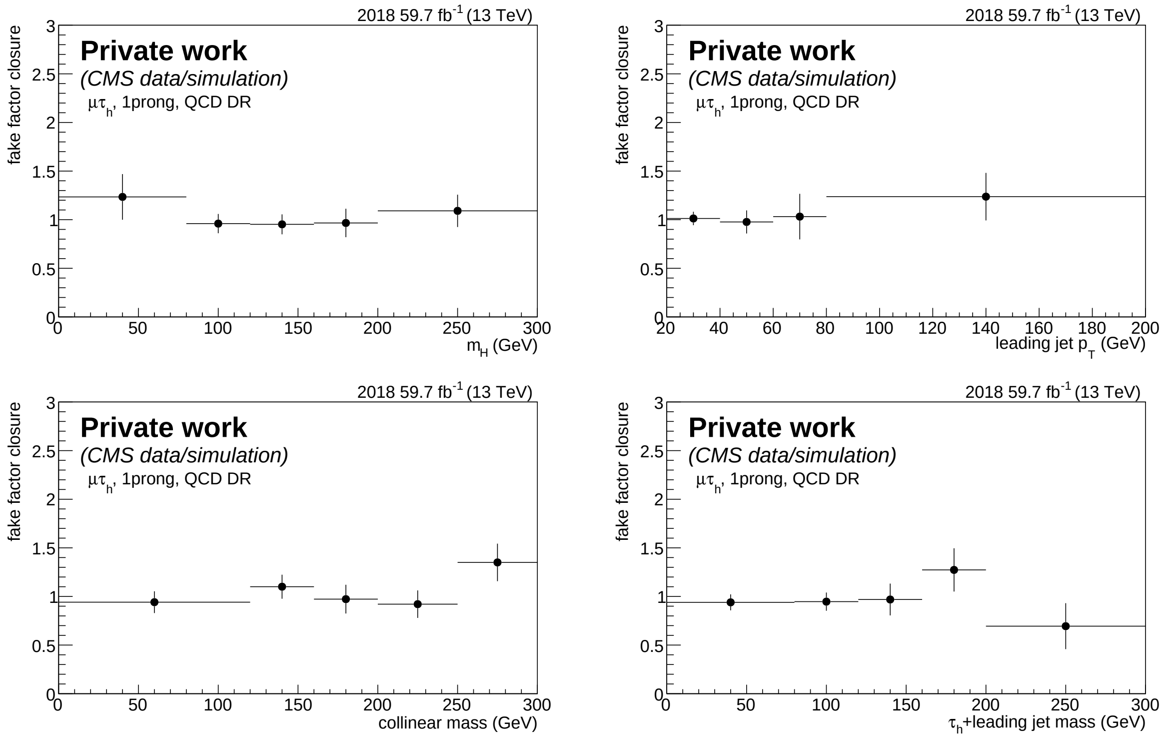


Figure 5.24: Closure of fake factors derived in the QCD multijet determination region for the 2018 $\mu\tau_\mathrm{h}$ 1-prong channel as function of Higgs boson mass (top left), p_T of jet with highest b-tagging discrimination score (top right), collinear mass (bottom left) and mass of τ_h and leading jet (bottom right).

for performing the closure test, and so on for the other variables.
In the W+jets determination region the closure test for the aforementioned variables showed good agreement, as shown in Fig. 5.25 as an example for the 2018 $\mu\tau_\mathrm{h}$ 1-prong channel, therefore no corrections are applied.

5.6.5 Uncertainties

The choice of the used fit function introduces an uncertainty that is estimated by varying each parameter of the fit function up and down.
The uncertainty arising from the selection of the determination regions is estimated by performing closure tests in control regions exclusive to each determination region. The control region for the QCD multijet determination region is defined by inverting the lepton isolation, for the W+jets determination region the transverse mass is required to be in the range 60 GeV $< m_T <$ 80 GeV and 60 GeV $< m_T <$ 70 for the $\mathrm{t\bar{t}}$ determination region.
An additional uncertainty is introduced to account for any mismodeling of the background contributions in the application region, that has an effect on the fractions used to produce the weighted average of the fake factor. The fractions from the application region are extracted with the renormalization and factorization scales (described in more detail in section 5.9) varied by 0.5 and 2 for the simulated samples.

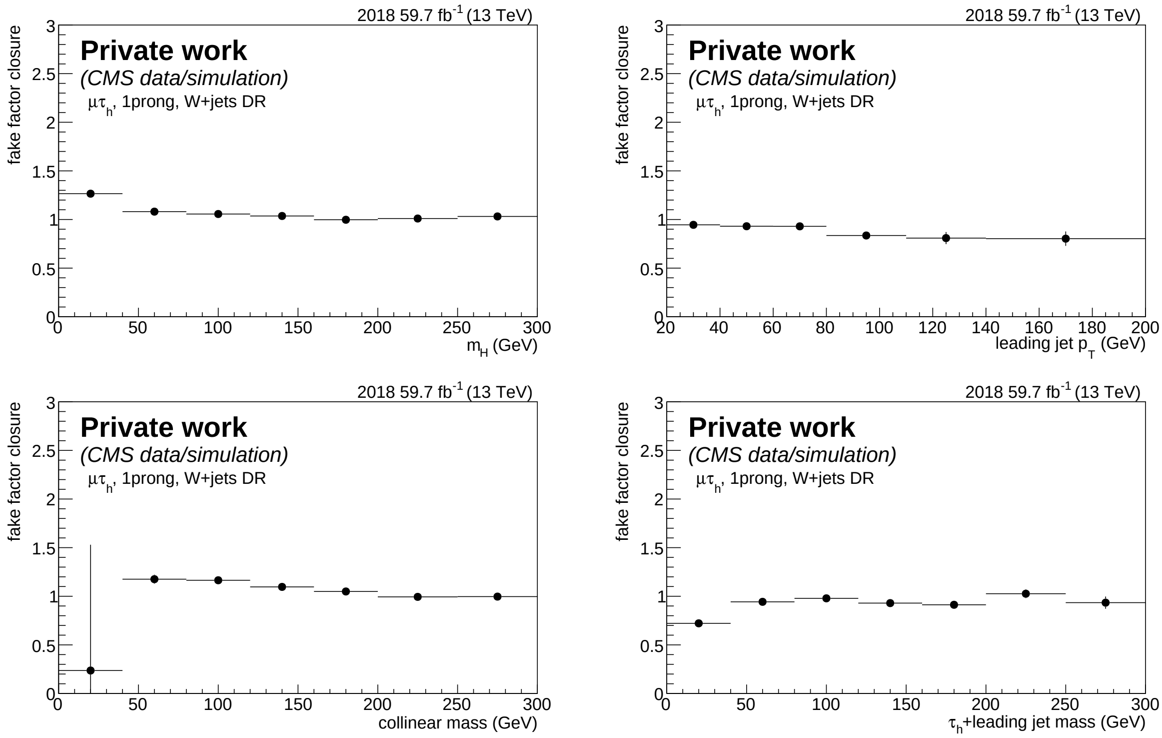


Figure 5.25: Closure of fake factors derived in the W+jets determination region for the 2018 $\mu\tau_\mathrm{h}$ 1-prong channel as function of Higgs boson mass (top left), p_T of jet with highest b-tagging discrimination score (top right), collinear mass (bottom left) and mass of τ_h and leading jet (bottom right). Given that the closure is nearly one, no correction factors are applied.

5.7 Discrimination of signal and background

The discrimination of the signal from the $\mathrm{t\bar{t}}$ and Drell-Yan+jets background is enhanced by using a multiclassifier boosted decision tree (BDT), trained separately for each year and the $\mu\tau_\mathrm{h}$ and $\mathrm{e}\tau_\mathrm{h}$ channel.

5.7.1 BDT Model

The BDT is trained with a gradient boosting library called XGBoost [96]. XGBoost stands for Extreme Gradient Boosting and is based on the gradient boosting framework, designed to build an ensemble of decision trees that corrects the errors of previous trees iteratively. The model is built with 700 decision trees, a learning rate of 0.01, softmax activation function and the categorical cross-entropy loss function. The learning rate defines the step size of the model training by scaling the contribution of each decision tree in the range from 0 to 1. A high learning rate leads to a faster learning process, but can cause overfitting and the model might converge to a suboptimal solution. Overfitting refers to the process when a model learns random fluctuations in the training sample instead of the underlying pattern. A low learning rate, as the chosen 0.01, requires more iterations to find an optimal solution, but gives more accurate results and is more stable. The softmax activation function is used to transform the output scores of the BDT into a probability distribution with the sum of the probabilities of all classes set to one. For a vector $z = [z_1, z_2, ..., z_C]$ of C BDT output scores the softmax function is defined as:

$$\sigma_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \tag{5.9}$$

where σ_i is the probability of class i. The loss function quantifies the difference between the predicted values by the BDT and the true values given in the training sample. The goal of the model is to minimize the loss function. The used categorical cross-entropy loss function is defined as:

$$\text{loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} log(\sigma_{ij}), \tag{5.10}$$

where N is the number of events, C is the number of classes, y_{ij} is the indicator function that is 1 if event i belongs to class j and 0 otherwise, and σ_{ij} is the predicted probability of event i belonging to class j, the output of the softmax activation function.

5.7.2 Input variables

The input variables to train the BDT include variables based on the tau lepton pair, kinematics from the jets with highest b-tagging discriminator value and variables combining the reconstructed Higgs boson and leading jet. The input variables are listed in Tab. 5.8.

The Higgs boson mass is reconstructed using SVFit [97, 98], an algorithm that is using dynamical likelihood techniques with the visible tau decay products and the missing transverse energy as inputs.

The collinear mass is the collinear approximation of the Higgs boson mass, assuming all the decay particles move in the same direction. In this way, only the magnitude of $p_\mathrm{T}^\mathrm{miss}$ needs to be

$\tau\tau$ variables	H_mass	Higgs boson mass
	collinear_mass	Collinear di-tau mass
	Dzeta	D_ζ
	vis_mass	mass of the visible $\tau\tau$ decay products
	Deta	$\Delta\eta$ between light lepton and τ_h
	H_pt	p_T of the reconstructed Higgs boson
	mt	transverse mass
Jet parameters	Jet1_pt	p_T of leading jet
	Jet1_btag	b-tag score of leading jet
	Jet2_pt	p_T of subleading jet
	Jet2_btag	b-tag score of subleading jet
H and jet properties	TauJ_mass	mass of τ_h+leading jet
	DRHJ	ΔR between Higgs boson and leading jet
	HJ_pt	p_T of Higgs boson+leading jet

Table 5.8: BDT inputs variables split into variables from the tau lepton pair, jet parameters and variables combining Higgs boson and jet properties. Leading jet refers to the jet with the highest b-tag score.

considered, not its direction, which is less well reconstructed. The collinear mass is defined as:

$$m_{\text{collinear}} = \frac{m_{vis}}{\sqrt{\frac{p_T^{lep}}{p_T^{lep}+p_T^{\text{miss}}} \cdot \frac{p_T^{\tau_h}}{p_T^{\tau_h}+p_T^{\text{miss}}}}} \tag{5.11}$$

The D_ζ variable indicates how the leptons and the direction of the p_T^{miss} are distributed. It is used to suppress $t\bar{t}$ and W+jets background, where the leptons and the p_T^{miss} are more isotropically distributed compared to the tau lepton pair from the Higgs boson decay. Isotropic means the uniform distribution in all directions. D_ζ is defined as the linear combination of the projection of the p_T^{miss} and the transverse momentum of the visible decay products along the bisector between the ℓ and the τ_h in the transverse plane, referred to as $\hat{\zeta}$:

$$D_\zeta = p_\zeta^{miss} - 0.85 \cdot p_\zeta^{vis}, \tag{5.12}$$

with the projections defined as:

$$p_\zeta^{miss} = \vec{p}_T^{\,miss} \cdot \hat{\zeta} \qquad p_\zeta^{vis} = (\vec{p}_T^{\,\ell} + \vec{p}_T^{\,\tau_h}) \cdot \hat{\zeta}, \tag{5.13}$$

The value of 0.85 was chosen to maximize the separation power between signal and background [99].
Each variable has a specific signal to background discrimination power, which is larger if the distributions are different for the signal and the background processes. The variables with a high discrimination power have the highest importance for training the BDT because of their ability to classify the correct class. The variable with the highest discrimination power is the reconstructed Higgs boson mass, followed by the p_T of the jet with the highest b-tagging discriminator value and the collinear mass. The importance is measured as number of times a variable is used in all the trees in the model. The weight importance ranking of the input variables is shown in Fig. 5.26.

The four input variables with the highest importance are shown in Fig. 5.27 for the 2018 $\mu\tau_h$ and $e\tau_h$ channel, in Fig. 5.28 for the 2017 $\mu\tau_h$ and $e\tau_h$ channel and in Fig. 5.29 for the 2016 $\mu\tau_h$ and $e\tau_h$ channel. The input variables show a good agreement between simulation and data, which is important to avoid any bias due to the evaluation with the BDT.

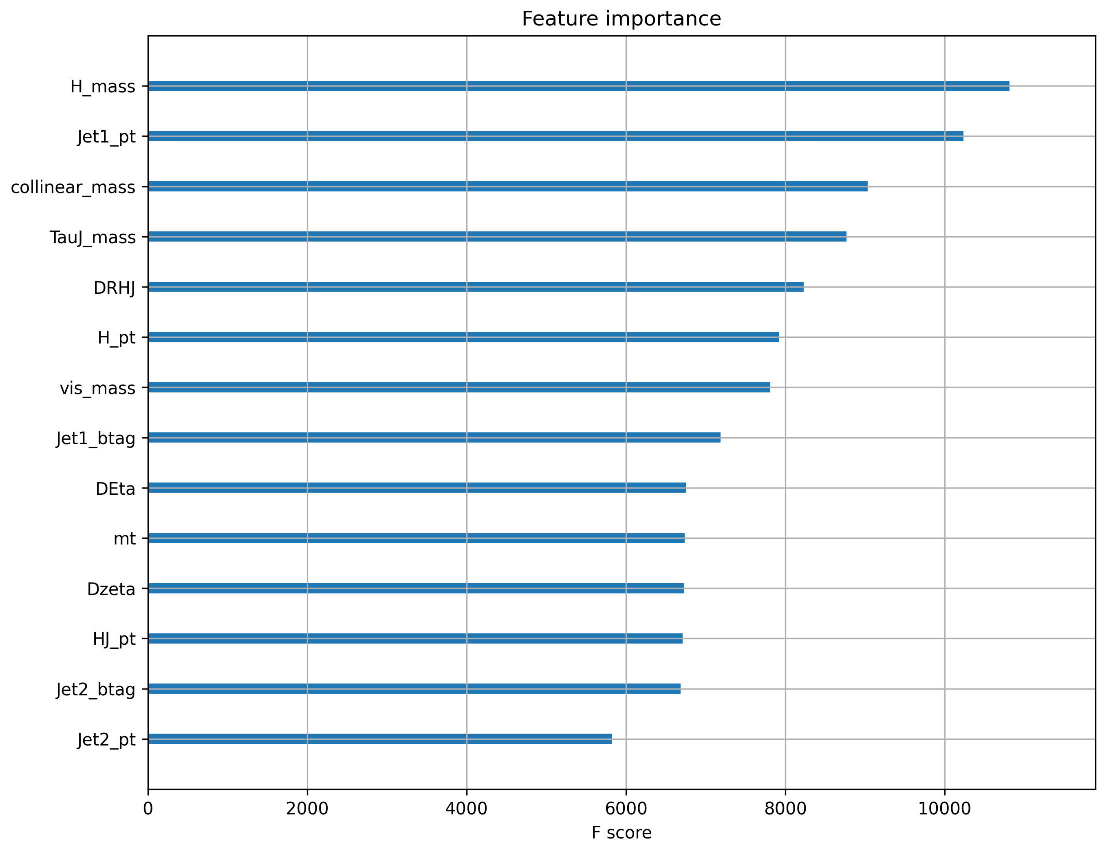


Figure 5.26: Weight importance of the BDT input variables. The F score describes how many times a variable was used in all the trees of the model.

5.7.3 Training and BDT scores

The BDT is trained as a multiclassifier with three classes, the signal combining b-fusion and gluon fusion process, $t\bar{t}$ and Drell-Yan+jets background. The background classes are selected due to their significant contributions. The fake τ_h background and other minor backgrounds are well identified as background by the BDT, so no additional classes were necessary. The b-fusion and gluon fusion signal processes are combined in one class because of similar shapes of the input variables. Two separate signal classes were found to lead to a smaller efficiency because of many b-fusion events getting misidentified as gluon fusion events and vice versa.
40% of the simulated signal and background events are used to train the BDT. Those events are excluded from the further analysis to avoid any bias. The training and testing sets are skimmed to have the same size for all classes to avoid any bias towards the background classes, of which significantly more events are generated.
The output of the BDT are three probabilities, one for each class. All events are evaluated and given a probability for each class. The BDT output scores are constructed for each class taking only events into account if they have the highest probability for the respective class. The BDT scores are fitted using a maximum likelihood fit to derive the limits, as described in the next section. They are shown combined for all data taking years in figure 5.30.

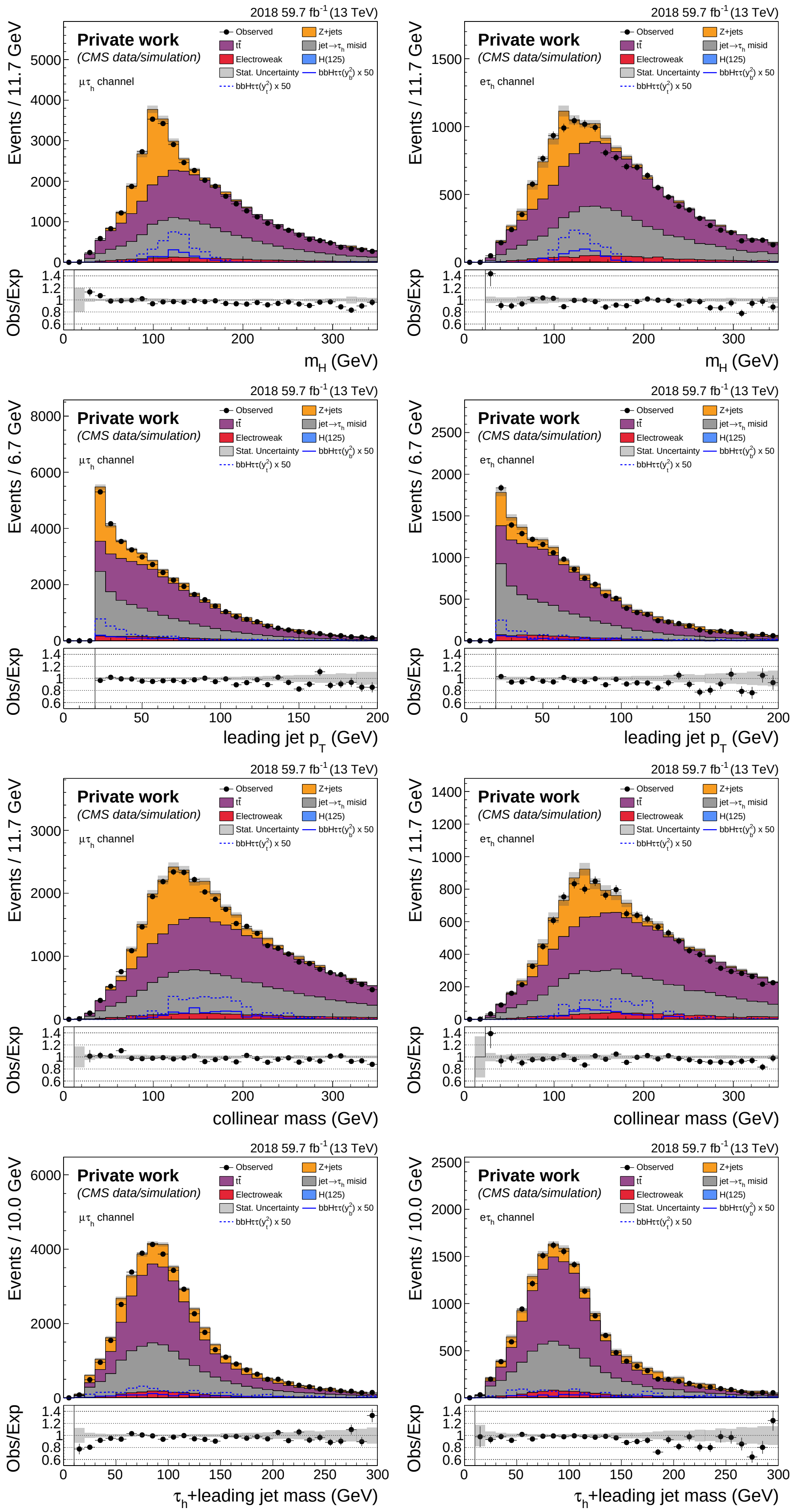


Figure 5.27: BDT inputs with highest importance for the 2018 $\mu\tau_\mathrm{h}$ channel (left) and $e\tau_\mathrm{h}$ channel (right). From top to bottom: Reconstructed Higgs boson mass, p_T of jet with highest b-tagging discriminator value, collinear mass and mass of τ_h plus leading jet. The uncertainty bands only include the statistical uncertainty of the simulated samples.

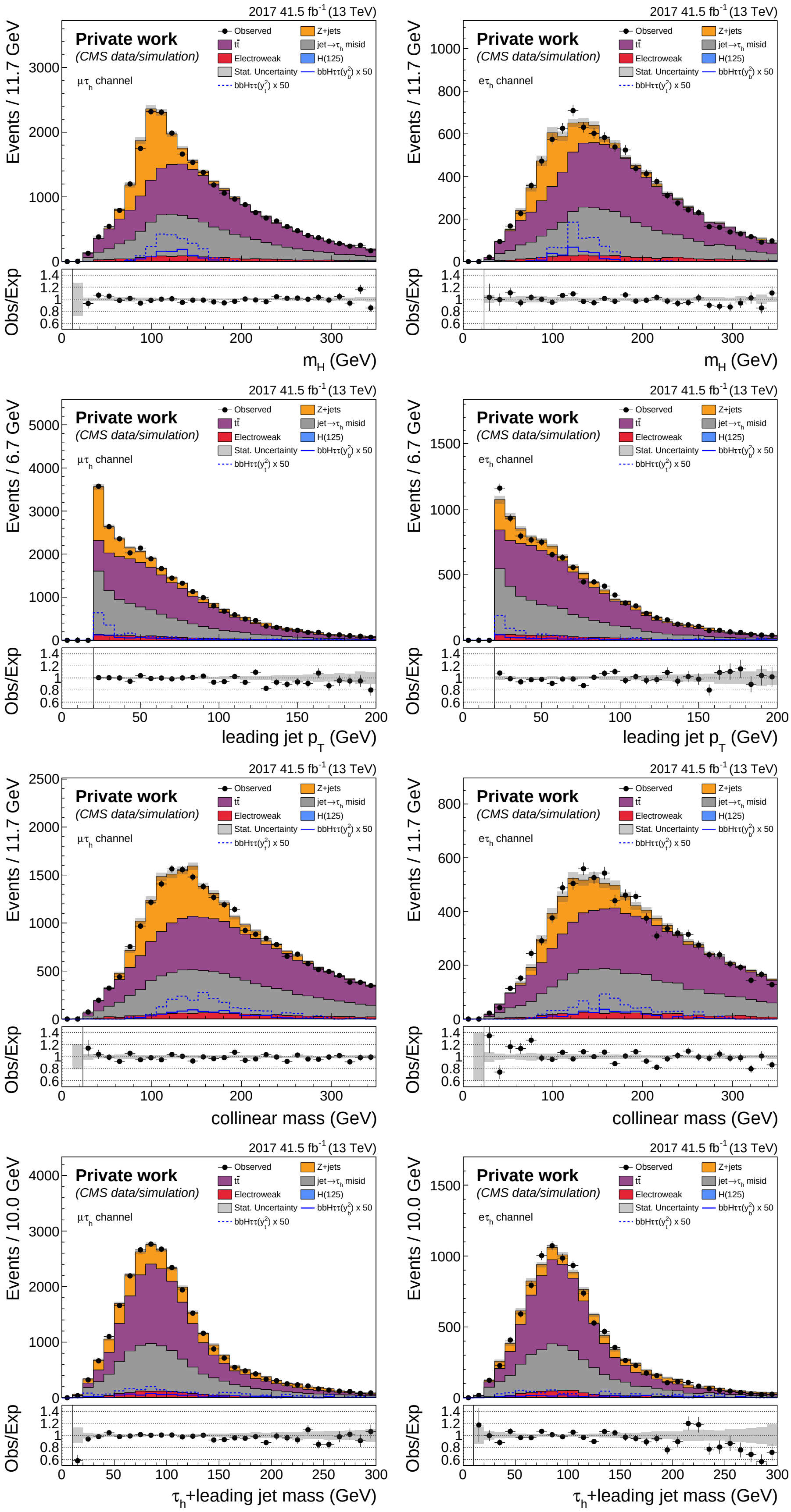


Figure 5.28: BDT inputs with highest importance for the 2017 $\mu\tau_\mathrm{h}$ channel (left) and $e\tau_\mathrm{h}$ channel (right). From top to bottom: Reconstructed Higgs boson mass, p_T of jet with highest b-tagging discriminator value, collinear mass and mass of τ_h plus leading jet. The uncertainty bands only include the statistical uncertainty of the simulated samples.

