Search for an additional neutral MSSM Higgs boson decaying to tau leptons with the CMS experiment

Gaël Touquet

N°d'ordre NNT : 2019LYSE1343

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

operée au sein de
L'Université Claude Bernard Lyon 1

Ecole Doctorale N°52
Ecole Doctorale de Physique et Astrophysique

Specialité du doctorat : Physique des particules

Soutenue publiquement le 19/12/2019 par :

Gael Touquet

Recherche d'un boson de Higgs supplémentaire du MSSM se désintégrant en deux leptons tau avec l'expérience CMS

---

devant le jury compose de :

Mme.	Collard	Caroline	Directrice de recherche	CNRS	Rapporteure
M.	Lafaye	Rémi	Directeur de recherche	CNRS	Rapporteur
Mme.	Petit	Elisabeth	Chargee de recherche	CNRS	Examinatrice
M.	Tsimpis	Dimitrios	Professeur des universités	UCBL	Examinateur
M.	Bernet	Colin	Charge de recherche	UCBL	Directeur de these

Résumé

Malgrès le fait que toutes les observations expérimentales faites dans des collisionneurs de particules sont théoriquement expliquées par le modèle standard, celui-ci n'explique pas des phénomènes naturels tels que la gravité, la matière noire, l'énergie noire, l'asymétrie matière-antimatière, etc. La super-symmetrie est une extension du modèle standard qui peut proposer des solutions à certaines de ces lacunes. La version minimale de cette extension, appelée MSSM, est un modèle a deux doublets de Higgs, ce qui propose l'existence de cinq bosons de Higgs observables. Trois de ces bosons de Higgs sont neutre, et le boson de Higgs découvert en 2012 pourrait être un d'entre eux. Le premier chapitre introduit les contextes théoriques que sont le modèle standard et l'extension super-symmetrique minimal du modèle standard.

Cette thèse présente l'analyse des données de collision proton-proton enregistrées par le détecteur CMS en 2017 pendant la deuxième période de prise de données du LHC au CERN, ce qui correspond a une luminosité intégrée de $41.5\,\mathrm{fb}^{-1}$. Cette analyse a pour but la recherche d'un boson de Higgs lourd et neutre du MSSM, qui se désintégrerai en une pair de leptons tau qui par la suite se désintégrerai eux-mêmes hadroniquement (τ_h). Le deuxième chapitre décrit le contexte expérimental, c'est-à-dire l'éxperience CMS du LHC.

Une des difficultés de l'étude de cet état final est liée au grand nombre de jets produit par QCD dans les collisions pp, car les jets peuvent facilement être mal identifies comme des τ_h. Pour mitiger cet effet, une nouvelle technique d'identification de τ_h basée sur un réseau de neurone profond récursif est présentée dans le troisième chapitre. Cette technique est ensuite comparée à la méthode d'identification standard utilisée à CMS, dans ce même chapitre.

Un boson de Higgs supplémentaire se manifesterai en tant qu'excès d'événements dans les distribution de la variable m_T^tot dans le canal $\tau_\mathrm{h}\tau_\mathrm{h}$. Dans le quatrième et dernier chapitre, des limites supérieurs dans l'intervalle de confiance a 95% pour le scenario $m_\mathrm{h}^\mathrm{max}$ sont déterminées dans l'espace des paramètres ${}_a - \tan\beta$. De plus, des limites sur le produit du rapport d'embranchement et de la section efficace de production sont déterminés pour la production par fusion de gluons et pour la production en association avec un quark b, pour des hypothèses de masses dans l'intervalle de 110 à 2900 GeV.

Acknowledgements

Plus que jamais, la science s'appuie aujourd'hui sur la collaboration. Je voudrais donc tout d'abord remercier tous les chercheurs de la collaboration CMS, ainsi que toutes les personnes y travaillant de près ou de loin. Je remercie aussi évidemment tous ceux qui m'ont prêté main forte pendant ces trois années, Colin bien sûr, qui m'a tellement appris, mais aussi Lucas, Ece, Yohann et Samuel, avec qui j'ai énormément apprécié travailler. J'ai aussi apprécié côtoyer l'ensemble des chercheurs de l'équipe CMS de Lyon, et tout particulièrement Viola, Stéphane et Nicolas.

Il n'y a rien de plus formateur que de former, et je me dois donc de remercier Lucas, Aurélien, Yohann ainsi que tous les élèves que j'ai eus pendant mon monitorat. J'espère leur avoir montré autant de patience qu'ils m'en ont montré, et j'espère leur avoir apporté autant qu'ils m'ont apporté. Un grand merci à Colin de non seulement m'avoir formé mais aussi fourni les conditions pour aller au bout de cette thèse! Mention spéciale aux courgettes du jardin!

Je ne pourrais jamais assez remercier mes confrères doctorants qui, pendant trois longues années, ont supporté mes incessantes envies de débats, entremêlées des habituelles complaintes, lors de leur pause déjeuner. Merci à Albert, Clémentine, Mehdi, Gaétan, Justine, Mathilde et Luis.

Un remerciement spécial à mon doppelganger, Hugues. Nos goûts en commun, et nos mauvaises fois respectives, nous ont permis d'aller jusqu'au bout de nos parcours parallèles et de cette épreuve avec le recul nécessaire, et toujours avec le sourire. J'espère sincèrement avoir la chance de travailler à nouveau avec toi, insupportable et extraordinaire rhétoricien. Une chose est sûre, tu vas me manquer!

J'en profite pour remercier tous les amis qui m'ont épaulés au long de cette aventure, mais aussi le long du chemin qui m'y a mené! Merci Nico et Arina, merci Poulet, merci PJ et Audrey, merci Albert, merci Jean-yves, merci Manu, merci Babiol, merci Emilie, merci Badis, merci Matthieu, merci aussi à tous ceux que je n'ai pas mis ici! Je vous aime!

Evidemment, merci à toute ma famille, vous qui restez le centre de mon monde et qui m'avez toujours poussé à être un scientifique et un homme accompli. Merci Maman, Papa, Raph, Sophie, Thibault et Martine. Merci aussi à Steve, Val, Paul, Helen, Alistair et Jane qui me donnent la chance d'avoir une famille en plus, et qui ont supporté des vacances avec un Gaël très stressé.

Enfin, merci Emma, toi qui m'a supporté pendant mes années les plus difficiles, qui a partagé mon stress à chaque instant, qui m'a remotivé un nombre incalculable de fois, et sans qui je n'aurai non seulement pas accompli ce travail, mais aussi sans qui je ne serai pas la moitié de l'homme que je suis.

Il ne me reste plus qu'à souhaiter un bon courage à ceux qui liront la suite!

Contents

List of Figures

List of Tables

Introduction

Physics is the branch of science that intends to study Nature at its most fundamental level. As the properties of any object in Nature derive from its constituents, the most fundamental understanding of our universe relies on understanding the basic constituents of matter and its interactions. The theoretical and experimental efforts of generations of scientists have allowed us to zoom into the structure of matter with more and more precision, and to find new fascinating structures at every level. We now understand that chemistry is driven by the properties of atoms, and that each atom is comprised of electrons and an atomic nucleus, which is in turn made up of protons and neutrons, that are themselves made of quarks and gluons. These objects are, for now, considered to be the fundamental constituents of matter.

Particle physics is the branch of physics that specialises in the description of these fundamental constituents and their interactions. At this level, quantum mechanics dictate the rules and properties of every object and interaction. In this framework, small distances correspond to large energy scales, so studying the smallest constituents of nature requires a theoretical understanding of high-energy processes and advanced technology to produce them in a laboratory. Thus the field of particle physics is also often called high-energy physics.

Indeed, to understand the properties and the laws governing the particles of Nature, any theory must be confronted with observations. Experiments have therefore been designed to test the properties of theories such as the standard model (SM) of particle physics, a unified quantum field theory which was formulated in the 1960s and early 1970s. This theory is one of the most rigorous and precise ever created, and it has passed innumerable experimental tests over the past decades. For instance, the SM predicted the existence of several elementary particles, such as the $W^{\pm}$ and Z bosons, gluons and the top quark, before they were experimentally discovered.

The last missing piece of the SM pending experimental confirmation was the existence of a Higgs boson. In the SM, elementary particles gain their masses by interacting with a field known as the Higgs field, manifesting itself as Higgs bosons. The Brout-Englert-Higgs mechanism that makes this possible via a spontaneous breaking of electroweak symmetry was predicted by three independent groups in 1964.

To test for the existence of this last piece of the SM, the Large Hadron Collider (LHC) was built. There, massive elementary particles such as these potential Higgs bosons can be produced in energetic particle collisions, converting the energy of the colliding particles, which is mostly kinetic energy, into mass. In 2012, after a few years of collision data gathering, the ATLAS and CMS experiments at the LHC discovered a new particle with a mass of approximately 125 GeV, which was later

confirmed to be a Higgs boson. The discovery completed the era of experimental searches for new particles guided by the SM.

While the SM is one of the most successful theories developed this far, it suffers from both experimental and theoretical shortcomings. These shortcomings suggest that the SM is not a complete description of nature, but rather a low-energy approximation of a more general theory. Many candidates for this beyond the Standard Model (BSM) theory have been proposed. The minimally supersymmetric extension of the standard model is one such BSM model that adds a new symmetry on top of the existing ones in the SM, leading to the prediction of extra particles. This model, along with most of the BSM theories, predicts an extended Higgs sector, with a spectrum of Higgs bosons with different masses, charges, and other properties. All these models are constrained, but not excluded, by the measured properties of the 125 GeV boson. Two-Higgs-doublet models (2HDMs) predict five different Higgs bosons: two neutral CP-even particles h and H, one neutral CP-odd particle A, and two charged Higgs bosons $H^{\pm}$. The observation of additional Higgs bosons would provide direct evidence for the existence of BSM physics, and could push the searches towards other predicted particles of the MSSM.

In this thesis, a theoretical context focusing on the SM and the MSSM is given in chapter 1. Chapter 2 provides the experimental context, namely the CMS experiment of the LHC. Chapter 3 details a new approach to hadronic tau identification, whose goal is to improve sensitivity in the search for extra heavy neutral Higgs bosons, detailed in chapter 4. This search is performed, based on proton-proton collisions provided by the LHC at a center-of-mass energy of 13 TeV and collected by the CMS experiment in 2017. The amount of data corresponds to an integrated luminosity of 41.5 fb^{-1}.

1

Theoretical context

The theoretical context of this research is presented in the following chapter. The Standard Model introduced in section 1.1, is the widely accepted basis of particle physics understanding, and its Higgs sector will be focused on in section 1.1.5. But the Standard Model as it is defined now has several limitations, which will be discussed in section 1.1.6. One of the promising extensions of the Standard Model, called the Minimal Supersymmetric extension of the Standard Model (MSSM) will then be described in section 1.2. Finally the phenomenology of the detection of the Higgs sector of the MSSM in proton-proton collisions will be presented in section 1.3.

1.1 Overview of the Standard Model

The Standard Model (SM) is the prevalent theoretical framework that describes fundamental elements, named particles, and their interactions. As a quantum field theory, this framework considers particles as excited states of the vacuum created by the fundamental quantum fields. The interactions between these particles can be assimilated to the classical forces of nature known as electromagnetic, weak and strong forces. Gravity is not included in the SM as it hasn't been successfully described by a quantised interaction that complies with general relativity. The SM also lacks explanation for several phenomena of nature, and is therefore considered incomplete. However, its validity is not questioned as it has provided confirmed predictions and has not yet been successfully contradicted by a particle physics experiment. An overview of the Standard Model content in terms of particles and their characteristics is shown in Figure 1.1.

Mathematically, the SM is described by a non-abelian gauge quantum field theory. A gauge theory is a type of field theory whose equations of motion are invariant under a continuous group of local transformations of the fields, which can be written $\psi \rightarrow \psi' = e^{i\lambda(x)}\psi$. The SM contains internal symmetries of the Lie's algebra unitary product group $SU(3)_C \times SU(2)_L \times U(1)_Y$, corresponding to the strong and electroweak forces. However, the electroweak symmetry is broken at low energies, and therefore the internal symmetries effectively become $SU(3)_C \times U(1)_Q$ in which the first term represents the strong and the second the electromagnetic symmetry.

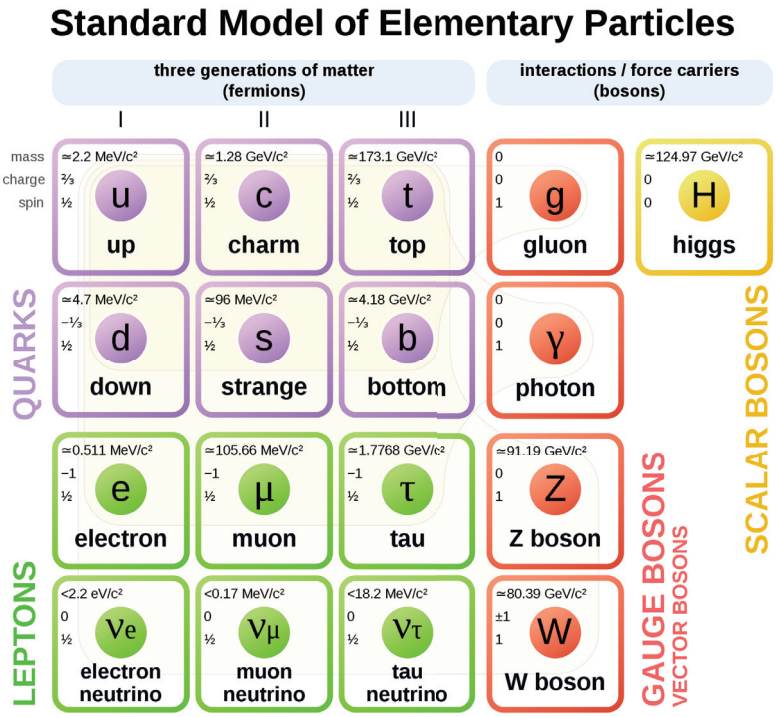


Figure 1.1: Overview of the Standard Model content in terms of particles and their characteristics from [1].

The SM is conventionally expressed using the lagrangian formalism. From this formalism, the principle of least action is equivalent to [3]

$$\delta\mathcal{S} = \int \mathrm{L}dt = \int \mathcal{L}(\phi, \partial_\mu\phi)d^4x = 0\,. \tag{1.1}$$

In this equation, ϕ represents a field and ∂_μ represent the space and time derivatives using Einstein's notation. The integration over all the space of the lagrangian density $\mathcal{L}$ in equation 1.1 is by convention usually implicit and the formulation of future lagrangian terms will be given in terms of lagrangian density.

Observable particles are considered to be excited states of the fundamental fields. Different types of fields, and therefore of particles, can be differentiated in the lagrangian terms by one defining property: the spin. Historically, this property allows particles to react to the presence of a magnetic field, in a sort of intrinsic angular momentum. Indeed the spin-statistics theorem allows to classify particles into two types from their spin value: the fermions, with half-integer valued spins, and the bosons with integer valued spins. Fermions are effectively the constituents of matter while bosons are the carriers of the forces.

1.1.1 Fermions

Fermions (ψ) are the fundamental particles that compose matter. They are described mathematically by the Fermi-Dirac statistics, which means that their spin

has a half-integer value and they therefore obey the Pauli exclusion principle and the canonical anti-commutation relations.

In the following equations, the convention $c = \hbar = 1$ is used, and will be used throughout this thesis. The notation $\not{\partial} = \gamma^\mu \partial_\mu$ is also used and the γ^μ matrices are defined in appendix A.

In order to keep the Lorentz invariance, a hermitian conjugate of the fermionic field has to be defined so that $\bar{\psi} \equiv \psi^\dagger \gamma^0$, which will then represent the antimatter particles. Antimatter particles, or antiparticles, have the same characteristics as their associated particles but opposite quantum numbers.

The fermionic term of the lagrangian can be written, for all fermionic fields:

$$\mathcal{L}_f = i\bar{\psi}\not{\partial}\psi - m\bar{\psi}\psi\,. \tag{1.2}$$

The fermionic field therefore takes the form of a plane wave

$$\psi(x) = \int \frac{d^3p}{(2\pi)^3} \frac{1}{\sqrt{2E_p}} \sum_s^2 \left(a_P^s u_s(p) e^{-ipx} + b_P^{s\dagger} v_s(p) e^{ipx} \right). \tag{1.3}$$

In this equation, $u_s(p)$ and $v_s(p)$ are spinors with momentum p and spin s; a_P^s and $b_P^{s\dagger}$ are the creation and annihilation operators respectively, which act as a base for the Fourier transform of the field. The creation operator raises the excitation of the state, creating new particles. On the other hand the annihilation operator de-excites the state, effectively lowering the number of particles.