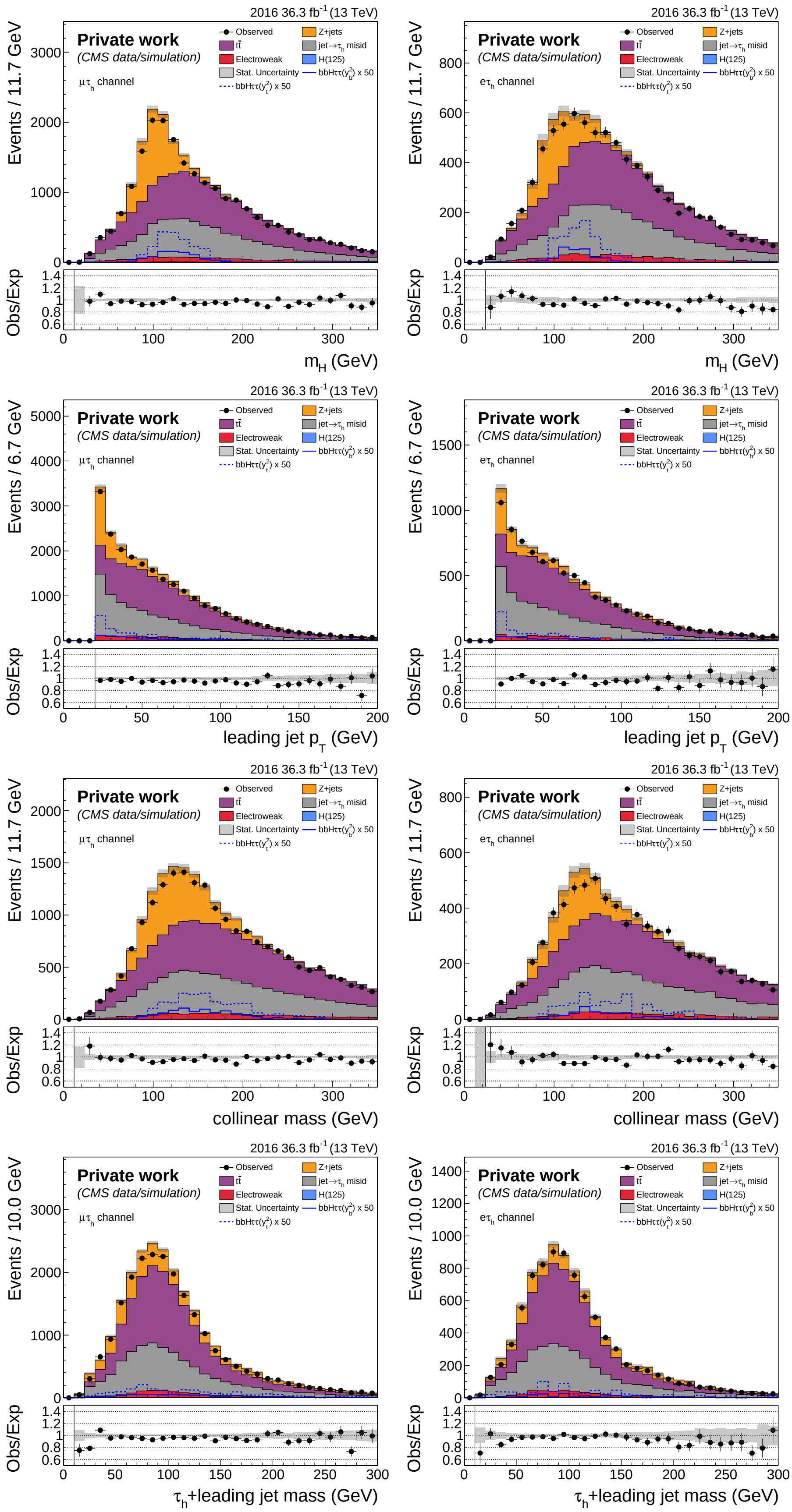


Figure 5.29: BDT inputs with highest importance for the 2016 $\mu\tau_h$ channel (left) and $e\tau_h$ channel (right). From top to bottom: Reconstructed Higgs boson mass, p_T of jet with highest b-tagging discriminator value, collinear mass and mass of τ_h plus leading jet. The uncertainty bands only include the statistical uncertainty of the simulated samples.

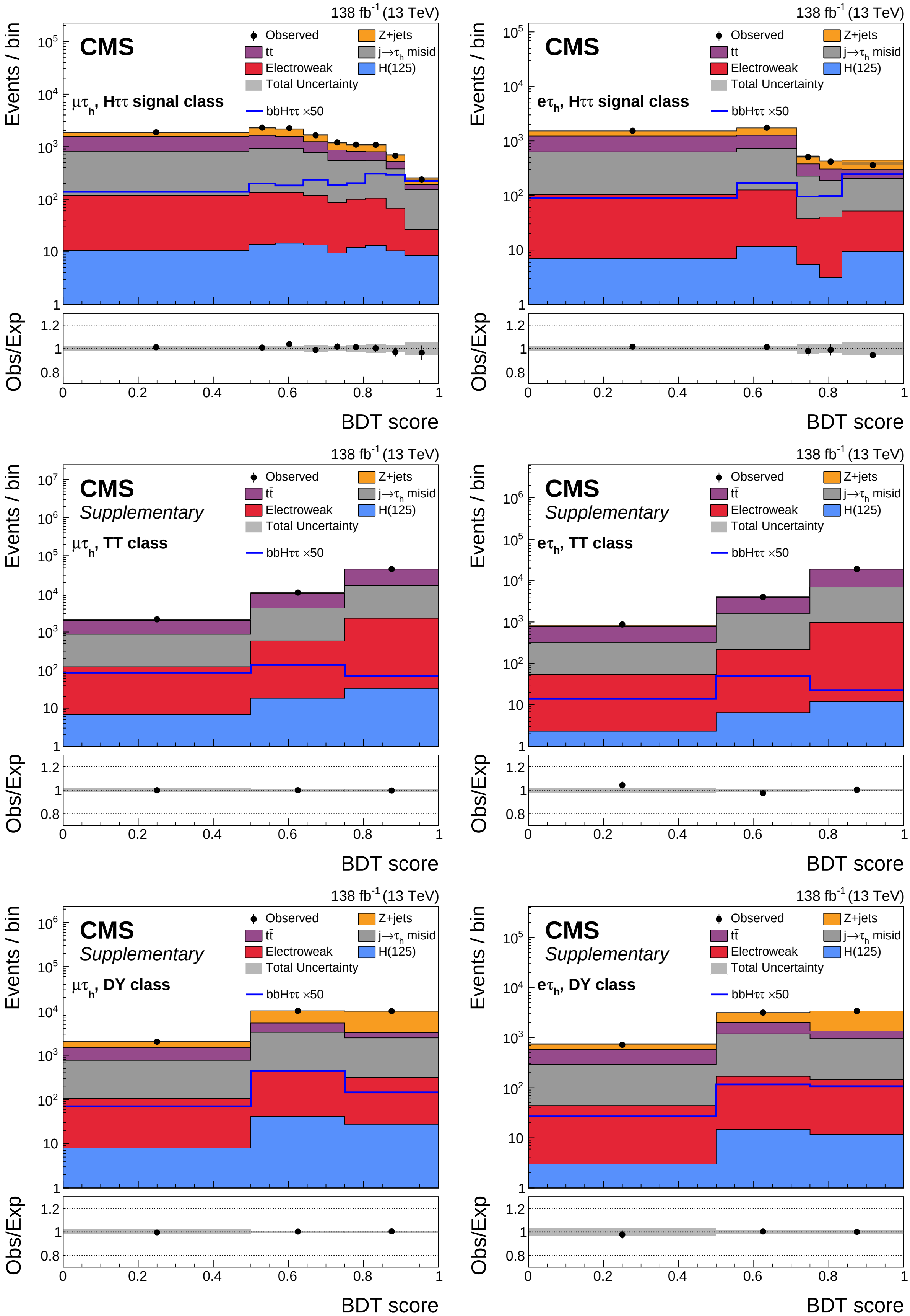


Figure 5.30: BDT output scores for simulation and data for the $\mu\tau_h$ channel (left) and $e\tau_h$ channel (right) for the signal class (top), $t\bar{t}$ background class (center) and Drell-Yan+jets background class (bottom). The plots combine data from all years, but each year is treated separately in the fit.

5.8 Interpretation of results

The bbH signal is accounted for by fitting the BDT output scores with a binned maximum likelihood fit. The likelihood function is constructed as the product of Poisson probabilities to observe n_i events in bin i:

$$\mathcal{L}(\text{data}|\mu,\theta) = \prod_i \frac{(\mu S_i(\theta) + B_i(\theta))^{n_i}}{n_i!} \cdot e^{-\mu S_i(\theta) - B_i(\theta)} \cdot p(\tilde{\theta}|\theta), \tag{5.14}$$

where $S_i(\theta)$ refers to the number of signal events and $B_i(\theta)$ the number of background events. They depend on the systematic uncertainties, introduced in the fit as nuisance parameters (θ). The number of signal events is scaled with the signal strength modifier $\mu = S_{obs}/S_{SM}$. $p(\tilde{\theta}|\theta)$ is the probability density function with $\tilde{\theta}$ the default value of the nuisance parameter and θ the true value.

The data is compared to the signal plus background and the background-only hypotheses using a parameter called the test statistic. The test statistic is based on the profile likelihood ratio, defined as:

$$\tilde{q}_\mu = -2\ln\frac{\mathcal{L}(\text{data}|\mu,\hat{\theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu},\hat{\theta})} \qquad \text{with } 0 \leq \hat{\mu} \leq \mu, \tag{5.15}$$

where $\hat{\theta}_\mu$ is the profiled value of θ that maximizes the likelihood for a given μ, $\hat{\mu}$ and $\hat{\theta}$ are the values of μ and θ that maximize the likelihood. Profiling means, that for every value of μ the nuisance parameters are set to the values that maximize the likelihood for that μ.

The compatibility of the data with the hypotheses is assessed with a p-value, which describes the probability to have a value of q with equal or lesser compatibility with the corresponding hypothesis than the observed $\tilde{q}_{\mu,\text{obs}}$:

$$p = \int_{\tilde{q}_{\mu,\text{obs}}}^{\infty} f(\tilde{q}_\mu|\mu,\theta)d\tilde{q}_\mu, \tag{5.16}$$

where f is the sampling (or probability) distribution of the test statistic $\tilde{q}_q$. The sampling distribution is not known analytically and in general needs to be generated from toy simulated data sets. But in the limit with large number of events f can be described by a formula. The usage of this formula is referred to as the asymptotic approximation [100]. A small p-value of the background-only hypothesis indicates that the observed value is very unlikely under the background-only hypothesis. The p-value is converted to a significance level using the inverse cumulative distribution of the Gaussian distribution. The significance level are given in terms of the standard deviation. The standard deviation (σ) is the square root of the variance, that is the average of the squared differences from the mean:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}, \tag{5.17}$$

where N is the number of events and $\overline{x}$ the mean of the distribution. It is a measure of how much the events deviate from the mean. In a Gaussian distribution one standard deviation ($\pm 1\sigma$) includes 68.27% of the events, two standard deviations 95.45% and three standard deviations 99.73% of the events. The higher the significance the less likely an excess is just from statistical

fluctuation. In particle physics, the threshold for an observation is usually taken to be 5 standard deviations with respect to the background-only hypothesis.
The upper limit on the signal strength is derived with the asymptotic 95% CL_s criterion [100–102], defined as:

$$CL_s = \frac{p_{s+b}}{1 - p_b}. \tag{5.18}$$

Asymptotic refers to assumption of the asymptotic approximations being true. The 95% confidence level means, that in 5% of the time a signal strength with a value larger than the upper limit is wrongly excluded.

5.9 Systematic uncertainties

Various sources of experimental and theoretical systematic uncertainties can change the results and need to be estimated. The uncertainties are included as nuisance parameters in the fit. Some uncertainties only affect the normalization of the BDT output scores, that are fit to derive the limits. Those uncertainties are given in percentage and are applied either to some specific signal or background contributions or to all of them. Other uncertainties affect the normalization and shape of the BDT output score. The shape uncertainties are estimated by constructing histograms with the uncertainty varied by one standard deviation from its nominal value. A function interpolates from the nominal and one standard deviation templates to all values of the uncertainty. If the source of an uncertainty impacts multiple data-taking years, processes or channels in a consistent way, the uncertainty is correlated across several years, processes or channels, otherwise uncorrelated. In table 5.9 all the uncertainties are summarized. The uncertainties only applied in the $\tau_h\tau_h$ and eμ channel are marked and discussed in section 5.10.

5.9.1 Normalization uncertainties

The uncertainties described in the following affect the normalization of the BDT output scores of the signal and background processes. A log-normal probability distribution is assigned to those uncertainties. The uncertainties are correlated between channels and data taking periods with exception of the luminosity uncertainty, which contains parts correlated and uncorrelated between different data taking periods.

Luminosity: The integrated luminosity, used to normalize the simulated signal and background samples, introduces a systematic uncertainty, that is estimated as 1.2% for 2016, 2.3% for 2017 and 2.5% for 2018 [103–105].

Electron and muon identification: The uncertainties arising from the electron and muon identification efficiency are estimated to be 2% [73].

Cross section of Higgs boson production: Uncertainties on the cross section for the Higgs boson production modes arise from higher order QCD and electroweak corrections. The uncertainties are estimated as -24% to 20.1% for the bottom quark fusion and the interference term and -7% to 4.6% for the gluon fusion. An additional uncertainty ranging from -31% to 47% is introduced to cover theoretical uncertainties in the gluon splitting to $b\bar{b}$ in the gluon fusion process. The uncertainties on the cross section of the Higgs boson production modes used as background in the search are estimated as 8% for top quark associated production, 0.9% for the production associated with a Z boson, 0.8% for the production associated with a W boson and 0.5% for the vector boson fusion [81, 84].

PDF+α_s in Higgs boson production: Uncertainties arising from the parton distribution function (PDF) and from the value of the strong coupling constant α_s, used in the evaluation of the PDF, are estimated as 3.6% for the top quark associated production, 3.2% for the gluon fusion production, 2.1% for the vector boson fusion, 1.8% and 1.3% for the production associated with a W boson and Z boson, respectively [84].

Cross section of major background processes: The uncertainty on the cross section is estimated as 6% for $t\bar{t}$, 5% for the single top quark and diboson, 4% for W+jets and 2% for Drell-Yan+jets production.

Branching ratio of the Higgs decay to tau lepton pair: The branching ratio of the Higgs boson to $\tau\tau$ has an uncertainty of 2.1% [84].

5.9.2 Shape uncertainties

The shape uncertainties affect both the normalization and the shape of the BDT output scores. If not mentioned otherwise, the uncertainties are correlated across channels and uncorrelated across data taking years.

τ_h identification: The uncertainties on the τ_h discriminator efficiency scale factors are applied as a function of the τ_h p_T and decay mode and have a range of 2% to 3%.

Tau energy scale: The uncertainty on the tau energy scale depends on the τ_h decay mode and alters the correction to the τ_h p_T and mass. In order to propagate the effect of the altered distributions to the BDT output scores, all the events are re-evaluated with each tau energy scale variation. The uncertainty has a range of up to 3% and is uncorrelated between τ_h decay modes and years.

Lepton to τ_h fake rate: Uncertainties in the lepton to τ_h fake rates are applied as function of the τ_h η and reach 2%.

Trigger efficiency: The uncertainties on the trigger efficiency scale factors depend on the lepton p_T and η, as described in section 5.4.2. The uncertainty is uncorrelated between channels and years.

Jet energy scale and resolution: The jet energy scale and resolution variations include a set of uncertainties based on sources such as the detector efficiency during data taking, the pileup interactions and the clustering efficiency. The uncertainties are applied to all the jets and p_T^{miss} of an event. The uncertainties alter the jet p_T and mass, and the transverse momentum and azimuthal angle of p_T^{miss}. An additional uncertainty is applied to the p_T^{miss} to account for particle flow candidates not clustered in jets. In order to propagate the effect of the altered distributions to the BDT output scores, all the events are re-evaluated with each jet energy scale variation. Some of the uncertainties are specific for each data taking year, others are correlated across years.

B-tagging efficiency: The uncertainty on the b-tagging efficiency is estimated from a set of b-tagging scale factor uncertainties. Uncertainties are derived for the limited statistics in the regions where the scale factors are derived and contamination of light jets in the b-enriched region as well as contamination of b-jets in the light jet enriched region. Uncertainties from the jet energy scale are propagated to the b-tagging scale factors. Some of the uncertainties are specific for a data taking year, others are correlated. The uncertainty has a range from 0.5% to 3%, depending on the year.

Fake factor method: The uncertainties based on the fake factor method or due to the choice of the fit function, the definition of the determination region and the background contributions in the application region, as discussed in section 5.6.5. The uncertainties have a range from 5% to 7%.

Top p_T reweighting: The uncertainty is related to the top p_T dependent correction applied to the $t\bar{t}$ simulated samples. The uncertainty is applied in a conservative manner, applying it twice for the upward variation and taking the nominal for the downward variation.

DY+jets reweighting: The normalization correction of the DY+jets simulated samples, discussed in section 5.4.8, introduces an uncertainty estimated by applying the correction twice for the upward variation and taking the nominal for the downward variation. The uncertainty due to the $t\bar{t}$ contribution is estimated by varying the cross section of the $t\bar{t}$ contribution by its uncertainty and measuring the effect on the reweighting. The uncertainty is uncorrelated between the one b-jet and two b-jets regions.

Prefiring: Uncertainties related to the effect of prefiring are applied to the 2016 and 2017 data taking periods and range up to 4% in the endcap region.

Renormalization and factorization scale: The choice of the renormalization (μ_R) and factorization scale (μ_F) in the generation of simulated samples introduces an uncertainty, that is estimated by constructing an envelope of scale variations. The μ_F and μ_R are varied by 0.5 and 2, excluding the combination where one scale is half and the other is double its nominal value. The uncertainty is applied to the signal, $t\bar{t}$ and Drell-Yan+jets samples.

Parton shower: Theoretical uncertainties in the parton shower, used in the generation of the simulated samples, are estimated by varying the scale at which the strong coupling constant is evaluated in the shower. Independent and uncorrelated uncertainties are derived for the initial state radiation and the final state radiation. The uncertainty is applied to the signal, $t\bar{t}$ and Drell-Yan+jets samples.

5.9.3 Impacts

The *impact* of a nuisance parameter on the measurement of the signal strength μ is defined as the difference in μ if the fit is performed with the nuisance parameter shifted by one standard deviation of the postfit uncertainty:

$$\text{impact} \;=\; \Delta\mu(\theta = \theta_{\text{postfit}} \pm \sigma_{\text{postfit}}) \tag{5.19}$$

The pull of a nuisance parameter indicates how compatible the preferred value by the data is to the nominal value. The pull is defined as the difference to the prefit value, normalized by the change in the uncertainty:

$$\text{pull} \;=\; \frac{\theta_{\text{postfit}} - \theta_{\text{prefit}}}{\sqrt{\sigma^2_{\text{prefit}} - \sigma^2_{\text{postfit}}}}. \tag{5.20}$$

In a similar way the fitted value of the nuisance parameter is defined as:

$$\text{fit} \;=\; \frac{\theta_{\text{postfit}} - \theta_{\text{prefit}}}{\sigma_{\text{prefit}}}. \tag{5.21}$$

If the uncertainty of the fit is smaller than ± 1 it means the nuisance parameter is constraint by the data.

In Fig. 5.31 the uncertainties with the largest impact, for all channels combined, are shown. The uncertainties with the largest impact are the theory uncertainty on the cross section of the gluon fusion and the bottom quark fusion process, an uncertainty introduced by the variation of the b-tagging scale factors and statistical uncertainties in the different bins of the BDT output scores.

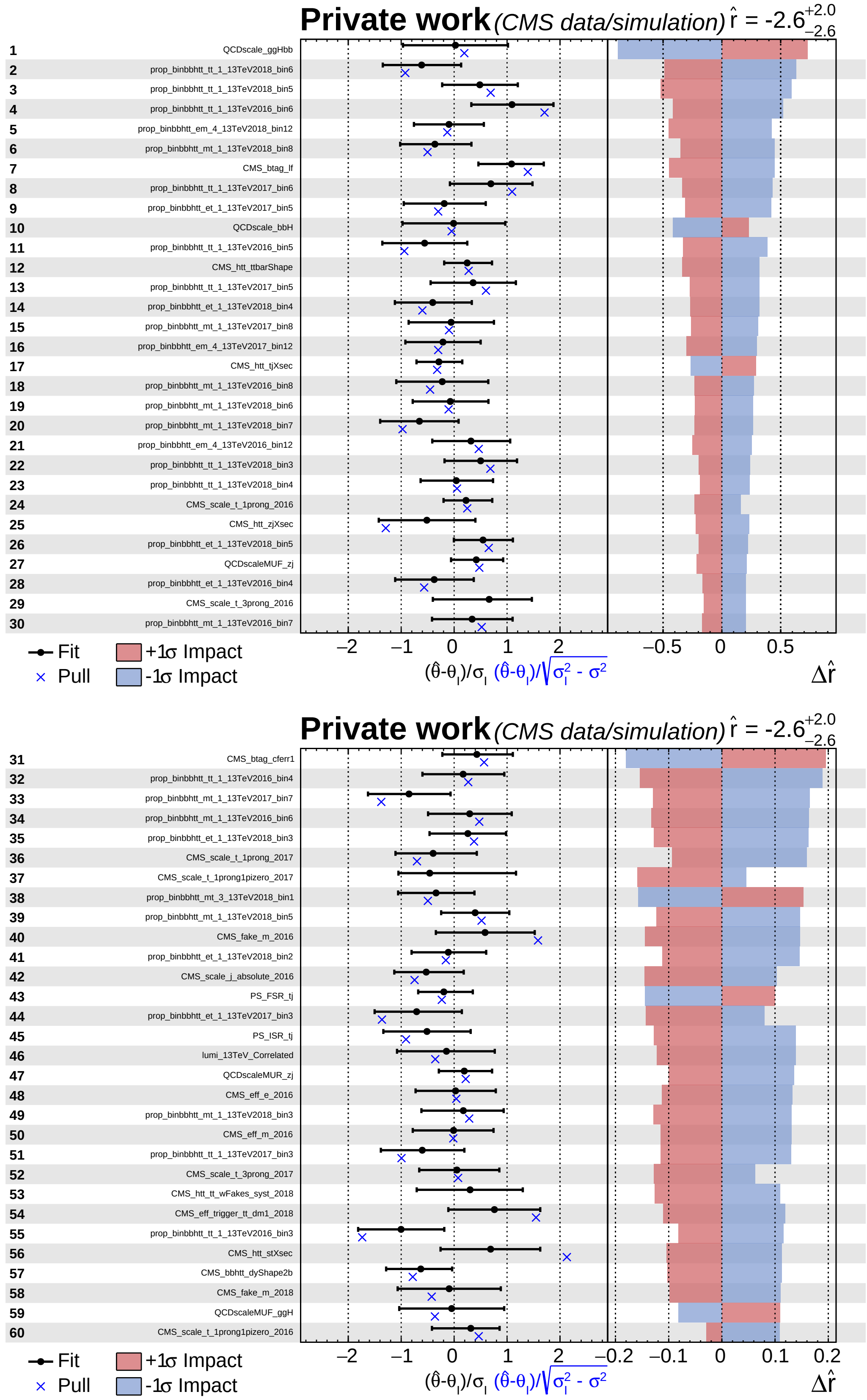


Figure 5.31: Pulls and impacts for the systematic uncertainties with the largest impact in the measurement of the signal strength with all channels combined. The uncertainties with the name *prob_binX* are statistical uncertainties of the BDT output score bins for the different channels and years.