The terms in equation 1.2 can also be rewritten so as to separate the fermion terms in their chiral components, as it is useful in cases where interactions are sensitive to chirality. The chiral operator $\gamma^5 \equiv i\gamma^0\gamma^1\gamma^2\gamma^3$ is used to write

$$\psi = \frac{1-\gamma^5}{2}\psi + \frac{1+\gamma^5}{2}\psi = P_L\psi + P_R\psi = \psi_L + \psi_R\,. \tag{1.4}$$

The orthogonality of the components leads to the mass term of equation 1.2 to become

$$\bar{\psi}\psi = (\bar{\psi}_R + \bar{\psi}_L)(\psi_L + \psi_R) = \bar{\psi}_R\psi_L + \bar{\psi}_L\psi_R\,. \tag{1.5}$$

The fermions of the SM can be split into two categories, namely quarks and leptons. The main difference is the fact that quarks carry colour charges, making them susceptible to interact through the strong force, while leptons do not. Also, leptons carry integer electric charges whereas quarks carry electric charges of either 1/3 or 2/3.

Both quarks and leptons can be classified into three generations depending on their mass. The first generation contains the most commonly encountered particles in our world: u and d quarks, the electron, and the electron neutrino. The two other generations contain particles with the same characteristics except their masses are higher, as illustrated in Figure 1.1. So far only three generations are known and the possibility of other generations is constrained by the Z boson decay [4], although

the validity of this affirmation depends on the mass and couplings of the neutrinos with the Z boson.

Quarks

Quarks (ψ_q) are by definition fermions carrying a colour charge. Colour is the strong force charge, and has three distinct types, namely red, green or blue. Anti-quarks have opposite quantum numbers, therefore they carry a charge of anti-red, anti-green or anti-blue. In each of the three generations, there are two types of quarks that can be described together as a weak isospin doublet. This leads to a total of 3 doublets of up-type and down-type quarks. The up-type quarks are the up (u), charm (c) and top (t) and have an electric charge of $+\frac{2}{3}$. Their weak isospin partners, the down-type quarks are named down (d), strange (s) and bottom (b) and have an electric charge of $-\frac{1}{3}$.

Since they carry a colour charge and are therefore subjected to the strong interaction, quarks are subjected to confinement. Confinement is a consequence of the nature of the strong interaction, which will be detailed in section 1.1.3. Indeed, in nature quarks are only observed when combined into composite particles of neutral colour charge (also referred to as white), which are called hadrons. Hadrons are made of different combinations of quarks, the simplest of which are mesons and baryons. Mesons are hadrons composed of a quark and an anti-quark with opposite colours (e.g. red and anti-red). Baryons are composed of three quarks of different colours, making them colour-neutral as a whole. Mesons have integer spin value, meaning they will behave effectively as bosons, whereas baryons have a half-integer value, behaving effectively as fermions. Ordinary matter nuclei are composed of proton and neutron, which are baryons of composition (uud) and (udd) respectively.

Leptons

Leptons (ψ_l) are fermions but contrary to quarks, they do not carry a colour charge, and therefore are not sensitive to the strong force. In a similar way to the quarks being ordered in three generations of weak isospin doublets, leptons can be split into doublets of weak isospin from an electron-like lepton and a neutrino. The electron-like group is composed of the electron (e), the muon (μ) and the tau lepton (τ), each of which has an electric charge of -1 ($+1$ for their antiparticles), thus they interact through the weak and electromagnetic forces. On the other hand the neutrino group is composed of electrically neutral particles, meaning they only interact through the weak force. Neutrinos are named from their weak isospin partners: electron neutrino (ν_e), muon neutrino (ν_μ) and tau neutrino (ν_τ). They have been proven to be massive, but their mass is so low that it is considered negligeable in accelerator experiments. In addition, they only interact through the weak force, making neutrinos elusive particles and hard to detect in practice.

1.1.2 Bosons

Equation 1.2 transforms under a phase rotation as

$$\psi \to \psi' = e^{i\lambda}\psi \tag{1.6}$$

while

$$\partial_\mu \psi \to e^{i\lambda}\partial_\mu \psi \tag{1.7}$$

and

$$\bar{\psi} = \psi^\dagger \gamma^0 \to e^{-i\lambda}\bar{\psi}\,, \tag{1.8}$$

which leads to

$$\mathcal{L} \to \mathcal{L}' = \mathcal{L}\,. \tag{1.9}$$

This Lagrangian is therefore said to conserve gauge invariance. The Noether theorem states that every differentiablE symmetry implies a conserved quantity. In the case of a simple electromagnetic interaction (QED), the electronic current j^μ must respect

$$\partial_\mu j^\mu = \partial_\mu(-e\bar{\psi}\gamma^\mu\psi) = 0\,. \tag{1.10}$$

The current j^μ is thus conserved, and so is the associated charge $Q = \int j^0 d^3x$.

For a local transformation, and introducing the charge operator Q, the fermionic fields transform as

$$\psi \to \psi' = e^{i\lambda(x)Q}\psi\,. \tag{1.11}$$

However, the standard derivatives, and hence the lagrangian, do not conserve the invariance for these rotations since

$$\partial_\mu \psi \to \partial_\mu \psi' = e^{iQ\lambda(x)}\partial_\mu\psi + iQe^{i\lambda(x)}\psi\partial_\mu\lambda(x)\,. \tag{1.12}$$

To recover gauge invariance, the covariant derivative must be introduced:

$$D_\mu \psi \to e^{i\lambda(x)Q} D_\mu \psi\,. \tag{1.13}$$

To be able to write such a derivative, the field A_μ must be introduced so that

$$D_\mu = \partial_\mu + ieQA_\mu \tag{1.14}$$

where A_μ transforms as

$$A_\mu \to A_\mu - \frac{1}{e}\partial_\mu\lambda(x)\,. \tag{1.15}$$

The A_μ field can be associated with the photon in the case of electromagnetism.

While electromagnetism is an abelian gauge theory based on the symmetry group $U(1)$, the other forces are based on more complex symmetries. Indeed, the Yang-Mills theory [5] is a non-abelian gauge field theory based on the internal continuous symmetries, the special unitary group $SU(N)$, of the Lie algebra. The $SU(N)$ is a symmetry group of $N \times N$ unitary matrices with determinant 1 and $N^2 - 1$ generators. The generators are then the different bosons associated with each field. More generally, bosons are particles associated with an excited state of the fundamental

forces in the SM. They are effectively considered as carriers or mediators for these fundamental forces. The forces in question are the strong interaction, responsible for nuclear cohesion, and the electroweak force (EW), which actually breaks down into two different interactions at low energies: the electromagnetic and the weak interactions. The strong and EW interaction specificities will be detailed in sections 1.1.3 and 1.1.4 respectively.

Mathematically, bosons have to obey the canonical commutation relations and they follow the Bose-Einstein statistics, which means their spin value is of integer value. The bosonic fields contribute to several terms in the lagrangian corresponding to the free propagation, the self-interaction and the interaction with other fields. The free propagation and the self-interaction terms are usually combined into a kinematic term

$$\mathcal{L}_{kin} = -\frac{1}{4} F^a_{\mu\nu} F_a^{\mu\nu} \,, \tag{1.16}$$

where

$$F^a_{\mu\nu} \equiv \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g f^a_{bc} A^b_\mu A^c_\nu \,. \tag{1.17}$$

In this expression, g is the coupling constant, a parameter associated with the relative strength of the force, and f^a_{bc} is the structure constant. This structure constant is defined in a Lie algebra from the commutator of two generators of the group as

$$\left[T_a, T_b\right] = i f^a_{bc} T_c \,. \tag{1.18}$$

1.1.3 Quantum Chromodynamics

In the Standard Model, the quantum field theory associated with the strong force is called Quantum Chromodynamics (QCD), as the conserved charge linked with this interaction is named colour. Colour was theorised from the discovery in 1951 of the Δ^{++} baryon. Indeed this baryon is composed of three up quarks (uuu). This baryon's spin is $\frac{3}{2}$ and up quarks are fermions of spin $\frac{1}{2}$. This would therefore lead to the conclusion that these up quarks are in the exact same quantum state, contradicting Pauli's exclusion principle. However, considering this new charge, all three quarks would each carry a different colour charge, avoiding Pauli's exclusion principle. Mesons and baryons are then colour-neutral, which has been confirmed by observations.

QCD is a non-abelian Yang-Mills gauge theory that corresponds to the $SU(3)_C$ sector of the SM. The bosonic field G^a_μ is called the gluon field. To express the free term of the interaction, the strength tensor field $G^a_{\mu\nu}$ is defined as

$$G^a_{\mu\nu} \equiv \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_s f^a_{bc} G^b_\mu G^c_\nu \,. \tag{1.19}$$

In this expression, the index a refers to the colour charge, g_s is the strong coupling constant, and f^a_{bc} is the result of the commutator of the generators of the group as described in equation 1.18. In the SU(3) group symmetry, there are eight different generators, meaning eight gluons, that mathematically take the form of the Gell-Mann matrices divided by two, as detailed in appendix A.

The colour-triplet of fermions and its derivative transform under the SU(3) group as

$$\partial_\mu \psi_Q \to \partial_\mu \psi'_Q = \partial_\mu \psi_Q + i g_s T_a G^a_\mu \tag{1.20}$$

leading to the covariant derivative

$$\partial_\mu \to D_\mu \equiv \partial_\mu - i g_s T_a G^a_\mu \,. \tag{1.21}$$

Finally, combining the kinematic term for both quarks and gluons to the interaction term of QCD leads to:

$$\mathcal{L}_{QCD} = \bar{\psi_Q}(i \not{D}_\mu - m)\psi_Q - \frac{1}{4} G^a_{\mu\nu} G^{\mu\nu a} \,. \tag{1.22}$$

Important specific properties of the QCD interaction are asymptotic freedom and confinement. Asymptotic freedom is the fact that QCD interaction grows weaker as the energy increases or distance decreases. This makes the quarks and gluons inside hadrons effectively free particles for short range interactions. On the other hand, confinement is a postulate of QCD driven by observations, stating that colour-charged particles cannot exist isolated in nature. In effect only colour-neutral composite particles can be stable, as separating them would lead the QCD potential energy to rapidly grow and create pairs of quark-antiquark. The newly created quarks will then start associating and creating new pairs of quarks until only colour-neutral hadrons are left. This process happens when quarks are created in high energy collisions such as at the LHC, and is called hadronization.

1.1.4 The electroweak theory

At low energies, two additional forces are known: the electromagnetic and weak interactions. The electromagnetic force is an infinite-range interaction responsible for the cohesion of atoms, whereas the weak force is historically the interaction responsible for various nuclear decays. In the early developments of particle physics, both forces were described as independent interactions, as these forces manifest themselves in very different ways. At the time the weak force was known as Fermi's interaction [6], a mechanism describing the interaction of four fermions as a point-like process, as shown in Figure 1.2a.

The electromagnetic force was successfully quantised by Schwinger, Tomonaga, Feynman and Dyso, leading to the birth of Quantum Electrodynamics (QED) [7–12]. It was described as an interaction of two fermions with the mediation of a boson, the photon (γ). QED was gauge invariant and overall renormalizable. These features pushed the physics community to consider gauge symmetries as fundamental symmetries of nature, and to try to formalise the other forces in the context of a gauge theory. However, these attempts failed due to the weak interaction's short range and the fact that it was not renormalizable.

To overcome this, Schwinger proposed a description of the weak interaction which is analogue to that of the electromagnetic force [13]. The weak force would not be a short-range interaction of four fermions but a long-range interaction of two fermions mediated by a charged vector boson, called $W^\pm$ as shown in Fig. 1.2b. The similarities between QED and this weak force, and the root of the gauge invariance,

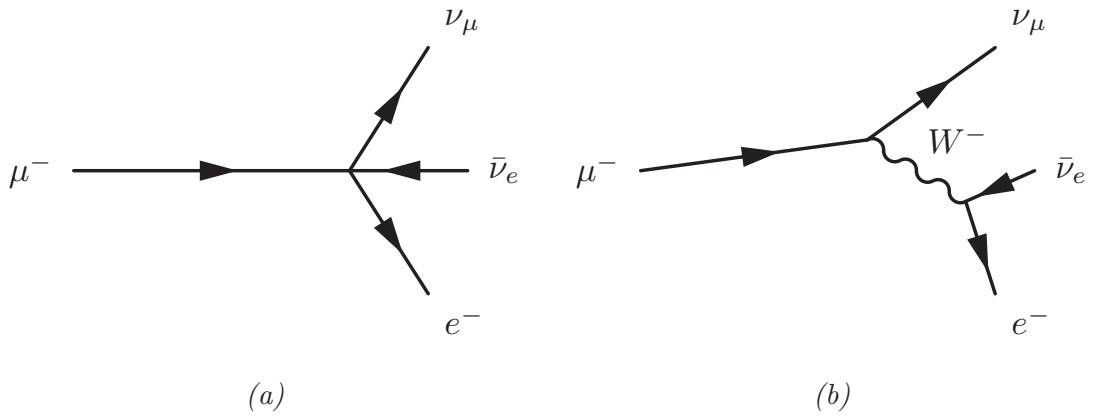


Figure 1.2: Feynman diagrams for the muon decay as (a) a Fermi's interaction, without a propagator and (b) a Yang-Mills theory mediated by a W^- *boson.*

led to the hope of unification of the two forces. Many attempts were done in the following years with little success, though small advances were achieved. One such advance was parity (P-symmetry) violation, meaning a behavioural difference of chiral components. This could be achieved by a Vector-Axial theory (V-A theory) [14, 15], in which QED would play the vectorial role.

In 1961, Sheldon Glashow proposed a $SU(2) \times U(1)$ Yang-Mills theory [16] to unify both interactions. The proposal defined two fields, B_μ and W^a_μ, with four fundamental bosons, B and W^i. The observable features, like the vector bosons $W^\pm$ and Z, were actually a linear combination of both fields with a mixing angle called θ_W. But the model had several flaws and was ignored at the time.

First, it predicted a new phenomenon for which there was no evidence at the time: neutral flavour-conserving currents, mediated by a neutral massive vector boson different from the photon. Secondly, its renormalizability was unclear. Lastly, the experimental results only showed evidence that the interaction was extremely short-ranged, pushing towards the existence of a massive, charged weak boson. But chiral Yang-Mills theories, being gauge invariant, did not allow the addition of a mass term.

The solution finally came from the field of condensed matter. Work on superconductivity led to the establishment of the Goldstone theorem [17], which states that scalar bosons, named Goldstone bosons, arise from the breaking of global continuous symmetries. However, the predicted bosons were not always observed. The models trying to employ the spontaneous symmetry breaking (SSB) encountered the same problem.

Philip Anderson pointed out in 1963 [18] that in a degenerated state within a gauge potential, the massless Goldstone bosons could combine with the massless propagators of the gauge field to become massive bosons. The Goldstone bosons would therefore not appear as observables and the massless vector bosons of the theory would acquire a mass, solving both issues. Indeed in superconductivity this phenomenon appears when photons interact with the electromagnetic potential in a superconducting electron gas: massless photons become massive plasmons, and no Goldstone boson appears.

In the second half of 1964, three groups of physicists independently developed such a mechanism based on SSB to give mass to gauge bosons in the QFT framework by adding a new scalar field: François Englert and Robert Brout in August 1964 [19], Peter Higgs in October 1964 [20, 21] and Gerald Guralnik, Carl Hagen and Tom Kibble in November 1964 [22]. This was the birth of the Brout-Englert-Higgs mechanism, more commonly known as the Higgs mechanism. Peter Higgs was the only one to explicitly remark upon a consequence of the addition of the new scalar field in the form of a new boson that could experimentally be observed to test for the mechanism.