Normalization uncertainties	Value	Process affected
Luminosity	2016: 1.0% 2017: 2.0% 2018: 1.5%	MC
μ/e identification	2%.	MC
Higgs boson cross section	-24–20.1%	$bbH(y_b^2)$
	-7–4.6%	$bbH(y_t^2)$
	0.5–8%	Higgs boson bkg.
Gluon splitting to $b\bar{b}$	-31–47%	$bbH(y_t^2)$
$\mathrm{t\bar{t}}$ cross section	6%	$\mathrm{t\bar{t}}$
DY+jets cross section	2%	DY+jets
W+jets cross section	4%	W+jets
VV cross section	5%	VV
Single top quark cross section	5%	single top
$H \to \tau\tau$ branching ratio	2.1%	$H \to \tau\tau$
PDF+α_s	1.3–3.6%	Higgs boson
$\mathrm{t\bar{t}V}$ +jets cross section (eμ channel)	15%	$\mathrm{t\bar{t}V}$ +jets
eμ trigger (eμ channel)	1.5%	MC
ABCD method (eμ channel)	event-dep.	non-prompt ℓ
$H \to WW$ BR (eμ channel)	1.5%	$H \to WW$
Shape uncertainties	Value	Process affected
Single μ/e trigger	p_T and η dep.	MC
τ_h identification	p_T and DM dep. (2–3%)	MC
τ_h energy scale	DM dep. (< 3%)	MC
b tagging	0.5–3%	MC
$\mu(e) \to \tau_\mathrm{h}$ fake rate	τ_h η dep. (< 2%)	MC
Jet energy scale	event-dep.	MC
Jet energy resolution	event-dep.	MC
$p_\mathrm{T}^\mathrm{miss}$ unclustered energy scale	event-dep.	MC
Top quark p_T reweighting	p_T dep.	$\mathrm{t\bar{t}}$
DY+jets reweighting	p_T and mass dep.	DY+jets
Fake factor method	event-dep. (5–7%)	$\mathrm{j} \to \tau_\mathrm{h}$ fakes
μ_R/μ_F scale	event-dep.	bbH, DY+jets, $\mathrm{t\bar{t}}$
Parton shower	event-dep.	bbH, DY+jets, $\mathrm{t\bar{t}}$
Pre-firing	event-dep. (< 4%)	MC
Bin-by-bin statistical uncertainty	$\sqrt{N_\mathrm{events}}$	All
$\mathrm{j} \to e/\mu$ fake rate (eμ channel)	jet p_T dep.	MC
τ_h trigger ($\tau_h\tau_h$ channel)	p_T and DM dep.	MC

Table 5.9: Summary of all the uncertainties, split into normalization uncertainties that only affect the normalization of the BDT scores used for the fit, and the shape uncertainties, which also affect the shape.

5.10 Fully leptonic and fully hadronic channel

The $\tau_h\tau_h$ and eμ channels described in this section were developed by other members of the CMS Collaboration.
The background composition in the $\tau_h\tau_h$ channel is dominated by the fake τ_h background, Drell-Yan plus additional jets, $t\bar{t}$ and diboson.
Similar to the semileptonic channel as described in section 5.3.3, the fake τ_h background is estimated using a data-driven fake factor method. A determination region enriched with QCD multijets events is defined, requiring the hadronic tau leptons to have the same sign charge (SS). The fake factor is derived in this determination region as ratio of τ_h candidates which pass the nominal WP of the DeepTau discriminator against jets to the τ_h which fail this requirement, but pass a looser working point of the same discriminator. The fake factor is derived as a function of the τ_h p_T, the τ_h decay mode and the jet multiplicity. The fake factor is applied to the data in the application region, exclusive to the signal region by reverting the DeepTau discriminator against jets, to estimate the fake contribution in the signal region.
In the eμ channel the dominant backgrounds are $t\bar{t}$, Drell-Yan, diboson and QCD multijet. The QCD multijet contribution is estimated with a data-driven method called ABCD method, discussed in more detail in Ref. [73]. The method requires four regions, the signal region and three control regions, as shown in Fig. 5.32. The control regions are defined with the same sign charge (SS) of the electron muon pair and a reverted muon isolation. A transfer factor is derived in the two regions with inverted muon isolation, from the same sign charge (SS) to the opposite sign charge (OS) region. The transfer factor is derived as function of the separation between electron and muon and jet multiplicity. The transfer factor is then applied to the SS region with the same muon isolation as the signal region (OS) to estimate the QCD multijet contribution in the signal region.

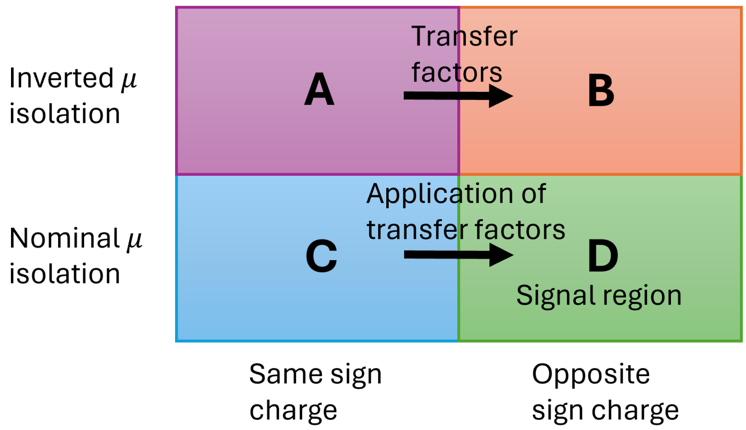


Figure 5.32: ABCD method used in the eμ channel with D being the signal region.

All background processes, aside from the fake τ_h background in the $\tau_h\tau_h$ channel and the QCD multijet background in the eμ channel, are simulated with Monte Carlo samples. Corrections to account for any differences between simulation and data are applied, similar than described for the semileptonic channels in section 5.4. In the $\tau_h\tau_h$ channel a $e/\mu \rightarrow \tau_\mathrm{h}$ fake rate is applied to the simulated samples to account for any leptons misidentified as τ_h and for the eμ channel the simulated samples are corrected for the fake rate of jet$\rightarrow e/\mu$, estimating the rate of non-prompt leptons.

In the $\tau_h\tau_h$ channel events are selected if they pass a double tau trigger and have two hadronic tau leptons and at least one b-tagged jet. The charge of the τ_h have to be opposite and they need to be separated by at least $\Delta R \geq 0.5$. A veto is set on any electrons and muons.
In the eμ channel the events must pass a trigger requiring an electron-muon pair and have one electron and one muon together with at least one b-tagged jet. The charge of the electron and muon have to be opposite and they are required to be separated by at least $\Delta R \geq 0.3$. A veto is set on electron and muon pairs. The key variable to separate the $H \to \tau\tau$ from the $H \to$WW signal is the D_ζ. The selection for the $\tau_h\tau_h$ and the eμ channel are listed in Tab. 5.10.

$\tau_h\tau_h$ channel	eμ channel
2 τ_h	1 e, 1μ
$\geq$ 1 b-jet	$\geq$ 1 b-jet
$q^i_{\tau_h} q^j_{\tau_h} = -1$	$q_e q_\mu = -1$
$\Delta R({\tau_h}^i, {\tau_h}^j) > 0.5$	$\Delta R(e, \mu) > 0.3$
veto on e and μ	veto on 2μ and 2e

Table 5.10: Signal region selection requirement for the $\tau_h\tau_h$ and eμ channel.

The sensitivity of the signal is increased using a multiclass Boosted Decision Tree with four classes. In the $\tau_h\tau_h$ channel the classes are Drell-Yan plus Higgs boson background, $t\bar{t}$ background, jet τ_h fakes and the signal class. In the eμ channel the classes are Drell-Yan, $t\bar{t}$ background and two signal classes, one for the $H \to WW$ and the other for the $H \to \tau\tau$ process. The input variables are similar than for the semileptonic channels, described in section 5.7, with a few notable differences to account for the different background contributions. In the eμ channel the p_T of the muon and electron are used, and in the $\tau_h\tau_h$ channel the p_T of both τ_h and the p_T^{miss}.
The BDT output scores for the $H \to \tau\tau$ signal class for the $\tau_h\tau_h$ channel and the $H \to WW$ signal class for the eμ channel are shown in Fig. 5.33.

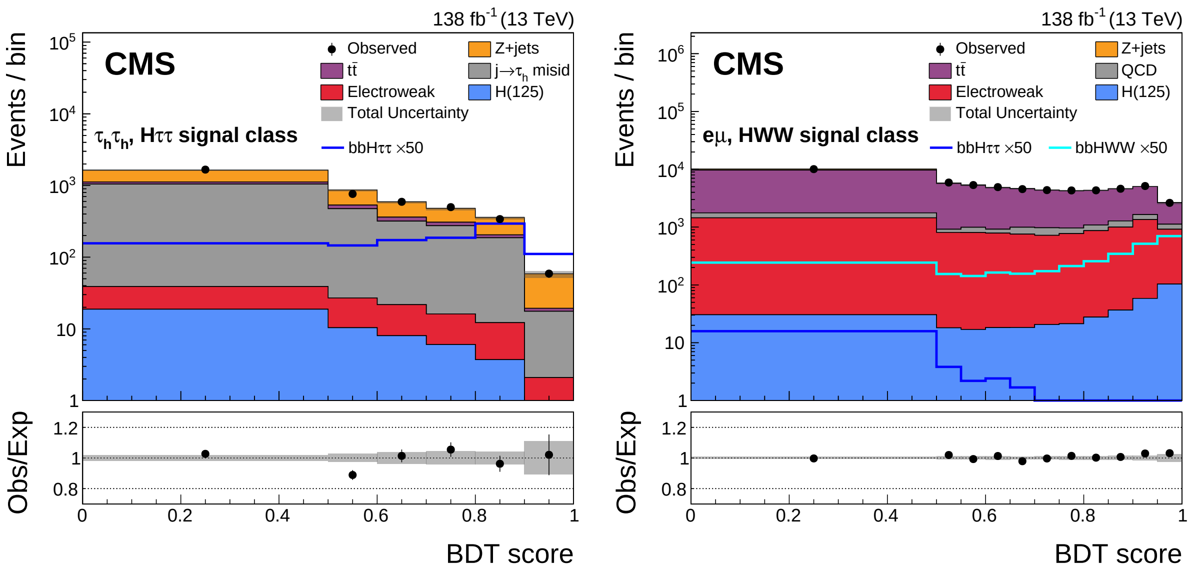


Figure 5.33: BDT output scores for the $H \to \tau\tau$ signal class for the $\tau_h\tau_h$ channel (left) and the $H \to$WW signal class for the eμ channel (right). The plots combine data from all years, but each year is treated separately in the fit. The background classes and for the eμ channel the less sensitive $H \to \tau\tau$ signal class can be found in the supplementary material of the paper [83].

The $\tau_h\tau_h$ and eμ channel have specific systematic uncertainties applied additional to the ones discussed in secion 5.9.

$\tau_h\tau_h$ trigger efficiency: The uncertainties on the trigger efficiency scale factors depend on the τ_h p_T and decay mode and are applied to each leg of the trigger. The uncertainties are provided by the TAU POG.

Fake factor method ($\tau_h\tau_h$): Uncertainties in the fake factor method used in the $\tau_h\tau_h$ channel include statistical uncertainties in the derivation of the fake factor and the non-closure correction, and an uncertainty on the correction applied because of the extrapolation of the fake factor from the determination region to the application region.

eμ trigger efficiency: For the efficiency of the eμ trigger a normalization uncertainty of 1.5% is applied, uncorrelated between years.

Branching ratio of the Higgs decay to WW (eμ): The branching ratio of the Higgs boson to WW has an uncertainty of 1.5%.

jet to e/μ fake rate (eμ): Scale factor uncertainties are applied as a function of the jet p_T for the light jets to light lepton fake rate and for the b-jet and c-jet to light lepton fake rate.

ABCD method (eμ): Uncertainties are applied on the extrapolation factors used in the ABCD method to estimate the QCD multijet background, including statistical and non-closure uncertainties.

Cross section of $\mathrm{t\bar{t}V}$ +jets (eμ): The uncertainty on the cross section is estimated as 15% for $\mathrm{t\bar{t}V}$ +jets

5.11 Results

As explained in section 5.8 all the BDT output scores are simultaneously fit in a maximum likelihood fit, combining the channels $\tau_h\tau_h$, $\mu\tau_h$, $e\tau_h$ and $e\mu$ and all data taking years. No significant excess over the background-only hypothesis was observed. The 95% confidence level upper limit on the signal strength is observed as 3.7 times the SM expected value of the production cross section, and expected was 6.1 times the SM value. In order to observe the process, an expected upper limit close to the SM expected value is necessary. The observed and expected limits for each channel and year are shown in Tab. 5.11.

channel	-2σ	-1σ	median	$+1\sigma$	$+2\sigma$	observed
$\tau_h\tau_h$	3.7	5.6	8.5	14.6	25.5	7.8
$\mu\tau_h$	5.1	7.3	11.6	19.8	34.4	8.6
$e\tau_h$	8.0	11.5	18.0	30.8	53.2	10.1
$e\mu$	8.6	12.3	19.1	31.8	53.0	18.7
2016	4.5	6.4	10.0	17.0	29.2	9.8
2017	4.9	7.1	11.3	19.5	34.2	8.5
2018	4.2	6.1	9.6	16.4	28.2	4.8
combined	2.7	3.9	6.1	10.4	18.0	3.7

Table 5.11: Observed and expected 95% confidence level upper limits on the signal strength.

The $\tau_h\tau_h$ channel is most sensitive to the signal, followed by the $\mu\tau_h$, $e\tau_h$ and $e\mu$ channels. The differences in the channels are mainly caused by the branching ratio and the different background compositions in each channel. The $\mu\tau_h$ channel is more sensitive than the $e\tau_h$ channel due to a higher selection efficiency of the muons compared to the electrons. Contrary to expectations based on the integrated luminosity, the data taking year 2016 is more sensitive than 2017. This difference can be explained by lower p_T thresholds of the triggers used in 2016 compared to 2017.

In Fig. 5.34 the result is visually represented with the one and two standard deviation expected uncertainty bands.

The observed best fit value is $\mu = -2.7^{+2.0}_{-2.6}$, with the uncertainty including the statistic and systematic uncertainties. The negative signal strength indicates an underfluctuation of data with respect to the background-only hypothesis. The observed result is compatible with the $\mu = 1$ hypothesis at the two standard deviation level.

Limits are derived on the Higgs boson coupling modifiers to bottom quarks (κ_b) and top quarks (κ_t). The coupling modifiers are derived as the ratio of the measured Yukawa coupling over the Yukawa coupling expected by the Standard Model:

$$\kappa_b = \frac{y_b}{y_{b,SM}} \quad \kappa_t = \frac{y_t}{y_{t,SM}} \tag{5.22}$$

The coupling modifiers scale the production cross sections and the partial decay widths, referring to the decay rate. The gluon fusion signal process ($bbH(y_t^2)$) has contributions from κ_b and κ_t and is therefore scaled with $1.04\kappa_t^2 + 0.002\kappa_b^2 - 0.038\kappa_t\kappa_b$ [84, 106]. The bottom quark fusion process ($bbH(y_b^2)$) is scaled with κ_b^2 and the interference term with $\kappa_t\kappa_b$. In order to constrain κ_t, the results of an inclusive SM $H \to \tau\tau$ measurement performed by the CMS Collaboration are used. This analysis rejects events containing b-jets, and is therefore exclusive to our analysis [107]. The coupling modifier of the Higgs boson to the tau lepton (κ_τ) is left floating in order to capture any modification in the branching fraction. The 95% and 68% confidence level simultaneous limits

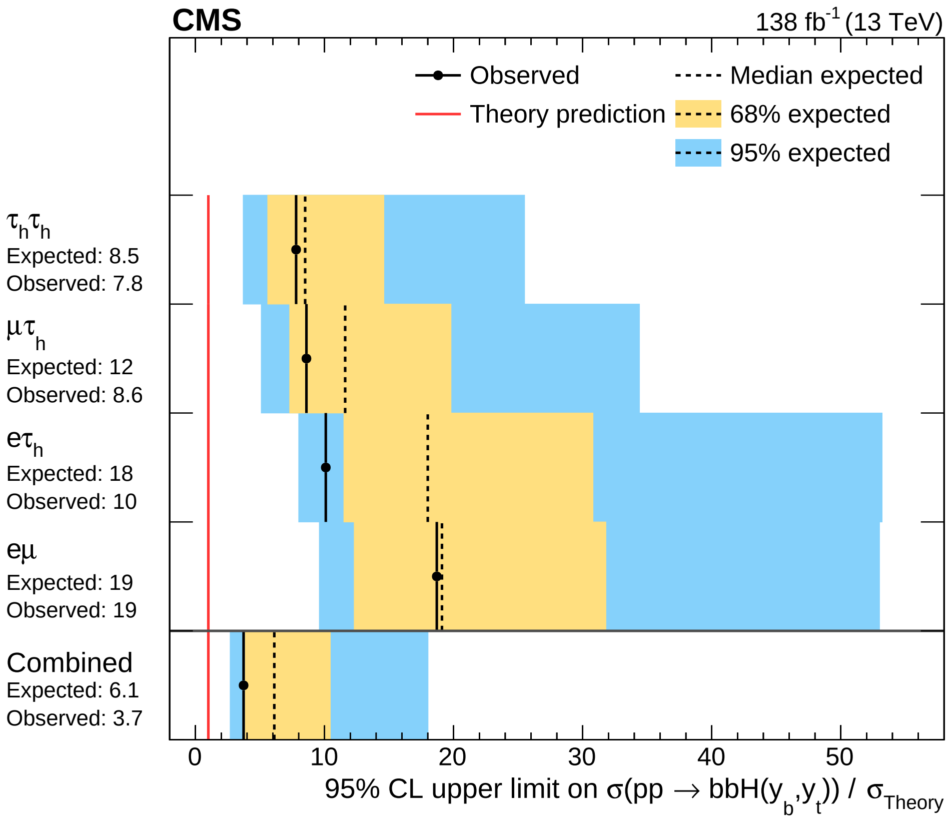


Figure 5.34: 95% confidence level upper limits on the signal strength for each channel and combined. The standard model expected value of $\mu = 1$ is included as theory prediction.

on the κ_b and κ_t are shown in Fig. 5.35. The combination of our analysis with the inclusive CMS $H \to \tau\tau$ search is able to constrain the limits on the κ_b while the constraints on the κ_t come mostly from the inclusive search. The best fit value is observed at $(\kappa_t, \kappa_b) = (-0.73, 1.58)$. The Standard Model expected value of $(\kappa_t, \kappa_b) = (1, 1)$ is included in the observed 95% confidence level limit.

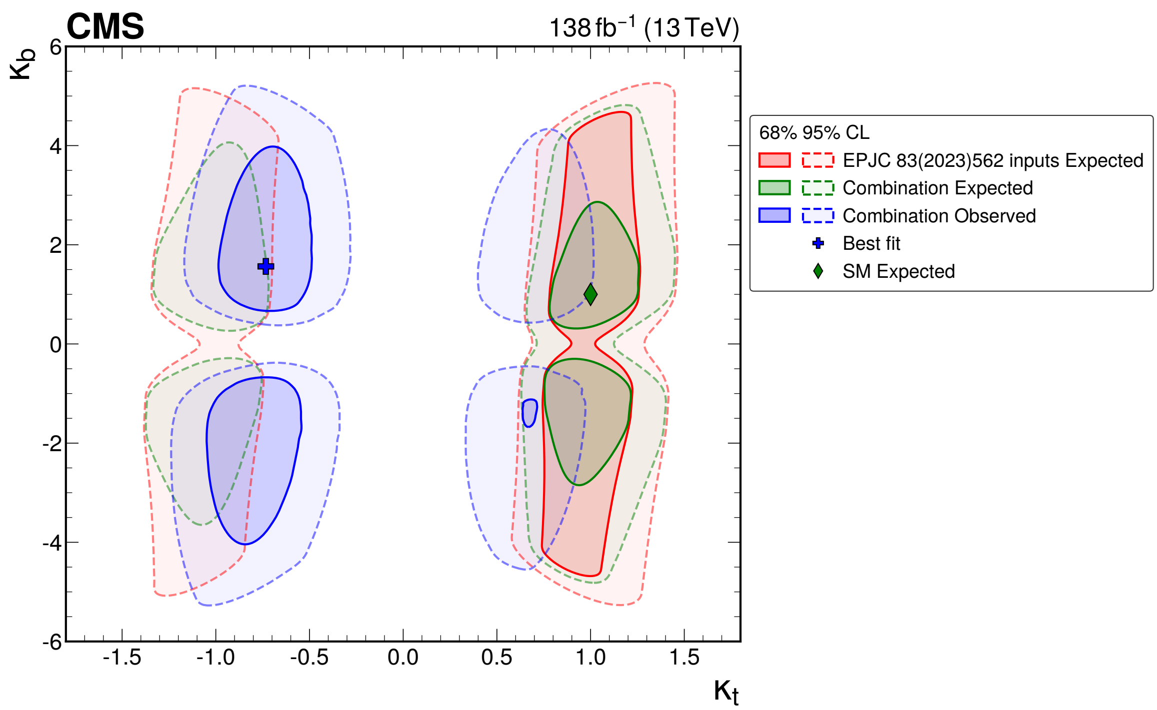


Figure 5.35: Limits on the Higgs boson coupling to bottom quarks (κ_b) as function of the Higgs boson coupling to top quarks (κ_t). The κ_t is constraint by using datacards from the CMS $H \to \tau\tau$ analysis published as *EPJC 83(2023)562*. The solid lines represent the 68% confidence level and the dashed lines the 95% confidence level. The contribution from *EPJC 83(2023)562* is shown in red, the combined expected limit in green and the combined observed limit in blue. The best fit value at $(\kappa_t, \kappa_b) = (-0.73, 1.58)$ is marked with the blue cross, the Standard Model expected value at $(\kappa_t, \kappa_b) = (1, 1)$ as green diamond.

5.12 Summary

The search for the bottom quark associated production of the Higgs boson is conducted using the data collected by the CMS experiment from 2016 to 2018, corresponding to an integrated luminosity of 138 fb^{-1}. The search is conducted in the final state with at least one bottom quark jet and with the Higgs boson decay to a tau lepton pair with the four final states $\tau_h\tau_h$, $\mu\tau_h$, $e\tau_h$ and $e\mu$, and in the Higgs boson decay to WW in the $e\mu$ final state. My contribution was the design of all stages of the analysis in the semileptonic final states, $\mu\tau_h$ and $e\tau_h$. The signal to background discrimination is enhanced by using machine learning. The 95% confidence level upper limit on the signal strength was observed as 3.7 times the expected Standard Model value and expected was 6.1 times the Standard Model value. The signal strength was observed as $\mu_{obs} = -2.7^{+2.0}_{-2.6}$, indicating an underfluctuation of the data with respect to the background-only hypothesis. The 68% and 95% confidence level simultaneous limits on the Higgs boson coupling to bottom quarks (κ_b) and to top quarks (κ_t) were derived, showing the potential of the analysis to constrain κ_b.

Our measurement of the bbH process is mainly constrained by large theoretical uncertainties and statistical uncertainties. If the theoretical uncertainty is assumed to become negligible, an estimate can be made based on the statistical uncertainty on the additional data required for an observation. The uncertainties for the expected signal strength are: $\mu_{exp} = 1.0^{+1.2}_{-1.2}\,(\text{stat})^{+2.5}_{-1.9}\,(\text{syst})$. In order to observe an excess with three standard deviation significance, defined as μ/σ_μ, the statistical uncertainty needs to decrease to $\sigma_\mu = 0.33$ and for an observation with five standard deviations to $\sigma_\mu = 0.2$, assuming the theoretical and systematic uncertainties become negligible. The statistical uncertainty scales with the square-root of the integrated luminosity, therefore the additional amount of data needed can be derived by $\sigma^2_{\text{stat}}/\sigma^2$. To obtain evidence of the process, 13 times the data set is required, corresponding to an integrated luminosity of 1800 fb^{-1} and 35 times the data or 4800 fb^{-1} for an observation. Further optimizations of the analysis, which would be possible with a larger dataset, can reduce this required amount significantly. The analysis showed that most variables have very similar shapes for the bottom quark fusion and the gluon fusion process with gluon splitting to bottom quarks, which makes it very challenging to differentiate between them. Novel techniques will be required in order to observe the bottom quark fusion Higgs boson production mode directly and to constrain the κ_b without any dependence on the κ_t.

The bbH analysis results were released in a Physics Analysis Summary [108] and submitted in a paper to the Physics Letters B journal [83].

6 Search for vector-like leptons

6.1 Motivation

This chapter describes the search for vector-like leptons (VLL), predicted by the 4321 model [26], that is proposing new couplings between quarks and leptons, with the strongest coupling for third generation fermions and weaker for second and first generation. The new couplings are introduced in the 4321 model by massive leptoquarks [27]. The 4321 model extends the SM gauge group to $SU(4)\times SU(3)'\times SU(2)\times U(1)'$ and is UV-complete, which means that the theory is predictive also at high energy scales. The 4321 model respects measurements that found good agreement of the lepton flavor universality with the SM expected value [21][22][23][24], while also allowing for the possibility of lepton flavor violation in the third generation of leptons. Mixing of fermion generations can be accomplished by the presence of vector-like fermions, which have a hierarchical mixing with the SM fermions. One particular interesting type for searches at the LHC are the vector-like leptons (VLL), that also contribute to $\mathrm{B}_s-\bar{\mathrm{B}}_s$ mixing, regulated by the VLL mass. The measurements of $\mathrm{B}_s-\bar{\mathrm{B}}_s$ mixing constrain the possible VLL mass range to less than 2 TeV [28], making it possible to search for them at the LHC.
In proton-proton collisions the VLL can be pair produced through a s-channel or a t-channel production, as shown in Fig. 6.1. *s-channel* refers to the fusion of two particles into an intermediate particle, that decays to a particle pair. In a *t-channel* production a particle mediates the intermediate particle, which is getting absorbed by another particle. The t-channel production of VLL pair is suppressed by the leptoquark couplings. The VLL pair production can be mediated by the SM electroweak bosons or by a Z', a new gauge boson introduced by the 4321 model. The production through the Z' is not considered in this search because of the unknown coupling strength and mass of the Z'. The kinematics in the Z' production were found to be the same as for the electroweak production, within the sensitivity of this search, therefore the results are expected to hold also if the Z' production is included.