The electroweak interaction has therefore distinct features such as the chiral asymmetry, the fact that electromagnetic and weak forces are distinct at low energies and finally that its gauge bosons are massive. The electroweak theory corresponds to the $SU(2)_L \times U(1)_Y$ sector of the SM and is formulated using two different fields (W^a_μ for $SU(2)_L$ and B_μ for $U(1)_Y$) along with their associated bosons and charges. At low energies, the Higgs field breaks the unification through SSB, leaving the electrodynamics symmetry $U(1)_Q$ of the electromagnetic field (A_μ)

$$\underbrace{SU(2)_L \times U(1)_Y}_{\text{EWT}} \xrightarrow{\text{SSB}} \underbrace{U(1)_Q}_{\text{EM}} . \tag{1.23}$$

The $SU(2)_L$ sector is described by the strength tensor field $W^a_{\mu\nu}$, constructed from the non-abelian field W^a_μ as

$$W^a_{\mu\nu} \equiv \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g_W f^a_{bc} W^b_\mu W^c_\nu . \tag{1.24}$$

In this expression, the parameter g_W corresponds to the coupling constant of the field W^a_μ and the structure constant f^a_{bc} is defined from the commutators of the generators of $SU(2)$, which take the form of the Pauli matrices divided by two as detailed in Appendix A,

$$[\tau_a, \tau_b] = i f^c_{ab} \tau_c . \tag{1.25}$$

The electroweak field is responsible for the chiral asymmetry as it only interacts with left-handed fermions, namely ψ_L, defined in equation 1.5. The associated charge of this field is the third component of the weak isospin (I_3) and the three generators of $SU(2)$ correspond to the fundamental bosons W^i, where $i = 1, 2, 3$. The third term in equation 1.24 represents the self-coupling of the field.

The $U(1)_L$ sector is described by the strength of the tensor field $B_{\mu\nu}$, constructed from the abelian field B_μ like

$$B_{\mu\nu} \equiv \partial_\mu B_\nu - \partial_\nu B_\mu . \tag{1.26}$$

The coupling constant associated with this field is g_B, the interaction is mediated by only one boson named B, and the associated charge is the weak hypercharge (Y), or just hypercharge. It is defined as the combination of the electromagnetic charge (Q) and the weak isospin (I_3) as

$$Y \equiv 2(Q - I_3) . \tag{1.27}$$

The kinematic term of the EWT, which includes the free propagator of both fields and the self-coupling of the non-abelian W_μ^a is

$$\mathcal{L}_{EW,\,kinematic} = -\frac{1}{4} W^a_{\mu\nu} W_a^{\mu\nu} - \frac{1}{4} B^{\mu\nu} B_{\nu\mu}\,. \tag{1.28}$$

The interaction with fermions depends on the chiral properties and the structure of the symmetry group, hence it is different for each field. The values of the charges associated to each field are summarised in table 1.1.

Table 1.1: Values of the electroweak charges (weak isospin I_3, hypercharge Y and electromagnetic charge Q) for the fermions, according to their type and chirality.

Interaction		**Left chirality** ψ_L				**Right chirality** ψ_R			
		Quarks		Leptons		Quarks		Leptons	
Charge	Group	u-type	d-type	ν-type	e-type	u-type	d-type	ν-type	e-type
Weak Isospin (I_3)	$SU(2)_L$	$+1/2$	$-1/2$	$+1/2$	$-1/2$	0	0	0	0
Hypercharge (Y)	$U(1)_Y$	$+1/3$	$+1/3$	-1	-1	$+4/3$	$-2/3$	0	-2
EM Charge (Q)	$U(1)_Q$	$+2/3$	$-1/3$	0	-1	$+2/3$	$-1/3$	0	-1

While the W_μ^a field only interacts with left-handed fermions, the B_μ field interacts with both chiralities indistinctly. The W_μ^a field, associated to $SU(2)_L$, requires that the fermions be organised using a $SU(2)$ doublet of isospin. A doublet is a two-component field, in which the components have opposite weak isospin and same hypercharge. Therefore the doublet transforms as a whole under $U(1)_Y$ but each component is different for $SU(2)$. Considering the chiral requirements as well, meaning only left-handed fermions transform under $SU(2)$, the doublet has to be composed of the left-handed fermions: a u-type and its respective d-type of the same generation. From these conditions, six doublets of left-handed fermions can be defined:

$$L_q \equiv P_L \begin{pmatrix} \psi_u \\ \psi_d \end{pmatrix} = \begin{pmatrix} \psi_u \\ \psi_d \end{pmatrix}_L = \left\{ \begin{pmatrix} u \\ d \end{pmatrix}_L, \begin{pmatrix} c \\ s \end{pmatrix}_L, \begin{pmatrix} t \\ b \end{pmatrix}_L \right\} \tag{1.29}$$

$$L_l \equiv P_L \begin{pmatrix} \psi_\nu \\ \psi_e \end{pmatrix} = \begin{pmatrix} \psi_\nu \\ \psi_e \end{pmatrix}_L = \left\{ \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L, \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L \right\} \tag{1.30}$$

The right-handed fermions (except neutrinos) only transform under $U(1)$ and thus have to be defined using singlets as

$$R \equiv P_R \psi = \psi_R = \left\{ u_R, c_R, t_R, d_R, s_R, b_R, e_R, \mu_R, \tau_R \right\} \tag{1.31}$$

The transformation of their derivatives, under the EW group is

$$\partial_\mu L \to \partial_\mu L' = \partial_\mu L + i g_B Y B_\mu + i g_W \tau_a W_\mu^a \tag{1.32}$$

$$\partial_\mu R \to \partial'_\mu = \partial_\mu R + i g_B Y B_\mu\,, \tag{1.33}$$

and therefore the covariant derivative is:

$$\text{for } L\text{: } \partial_\mu \to D_\mu \equiv \partial_\mu - ig_B Y B_\mu - ig_W \tau_a W_\mu^a \tag{1.34}$$

$$\text{for } R\text{: } \partial_\mu \to D_\mu \equiv \partial_\mu - ig_B Y B_\mu \tag{1.35}$$

Before the symmetry breaking, the lagrangian of the Electroweak interaction is therefore:

$$\mathcal{L}_{EW} = i\bar{L}(\not{\partial}_\mu)L + i\bar{R}(\not{\partial}_\mu)R - \frac{1}{4}W^a_{\mu\nu}W_a^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \tag{1.36}$$

1.1.5 The Higgs sector of the Standard Model

As described in the previous section, the observable gauge bosons, $W^\pm$ and Z, are massive and therefore their masses have to be included in the lagrangian. Since the EW theory is not chiral invariant, a mass term cannot be included explicitly since it would break the gauge invariance of the SM lagrangian. The solution [21] is the addition of a new complex scalar field ϕ, commonly named the Higgs field, as a $SU(2)$ doublet:

$$\phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi^+\\ \phi^0\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_3 + i\phi_4\\ \phi_1 + i\phi_2\end{pmatrix} \tag{1.37}$$

The lagrangian term associated with this field is composed of a potential term created by the field and a kinematic term, which includes the free propagator of the field and the interaction with the weak fields as

$$\mathcal{L}_H = |D_\mu\phi|^2 - V(\phi)\,. \tag{1.38}$$

Since this interaction breaks the gauge invariance of the Higgs derivative, the covariant derivative has to be defined as:

$$\partial_\mu \to D_\mu \equiv \partial_\mu - ig_W\tau_a W_\mu^a - ig_B Y B_\mu\,. \tag{1.39}$$

The potential energy of the field $V(\phi)$ is chosen *ad hoc* to justify the spontaneous symmetry breaking, so it should have a degenerated vacuum state and a local maximum. The simplest form of such a potential is

$$V(\phi) \equiv \lambda(\phi^\dagger\phi)^2 - \mu^2(\phi^\dagger\phi)\,. \tag{1.40}$$

The λ parameter is chosen to be positive, so that $\lim_{\phi\to+\inf} V(\phi) = +\inf$, and in combination with the quadratic term it creates a minima, and the Mexican hat shape, as illustrated on fig 1.3.

The expected value of the field in the vacuum (vev) is therefore the minimum of this potential. For the Higgs field this value is

$$\langle\phi\rangle_0 = \frac{1}{\sqrt{2}}\sqrt{\frac{\mu^2}{\lambda}} \equiv \frac{v}{\sqrt{2}}\,. \tag{1.41}$$

Physically, this means that the vacuum corresponds to a non-zero expectation value for the Higgs field, which causes the spontaneous symmetry breaking, leading to massive weak bosons.

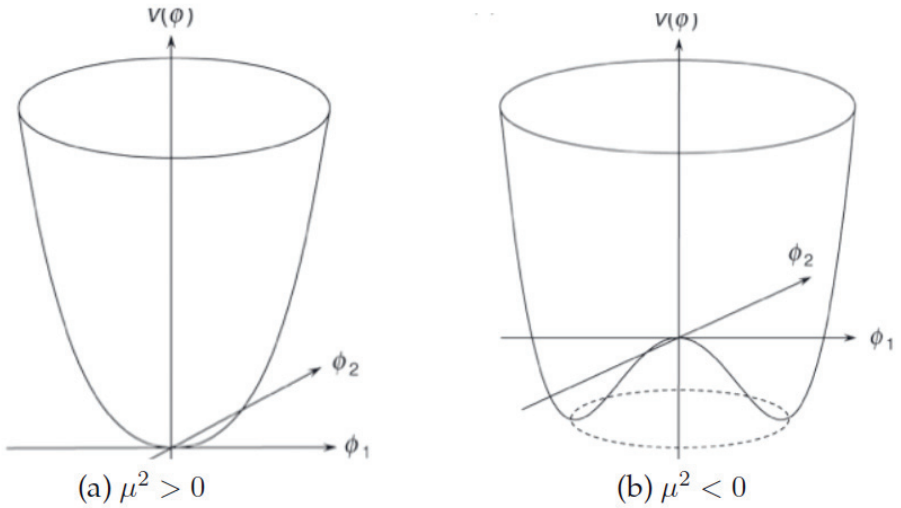


Figure 1.3: Illustration of the form of the Higgs potential depending on the sign of μ^2.

Since the Lagrangian is gauge invariant, the Higgs field can be described from its minimum without loss of generality by applying a gauge transformation, which conserves the number of degrees of freedom. It is fitting to re-express the field using an exponential decomposition:

$$\phi(x) = \frac{1}{\sqrt{2}} e^{e\tau_a \theta^a(x)/f} \begin{pmatrix} 0 \\ \rho(x) \end{pmatrix} \tag{1.42}$$

where $\theta^a(x)$ and $\rho(x)$ are real fields, τ_a corresponds to the generators of $SU(2)$ and f is a unit normalization constant. θ^a contains three of the four degrees of freedom of the Higgs doublet, while ρ conserves the remaining one, as a sort of module of the field.

By expanding the Higgs field in a particular set of coordinates, one can break the vacuum symmetry. To get the physical observables, the expansion will be made around the position of the minimum of the field v. The new real field h is then defined by translation as

$$h(x) \equiv \rho(x) - \langle\phi\rangle_0 = \rho(x) - v\,. \tag{1.43}$$

However this minimum is degenerated, meaning there are an infinite number of points satisfying the condition. The symmetry will be broken by choosing one point and developing the Higgs field around it. The simplest choice is the unitary gauge, in which the degrees of freedom are minimised. In this case it means setting all θ^a to 0, which is analogous to setting $\phi_2 = \phi_3 = \phi_4$. The Higgs field then becomes

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}\,. \tag{1.44}$$

Developing the Higgs lagrangian from equation 1.38 using this Higgs field form gives

$$\mathcal{L}_H = \frac{1}{2}(\partial_\mu h)(\partial^\mu h) + \frac{1}{2}(2\mu^2)h^2 + \frac{1}{2}\big(\frac{g_w^2 v^2}{4}(W_\mu^1 W^{1\mu} + W_\mu^2 W^{2\mu}) \tag{1.45}$$

$$+\frac{1}{8}v^2(g_W W_\mu^3 - g_B B_\mu)(g_W W^{3\mu} - g_B B^\mu) + \mathcal{O}(h^2)\big)\,, \tag{1.46}$$

where $\mathcal{O}(h^2)$ refers to higher orders of the lagrangian, which include the coupling with the vector bosons and the self-coupling of the Higgs boson.

To obtain the physical bosons, the fields have to be rewritten in such a way that their mass terms are independent. For that purpose, new fields A_μ, W_μ and Z_μ are defined as

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp i W_\mu^2)\,, \tag{1.47}$$

$$Z_\mu = \cos(\theta_W) W_\mu^3 - \sin(\theta_{W_\mu}) B_\mu\,, \tag{1.48}$$

$$A_\mu = \cos(\theta_W) B_\mu + \sin(\theta_W) W_\mu^3\,, \tag{1.49}$$

where the parameter θ_W is the Weinberg angle [16] defined from the ratio of the coupling constants

$$\tan\theta_W \equiv \frac{g_B}{g_W}\,. \tag{1.50}$$

The coupling constants are related to the electric charge as:

$$\frac{g_W g_B}{\sqrt{g_W^2 + g_B^2}} \equiv e \tag{1.51}$$

Equation 1.46 then becomes:

$$\mathcal{L}_H = \frac{1}{2}(\partial_\mu h)(\partial^\mu h) + \frac{1}{2}\underbrace{(2\mu^2)}_{m_H^2} h^2 \tag{1.52}$$

$$+ \frac{1}{2}\underbrace{(\frac{g_w^2 v^2}{4})}_{m_{W+}^2} W_\mu^+ W^{+\mu} + \frac{1}{2}\underbrace{(\frac{g_w^2 v^2}{4})}_{m_{W-}^2} W_\mu^- W^{-\mu} \tag{1.53}$$

$$+ \frac{1}{2}\underbrace{(\frac{g_W^2 v^2}{4\cos\theta_W})}_{m_Z^2} Z_\mu Z^\mu + \underbrace{0}_{m_\gamma^2} \times A_\mu A^\mu \tag{1.54}$$

where mass terms appear for the different bosons. As one degree of freedom of the Higgs field is used to build the physical scalar Higgs field, one of the generators remains unbroken, which gives a massless boson, the photon.

The theoretical masses of the gauge bosons, obtained from the Fermi constant(G_F) and the EW coupling constants g_W and g_B were confirmed experimentally [23–27]. But as the Higgs boson mass depends on a free parameter of the theory, μ^2, it could

only be determined experimentally. It was eventually measured at $m_H \sim 125\,\text{GeV}$ [28].

The Higgs field can also interact with fermions. This interaction between a scalar field (ϕ) and a Dirac field (ψ), called the Yukawa interaction, is a way to introduce gauge-invariant mass terms for fermions in the lagrangian. Before SSB, the Yukawa lagrangian is expressed as

$$\mathcal{L}_Y = -i\lambda_f \bar{L}\phi R_f - i\lambda_f \bar{R}_f \phi L = -i\lambda_f(\bar{L}\phi R_f + \bar{R}_f \phi L) \tag{1.55}$$

where the λ_f terms are the coupling constants of the Higgs fields to the respective fermion R_f.

If the symmetry of the Higgs scalar doublet is spontaneously broken in the form of Equation 1.44, a mass term for the down-type component of the fermion doublet appears. This would work for the lepton sector, as the neutrino-type is considered to be massless in the SM lagrangian, but it fails for up-type quarks, which also require a mass.

The Higgs field can be rewritten in the charge-conjugated of the $SU(2)$ framework:

$$\phi^c \equiv i\sigma_2 \phi^* = \begin{pmatrix} \phi^{(0)*} \\ -\phi^{(-)} \end{pmatrix} \xrightarrow{\text{SSB}} \frac{1}{\sqrt{2}} \begin{pmatrix} -(v + h(x)) \\ 0 \end{pmatrix} \tag{1.56}$$

where σ_2 is the second Pauli matrix, generator of $SU(2)$. The Higgs field in this form provides the symmetry-breaking term as the upper component, meaning it couples with up-type quarks to provide a mass term.

The Yukawa lagrangian becomes:

$$\begin{aligned} \mathcal{L}_Y &= -i\lambda_e(\bar{\nu}\bar{e})_L \phi e_R - i\lambda_d(\bar{u}\bar{d})_L \phi d_R - i\lambda_u(\bar{u}\bar{d})_L \phi^c u_R && (1.57)\\ &= -i\lambda_f(\bar{L}\phi R_f + \bar{R}\phi L) - i\lambda_f(\bar{L}\phi^c R_f + \bar{R}\phi^c L) && (1.58)\\ &= -i\lambda_f(\bar{L}\phi R_f + \bar{L}\phi^c R_f) + h.c. && (1.59) \end{aligned}$$

After picking the gauge, by arranging the chiralities, the lagrangian simplifies to:

$$\mathcal{L}_Y = \frac{-\lambda_f(v + h)}{\sqrt{2}}(\bar{\psi}_R \psi_L + \bar{\psi}_L \psi_R) \tag{1.60}$$

therefore the mass terms of the fermions are:

$$m_f \equiv \lambda_f \frac{v}{\sqrt{2}} \tag{1.61}$$

Hence the coupling of the Higgs to a fermion is proportional to its mass, making heavier fermions, such as τ leptons, have an enhanced coupling to the Higgs boson.