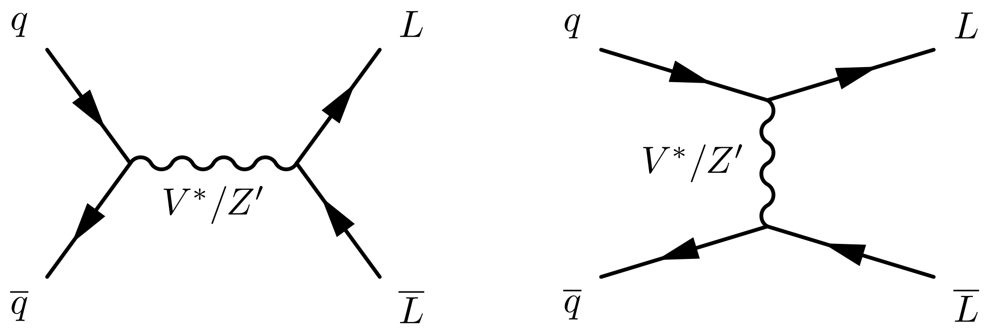


Figure 6.1: Feynman diagrams for VLL production through a s-channel production (left) and a t-channel production (right).

There are charged and neutral VLL, denoted with E and N, respectively. Like the SM leptons they form a $SU(2)_L$ doublet and can be produced as a pair of charged or neutral VLL through a Z boson, $Z \to E^+E^-$, $Z \to N\overline{N}$ or through a W boson as pairs of one charged and one neutral VLL, $W^+ \to E^+N$, $W^- \to E^-\overline{N}$. The Feynman diagrams for these production modes are shown in Fig. 6.2.

The VLL decays through the leptoquark (U) into SM fermions. Because the leptoquarks couple most strongly to the third generation fermions, the VLL decays almost entirely to third generation SM fermions. The VLL form a $SU(2)_L$ doublet and therefore the neutral VLL N decays through a leptoquark to a top quark and the charged VLL E through a leptoquark to a bottom quark. The leptoquarks decay to a third generation quark (top or bottom quark) and a third generation lepton (tau lepton or tau lepton neutrino). The decays of the charged and the neutral VLL are shown in Fig. 6.3.

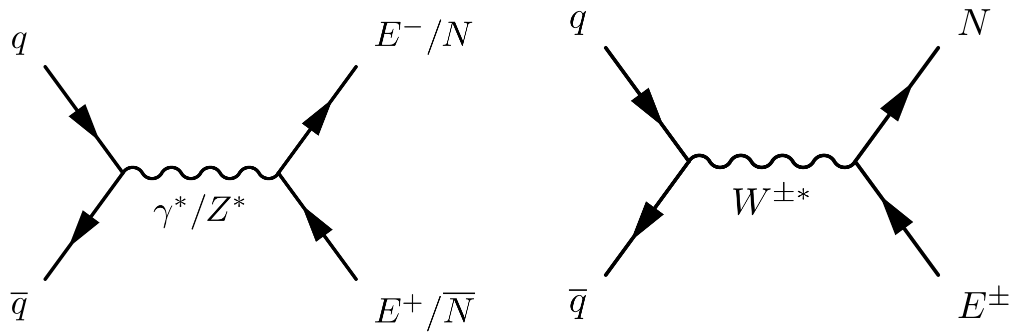


Figure 6.2: Feynman diagrams for the production of a charged-charged (EE) or neutral-neutral (NN) VLL pair through a off-shell gamma or Z boson (left) and the production of the charged-neutral (EN) VLL pair through the W boson (right).

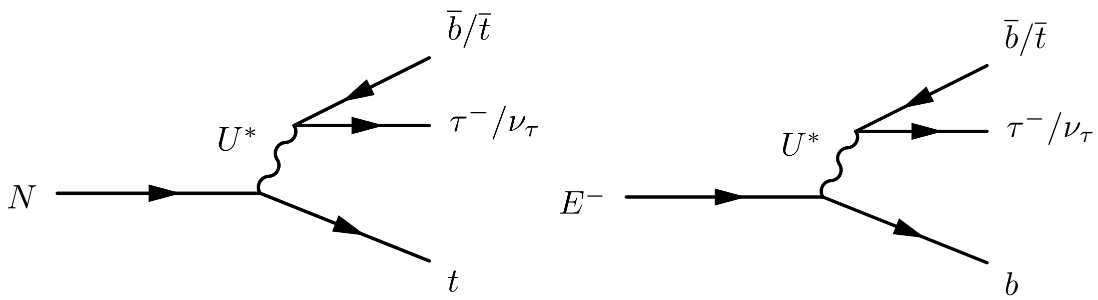


Figure 6.3: Decay of a neutral VLL (N) through a leptoquark into a top quark with the leptoquark decaying to third generation fermions (left) and the decay of a charged VLL (E) through the leptoquark into a bottom quark (right).

As explained already in section 2.3, a search by the CMS Collaboration for VLL in an all hadronic final state found an excess for mass points ranging from 500 to 1000 GeV, with a significance of 2.8 standard deviations at a representative mass point of 600 GeV [29]. The search presented in this chapter includes electrons and muons in the final state and acts as an independent test to check the excess.

6.2 Overview

The analysis searches for VLL in the final state with one electron or muon, one hadronic tau lepton and at least three bottom quarks. The signal includes the VLL as charged-charged (EE), charged-neutral (EN) and neutral-neutral (NN) production modes for a mass range from 500 to 1000 GeV in steps of 100 GeV. The final states are shown in Table 6.1, with the light lepton (electron or muon) originating from a top quark decay or from a tau lepton decay. An example of a charged-neutral VLL pair decaying to $EN \to b(t\nu_\tau)t(b\tau)$ with the light lepton coming from a top quark decay is shown in Fig. 6.4.

production + decay mode	origin of τ	origin of ℓ	final state
$EE \to b(t\nu_\tau)b(t\nu_\tau)$			$4b + \ell + \tau_h + 4\nu$
$EN \to b(t\nu_\tau)t(t\nu_\tau)$	$t \to \tau\nu_\tau b$	$t \to \ell\nu_\ell b$	$4b + 2j + \ell + \tau_h + 4\nu$
$NN \to t(t\nu_\tau)t(t\nu_\tau)$			$4b + 4j + \ell + \tau_h + 4\nu$
$EE \to b(b\tau)b(t\nu_\tau)$			$4b + \ell + \tau_h + 2\nu$
$EN \to b(t\nu_\tau)t(b\tau)$	$U \to b\tau$	$t \to \ell\nu_\ell b$	$4b + 2j + \ell + \tau_h + 2\nu$
$EN \to b(b\tau)t(t\nu_\tau)$			$4b + 2j + \ell + \tau_h + 2\nu$
$NN \to t(b\tau)t(t\nu_\tau)$			$4b + 4j + \ell + \tau_h + 2\nu$
$EE \to b(b\tau)b(b\tau)$			$4b + \ell + \tau_h + 2\nu$
$EN \to b(b\tau)t(b\tau)$	$U \to b\tau$	$\tau \to \ell\nu_\ell\nu_\tau$	$4b + 2j + \ell + \tau_h + 2\nu$
$NN \to t(b\tau)t(b\tau)$			$4b + 4j + \ell + \tau_h + 2\nu$

Table 6.1: VLL production and decay modes used in the search. Depending on the VLL charge, the pairs can be charged-charged (EE), charged-neutral (EN) or neutral-neutral (NN). The light leptons, referring to the electrons and muons, are labeled as ℓ, the hadronic decaying tau leptons as τ_h and the light jets excluding top and bottom quarks as j. The decay products in parentheses represent the objects coming from the intermediate Leptoquark, U. In the upper final states the light lepton and hadronic tau originate from the top quark decay, in the final states in the center the light lepton originates from the top quark decay and the hadronic tau lepton from a leptoquark decay. In the lower final states two tau leptons are present, with one decaying leptonically to a light lepton and the other decaying to hadrons.

Events are selected in the electron channel by a single electron trigger and in the muon channel by single muon trigger. The basic selection requirements include at least four jets and H_T of at least 400 GeV, where H_T is referring to the transverse momentum scalar sum of the four leading jets. In the signal region one electron or muon, one hadronic decaying tau lepton and at least three bottom quark jets are required. The signal region refers to the region where we expect to observe the signal. At the time of writing, the analysis is blinded, which means that no data is shown in the signal region, and the limits are derived only based on simulation, called *expected limits.* The blinding procedure helps avoid bias towards a desired result until the simulation and the analysis methods are verified. The verification is conducted in validation regions, which are exclusive regions with different selection criteria than the signal region. The selection requirements for the signal and the validation regions are shown in Table 6.2.

Besides the signal, other processes with the same final state also contribute to the signal region, referred to as *background.*

The dominant backgrounds are from $\mathrm{t\bar{t}}$ production, because the top quark decays predominantly to a bottom quark and a W boson, which can then decay to leptons, giving the same signature we search for. The $\mathrm{t\bar{t}}$ backgrounds are split into $\mathrm{t\bar{t}jj}$, $\mathrm{t\bar{t}b\bar{b}}$ and fake τ_h background, which is the contribution from jets initiated by quark or gluons and misidentified as hadronic tau lepton decay. Minor backgrounds are $\mathrm{t\bar{t}X}$, V+jets, diboson and $\mathrm{t\bar{t}t\bar{t}}$. The background processes are

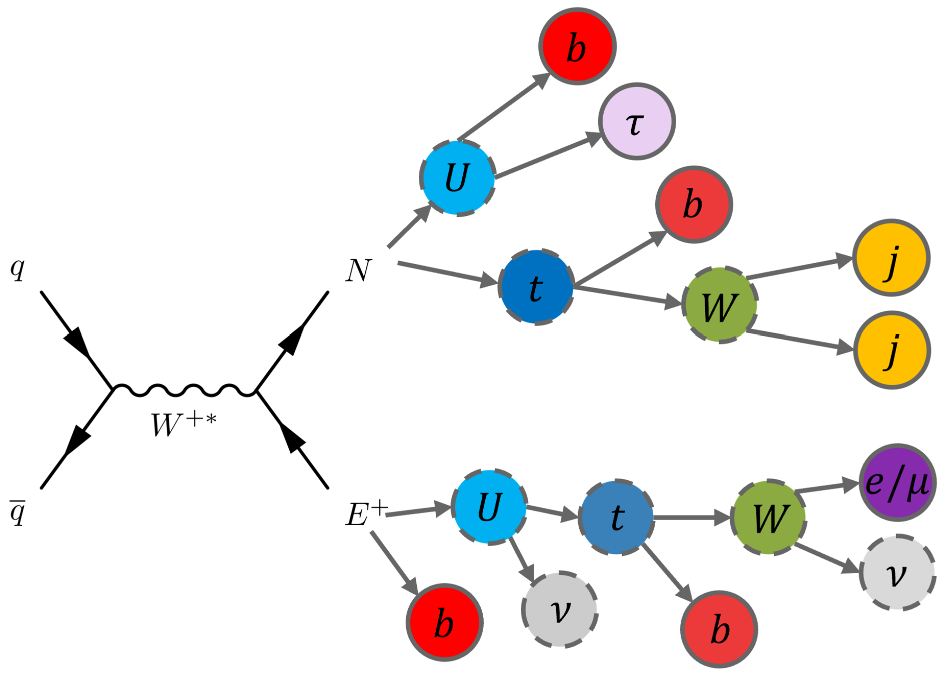


Figure 6.4: Decay product of neutral-charged VLL pairs with the decay $EN \to b(t\nu_\tau)t(b\tau)$ and the final state with four bottom quark jets, two light jets, one light lepton (electron or muon), one hadronic decaying tau lepton and two neutrinos. Dashed circles represent the intermediate particles plus neutrinos, which are not detected.

Basic Selection Requirements							
Njets ≥ 4 $H_T \geq 400\,\text{GeV}$							
Signal Region Selections		Validation Region Selections					
$1\tau_\text{h}$ 3 b-jets 1e/μ	$1\tau_\text{h}$ 4 b-jets 1e/μ	$0\tau_\text{h}$ 3 b-jets 1e/μ	$0\tau_\text{h}$ 4 b-jets 1e/μ	$1\tau_\text{h}$ 2 b-jets 1e/μ	$1\tau_\text{h}{}^\text{loose}$ 2 b-jets 1e/μ	$1\tau_\text{h}{}^\text{loose}$ 3 b-jets 1e/μ	$1\tau_\text{h}{}^\text{loose}$ 4 b-jets 1e/μ

Table 6.2: Selection requirements for the signal region and validation regions.

simulated with Monte Carlo samples, and the modeling is verified in the validation regions. Because one isolated light lepton is required in the final state no QCD multijet background is expected in the signal region. The event simulation is described in section 4.9.

After the selection, a Deep Neutral Network (DNN) is trained on a subset of the $t\bar{t}$ and signal events to separate the signal from the $t\bar{t}$ background process. The signal is expected in the highest bins of the DNN output and the $t\bar{t}$ background in the lower bins.

The DNN outputs are fit simultaneously for the muon and the electron channel for each signal mass point with the VLL production cross-section as parameter of interest. Expected limits on the VLL production cross-section are derived as a function of the VLL mass.

My contribution was to update the framework and deep neural network, previously used in the fully hadronic analysis, to include one electron or muon in the final state, perform validations of the simulation, improve the uncertainty estimates and derive expected upper limits on the signal cross section.

6.3 Data and simulation

This search uses data collected by the CMS experiment in the years from 2016 to 2018. The data is compared with Monte Carlo simulation, described in section 4.9.

6.3.1 Data

The data used in the analysis was collected by the CMS experiment from 2016 to 2018 with proton-proton collisions at a center of mass energy of 13 TeV. The integrated luminosity for 2016 amounts to 35.9 fb^{-1}, 41.5 fb^{-1} for 2017 and 59.97 fb^{-1} for 2018, which gives a combined integrated luminosity of 138 fb^{-1}. 2016 is split in two different eras, pre and postVFP (Vertex Fitting Procedure) because of a substantial change of the conditions in the strip tracker detector during the 2016 data taking. Data events are selected by a set of single muon and single electron triggers shown in Tab. 6.3. Multiple triggers are used in order to achieve optimal efficiency across the p_{T} and pseudorapidity range, and to ensure redundancy.

channel	Trigger type	p_{T} thresholds [GeV]
muon	Single isolated muon	24 in 2016 & 2018 27 in 2017
electron	Single electron	27 in 2016 28 or 32 in 2017 & 2018

Table 6.3: Trigger requirements in the electron and muon channel.

6.3.2 Signal simulated samples

As introduced in section 6.1 the considered signal consists of VLL pairs produced by intermediate W or Z bosons. The VLL pairs can be either charged-charged (EE), charged-neutral (EN) or neutral-neutral (NN). The signal samples are simulated at leading order with the MadGraph5_aMC@NLO generator. The charged VLL can decay to $E \to b(b\tau)$ or $E \to b(t\nu)$ and the neutral VLL to $N \to t(t\nu)$ or $N \to t(b\tau)$ with the decay particles in the parenthesis coming from an intermediate leptoquark. The decay of the VLL is simulated with MadGraph, for the top quarks pythia is used. The samples are produced for a VLL mass range of 500 to 1000 GeV with 50 GeV steps. The masses of the charged and neutral VLL are constraint to be equal as they are expected to be almost the same in the model. The leptoquark mass is larger than the VLL mass, causing the leptoquark to be off-shell. As a result, the leptoquark mass does not affect the kinematics of the VLL [31, 109].
The simulated signal samples and their cross sections are shown in Tab. 6.4.

6.3.3 Background simulated samples

Several processes have the same final state as our signal with one light lepton, one hadronic tau lepton and several bottom quark jets. The dominant backgrounds are $\mathrm{t\bar{t}}$ and jets misidentified as hadronic tau leptons.
$\mathrm{t\bar{t}}$: This process decays predominantly to $\mathrm{t\bar{t}} \to b\bar{b}W^{+}W^{-}$. The final state can contain genuine tau leptons from the leptonic decay of the W boson or jets misidentified as τ_{h}. The $\mathrm{t\bar{t}}$ background process is generated with the powheg generator at NLO. The $\mathrm{t\bar{t}}$ background process is split into the $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}jj}$ processes. If there is at least one additional bottom quark jet, not coming

Mass [GeV]	EN σ [pb]	EE σ [pb]	NN σ [pb]
500	0.01745	0.00493	0.00471
550	0.01136	0.00315	0.00301
600	0.00741	0.00208	0.00196
650	0.00505	0.00140	0.00133
700	0.00346	0.00095	0.00090
750	0.00241	0.00066	0.00062
800	0.00170	0.00046	0.00043
850	0.00120	0.00033	0.00031
900	0.00088	0.00024	0.00022
950	0.00063	0.00017	0.00016
1000	0.00046	0.00012	0.00011
1050	0.00034	8.995e-05	8.490e-05

Table 6.4: VLL signal samples including the charged-neutral VLL pair (EN), the charged-charged VLL pair (EE) and the neutral-neutral VLL pair (NN) for the considered VLL mass points and the corresponding production cross sections. Values are calculated based on the Universal FeynRules Output (UFO) model, described in Ref. [31].

from the top quark decay, the process is considered $\mathrm{t\bar{t}b\bar{b}}$, otherwise $\mathrm{t\bar{t}jj}$. Measurements of the CMS and ATLAS Collaborations [110–112] show that the predicted $\mathrm{t\bar{t}b\bar{b}}$ cross section is below the measured one. To account for this difference a scale factor of 1.2 is applied to the $\mathrm{t\bar{t}b\bar{b}}$ process.

Fake τ_h: The background from jets misidentified as hadronic tau lepton decays arises from the $\mathrm{t\bar{t}}$ background process. The signal final states are produced by the leptonic decay of the W boson and one jet misidentified as hadronic tau. The fake τ_h background is estimated from the simulated $\mathrm{t\bar{t}}$ samples by requiring at least one tau lepton not matched to a generator level hadronic tau decay.

Minor background arise from $\mathrm{t\bar{t}X}$, Drell-Yan+jets, W+jets as well as double and triple-bosons.

$\mathrm{t\bar{t}X}$: A $\mathrm{t\bar{t}}$ pair plus a Higgs boson or vector boson, or $\mathrm{t\bar{t}t\bar{t}}$ have minor contributions. For the $\mathrm{t\bar{t}Z}$ the $Z \rightarrow q\bar{q}$ and the $Z \rightarrow \ell^-\ell^+$ decay, for the $\mathrm{t\bar{t}W}$ production the $W \rightarrow \ell\nu$ and in the $\mathrm{t\bar{t}H}$ production only the $H \rightarrow b\bar{b}$ decay is considered. The samples are generated with the MadGraph5_aMC@NLO generator at NLO in QCD.

Drell-Yan+jets: The Drell-Yan process $Z/\gamma* \rightarrow \ell^+\ell^-$ can be accompanied by jets from initial or final state radiation, giving the same final state as the signal. Because at least four jets are required, this background only has a small contribution. The Drell-Yan+jets background process is simulated in the jet binned samples with 3 or 4 jets at LO with the MadGraph generator.

W+jets: The leptonic decay of the W boson accompanied with at least four jets can also mimic the signal final state. A jet binned sample with 4 jets is generated at LO with the MadGraph generator.

VV(V): The double- and triple boson background includes WW,WZ, ZZ, WWZ and ZZZ with their semi-leptonic final states. WW, WZ and ZZ are simulated using pythia, WWZ and ZZZ are generated at NLO in QCD with the MadGraph5_aMC@NLO generator with a 4-flavor scheme for WWZ and a 5-flavor scheme for ZZZ.

The simulated background samples with the generators used and their cross sections are shown in Tab.6.5.

Process	Generators	Cross section σ [pb]
$\mathbf{t\bar{t}+jets}$		
Fully leptonic	POWHEG, PYTHIA	89.048 (NNLO)
Semi-leptonic	POWHEG, PYTHIA	366.914 (NNLO)
Fully Hadronic	POWHEG, PYTHIA	377.960 (NNLO)
DY+jets, $\mathbf{Z/\gamma^* \to \ell^+\ell^-}$		
$+$3jets, $m_{\ell\ell} = 50\,\text{GeV}$	MadGraph, PYTHIA	111.5 (LO)
$+$4 jets, $m_{\ell\ell} = 50\,\text{GeV}$	MadGraph, PYTHIA	44.05 (LO)
$\mathbf{W+jets,\ W \to \ell\nu}$		
$+$4 jets	MadGraph, PYTHIA	544.3 (LO)
$\mathbf{t\bar{t}X}$		
ttH	aMC@NLO, PYTHIA	0.5086 (NLO)
$\mathrm{t\bar{t}Z}$	aMC@NLO, PYTHIA	0.5297 (NLO)
$\mathrm{t\bar{t}W}$	aMC@NLO, PYTHIA	0.2043 (NLO)
$\mathrm{t\bar{t}t\bar{t}}$	aMC@NLO, PYTHIA	0.00821 (NLO)
$\boldsymbol{VV(V)}$		
WW	PYTHIA	118.7 (NLO)
WWZ	aMC@NLO, PYTHIA	0.1651 (NLO)
WZ	PYTHIA	47.13 (NLO)
ZZ	PYTHIA	16.523 (NLO)
ZZZ	aMC@NLO, PYTHIA	0.01398 (NLO)

Table 6.5: Summary of simulated background processes with the generators and the corresponding production cross sections.

6.4 Data and MC comparison and corrections

The simulated samples can differ from data because of mismodeling of the reconstruction and identification of particles, of the detector noise or because of mismodeled physics processes. Corrections are applied to the simulated samples to improve the agreement with the observed data. The corrections are applied either as scale factors or as energy scales. The scale factor is defined as the ratio of the efficiency in data ϵ_{Data} over the efficiency in the simulated Monte Carlo (MC) samples ϵ_{MC}: $\mathrm{SF} = \epsilon_{\mathrm{Data}}/\epsilon_{\mathrm{MC}}$. The lepton efficiency is measured using a tag-and-probe technique, which selects lepton pairs based on the tag lepton and measures the efficiency with the probe lepton. This ensures that the probe leptons are not biased by any other selection other than the one for which the efficiency is measured. The efficiency is measured by comparing the number of probe leptons that pass the identification requirements, for which the efficiency is evaluated, to the total number of probe lepton. The numbers are extracted from a fit to the reconstructed lepton pair mass around the Z boson mass peak. This ensures that the probe lepton is indeed the lepton for which we want to measure the efficiency. The energy scale corrections are applied to the p_{T} and mass of the jets, as discussed in section 4.7, or as a multiplicative factor to the p_{T} and mass of the τ_{h}. Most of the corrections used in this search are provided by the respective Physics Object Group (POG) of the CMS Collaboration.

6.4.1 Pileup Reweighting

The simulated samples are generated with a pileup distribution that approximates the condition during data taking. In order to ensure an accurate modeling of the pileup effects, it is important to apply pileup reweighting to the simulated samples. The mean number of interactions per bunch crossing is rebinned in 100 bins and reweighted in each bin to match the distribution in data. The number of interactions per bunch crossing in data is estimated as the product of the instantaneous luminosity and assuming a total inelastic cross section of $\sigma_{pp} = 69.2$ mb.

6.4.2 Trigger corrections

The scale factors for the single muon and the single electron triggers are derived by comparing the efficiency in data with the efficiency in simulated samples. For the measurement of the efficiency of the single muon trigger an inclusive Drell-Yan plus addition jets simulated sample is used. Two muons are required, which need to satisfy the same selection as in the signal region. The tag muon must be matched to the trigger while the probe muon does not have any additional selection requirements. The efficiency is measured as function of the muon p_{T}, with small bins close to the trigger threshold, and the muon η in the di-muon mass distribution in the range of the Z boson mass from 70 to 130 GeV. The measurement is conducted in a *pass* and a *fail* region. The *pass* region includes all events where the probe lepton activated the trigger, the *fail* region all the events where the probe lepton did not activate the trigger. A maximum likelihood fit is performed simultaneously on the pass and on the fail region.
The single muon trigger scale factors are provided by the muon POG of the CMS Collaboration. The measurement of the single electron trigger scale factors is performed with dileptonic and semileptonic $\mathrm{t\bar{t}}$ plus additional jets simulated samples, requiring an $\mathrm{e}\mu$ pair in the final state. Single muon triggers are used as exclusive reference triggers. The muon is used as tag, matched to a single muon trigger, and the electron as probe. The efficiency is extracted for each bin of a two dimensional distribution of the electron p_{T} and electron supercluster η. The single electron

trigger scale factors are taken from the ttbb cross section measurement analysis of the CMS Collaboration due to similar electron selection criteria compared to our analysis [113].

6.4.3 Muon and electron corrections

Lepton identification, isolation and reconstruction efficiencies are measured in data and simulation using a tag-and-probe technique and applied as scale factors for each event to correct the simulation. Efficiencies are measured in $Z \to ee$ and $Z \to \mu\mu$ events as function of the lepton p_{T} and $|\eta|$. The tag lepton must pass the kinematic selection, identification and isolation while the probe lepton must only pass the kinematic selection. The efficiency is defined as the ratio of number of probe leptons, which pass the identification and isolation, to the total number of probe leptons, and is extracted by a fit to the dilepton mass distribution around the Z boson mass [47, 55]. The lepton identification, isolation and reconstruction scale factors are provided by the respective POG of the CMS Collaboration.