In conclusion, the overall SM lagrangian can be written by combining all terms as:

Table 1.2: Experimental value of the 19 free parameters of the Standard Model.

Name	Symbol	Value	
Up quark mass	m_u	$2.2^{+0.6}_{-0.4}$	MeV
Charm quark mass	m_c	1.27 ± 0.03	GeV
Top quark mass	m_t	173.1 ± 0.6	GeV
Down quark mass	m_d	$4.7^{+0.5}_{-0.4}$	MeV
Strange quark mass	m_s	96^{+8}_{-4}	MeV
Bottom quark mass	m_b	$4.18^{+0.04}_{-0.03}$	GeV
Electron mass	m_e	$0.511 \pm (0.31 \times 10^{-8})$	MeV
Muon mass	m_μ	$105.66 \pm (0.24 \times 10^{-5})$	MeV
Tau mass	m_τ	1776.86 ± 0.12	MeV
CKM I-II mixing angle	θ_{12}	$(13.01 \pm 0.03)^\circ$	
CKM II-III mixing angle	θ_{23}	$(2.35 \pm 0.09)^\circ$	
CKM I-III mixing angle	θ_{13}	$(0.20 \pm 0.04)^\circ$	
CKM CP-violating phase	δ_{CKM}	$(70 \pm 3)^\circ$	
$U(1)_Y$ gauge coupling	g_B	0.34970 ± 0.00019	
$SU(2)_L$ gauge coupling	g_W	0.65295 ± 0.00012	
$SU(3)_C$ gauge coupling	g_s	0.1182 ± 0.00012	
QCD vacuum angle	θ_{QCD}	$< 10^{-10}$	~ 0
Higgs v.e.v.	v	$246 \pm (6 \times 10^{-5})$	GeV
Higgs boson mass	m_H	125.09 ± 0.24	GeV

$$\mathcal{L}_{SM} = \underbrace{-\frac{1}{4}F^a_{\mu\nu}F^{\mu\nu}_a}_{\text{Free bosons}} + \underbrace{i\bar{\psi}\not{D}\psi}_{\text{Fermion term}} + \underbrace{i\lambda_f(\bar{L}\phi R_f + \bar{L}\phi^c R_f) + h.c.}_{\text{Yukawa interaction}} + \underbrace{|D_\mu\phi|^2 - V(\phi)}_{\text{Higgs mechanism}} \tag{1.62}$$

where:

$$\begin{aligned}
&F^a_{\mu\nu}F^{\mu\nu}_a \equiv G^a_{\mu\nu}G^{\mu\nu}_a + W^a_{\mu\nu}W^{\mu\nu}_a + B^a_{\mu\nu}B^{\mu\nu}_a \\
&i\bar{\psi}\not{D}\psi \equiv i\bar{\psi}\gamma^\mu(\partial_\mu - g_s T_a G^a_\mu - g_W T_a W^a_\mu - g_B Y B_\mu)\psi \\
&D_\mu\phi \equiv (\partial_\mu - ig_W T_a W^a_\mu - ig_B B_\mu)\phi \\
&V(\phi) \equiv \lambda(\phi^\dagger\phi)^2 - \mu^2(\phi^\dagger\phi)
\end{aligned}$$

The experimental value of the nineteen free parameters of the SM are given in Table 1.2.

1.1.6 Issues of the Standard Model

While the SM is a huge breakthrough, it still provides an incomplete description of nature, as it lacks explanation for several phenomena.

Gravity The gravitational force has not been successfully included in a QFT such as the SM. However, in an experimental particle physics context, gravitation is negligible. In fact, no experiment within the field has been accurate enough to pinpoint the effect of gravity at a particle level.

Neutrino masses and right-handed neutrinos Neutrinos do not interact with the Higgs boson, since they are massless fermions in the current SM lagrangian. However, experimental observations such as the neutrino oscillations [29, 30] prove that they are massive particles.

Another issue rises from the neutrinos being massive: the EWT is a chiral theory that violates parity maximally, and only left-handed neutrinos are required in its formulation. Although, being massive, their right-chirality particles must exist in nature but do not interact, and therefore cannot be included in the SM. Several hypotheses, such as the seesaw mechanism [31–36] have been proposed to describe these sterile neutrinos but these hypotheses are not supported by any experimental evidence.

Dark matter Dark matter is postulated to be a type of matter which interacts gravitationally but not electromagnetically. Its existence can be inferred from the shape and rotation speed of galaxies, but this matter does not give off or interact with light. The SM does not account for a candidate for this dark matter.

Dark energy Dark energy is also a postulate from astro-physics observations. Indeed, the measurements of the expansion of the universe imply acceleration. The energy that would correspond to this phenomenon would account for 73% of the energetic content of the universe.

Anti-matter asymmetry The Dirac equation predicts that each particle is created with an anti-matter partner with opposite quantum numbers. But the observed universe is mainly made up of regular matter, and even though EWT provides a CP-symmetry violating mechanism that can explain such asymmetry, the scale of the asymmetry does not match the effect of such a phenomenon alone.

Although the SM lacks explanations for all these phenomena, it has provided extremely precise predictions. Therefore the SM can be considered as an effective theory, while a deeper, more complete description of the subatomic world exists. Such theories are said to be beyond the Standard Model (BSM), with an example being the minimal supersymmetric extension of the Standard Model (MSSM).

1.2 The Minimal Supersymmetric extension of the Standard Model and its Higgs sector

1.2.1 Introduction to the Minimal supersymmetric extension of the Standard Model

One of the promising BSM models is Supersymmetry (SUSY) as it extends the symmetries of the Standard Model, solving many issues and providing a Dark Matter candidate. SUSY [37] introduces a new symmetry between fermions and bosons, making them effectively not independent objects but flavours of a more fundamental field. From this symmetry the existence of superpartners of each of the SM fermions

Table 1.3: Relations between the SM particles and their superpartners, before the EW symmetry-breaking, according to the MSSM. In the MSSM, the ordinary SM Higgs sector requires four additional Higgs bosons H, A and $H^{\pm}$ in addition to the SM Higgs, called h here.

SM particle (R = +1)				**Superpartner** (R = -1)			
Type	Spin	Particle	Symbol	Symbol	Particle	Spin	Type
Fermions	1/2	Quark	ψ_f	$\tilde{\psi}_f$	Squark	0	Sfermions
		Lepton	ψ_l	$\tilde{\psi}_l$	Slepton		
Bosons	1	Gluon	g	$\tilde{g}$	Gluino	1/2	Bosinos
		W	W^i	$\tilde{W}$	Wino		
		B	B	$\tilde{B}$	Bino		
	0	Higgs	h, H, $H^{\pm}$, A	$\tilde{h}$, $\tilde{H}$, $\tilde{H}^{\pm}$, $\tilde{A}$	Higgsinos		

and bosons is inferred. Each fermion therefore has a bosonic partner, called sfermion, which carries an integer spin. Conversely each boson has a fermionic partner called bosino, which carries a semi-integer spin. Both objects belong to a multiplet with the same quantum numbers except the spin value.

However, none of the superpartners have been observed so far, leading to the conclusion that this symmetry must be broken at the current energy scale. The SUSY theory implies the addition of a big set of new, undiscovered, observable particles, along with more free parameters in the theoretical formulation, in particular the masses of the new particles.

A consequence of the SUSY models could be the unification of the three forces of nature, namely electromagnetic, weak and strong interactions, into one, making such a model of a Grand Unified Theory (GUT). Indeed the unification of the weak and electromagnetic forces is already achieved in the SM with EWT, but the QCD and EW theories do not seem to converge. If interactions with the SUSY particles were added, the running couplings of the forces would be modified in such a way that they could converge at a large energy scale, thus providing a natural mechanism for a GUT.

Since SUSY is a general framework which depends on many unknown parameters, it can be implemented in different forms. The simplest model that realises SUSY while being compatible with the current observations is called the Minimal Supersymmetric Standard Model (MSSM). Its aim is to add the minimal amount of new parameters, particles and interactions, while keeping all the current symmetries and observables of the SM. Table 1.3 summarises the symmetry between ordinary particles and their superpartners in the MSSM.