6.4.4 Tau corrections

Scale factors for the DeepTau discriminator against jets are derived in the $Z \to \tau\tau \to \mu\tau_{\mathrm{h}}$ process. The muon is tagged as one of the tau leptons decaying leptonically and the probe side that reconstructs a Z boson is assumed to be the τ_{h} with only a loose requirement on the identification. The *pass* region includes all probes which pass the working point of the DeepTau identification against jets, for which the scale factors are derived. Probes that do not pass the working point are included in the *fail* region. The efficiency is extracted in bins of the τ_{h} p_{T} from a fit to the visible invariant mass of the $\mu\tau_{\mathrm{h}}$ pair.
The scale factors for the DeepTau electron and muon discriminators are derived in a $Z \to ee/Z \to \mu\mu$ process with one electron/muon misidentified as τ_{h}, considered as probe. If the probe passes the working point of the DeepTau identification against electrons or DeepTau identification against muons it is selected for the *pass* region, otherwise for the *fail* region. The efficiencies are extracted in several bins of the τ_{h} η as fit on the visible invariant mass.
The τ_{h} energy scale correction is derived with a tag-and-probe method in the same $\mu\tau_{\mathrm{h}}$ event sample as described before, but using the mass of the reconstructed hadronic tau and the visible invariant mass of the $\mu\tau_{\mathrm{h}}$ pair as distributions. A maximum likelihood fit is performed for each τ_{h} decay mode to extract the τ_{h} energy scale. The relative difference in the τ_{h} energy scale between data and simulation is applied as a multiplicative factor to the τ_{h} p_{T} and mass in simulation.
The scale factors are provided by the Tau POG of the CMS Collaboration and are described in more detail in Refs. [57, 58].

6.4.5 Bottom quark tagging corrections

The difference in the efficiency of the b-tagging in simulation and data is corrected by applying scale factors. Since the b-tagging discriminator value is used as input to the DNN, the entire discriminator distribution is corrected, not just at a specific working point as is done for other objects. The scale factors are derived with an iterative fit to the discriminator value to match the distribution in simulation with the one in data. A $\mathrm{t\bar{t}}$ sample with a dilepton final state is used as b-jet enriched region and a Drell-Yan+jets sample with a dilepton final state as light-jet enriched region, including the up, down, strange and charm quarks. The fit is performed for

every bin of the discriminator value in the b-jet enriched region, assuming the nominal light-jet distribution. Then the discriminator distribution in the light-jet enriched region is fit, taking into account the just derived b-jet scale factors. The fit is iterated several times, at each stage taking into account the scale factors derived in the previous stage [92]. The b-tagging scale factors are provided by the BTV POG of the CMS Collaboration.
The scale factors are applied to each jet in an event depending on its flavor (light-jet or b-jet), p_{T}, η and b-tagging discriminator value. The product of all the jet scale factors yields the per event scale factor applied to the simulated samples. This correction is not intended to alter the overall normalization of the simulated samples, only the discriminator distribution. Therefore the normalization of each sample is corrected as a two dimensional function of the number of jets and H_T (referring to the p_{T} of the four leading jets) after making the signal region selection, except the requirement on the number of b-jets. The normalization correction weight is extracted from the ratio of the 2D distribution without the b-tag scale factors applied over the 2D distribution with the b-tag scale factors applied.

6.4.6 L1 Prefiring

In 2016 and 2017 the gradual timing shift of the ECAL crystals was not properly propagated to the level-1 trigger primitives. This had the effect, that the trigger primitives in the forward region of $2 < |\eta| < 3$ could be incorrectly assigned to the previous bunch crossing, which can lead to an incorrect self-veto of the level-1 trigger in the correct bunch crossing [93]. In simulation this effect was not taken into account. A correction is derived with 2017 simulated samples by assigning all the jets and photons in an event a prefiring probability. The probability that no self-veto occurred in an event is derived from taking the product of all the particles not prefiring. This probability of no self-veto is applied as a weight to each event in the 2016 and 2017 simulated samples.

6.4.7 Top-quark pt correction

The p_{T} distribution of the top quarks in the $\mathrm{t\bar{t}}$ events is observed to be softer in data than in simulation in many measurements. Corrections were derived based on measurements of the $\mathrm{t\bar{t}}$ cross section [94, 95]. The difference between data and simulation was fitted and a correction weight depending on the generator level top quark p_{T} was derived. For each top quark a weight is calculated:

$$\mathrm{weight}_t = \exp(0.0615 - 0.0005 \cdot p_{\mathrm{T}}). \tag{6.1}$$

The weight is then derived as the square root of the product of the weight for the top quark and the antitop quark:

$$\mathrm{weight}_{\mathrm{t\bar{t}}} = \sqrt{w_t \cdot w_{\bar{t}}} \tag{6.2}$$

6.5 Selection of events

Events have to pass a certain number of selection criteria before being considered signal candidates, identically in both data and simulation. The selections are reported as follows for each object.

Muon and electron: The muon is required to have $p_T > 30$ GeV, $|\eta| < 2.5$ (2.4 for 2016), identified with the *tight* WP and isolated with $I^{\mu}_{rel} < 0.15$. More detail can be found in section 4.5. The electron must pass $p_T > 35$ GeV, $|\eta| < 2.5$ (2.4 for 2016), identified by the MVA based identification at the 80% efficiency WP with isolation components included. Details about the reconstruction are described in section 4.4.

Hadronic tau lepton: The τ_h is required to have $p_T > 20$ GeV, $|\eta| < 2.1$ and has to be identified with the *tight* WP of the DeepTau discriminator against jets, and *loose* WP of the DeepTau discriminator against electrons and against muons.

Jets: Jets are included in the events if they have p_T >40 GeV, $|\eta| < 2.4$ and the *tight* WP jet identification. Jets must be at least $\Delta R > 0.4$ away from any τ_h with *very loose* WP for the DeepTau discriminator against jets. An event is selected if it has at least four jets and in the signal region three or four jets must be identified as bottom quark jets by the *medium* WP of the DeepJet b-tagging algorithm.

Veto on additional leptons: In the muon channel only one muon is allowed and no electrons with $p_T > 15$ GeV, $I^{e}_{rel} < 0.15$ and *loose* WP. In the electron channel only one electron is allowed and no muons with $p_T > 13$ GeV, $I^{e}_{rel} < 0.25$ and identified with the *loose* WP.

HT: HT, the p_T sum of the four leading jets, is required to be larger than 400 GeV.

Validation regions: Six validation regions are defined, which are exclusive to the signal region. The τ_h have either the same discrimination against jets as in the signal region, which is the *tight* WP, or a looser requirement that is exclusive by vetoing the *tight* WP of the DeepTau discriminator against jets.

Common selection: one electron or muon, at least four jets and $HT > 400$ GeV

- $0\tau_h$3b: Zero τ_h with the *tight* WP of the DeepTauVSjet and three b-jets
- $0\tau_h$4b: Zero τ_h with the *tight* WP of the DeepTauVSjet and four b-jets
- ${1\tau_h}^{loose}$2b: One τ_h with the *very loose* or *loose* WP of the DeepTauVSjet and two b-jets
- $1\tau_h$2b: One τ_h with the *tight* WP of the DeepTauVSjet, two b-jets and DNN output (described in section 6.6) lower than 0.95
- ${1\tau_h}^{loose}$3b: One τ_h with the *very loose* or *loose* WP of the DeepTauVSjet, three b-jets and DNN output lower than 0.95
- ${1\tau_h}^{loose}$4b: One τ_h with the *very loose* or *loose* WP of the DeepTauVSjet, four b-jets and DNN output lower than 0.95

The selections are summarized in Tab. 6.6.

	muon channel	electron channel
$\mu/e\ p_{\mathrm{T}}$	$\mu\ p_{\mathrm{T}} > 30$ GeV	$e\ p_{\mathrm{T}} > 35$ GeV
$\mu/e\ \eta$	$\mu\ \lvert\eta\rvert < 2.5$ (2.4 for 2016)	$e\ \lvert\eta\rvert < 2.5$ (2.4 for 2016)
μ/e Id	μ *tight* WP	e 80% WP (iso)
μ/e Iso	$I^{\mu}_{\mathrm{rel}} < 0.15$	-
$\tau_{\mathrm{h}}\ p_{\mathrm{T}}$	$\tau_h\ p_{\mathrm{T}} > 20\,\mathrm{GeV}$	
$\tau_{\mathrm{h}}\ \eta$	$\tau_h\ \lvert\eta\rvert < 2.1$	
τ_{h} Id	τ_h *tight* WP DeepTauVSjet *vvloose* DeepTauVSe *vloose* DeepTauVSmu	
Jet p_{T}	Jet $p_{\mathrm{T}} > 40$ GeV	
Jet η	Jet $\lvert\eta\rvert < 2.1$	
Jet Id	*tight* WP jet Id	
Jet separation	$\Delta R(j, \tau_{\mathrm{h,loose}}) > 0.4$	
Jet b-tagging	*medium* WP DeepJet algorithm	
Veto on μ/e	no electrons	no muons
H_T	$H_T > 400$ GeV	

Table 6.6: Summary of the selection requirements for the muon and the electron channel.

6.6 Discrimination of signal and ttbar background

The signal discrimination to the main background $t\bar{t}$ is enhanced by training a deep neutral network (DNN). The final state with many different objects requires the use of a graph neural network with each final state particle represented by a node in a graph. The relationships between nodes are called edges. The graph neural network consists of several layers. In each layer the node features are aggregated from their neighbors, updated and passed to the next layer. The relationships between particles are incorporated with an attention mechanism. Each node gets an attention coefficient, that defines its importance in the graph structure. The DNN model used in this search is a attention-based cloud network called ABCNet [114].

6.6.1 ABCNet

ABCNet uses graph attention methods together with graph pooling layers (GAPLayers). The GAPLayers are used to reduce the size of the graph by applying attention coefficients to a set of nodes and edges in the graph, only considering the k-nearest neighbors. The edge features are constructed as $y_{ij} = (x_i - x_{ij})$ with the point x_i and its k-nearest neighbors x_{ij}. The GAPLayers are defined by encoding nodes and edges as:

$$x'_i = h(x_i, \theta_i, F) \quad y'_{ij} = h(y_{ij}, \theta_{ij}, F) \tag{6.3}$$

with h being a single layer neutral network with the rectified linear unit (ReLu) activation function, θ a set of learnable parameters and a output node of size F. The ReLu activation function is defined as:

$$\text{ReLu}(x) = \max(0, x) \tag{6.4}$$

The encoded nodes and edges are combined in the coefficients c_{ij} as:

$$c_{ij} = \text{ReLu}\left(h(x'_i, \theta'_i, 1) + h(y'_{ij}, \theta'_{ij}, 1)\right), \tag{6.5}$$

with the ReLu activation function applied to a neutral network with output node size of one for the encoded nodes and edges. The attention coefficients are aligned between different points by normalizing them with the softmax function:

$$\alpha_{ij} = \frac{e^{c_{ij}}}{\sum_k e^{c_{ik}}} \tag{6.6}$$

The output of the GAPLayer includes updated node values x_i and graph features y_i defined as:

$$x_i = \text{ReLu}\left(\sum_j \alpha_{ij} y'_{ij}\right) \quad y'_i = \frac{1}{K}\sum_k y_{ij} \tag{6.7}$$

The updated node values are used in the subsequent GAPLayers instead of the input variables. The numbers of neighbors is chosen as 8. The 8-nearest neighbors are considered by deriving the distance between particles as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

6.6.2 Input variables

The DNN is trained with events that include one electron or muon, one hadronic tau lepton and at least four jets, with at least two of those jets being b-tagged. The kinematics from the leptons and up to eight jets are used per event. If an event contains more than eight jets, the jets with lowest momentum are dropped. The input variables common to all objects are the logarithm of the p_T and the mass, the pseudorapidity η and the azimuthal angle ϕ. For the leptons the charge and for the jets the b-tagging discriminator value are used as input. The charge entry is set to zero for jets and the discriminator entry is set to zero for leptons, enabling the DNN to identify leptons and jets during the training. In Fig. 6.5 several input variables are displayed, including the p_T of the leading jet and subleading jet, τ_h p_T, muon and electron p_T and the b-tagging discriminator value of the leading jet. The input variables show good agreement between simulation and data and a small signal sensitivity, as shown by the signal at a VLL mass of 600 GeV.

6.6.3 Training and output

The DNN is trained and tested with only 10% of the available signal and $t\bar{t}$ samples, which is not used further in the analysis to avoid any bias in the distributions. For the signal samples the mass range from 500 to 1000 GeV in 50 GeV steps is used. During the training the performance of the model is tested with a separate test sample. The training is performed over many epochs, with each epoch being a complete pass through the entire training dataset and updating the model parameters. If the value of the loss function does not decrease for 10 consecutive epochs, the training is stopped and the model with the parameters found 10 epochs earlier is used. The output of the DNN is one score with a range between zero and one. A score of zero identifies events as $t\bar{t}$ background, a score of one as signal. The DNN output scores are fitted using a maximum likelihood fit to derive the limits, with a binning of the scores that maximizes the signal sensitivity. In Fig. 6.6 the expected DNN output scores are shown for the electron and muon channel in the three b-jet and four b-jet region. The signal is shown for the VLL mass of 600 GeV, 800 GeV and 1000 GeV.

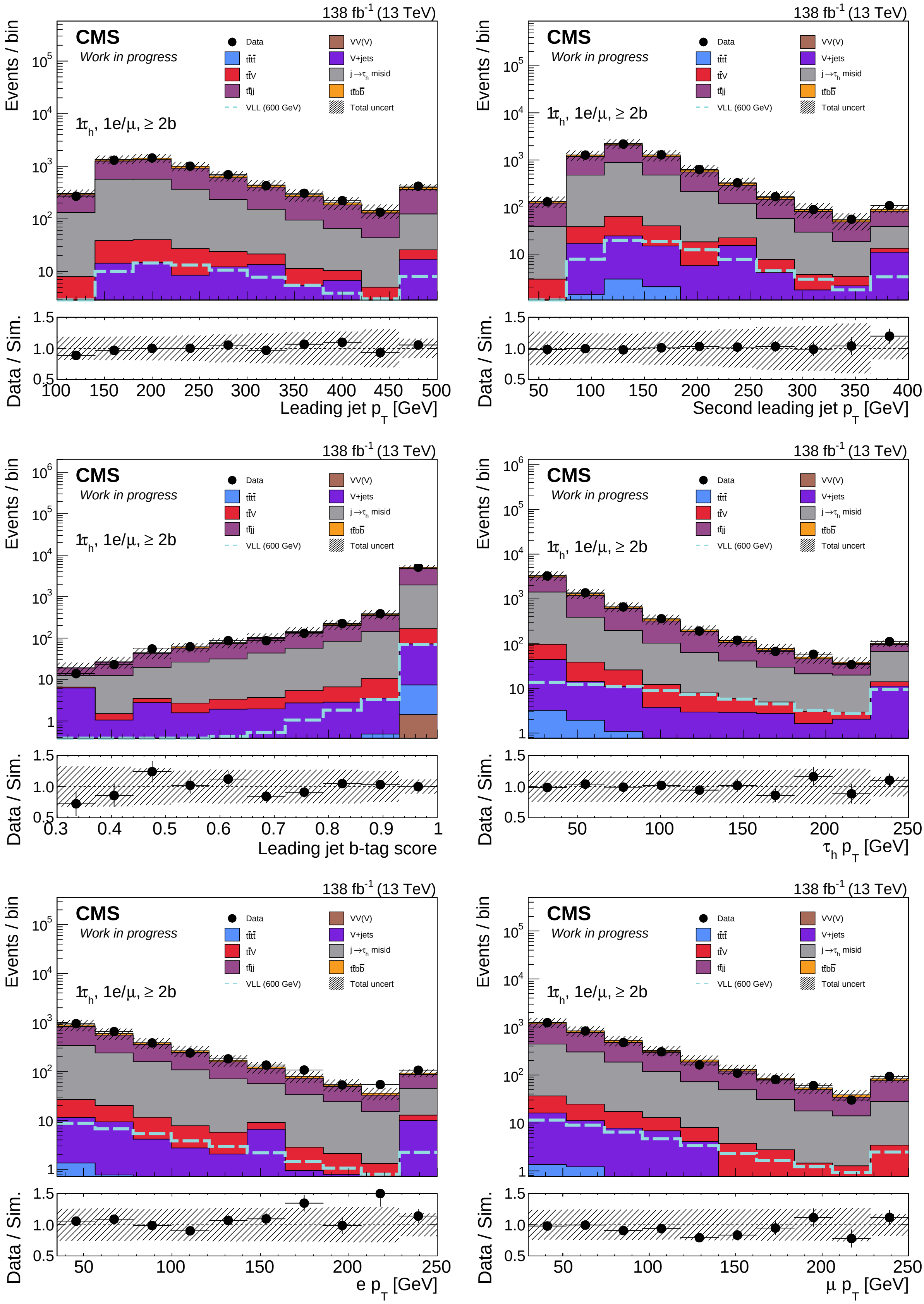


Figure 6.5: DNN input variables, including the leading jet p_T (top left), subleading jet p_T (top right), leading jet b-tagging discriminator value (center left), τ_h p_T (center right), electron p_T (bottom left) and muon p_T (bottom right).

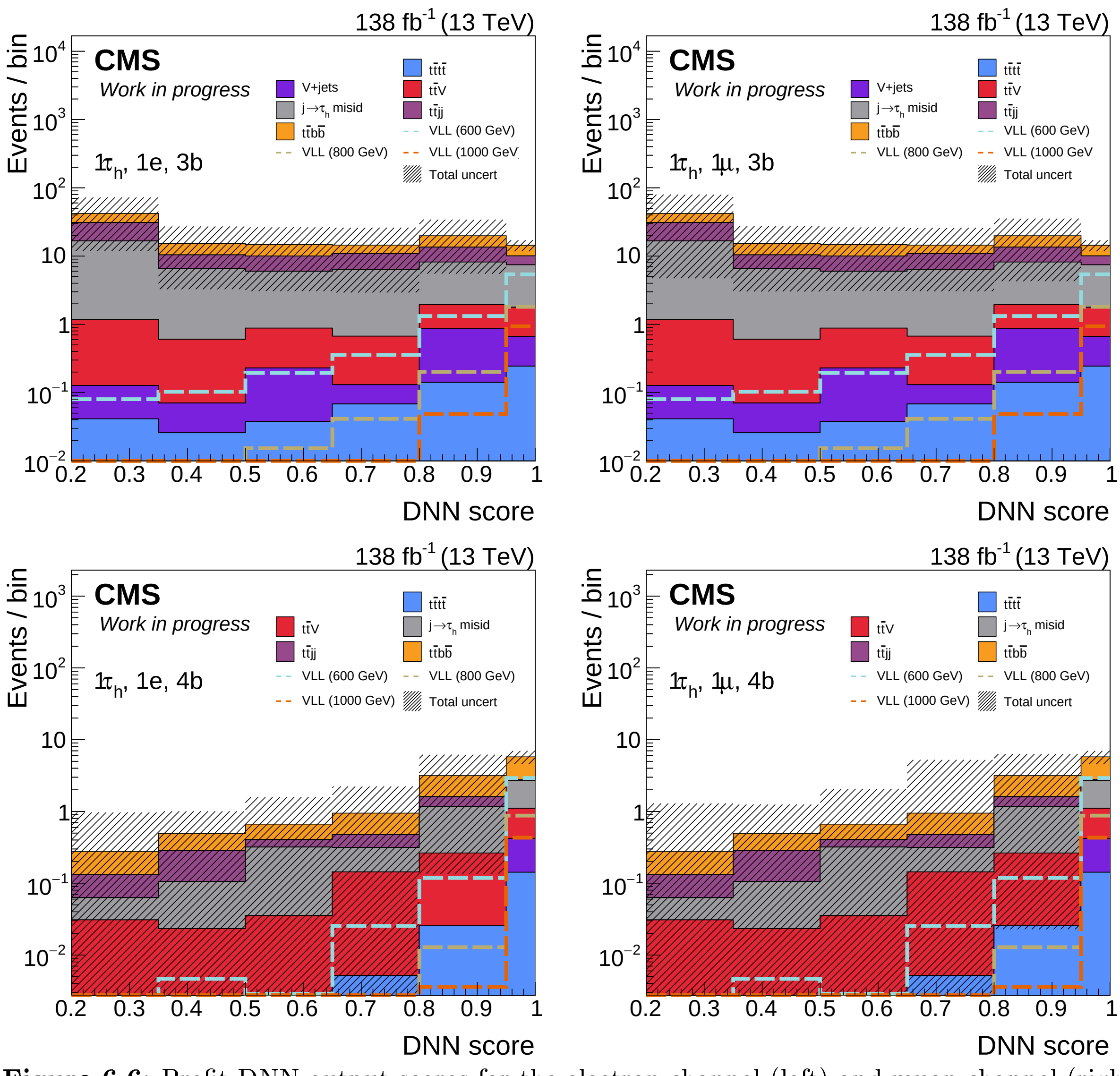


Figure 6.6: Prefit DNN output scores for the electron channel (left) and muon channel (right) for three b-jets (top) and four b-jets (bottom) with all years combined, but treated separately in the fit. The analysis is still blinded, therefore only the expected scores are shown. The signal is shown at the VLL mass of 600 GeV, 800 GeV and 1000 GeV.

6.7 Systematic Uncertainties

The result of the search depends on various sources of experimental and theoretical systematic uncertainties, which have to be estimated. The uncertainties are included as nuisance parameters in the fit, described in more detail in section 5.8. Some uncertainties only affect the normalization of the DNN output score, that is fit to derive the limits. A log-normal probability distribution is assigned to the normalization uncertainties. Other uncertainties affect the normalization and shape of the DNN output score. The shape uncertainties are estimated by constructing histograms with the uncertainty varied by one standard deviation from its nominal value. A function interpolates from the nominal and one standard deviation templates to all values of the systematic uncertainties. If the source of an uncertainty impacts multiple data-taking years, processes or light lepton types in a consistent way, the uncertainty is correlated across several years, processes or light lepton types, otherwise uncorrelated.
In table 6.7 all the uncertainties are summarized.

6.7.1 Experimental uncertainties

The systematic uncertainties from experimental sources, such as the identification of the objects and the determination of the luminosity, are described in the following. If not mentioned otherwise, the uncertainties are treated as correlated between light lepton types, processes and years.

Luminosity: The integrated luminosity, used to normalize the simulated signal and background samples, introduces a systematic uncertainty, that is estimated as a normalization uncertainty of 1.2% for 2016, 2.3% for 2017 and 2.5% for 2018 [103–105].

Lepton identification: The muon identification and isolation efficiency scale factor uncertainties are applied as function of the muon p_{T} and η. For the electron, the identification and reconstruction efficiency scale factor uncertainties are applied as function of the electron p_{T} and supercluster η.

τ_{h} identification against jets and energy scale: The uncertainty on the τ_{h} discriminator against jets is applied as function of the τ_{h} p_{T} and decay mode. The uncertainties on the τ_{h} discriminator against muons and electrons are expected to be negligible.
The uncertainty on the tau energy scale depends on the τ_{h} decay mode and alters the correction to the τ_{h} p_{T} and mass. The uncertainty is uncorrelated between years.

Trigger efficiency: The uncertainties on the trigger efficiency scale factors depend on the lepton p_{T} and η. The uncertainty is uncorrelated between light lepton types and years.

Jet energy scale and resolution: The jet energy scale and resolution variations include a set of uncertainties based on sources such as the detector efficiency during data taking, the pileup interactions and the clustering efficiency. Some of the uncertainties are specific for each data taking year, others are correlated across years. The uncertainties are applied to all jets and alter the jet p_{T} and mass.

B-tagging efficiency: The uncertainty on the b-tagging efficiency is estimated from a set of b-tagging scale factor uncertainties. Uncertainties are derived for the limited statistics in the regions where the scale factors are derived, and contamination of light jets in the b-enriched region as well as contamination of b-jets in the light jet enriched region. Uncertainties from the jet energy scale are propagated to the b-tagging scale factors. Some of the uncertainties are specific for a data taking year, others are correlated.

Prefiring: Uncertainties related to the effect of prefiring are applied to the 2016 and 2017 data taking periods.

Pile up: The uncertainty introduced by the simulation of pileup is estimated by varying the inelastic parton-parton cross-section up an down by 4.6%, which corresponds to the uncertainty in the inelastic cross-section measurement.

6.7.2 Theory uncertainties

The theoretical systematic uncertainties account for any uncertainties in the simulation of the signal and background processes, including the proton composition and approximations used in the matrix element and parton shower calculations. If not mentioned otherwise, the uncertainties are correlated between $\mathrm{t\bar{t}b\bar{b}}$, $\mathrm{t\bar{t}jj}$ and fake τ_h and uncorrelated for signal and $\mathrm{t\bar{t}V}$. The uncertainties are not applied for the minor backgrounds.

PDF+α_s: Uncertainties arising from the parton distribution function (PDF) and from the value of the strong coupling constant α_s, used in the evaluation of the PDF, are applied. The difference between the nominal weight and 102 alternative weights is added in quadrature to define the uncertainty.

Renormalization and factorization scale: The choice of the renormalization (μ_R) and factorization scale (μ_F) in the generation of simulated samples introduces an uncertainty, that is estimated by constructing an envelope of scale variations. The μ_F and μ_R are varied by 0.5 and 2, excluding the combination where one scale is half and the other is double its nominal value.

Parton shower: Theory uncertainties in the parton shower, used in the generation of the simulated samples, are estimated by varying the scale at which the strong coupling constant is evaluated in the shower. Independent and uncorrelated uncertainties are derived for the initial state radiation and the final state radiation.

Top p_T reweighting: In order to account for any mismodeling in the $\mathrm{t\bar{t}}$ simulated samples an uncertainty based on the top p_T reweighting is constructed, with the reweighting as the upward variation and the nominal template as downward variation. The uncertainty is applied to fake τ_h, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}jj}$, treating the uncertainty as correlated.

ttbb cross section: The $\mathrm{t\bar{t}b\bar{b}}$ cross section is scaled by 1.2 to account for a disagreement between simulation and data found in measurement, as discussed in section 6.3.3. A 20% normalization uncertainty is applied to the $\mathrm{t\bar{t}b\bar{b}}$ process to account for any mismodeling.

Fake rate: The fake contribution is estimated from simulation. In order to cover any mismodeling a conservative normalization uncertainty of 50% is applied to the fake τ_h process.

6.7.3 Impacts

The definitions of the impacts and of the pulls have been described in section 5.9.3. In figure 6.7 the uncertainties are listed according to their impact on the measurement. The uncertainties with the largest impacts are theory uncertainties based on the top p_T reweighting and scaling of the $\mathrm{t\bar{t}b\bar{b}}$ contribution cross section.

Experimental uncertainties	Process affected	Correlations across processes	Correlations across years
e/μ identification	MC	✓	✓
Trigger Efficiency	MC	✓	×
τ_{h} identification against jets	MC	✓	×
τ_{h} energy scale	MC	✓	×
b-tagging	MC	✓	partially
Jet energy scale	MC	✓	partially
Jet energy resolution	MC	✓	×
b-tagging	MC	✓	partially
Luminosity	MC	✓	partially
Pile up	MC	✓	✓
Pre-firing	MC	✓	-
Theoretical uncertainties	Process affected	Correlations across processes	Correlations across years
PDF $+\alpha_s$	$\mathrm{t\bar{t}b\bar{b}}$,$\mathrm{t\bar{t}jj}$, fake τ_{h}, $\mathrm{t\bar{t}V}$, VLL	×	✓
μ_F and μ_R scale	$\mathrm{t\bar{t}b\bar{b}}$,$\mathrm{t\bar{t}jj}$, fake τ_{h}, $\mathrm{t\bar{t}V}$, VLL	×	✓
Parton Shower	$\mathrm{t\bar{t}b\bar{b}}$,$\mathrm{t\bar{t}jj}$, fake τ_{h}, $\mathrm{t\bar{t}V}$, VLL	×	✓
top p_{T} reweighting	$\mathrm{t\bar{t}b\bar{b}}$,$\mathrm{t\bar{t}jj}$, fake τ_{h}	✓	✓
$\mathrm{t\bar{t}b\bar{b}}$ Cross-Section	$\mathrm{t\bar{t}b\bar{b}}$	-	✓
Fake rate	fake τ_{h}	-	×
Limited Simulation Statistics	MC	-	×

Table 6.7: Summary of all the uncertainties affecting the parameter of interest. The PDF, μ_F/μ_R scale and Parton shower uncertainties are correlated between $\mathrm{t\bar{t}b\bar{b}}$, $\mathrm{t\bar{t}jj}$ and fake τ_{h}, but uncorrelated to the $\mathrm{t\bar{t}V}$ and VLL signal process.

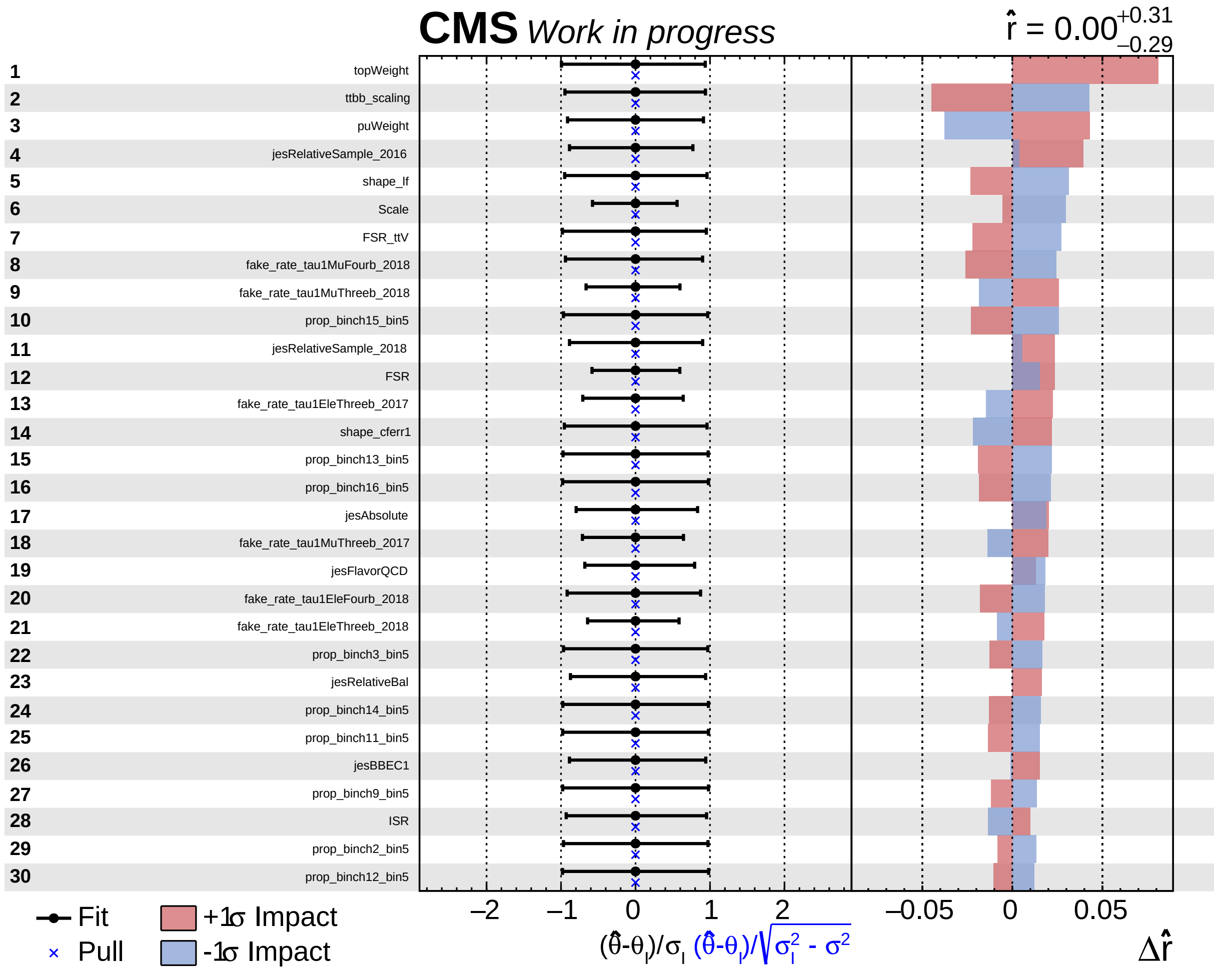


Figure 6.7: Pulls and impacts for the systematic uncertainties with the largest impact in the measurement of the signal strength. The uncertainties with the name *prob_ binX* are statistical uncertainties of the DNN output score bins for the different channels and years. The uncertainties with one-sided impacts only affect one direction of the signal strength.

6.8 Results

The search is conducted with a log-likelihood fit as for the bbH search, described in section 5.8. In contrast to the bbH search the DNN has only one output, the discrimination of signal and $t\bar{t}$ background, which is fit to derive the limits. The fit is performed only with bins above 0.2 to exclude a large background contribution at low DNN values. This prevents the low-DNN region, which has high statistics but different kinematics, from overconstraining the signal sensitive high-DNN region.

6.8.1 Validation region studies

The validation regions are used to check the agreement of the data with simulation and therefore verify the simulation of the major backgrounds and the DNN. In the validation regions only a very small signal contribution is expected. The prefit and background-only postfit plots in Fig. 6.8 for the validation regions with two b-jets, Fig. 6.9 for the three b-jets regions and in Fig. 6.10 for the four b-jets regions show a good agreement of the data with the background model.

6.8.2 Blinded limits

At the time of writing, the results of the analysis are still blinded. This means that in the signal region the data is not yet compared to the signal plus background hypothesis or the background-only hypothesis. The expected 95% confidence level upper limit on the signal strength are shown in Fig. 6.11 combining the electron and muon channel and all years. The expected sensitivity of the analysis is similar to the results from the fully hadronic analysis. In the limit plots the theory prediction for a electroweak VLL production is shown. If the observed upper limit is below the theory prediction for any mass point the VLL is considered excluded for this specific mass point. In the absence of a signal, VLL masses up to 680 GeV are expected to be excluded at a 95% confidence level.

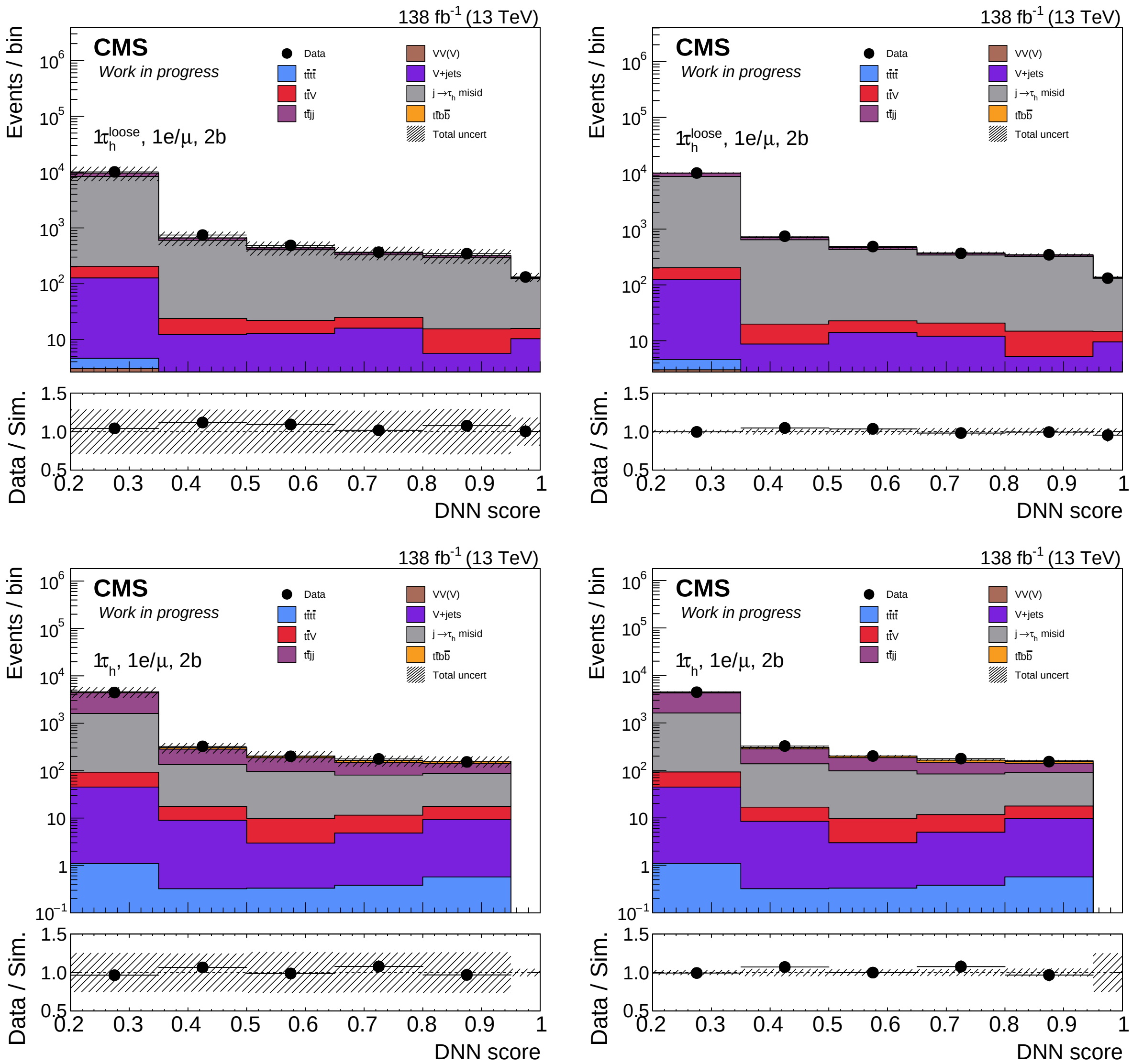


Figure 6.8: Prefit (left) and background only postfit plots (right) of the DNN output scores for the validation region with one τ_h with loose isolation and two b-jets (top) and for the validation region with one τ_h with tight isolation and two b-jets (bottom). The plots combine the electron and muon channel as well as all years, but they are treated separately in the fit.

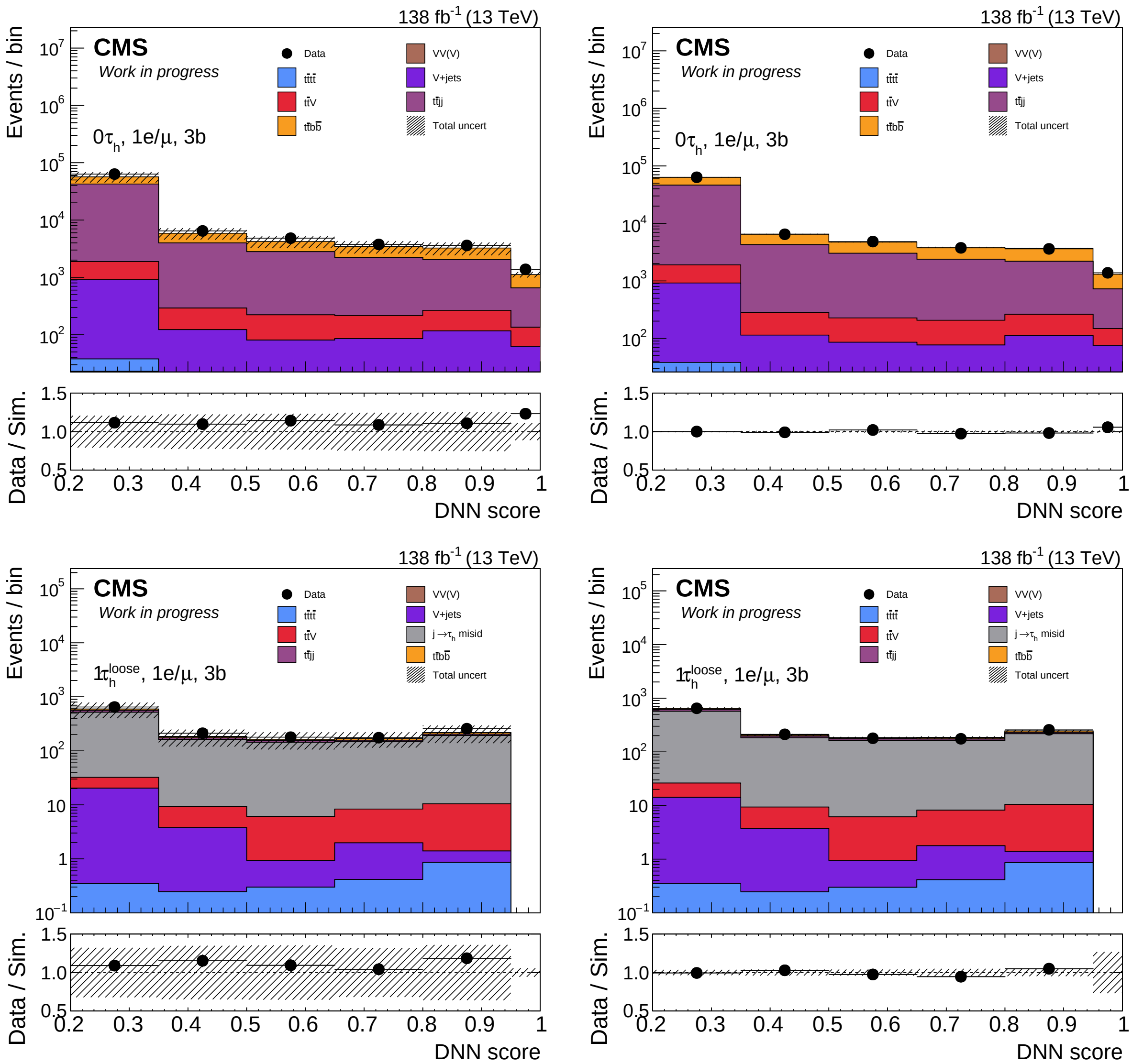


Figure 6.9: Prefit (left) and background only postfit plots (right) of the DNN output scores for the validation region with zero τ_h and three b-jets (top) and for the validation region with one τ_h with loose isolation and three b-jets (bottom). The plots combine the electron and muon channel as well as all years, but they are treated separately in the fit.

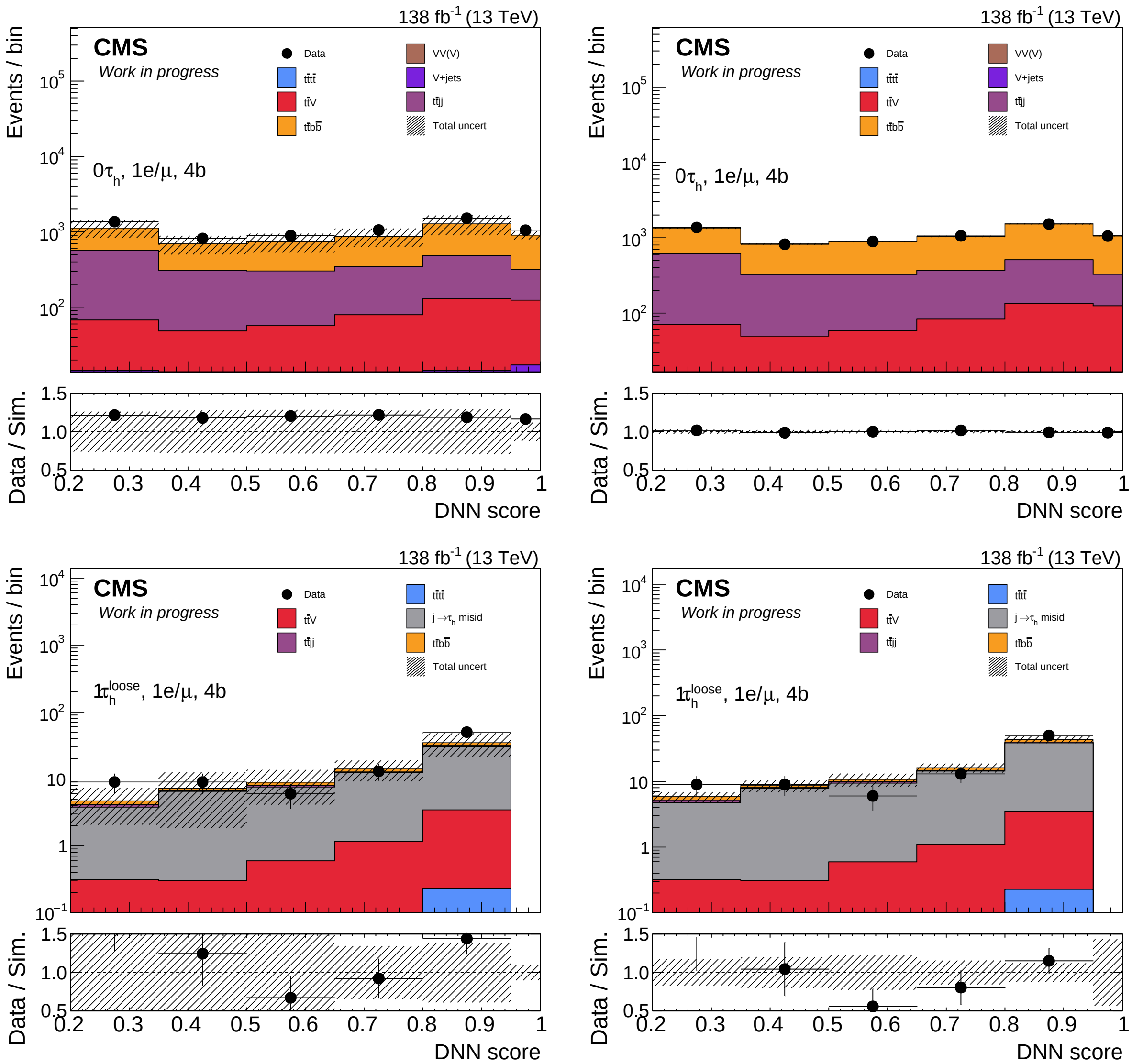


Figure 6.10: Prefit (left) and background only postfit plots (right) of the DNN output scores for the validation region with zero τ_h and four b-jets (top) and for the validation region with one τ_h with loose isolation and four b-jets (bottom). The plots combine the electron and muon channel as well as all years, but they are treated separately in the fit.

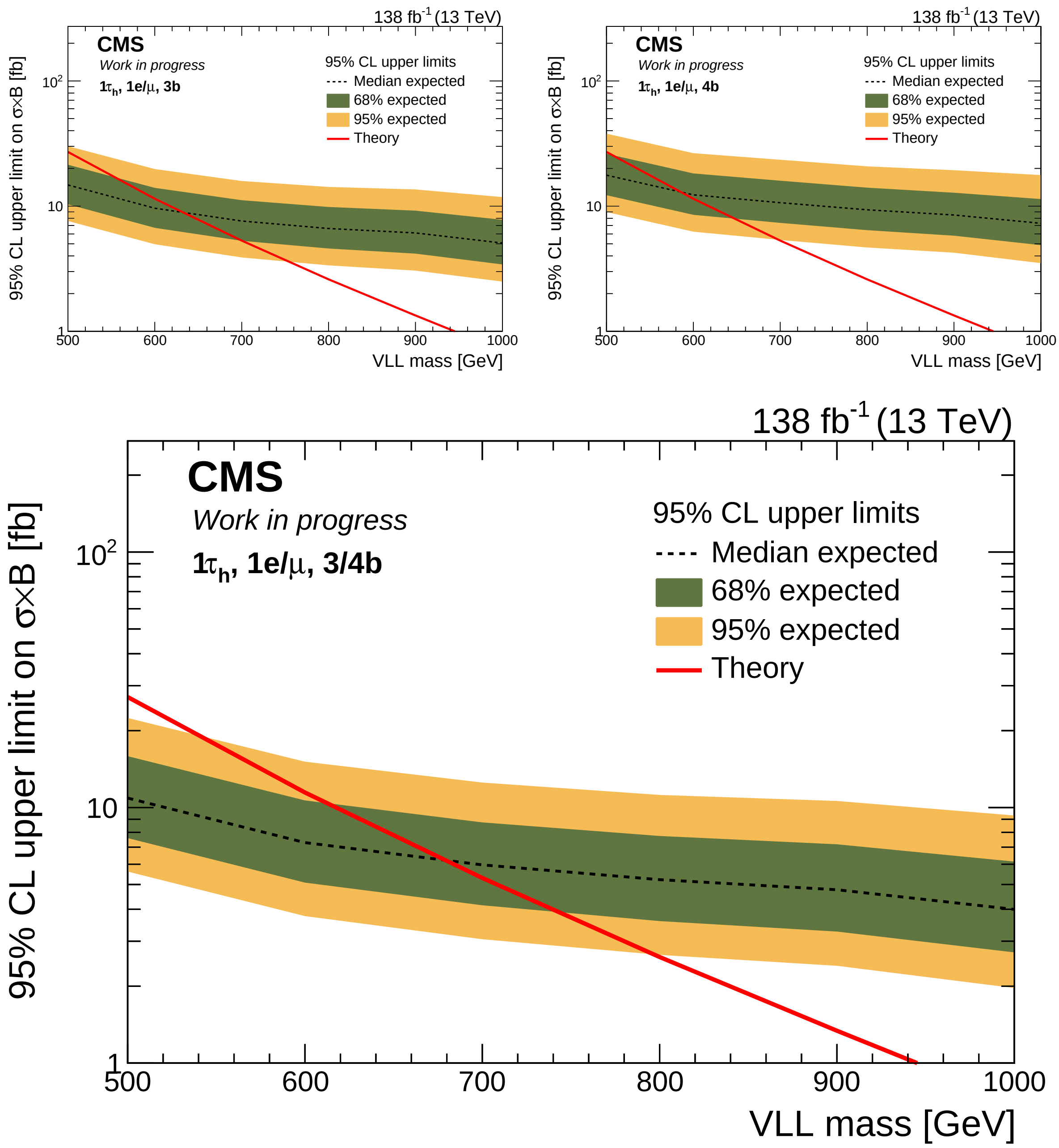


Figure 6.11: Expected 95% confidence upper limits on the signal strength, combining all years for the three b-jets (top left) and four b-jets channel (top right), and the combination of the channels (bottom). VLL masses up to 680 GeV are expected to be excluded at a 95% confidence level.

6.9 Summary

The search for vector-like leptons is conducted in a mass range of 500 to 1000 GeV using the data collected by the CMS experiment from 2016 to 2018, corresponding to an integrated luminosity of $138\,\mathrm{fb}^{-1}$. The VLL are predicted by the 4321 model to address flavor hierarchies and potential lepton flavor violations. The VLL couple to SM fermions via massive leptoquarks, that couple leptons to quarks with the strongest coupling to the third generation. The excess of data events over the background-only hypothesis found in the search for VLL in the fully hadronic final state is checked with the presented search in the final state with one electron or muon, one hadronic tau decay and three or four bottom quark jets. The signal is discriminated from the $\mathrm{t\bar{t}}$ background by training a deep neutral network. The analysis is still blinded at the time of writing, therefore only the expected limits can be evaluated. In the absence of a signal, VLL masses up to 680 GeV are expected to be excluded at a 95% confidence level. The sensitivity of the limit is similar to the search with the fully hadronic final state. If the excess is confirmed, it would be an exciting hint of the existence of vector-like leptons.

7 CMS Pixel detector upgrade

In this chapter my contribution to the pixel sensor testbeam data analysis is described. The testbeam measurements were conducted in 2020 and 2021 at a testbeam at DESY by a team of UZH, ETH and PSI.

7.1 Introduction

In the future, planned starting date is 2029, the large hadron collier (LHC) will increase instantaneous luminosity to enable the experiment to collect more data. This phase of the LHC is called high-luminosity LHC. The increase in the proton collision rate leads to higher radiation in the detector and also to more pile up. This requires, among different upgrades of the CMS subdetectors, a complete replacement of the current tracker detector. In section 3.2.2 the current pixel detector is described. The radiation close to the beam pipe, where the CMS pixel detector is located, is expected to increase to 5 times the current rate. The pile up per bunch crossing is expected to increase to 200. The higher hit rate requires pixel sensors with smaller cell size. The upgraded pixel detector will have four barrel layers as in the current detector design. On either side of the barrel the detector will have eight endcap discs and four discs further from the interaction point, to increase the coverage from $\eta = 2.5$ to $\eta = 4$. The layout of the Phase 2 pixel detector is shown in Fig. 7.1.

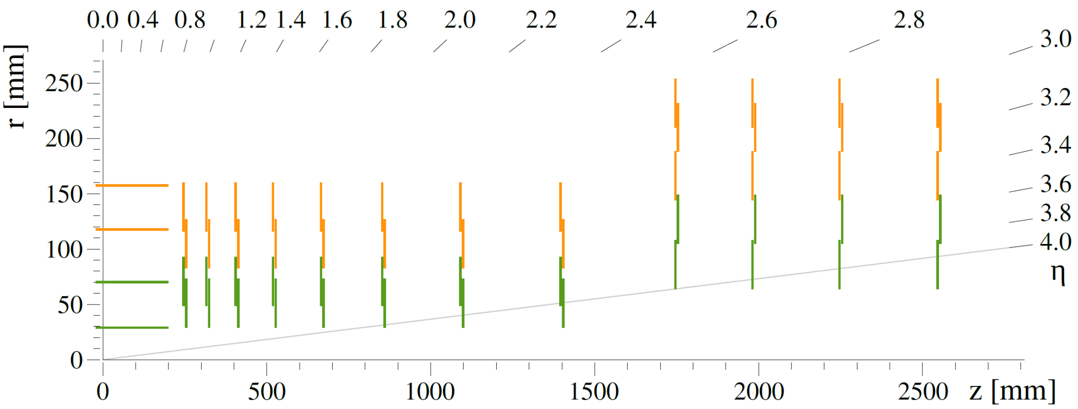


Figure 7.1: Layout of one quarter of the Phase 2 pixel tracker detector. The modules with two readout chips (1×2) are marked in green, the modules with four readout chips (2×2) in orange. Figure taken from [115].

The barrel layers and disks are built from modules. Each module consists of pixel sensors with readout chips bump-bonded to the pixels. A flex circuit is glued to the sensors and wire-bonded to the readout chips for transmitting data and providing power. Each module will have either two CMS readout chips, called CROC, in a 1x2 format or 4 chips in a 2x2 format, depending on the distance to the interaction point as shown in Fig. 7.1. The pixel sensor with two planar pixel cells is shown in Fig. 7.2.

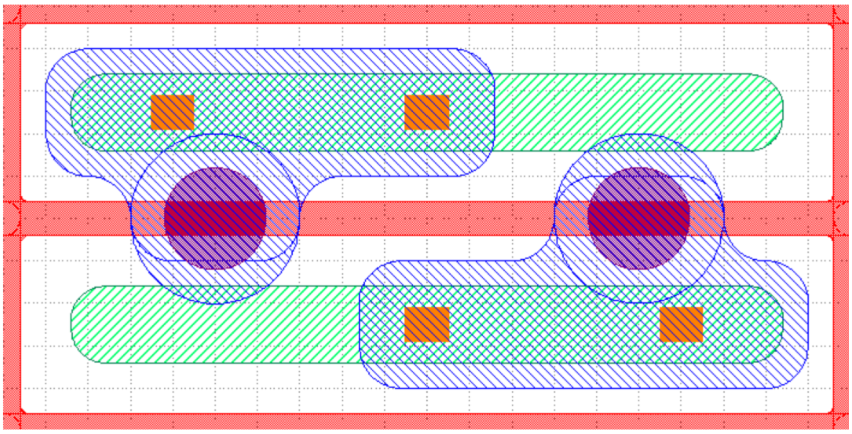


Figure 7.2: Pixel module with metal layers in blue, the n^+ implants in green and the p-stop region in red. The bump-bond between sensors and readout chip are shown in purple. Figure from [115].

The sensors have been tested during the prototyping phase at the Super Proton Synchrotron (SPS) at CERN or the electron testbeam at DESY, short for *Deutsches Elektronen Synchrotron*. In this chapter the data analysis of the testbeam data measured at DESY is described.

7.2 Silicon sensor

The silicon sensor consists of a p-type and a n-type region. P-type refers to the excess of free positive charged holes, and n-type to the excess of free negative charged electrons. The excess can be achieved by doping the silicon with impurity atoms. The electrons from the n-type region recombine with the holes from the p-type region, creating a volume depleted of free charge carriers between the n-type and p-type region. The recombination of electrons also results in an electric field across the depletion zone due to the fixed charges of opposite signs left on both sides. A reverse bias voltage is applied in order to widen the depletion zone and to prevent any current from flowing. If a charged particle passes the silicon pixel, it ionizes the silicon by generation of electron-hole pairs. Due to the electric field, the electrons drift towards the n-type region and the holes towards the p-type region, where the hit is registered and passed to the readout chip. In order to deal with the higher hit rate, the pixel sensor cell size is reduced from $100 \times 150\,\mu\text{m}^2$ to $25 \times 100\,\mu\text{m}^2$. The long side of the pixels is along the z axis for the barrel and along r in the endcap discs. Pixel sensors with a size of $50 \times 50\,\mu\text{m}^2$ were also tested. Those pixel sensors have a better primary vertex discrimination but a worse resolution in the impact parameter. In the end it has been decided to adopt a pixel cell size of $25 \times 100\,\mu\text{m}^2$.
Two different designs of silicon sensors are tested, planar sensors and 3D sensors. The planar sensors consist of n^+-in-p implants and of p-stop regions to isolate the n^+ wells. The planar sensors need a bias voltage of up to 800 GeV to ensure a good enough efficiency after irradiation. The planar sensors of the companies Hamamatsu Photonics K.K. (HPK) and LFoundry are tested. The HPK planar sensors are built from 6-inch wafers and their design is shown in Fig. 7.3. The planar sensors from LFoundry are 8-inch wafers built with the Complementary Metal-Oxide-Semiconductor (CMOS) technology.
The 3D sensors have n^+ vertical columns for collecting the charge carriers created by ionization and p^+ vertical columns that reach the backside of the sensor, at which the bias voltage is applied. This design reduces the distance that the charge carriers have to drift through the silicon and requires less bias voltage and therefore also less power consumption. The 3D sensors

are more radiation hard than the planar sensors, which means they can withstand a larger dose of radiation without or with only minor defects. The complicated design makes them more expensive than the planar sensors, therefore they are only used in the first layer, where the radiation is highest. The 3D sensors from the company Fondazione Bruno Kessler (FBK) are built as 6-inch wafers, with the design shown in Fig. 7.4.

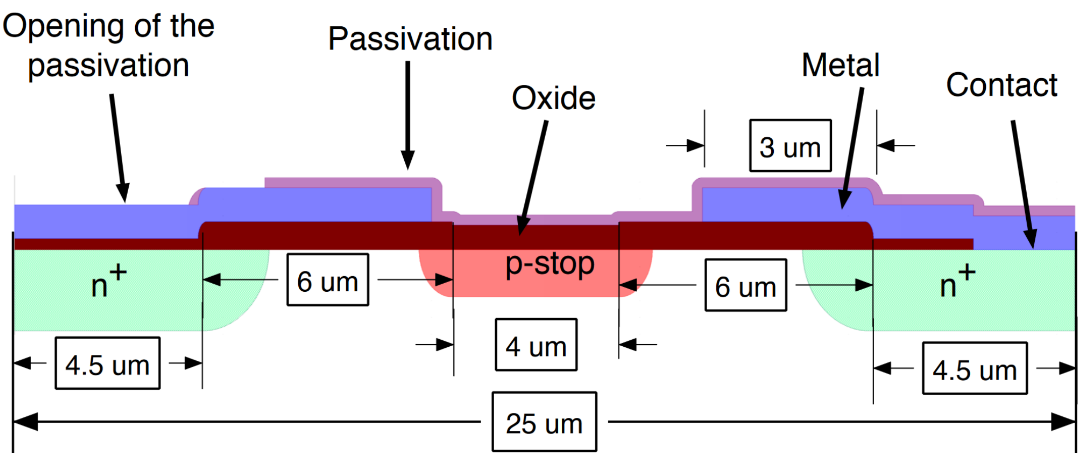


Figure 7.3: HPK planar sensor with the n^+-in-p implants isolated by the p-stop region. Figure taken from [116].

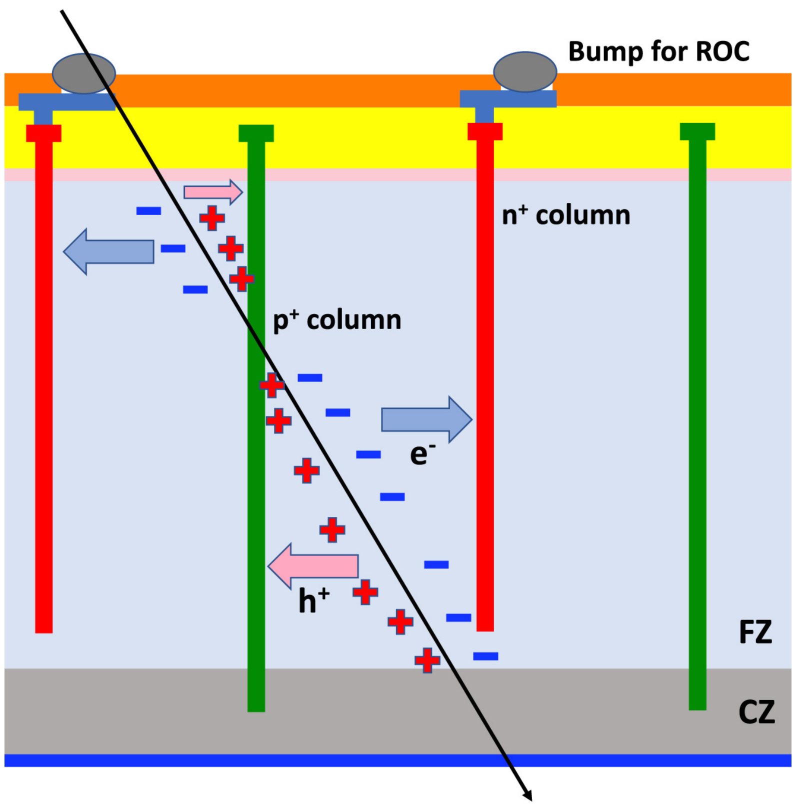


Figure 7.4: FBK 3D sensor with n^+ and p^+ columns inside the silicon volume. Figure taken from [117].

7.3 Readout Chip

The CMS readout chip (CROC) is bump-bonded to the sensors and measures the signal induced by the charge carriers. The analog signal is amplified and the time during which the signal is over the threshold (ToT) is digitized. The ToT is used to differentiate signal from noise and indicates the time the signal is over a specified threshold.
The Phase 2 readout chip for CMS and ATLAS is developed by the RD53 collaboration. The first prototype of the chip called RD53A has three front ends, the synchronous, linear and differential [118]. CMS decided to use the linear front end, which was used for the CROC aside from other modifications specific to the conditions inside the CMS detector. The chip will be serially powered in the detector by applying a constant current flowing across a module chain [119].

7.4 Testbeam data analysis

The sensors are tested in a test beam at DESY with 5.2 GeV electrons to measure their attributes. In order to mimic the radiation the sensors accumulate over time, some of the sensors are irradiated. The sensors are irradiated with protons and the doses are expressed in terms of fluence of protons per square centimeter. The radiation causes displacement damage in the silicon lattice, inducing an increased leakage current, a higher depletion voltage and trapping of the charge carriers.
The sensor to be tested is called Device Under Test (DUT) and is cooled to around -30 degree Celsius during the testing. Down- and up-stream of the DUT three telescope planes are placed in order to reconstruct the particle track. A time reference plane is used because of the long integration time of the telescope planes with respect to the 25 ns integration time of the readout chip. A scintillator is placed upstream of the beam telescope to provide a trigger [120]. The setup of the testbeam is shown in Fig. 7.5.

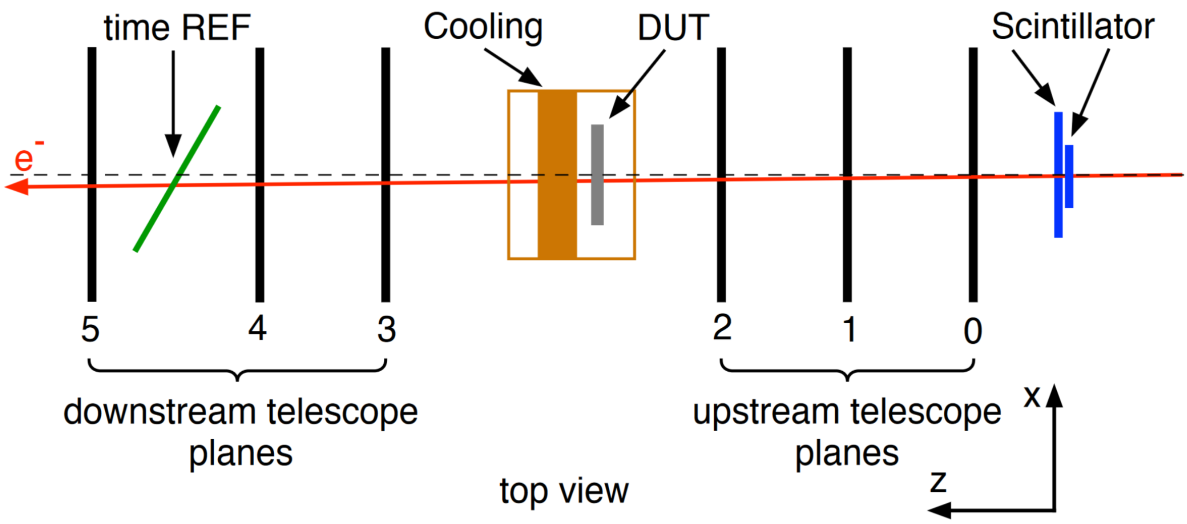


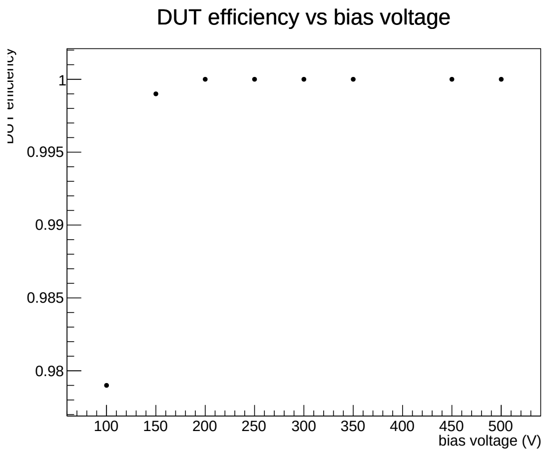
Figure 7.5: Testbeam setup with the device under test (DUT) in a cooling box, three telescope planes down- and up-stream of the DUT, a time reference plane and the scintillator for triggering. Figure taken from [116].

The tested modules include sensors from different companies, some irradiated, as shown in the table 7.1. The sensors were interconnected with the RD53A readout chip.

Vendor	size (μm^2)	design	irradiated (n_{eq}/cm^2)
HPK (M595)	25×100	bitten	0.8×10^{16}
HPK (M596)	25×100	bitten	1.2×10^{16}
HPK	25×100	bitten	-
FBK	25×100	non-bitten	-
FBK	25×100	bitten	-
LFoundry	25×100	bitten	10×10^{15}
LFoundry	25×100	bitten	2.1×10^{15}
LFoundry	25×100	bitten	9.2×10^{15}
LFoundry	50×50	bitten	5×10^{15}

Table 7.1: Tested modules in the testbeam at DESY in 2020 and 2021.

In the bitten design the aluminum layer in each pixel cell has been cut out to avoid any overlay with the bump-bond of the neighboring pixel. The reason is that crosstalk effects have been observed with the design shown in Fig. 7.2. Crosstalk refers to the capacitive induction of charge from a pixel into its paired neighboring pixel, that leads to a deterioration of the spatial resolution [121].
The signal strength of the collected charge is given by the time over threshold (ToT). More collected charge leads to a larger ToT, but not necessarily in a linear way. The ToT values for each pixel need to be calibrated by injecting a certain amount of charge to the pixel. The ToT is mapped to charge, given in ΔVCAL units, where 1 ΔVCAL corresponds to 10 electrons. [122].
The efficiency of the sensor is derived as the ratio of the number of DUT hits matched to a reconstructed track within 0.2 mm over the number of reconstructed tracks passing the DUT. The tracks are reconstructed from the telescope planes. The DUT efficiency depends on the bias voltage, therefore the efficiency is measured as function of the applied bias voltage. In figure 7.6 the efficiency is shown for two LFoundry modules, one irradiated with 2.1×10^{15}cm^{-2} and the other with 9.2×10^{15}cm^{-2}. In the irradiated module a fraction of the charge carriers get trapped and are not collected, leading to a lower efficiency. In order to have an acceptable efficiency, the bias voltage needs to be increased after irradiation to create a higher electric field and a wider depletion volume. The bias voltage is also increased from time to time in the current CMS pixel detector to balance out any defects caused by radiation.



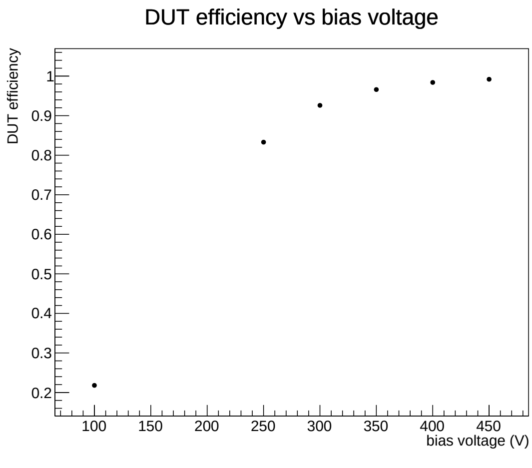


Figure 7.6: Efficiency measured as function of the applied bias voltage for the LFoundry Module irradiated with $\phi_{eq} = 2.1 \times 10^{15}$cm^{-2} (left) and $\phi_{eq} = 9.2 \times 10^{15}$cm^{-2} (right).

The spatial resolution of the sensor is derived by the standard deviation of the difference between the reconstructed position from the telescope planes and the charge-weighted center of the cluster in the DUT. The telescope also has a resolution that needs to be measured and subtracted from the total resolution in order to get the resolution from the DUT. The contribution from the telescope resolution depends on the incident angle:

$$\sigma_{\mathrm{DUT}} = \sqrt{\sigma^2 - \left(\frac{\sigma_{\mathrm{telescope}}}{\cos(\theta)}\right)}, \tag{7.1}$$

with θ the incident angle of the beam.

The spatial resolution along the short, 25 µm side of the sensor, is measured as a function of the applied bias voltage, as shown in Fig. 7.7. The irradiated modules need a much higher bias voltage and have a larger spatial resolution than the non-irradiated modules. The resolution decreases with the increasing bias voltage. For the non-irradiated modules the difference between planar and 3D sensors is visible. The planar HPK sensor has a better resolution with larger bias voltage, for the 3D sensors the best resolution is with a bias voltage of only 10 V. The DUT is rotated to measure the resolution at different incident angles of the beam. The angular scan was taken at a bias voltage of 120 V for the non-irradiated modules. The best resolution is at an angle of -10 degrees, where the number of 2-hit clusters is maximized, improving the precision of the position determined by charge weighting. For larger incident angles the charge is spread over several pixels, leading to a decrease of the resolution. The resolution as function of the incident angle is shown in Fig. 7.8.

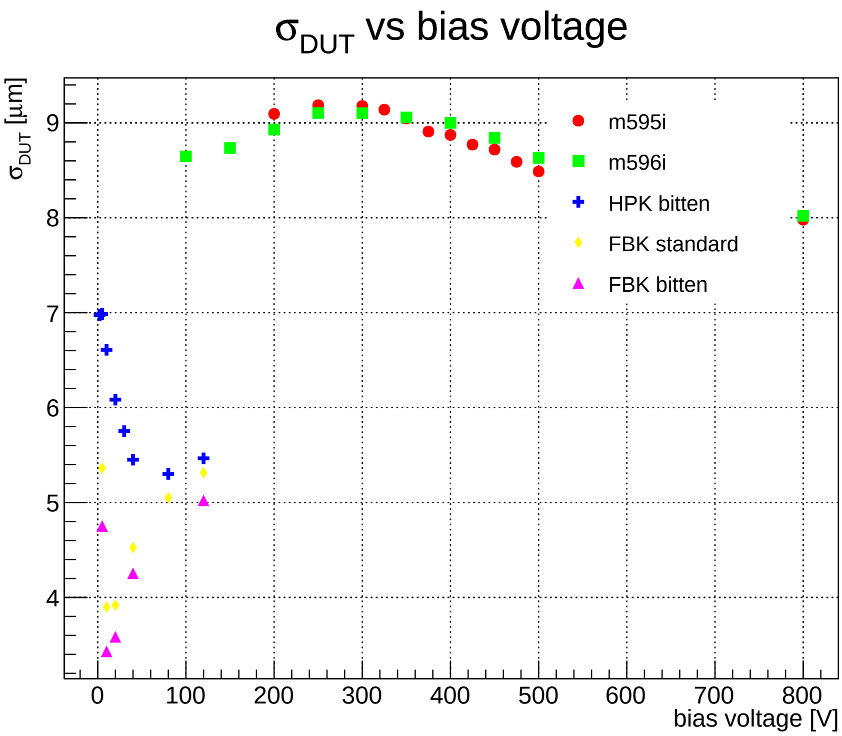


Figure 7.7: Spatial resolution along the shorter pixel cell side of the irradiated (m595i,m596i) and non-irradiated (HPK, FBK) planar HPK and 3D FBK sensors as function of the bias voltage. The 3D FBK sensors are tested in the standard (as shown in Fig. 7.2) and in the bitten design. Measurement was taken in the testbeam at DESY.

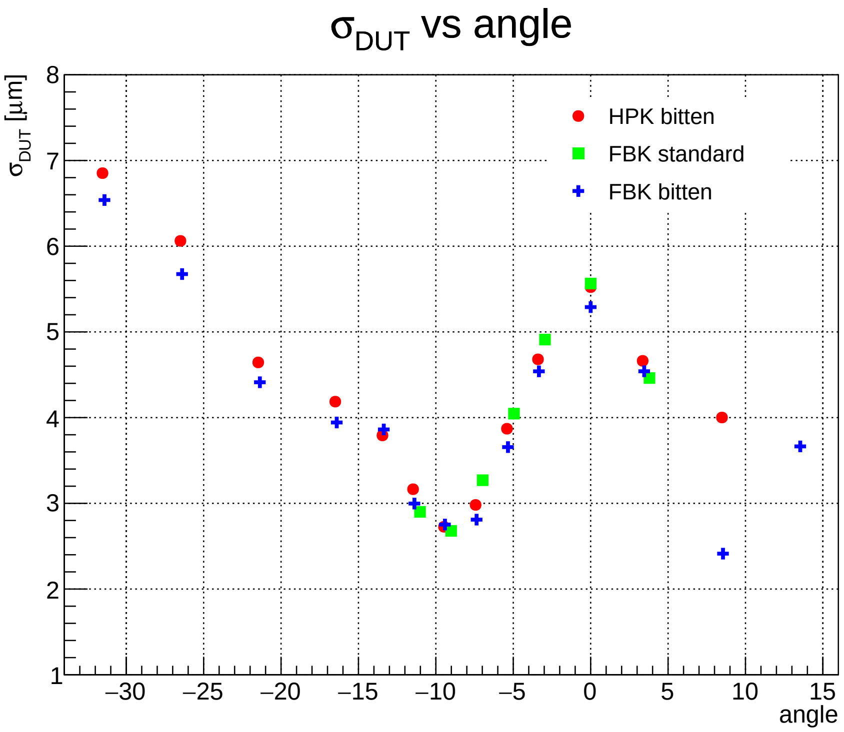


Figure 7.8: Spatial resolution of the non-irradiated planar HPK and 3D FBK sensors as function of the incident angle with a bias voltage of 120 V. Measurement was taken in the testbeam at DESY.

8 Conclusions

In this thesis two searches with bottom quarks and tau leptons in the final state are presented, one for the Standard Model process of the bottom quark associated production of the Higgs boson (referred to as bbH), the other for vector-like leptons predicted by beyond the Standard Model theories (referred to as VLL). Both searches used the data collected by the CMS experiment from 2016 to 2018, corresponding to an integrated luminosity of $138\,\mathrm{fb}^{-1}$.

The bbH process is being searched for the first time. The search is conducted in the final state with at least one bottom quark jet and with the Higgs boson decay to a tau lepton pair with the four final states $\tau_h\tau_h$, $\mu\tau_\mathrm{h}$, $\mathrm{e}\tau_\mathrm{h}$ and $\mathrm{e}\mu$, and in the Higgs boson decay to WW in the $\mathrm{e}\mu$ final state. My contribution was the design of all stages of the analysis in the semileptonic final states, $\mu\tau_\mathrm{h}$ and $\mathrm{e}\tau_\mathrm{h}$. The signal to background discrimination is enhanced by using machine learning algorithms, and their output is simultaneously fitted for all channels to derive upper limits on the signal process cross section. The 95% confidence level upper limit on the signal strength was observed as 3.7 times the expected Standard Model value and the expected upper limit was 6.1 times the Standard Model value. The signal strength was observed as $\mu_\mathrm{obs} = -2.7^{+2.0}_{-2.6}$, indicating an underfluctuation of the data with respect to the background-only hypothesis. The 68% and 95% confidence level simultaneous limits on the Higgs boson coupling to bottom quarks (κ_b) and to top quarks (κ_t) were derived, showing the potential of the analysis to constrain κ_b. Our measurement of the bbH process is mainly constrained by large theoretical uncertainties and statistical uncertainties. If the theoretical uncertainty can be reduced significantly, it will become possible to find evidence, defined as an observed excess over the expected background with a significance of three standard deviations, during the High Luminosity Large Hadron Collider phase. The analysis showed that most variables have very similar shapes for the bottom quark fusion and the gluon fusion process with gluon splitting to bottom quarks, which makes it very challenging to differentiate between them. Novel techniques will be required in order to observe the bottom quark fusion Higgs boson production mode directly and to constrain the κ_b without any dependence on the κ_t.

The bbH analysis results were released in a Physics Analysis Summary [108] and submitted in a paper to the Physics Letters B journal [83].

The VLL are predicted by the 4321 model to address flavor hierarchies and potential lepton flavor violations. The excess of data events over the background-only hypothesis found in the search for VLL in the fully hadronic final state is checked with the presented search in the final state with one electron or muon, one hadronic tau decay and three or four bottom quark jets. The signal is discriminated from the $\mathrm{t\bar{t}}$ background by training a deep neural network. My contribution was to update the framework and deep neural network, previously used in the fully hadronic analysis, to include one electron or muon in the final state, perform validations of the simulation, improve the uncertainty estimates and derive expected upper limits. The analysis is still blinded at the time of writing, therefore only the expected upper limits on the signal cross section can be evaluated. In the absence of a signal, VLL masses up to 680 GeV are expected to be excluded at a 95% confidence level. The sensitivity of the limit is similar to the search with the fully hadronic final state. If the excess is confirmed, it would be an exciting hint of the existence of vector-like leptons.

I also presented my contribution to the CMS pixel detector upgrade by analyzing the properties of pixel sensors measured at testbeams.

With the upcoming runs of the Large Hadron Collider the CMS experiment will collect substantial more data than used in the presented searches. With the increased size of data events the statistical uncertainty on the measurements become smaller, giving the opportunity to observe rare Standard Model processes like the bottom quark associated production of the Higgs boson, and improve the sensitivity for particles predicted by beyond Standard Model theories like vector-like leptons.

References

[1] M. Thomson, "Modern particle physics". Cambridge University Press, New York, 2013. doi:10.1017/CBO9781139525367, ISBN 978-1-107-03426-6.

[2] D. J. Griffiths, "Introduction to elementary particles; 2nd rev. version". Physics textbook. Wiley, New York, 2008. doi:10.1002/9783527618460, ISBN 978-3-527-40601-2.

[3] Wikimedia Commons, "Standard Model of Elementary Particles", https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg, 2024. (Retrieved 25 March 2024).

[4] S. L. Glashow, "The renormalizability of vector meson interactions", *Nuclear Physics* **10** (1959) 107, doi:10.1016/0029-5582(59)90196-8.

[5] A. Salam and W. J. C., "Weak and electromagnetic interactions", *Il Nuovo Cimento* **11** (1959) 568, doi:10.1007/BF02726525.

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

[7] S. L. Glashow, "Partial-symmetries of weak interactions", *Nuclear Physics* **22** (1961), no. 4, 579, doi:https://doi.org/10.1016/0029-5582(61)90469-2.

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

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

[10] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles", *Phys. Rev. Lett.* **13** (1964) 585, doi:10.1103/PhysRevLett.13.585.

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

[12] **CMS** Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys. Lett. B* **716** (2012) 30, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

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

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

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

[16] **Particle Data** Group, "Review of Particle Physics, 2016-2017", *Chin. Phys. C* **40** (2016), no. 10, 100001, doi:10.1088/1674-1137/40/10/100001.

[17] **CMS** Collaboration, "Stairway to discovery: a report on the CMS programme of cross section measurements from millibarns to femtobarns", arXiv:2405.18661. Submitted to Phys. Rep.

[18] **ATLAS** Collaboration, "A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery", *Nature* **607** (2022), no. 7917, 52, doi:10.1038/s41586-022-04893-w, arXiv:2207.00092. [Erratum: Nature 612, E24 (2022)].

[19] M. Spira, "Higgs Boson Decays: Theoretical Status", *CERN Yellow Rep. Monogr.* **3** (2020) 123, doi:10.23731/CYRM-2020-003.123.

[20] I. Neutelings, "Piechart of SM decays", https://tikz.net/sm_decay_piechart, 2021. (Retrieved 16 May 2024).

[21] **CMS** Collaboration, "Precision measurement of the W boson decay branching fractions in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. D* **105** (2022), no. 7, 072008, doi:10.1103/PhysRevD.105.072008, arXiv:2201.07861.

[22] **ATLAS** Collaboration, "Precise test of lepton flavour universality in W-boson decays into muons and electrons in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", arXiv:2403.02133. Submitted to Eur. Phys. J. C.

[23] **LHCb** Collaboration, "Measurement of forward $W \to e\nu$ production in pp collisions at $\sqrt{s} = 8\,\text{TeV}$", *JHEP* **10** (2016) 030, doi:10.1007/JHEP10(2016)030, arXiv:1608.01484.

[24] **ATLAS** Collaboration, "Test of the universality of τ and μ lepton couplings in W-boson decays with the ATLAS detector. Test of the universality of τ and μ lepton couplings in W-boson decays from $t\bar{t}$ events with the ATLAS detector", *Nature Phys.* **17** (2021), no. 7, 813, doi:10.1038/s41567-021-01236-w, arXiv:2007.14040.

[25] A. Greljo, G. Isidori, and D. Marzocca, "On the breaking of Lepton Flavor Universality in B decays", *JHEP* **07** (2015) 142, doi:10.1007/JHEP07(2015)142, arXiv:1506.01705.

[26] L. Di Luzio, A. Greljo, and M. Nardecchia, "Gauge leptoquark as the origin of B-physics anomalies", *Phys. Rev. D* **96** (2017), no. 11, 115011, doi:10.1103/PhysRevD.96.115011, arXiv:1708.08450.

[27] L. Di Luzio et al., "Maximal Flavour Violation: a Cabibbo mechanism for leptoquarks", *JHEP* **11** (2018) 081, doi:10.1007/JHEP11(2018)081, arXiv:1808.00942.

[28] L. Di Luzio, M. Kirk, A. Lenz, and T. Rauh, "ΔM_s theory precision confronts flavour anomalies", *JHEP* **12** (2019) 009, doi:10.1007/JHEP12(2019)009, arXiv:1909.11087.

[29] **CMS** Collaboration, "Search for pair-produced vector-like leptons in final states with third-generation leptons and at least three b quark jets in proton-proton collisions at $\sqrt{s}$=13TeV", *Phys. Lett. B* **846** (2023) 137713, doi:10.1016/j.physletb.2023.137713, arXiv:2208.09700.

[30] CMS Higgs group Conveners, "CMS Higgs Physics Results", https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsHIG, 2023.

(Retrieved 16 May 2024).

[31] C. Cornella et al., “Reading the footprints of the B-meson flavor anomalies”, *JHEP* **08** (2021) 050, doi:10.1007/JHEP08(2021)050, arXiv:2103.16558.

[32] I. Neutelings, “B tagging jets”, https://tikz.net/jet_btag, 2021. (Retrieved 9 May 2024).

[33] I. Neutelings, “Hadronically decayed tau jet”, https://tikz.net/jet_tau, 2021. (Retrieved 19 May 2024).

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

[35] I. Efthymiopoulos et al., “Bunch Luminosity Variations in LHC Run 2”, *JACoW* **IPAC2021** (2021) 4094, doi:10.18429/JACoW-IPAC2021-THPAB172.

[36] CERN Accelerator Coordination and Engineering group, “CERN accelerator complex”, https://home.web.cern.ch/science/accelerators/accelerator-complex/panoramas, 2019. (Retrieved 5 January 2024).

[37] **CMS** Collaboration, “Public CMS Luminosity Information”, https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults, 2024. (Retrieved 15 April 2024).

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

[39] I. Neutelings, “CMS Coordinate System”, https://tikz.net/axis3d_cms, 2021. (Retrieved 4 April 2024).

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

[41] **CMS** Collaboration, “The CMS tracker system project: Technical Design Report”, Technical Report CERN-LHCC-98-006, CMS-TDR-5, 1997.

[42] CMS Collaboration, “CMS Tracker Detector Performance Results”, https://twiki.cern.ch/twiki/bin/view/CMSPublic/DPGResultsTRK, 2020. (Retrieved 14 April 2024).

[43] **CMS** Collaboration, “The CMS electromagnetic calorimeter project: Technical Design Report”, Technical Report CERN-LHCC-97-033, CMS-TDR-4, Geneva, 1997.

[44] **CMS** Collaboration, “The CMS ECAL performance with examples”, *JINST* **9** (2014) C02008, doi:10.1088/1748-0221/9/02/C02008.

[45] **CMS** Collaboration, “The CMS hadron calorimeter project: Technical Design Report”, Technical Report CERN-LHCC-97-031, CMS-TDR-2, Geneva, 1997.

[46] **CMS** Collaboration, “Calibration of the CMS hadron calorimeters using proton-proton collision data at $\sqrt{s} = 13$ TeV”, *JINST* **15** (2020), no. 05, P05002, doi:10.1088/1748-0221/15/05/P05002, arXiv:1910.00079.

[47] **CMS** Collaboration, “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV”, *JINST* **13** (2018) P06015,

doi:10.1088/1748-0221/13/06/P06015, arXiv:1804.04528.

[48] **CMS** Collaboration, "The CMS muon project : Technical Design Report", Technical Report CERN-LHCC-97-032, CMS-TDR-3, Geneva, 1997.

[49] C. Collaboration, "The CMS trigger system", *JINST* **12** (2017) P01020, doi:10.1088/1748-0221/12/01/p01020, arXiv:0312279.

[50] D. Barney, "CMS Slice", https://cds.cern.ch/record/2628641, 2015. (Retrieved 11 April 2024).

[51] **CMS** Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* **12** (2017) P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:1706.04965.

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

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

[54] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov, "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC", *J. Phys. G: Nucl. Part. Phys.* **31** (2005), no. 9, N9, doi:10.1088/0954-3899/31/9/n01, arXiv:0306087.

[55] **CMS** Collaboration, "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC", *JINST* **16** (2021) P05014, doi:10.1088/1748-0221/16/05/P05014, arXiv:2012.06888.

[56] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas", *Phys. Lett. B* **659** (2008) 119, doi:10.1016/j.physletb.2007.09.077, arXiv:0707.1378.

[57] **CMS** Collaboration, "Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV", *JINST* **13** (2018), no. 10, P10005, doi:10.1088/1748-0221/13/10/P10005, arXiv:1809.02816.

[58] **CMS** Collaboration, "Identification of hadronic tau lepton decays using a deep neural network", *JINST* **17** (2022), no. 07, P07023, doi:10.1088/1748-0221/17/07/p07023, arXiv:2201.08458.

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

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

[61] **CMS** Collaboration, "Jet Energy Scale and Resolution Measurements in CMS", *PoS* **ICHEP2022** (2022) 652, doi:10.22323/1.414.0652, arXiv:2301.02175.

[62] **CMS** Collaboration, "Pileup Jet Identification", CMS Physics Analysis Summary CMS-PAS-JME-13-005, CERN, 2013.

[63] **CMS** Collaboration, "Jet algorithms performance in 13 TeV data", CMS Physics Analysis Summary CMS-PAS-JME-16-003, CERN, 2017.

[64] E. Bols et al., "Jet Flavour Classification Using DeepJet", *JINST* **15** (2020), no. 12,

P12012, doi:10.1088/1748-0221/15/12/P12012, arXiv:2008.10519.

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

[66] J. Alwall et al., “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”, *JHEP* **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

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

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

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

[70] J. Alwall et al., “MadGraph 5 : Going Beyond”, *JHEP* **06** (2011) 128, doi:10.1007/JHEP06(2011)128, arXiv:1106.0522.

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

[72] S. Agostinelli et al., “Geant4—a simulation toolkit”, *Nucl. Instrum. Meth. A* **506** (2003), no. 3, 250, doi:https://doi.org/10.1016/S0168-9002(03)01368-8.

[73] **CMS** Collaboration, “Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector”, *Phys. Lett. B* **779** (2018) 283, doi:10.1016/j.physletb.2018.02.004, arXiv:1708.00373.

[74] **ATLAS** Collaboration, “Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”, *Phys. Rev. D* **99** (2019) 072001, doi:10.1103/PhysRevD.99.072001, arXiv:1811.08856.

[75] **CMS** Collaboration, “Observation of Higgs boson decay to bottom quarks”, *Phys. Rev. Lett.* **121** (2018), no. 12, 121801, doi:10.1103/PhysRevLett.121.121801, arXiv:1808.08242.

[76] **ATLAS** Collaboration, “Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector”, *Phys. Lett. B* **786** (2018) 59, doi:10.1016/j.physletb.2018.09.013, arXiv:1808.08238.

[77] **CMS** Collaboration, “Measurement of the Higgs boson production rate in association with top quarks in final states with electrons, muons, and hadronically decaying tau leptons at $\sqrt{s} = 13$ TeV”, *Eur. Phys. J. C* **81** (2021), no. 4, 378, doi:10.1140/epjc/s10052-021-09014-x, arXiv:2011.03652.

[78] **ATLAS** Collaboration, “Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector”, *Phys. Lett. B* **784** (2018) 173,

doi:10.1016/j.physletb.2018.07.035, arXiv:1806.00425.

[79] **CMS** Collaboration, “A portrait of the Higgs boson by the CMS experiment ten years after the discovery.”, *Nature* **607** (2022), no. 7917, 60, doi:10.1038/s41586-022-04892-x, arXiv:2207.00043. [Erratum: Nature 623, (2023)].

[80] **ATLAS** Collaboration, “A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery”, *Nature* **607** (2022), no. 7917, 52, doi:10.1038/s41586-022-04893-w, arXiv:2207.00092. [Erratum: Nature 612, E24 (2022)].

[81] N. Deutschmann, F. Maltoni, M. Wiesemann, and M. Zaro, “Top-Yukawa contributions to bbH production at the LHC”, *JHEP* **07** (2019) 054, doi:10.1007/JHEP07(2019)054, arXiv:1808.01660.

[82] D. Pagani, H.-S. Shao, and M. Zaro, “RIP $Hb\bar{b}$: how other Higgs production modes conspire to kill a rare signal at the LHC”, *JHEP* **11** (2020) 036, doi:10.1007/JHEP11(2020)036, arXiv:2005.10277.

[83] **CMS** Collaboration, “Search for bottom quark associated production of the standard model Higgs boson in final states with leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV”, arXiv:2408.01344. Submitted to Phys. Lett. B.

[84] **LHC Higgs Cross Section Working** Group, “Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector”, *CERN Yellow Rep. Monogr.* **2/2017** (2016) doi:10.23731/CYRM-2017-002, arXiv:1610.07922.

[85] S. Dawson, C. B. Jackson, L. Reina, and D. Wackeroth, “Exclusive Higgs boson production with bottom quarks at hadron colliders”, *Phys. Rev. D* **69** (2004) 074027, doi:10.1103/PhysRevD.69.074027, arXiv:hep-ph/0311067.

[86] S. Dittmaier, M. Krämer, and M. Spira, “Higgs radiation off bottom quarks at the Tevatron and the CERN LHC”, *Phys. Rev. D* **70** (2004) 074010, doi:10.1103/PhysRevD.70.074010, arXiv:hep-ph/0309204.

[87] R. V. Harlander and W. B. Kilgore, “Higgs boson production in bottom quark fusion at next-to-next-to leading order”, *Phys. Rev. D* **68** (2003) 013001, doi:10.1103/PhysRevD.68.013001, arXiv:hep-ph/0304035.

[88] S. Forte, D. Napoletano, and M. Ubiali, “Higgs production in bottom-quark fusion in a matched scheme”, *Phys. Lett. B* **751** (2015) 331, doi:10.1016/j.physletb.2015.10.051, arXiv:1508.01529.

[89] S. Forte, D. Napoletano, and M. Ubiali, “Higgs production in bottom-quark fusion: matching beyond leading order”, *Phys. Lett. B* **763** (2016) 190, doi:10.1016/j.physletb.2016.10.040, arXiv:1607.00389.

[90] M. Bonvini, A. S. Papanastasiou, and F. J. Tackmann, “Resummation and matching of b-quark mass effects in $b\bar{b}H$ production”, *JHEP* **11** (2015) 196, doi:10.1007/JHEP11(2015)196, arXiv:1508.03288.

[91] M. Bonvini, A. S. Papanastasiou, and F. J. Tackmann, “Matched predictions for the $b\bar{b}H$ cross section at the 13 TeV LHC”, *JHEP* **10** (2016) 053, doi:10.1007/JHEP10(2016)053, arXiv:1605.01733.

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

[93] **CMS** Collaboration, "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **15** (2020), no. 10, P10017, doi:10.1088/1748-0221/15/10/P10017, arXiv:2006.10165.

[94] **CMS** Collaboration, "Measurement of differential cross sections for the production of top quark pairs and of additional jets in lepton+jets events from pp collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. D* **97** (2018), no. 11, 112003, doi:10.1103/PhysRevD.97.112003, arXiv:1803.08856.

[95] **CMS** Collaboration, "Measurement of normalized differential $t\bar{t}$ cross sections in the dilepton channel from pp collisions at $\sqrt{s} = 13$ TeV", *JHEP* **04** (2018) 060, doi:10.1007/JHEP04(2018)060, arXiv:1708.07638.

[96] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016. arXiv:cs/1603.02754. doi:10.1145/2939672.2939785.

[97] L. Bianchini, J. Conway, E. K. Friis, and C. Veelken, "Reconstruction of the Higgs mass in $H \to \tau\tau$ Events by Dynamical Likelihood techniques", *J. Phys. Conf. Ser.* **513** (2014) 022035, doi:10.1088/1742-6596/513/2/022035.

[98] L. Bianchini et al., "Reconstruction of the Higgs mass in events with Higgs bosons decaying into a pair of τ leptons using matrix element techniques", *Nucl. Instrum. Meth. A* **862** (2017) 54, doi:10.1016/j.nima.2017.05.001, arXiv:1603.05910.

[99] **CDF** Collaboration, "Search for neutral MSSM Higgs bosons decaying to tau pairs in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV", *Phys. Rev. Lett.* **96** (2006) 011802, doi:10.1103/PhysRevLett.96.011802, arXiv:hep-ex/0508051.

[100] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *Eur. Phys. J. C* **71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727. [Erratum: Eur.Phys.J.C 73, 2501 (2013)].

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

[102] ATLAS and CMS Collaborations and LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011", Technical Report CMS-NOTE-2011-005, ATL-PHYS-PUB-2011-011, CERN, 8, 2011.

[103] **CMS** Collaboration, "Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS", *Eur. Phys. J. C* **81** (2021) 800, doi:10.1140/epjc/s10052-021-09538-2, arXiv:2104.01927.

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

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

[106] **LHC Higgs Cross Section Working** Group, T. Rei, "κ coupling modifiers (update in 2017 for CERN Yellow Report 4)", `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWG2KAPPA`, 2017. (Retrieved 26 June 2024).

[107] **CMS** Collaboration, "Measurements of Higgs boson production in the decay channel with a pair of τ leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Eur. Phys. J. C* **83** (2023), no. 7, 562, `doi:10.1140/epjc/s10052-023-11452-8`, `arXiv:2204.12957`.

[108] **CMS** Collaboration, "Search for the SM Higgs boson produced in association with bottom quarks in final states with leptons", CMS Physics Analysis Summary CMS-PAS-HIG-23-003, CERN, Geneva, 2024.

[109] **CMS** Collaboration, "Search for a third-generation leptoquark coupled to a τ lepton and a b quark through single, pair, and nonresonant production in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **05** (2024) 311, `doi:10.1007/JHEP05(2024)311`, `arXiv:2308.07826`.

[110] **CMS** Collaboration, "Measurement of the cross section for $\mathrm{t\bar{t}}$ production with additional jets and b jets in pp collisions at $\sqrt{s} = 13$ TeV", *JHEP* **07** (2020) 125, `doi:10.1007/JHEP07(2020)125`, `arXiv:2003.06467`.

[111] **CMS** Collaboration, "Measurement of the $\mathrm{t\bar{t}b\bar{b}}$ production cross section in the all-jet final state in pp collisions at $\sqrt{s} = 13$ TeV", *Phys. Lett. B* **803** (2020) 135285, `doi:10.1016/j.physletb.2020.135285`, `arXiv:1909.05306`.

[112] **ATLAS** Collaboration, "Measurements of inclusive and differential fiducial 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", *JHEP* **04** (2019) 046, `doi:10.1007/JHEP04(2019)046`, `arXiv:1811.12113`.

[113] **CMS** Collaboration, "Inclusive and differential cross section measurements of $\mathrm{t\bar{t}b\bar{b}}$ production in the lepton+jets channel at $\sqrt{s} = 13$ TeV", *JHEP* **05** (2024) 042, `doi:10.1007/JHEP05(2024)042`, `arXiv:2309.14442`.

[114] V. Mikuni and F. Canelli, "ABCNet: An attention-based method for particle tagging", *Eur. Phys. J. Plus* **135** (2020), no. 6, 463, `doi:10.1140/epjp/s13360-020-00497-3`, `arXiv:2001.05311`.

[115] **CMS** Collaboration, "The Phase-2 Upgrade of the CMS Tracker", Technical Report CERN-LHCC-2017-009, CMS-TDR-014, CERN, Geneva, 2017. `doi:10.17181/CERN.QZ28.FLHW`.

[116] **Tracker Group** Collaboration, "Evaluation of HPK n^{+}-p planar pixel sensors for the CMS Phase-2 upgrade", *Nucl. Instrum. Meth. A* **1053** (2023) 168326, `doi:10.1016/j.nima.2023.168326`, `arXiv:2212.04793`.

[117] M. Meschini et al., "Radiation resistant innovative 3D pixel sensors for the CMS upgrade at the High Luminosity LHC", *Nucl. Instrum. Meth. A* **978** (2020) 164429, `doi:10.1016/j.nima.2020.164429`.

[118] **RD53** Collaboration, "The RD53A Integrated Circuit", Technical Report CERN-RD53-PUB-17-001, CERN, Geneva, 2017.

[119] **RD53** Collaboration, “RD53B users guide”, Technical Report CERN-RD53-PUB-21-001, CERN, Geneva, 2020.

[120] R. Diener et al., “The DESY II Test Beam Facility”, *Nucl. Instrum. Meth. A* **922** (2019) 265, doi:10.1016/j.nima.2018.11.133, arXiv:1807.09328.

[121] A. Jofrehei et al., “Investigation of crosstalk effects in RD53A modules with 100 and 150 μm thick n-in-p planar sensors”, *PoS* **EPS-HEP2019** (2020) 151, doi:10.22323/1.364.0151.

[122] **CMS** Collaboration, “Characterization of irradiated RD53A pixel modules with passive CMOS sensors”, *JINST* **17** (2022), no. 09, C09004, doi:10.1088/1748-0221/17/09/C09004, arXiv:2203.11376.