This approach would allow certain interactions which have not been observed in the SM. In particular, SUSY models could allow processes where the baryon (B) and lepton (L) numbers are not conserved, and by extension $B - L$ either.

Because processes violating these numbers would make the proton unstable, and this instability has not been observed, a new symmetry has to be added to the MSSM to suppress the $B - L$ violating process: the $R - parity$. The operator of said R-parity, which is discrete, is defined for each particle in an interaction as:

$$P_R = (-1)^{3(B-L)-2s} \tag{1.63}$$

where s stands for the spin of the particle, B and L are the baryon and lepton numbers, respectively. The SM particles are defined as having $P_R = 1$ while superpartners have $P_R = -1$. To conserve R-parity the combined P_R has to be positive. SUSY models conserving R-parity have an additional consequence: the lightest supersymmetric particle (LSP) is stable, thus chains of heavier SUSY particles end with the LSP. If the LSP is, in addition, electrically neutral, it would be a candidate for the composition of Dark Matter. Multiple searches looking for an LSP are being performed but no evidence of its existence has been observed so far.

Contrary to what is done in the MSSM, adding a unique fermionic superpartner for the Higgs boson (named higgsino) would have had several implications. First, a chiral anomaly would appear, which means the generation of low mass-states due to the non-conservation of a chiral current. These states have not been observed in the Higgs sector, so a mechanism to suppress such states must be present. Second, the suppression of the flavour-changing neutral currents, which are not observed in nature either, would not be granted. Finally, the ratio between the neutral (G_n) and charged (G_c) currents in the EWT could not be of the order of unity, in contradiction with observations. The simplest solution which avoids these issues, recovering the observations of the SM, is the addition of a second Higgs field doublet, as performed in the MSSM.

1.2.2 The Two-Higgs-Doublet Model

The MSSM is therefore a specific case of a Two-Higgs-Doublet Model (2HDM). The 2HDM can be separated in different cases, called types, and the MSSM is a Type-II [38, 39]. In Type-II models, one of the doublets couples with up-type quarks (ϕ_u) while the second one couples with down-type quarks and leptons (ϕ_d). The two Higgs doublets are defined as

$$\phi_d \equiv \begin{pmatrix} \phi^{(+)} \\ \phi^{(0)} \end{pmatrix} \tag{1.64}$$

and

$$\phi_d \equiv \begin{pmatrix} \phi^{(0)*} \\ -\phi^{(-)} \end{pmatrix} . \tag{1.65}$$

The total number of degrees of freedom of the pair is eight. As in the SM, three of them are taken by the vector bosons in the SSB, leading to five massive observable bosons arising from the remaining ones. In this thesis, the lightest neutral scalar boson is denoted h and is taken to be the SM Higgs boson discovered in 2012. The heavier neutral scalar Higgs boson is denoted H, while a pseudo-scalar neutral Higgs boson is called A. The last two form a charged pair called $H^{\pm}$. As seen in Table

1.3, those particles are not superpartners but conventional bosons. Therefore, each of the five bosons would have its own fermionic partner, called higgsinos:

$$\underbrace{(h, H, A, H^{\pm})}_{ordinary} \rightarrow \underbrace{(\tilde{h}, \tilde{H}, \tilde{A}, \tilde{H}^{\pm})}_{supersymmetric} \tag{1.66}$$

1.2.3 Higgs sector in the MSSM

Mathematically the simplified Higgs potential can be written as [40]:

$$V(\phi_d, \phi_u) = \mu_u^2(\phi_u^\dagger \phi_u) + \mu_d^2(\phi_d^\dagger \phi_d) - \mu^4(\epsilon_{ij}\phi_d^i \phi_u^j + h.c.) \tag{1.67}$$

$$+ \frac{g_W^2 + g_B^2}{8}(\phi_d^\dagger \phi_d - \phi_u^\dagger \phi_u) + \frac{g_W^2}{2}|\phi_d^\dagger \phi_u|^2 \tag{1.68}$$

where $\epsilon_{ij} = 0$ if $i = j$ and $\epsilon_{ji} = -\epsilon_{ij} = 1$. To ensure vacuum stability, the potential must be bound from below, meaning $\mu_d^2 + \mu_u^2 > 2\mu^2$. The requirement for the spontaneous symmetry breaking becomes $\mu^4 > \mu_u^2\mu_d^2$, and the symmetry is spontaneously broken by the choice of the non-zero vacuum at

$$\langle \phi_d \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu_d \end{pmatrix} \tag{1.69}$$

and

$$\langle \phi_d \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} \nu_u \\ 0 \end{pmatrix} . \tag{1.70}$$

The vacuum expectation value of the SM , ν is recovered by

$$\nu^2 \equiv \nu_d^2 + \nu_u^2 . \tag{1.71}$$

This allows to introduce the parameter β defined by

$$\tan\beta \equiv \frac{\nu_u}{\nu_d} . \tag{1.72}$$

The masses of the $W^\pm$ and Z bosons are now defined as

$$m_W \equiv \frac{\nu . g_W}{2} \tag{1.73}$$

and

$$m_Z \equiv \frac{\mu_d^2\mu_u^2 \times \tan^2\beta}{\tan^2\beta - 1} . \tag{1.74}$$

The Yukawa couplings of the different Higgs bosons to the quarks are also modified with respect to the SM. They can be expressed as a correction of the coupling of the SM boson, depending on $\tan\beta$ and another angle α defined as

$$\tan 2\alpha \equiv \frac{m_A^2 + m_Z^2}{m_A^2 - m_Z^2} \times \tan 2\beta . \tag{1.75}$$

These Yukawa couplings expressed in terms of the α and β parameters can be found in table 1.4.

Table 1.4: Relation of the Yukawa coupling parameters (λ_{ii}) with respect to the SM coupling (λ_{SM}) for the neutral MSSM Higgs bosons (h, H and A) to the vector bosons (λ_{VV}), and to the different fermions, split into u-type (only quarks, λ_{uu}) and d-type (quarks and charged leptons $\lambda_{dd,ll}$) as a function of the angles α and β.

λ_{ij} / λ_{SM}	λ_{VV}	λ_{uu}	$\lambda_{dd,ll}$
h	$\sin(\beta-\alpha)$	$\cos\alpha/sin_\beta$	$-\sin\alpha/cos\beta$
H	$\cos(\beta-\alpha)$	$\sin\alpha/\sin\beta$	$\cos\alpha/\cos\beta$
A	0	$\cot\beta$	$\tan\beta$

The masses of the five Higgs bosons at tree level are then:

$$m_A^2 = \frac{2\mu^2}{sin2\beta} \tag{1.76}$$

$$m_{H^\pm}^2 = m_A^2 + m_W^2 \tag{1.77}$$

$$m_{H,h}^2 = \frac{1}{2}\big(m_A^2 + m_Z^2 \pm \sqrt{(m_A^2+m_Z^2)^2 - 4m_Z^2 m_A^2 cos^2 2\beta}\big) \tag{1.78}$$

These equations show that, at lowest order, the masses of the Higgs bosons depend on two free unknown parameters, namely $tan\beta$ and m_A. However, higher-order corrections depend on several additional parameters. One such parameter is the stop (top quark superpartner) mixing parameter

$$X_t \equiv A_t - \mu cot\beta \tag{1.79}$$

which depends on the soft SUSY-breaking Higgs-stop coupling A_t. The other parameter is the average scale of SUSY, defined as the average scale of the stop masses as

$$m_{SUSY} \equiv \sqrt{m_{\tilde{t_1}} m_{\tilde{t_2}}}\,. \tag{1.80}$$

Indeed, at higher orders the correction on the Higgs mass becomes [40]

$$\delta m_h^2 \approx \frac{3m_t^4}{2\pi^2 v^2}\Big[ln\frac{m_{SUSY}^2}{m_t^2} + \frac{X_t^2}{m_{SUSY}^2}\big(1 - \frac{X_t^2}{12m_{SUSY}^2}\big)\Big]\,. \tag{1.81}$$

As a huge number of free parameters is impractical for experimental tests, a common procedure is to set high-order parameters to a particular value, aiming to focus on specific MSSM phenomenologies, called scenarios, and then set experimental limits on the $tan\beta$ vs m_A parameter space. Such scenarios are detailed in section 1.3.

The search for the additional neutral heavy Higgs boson of the MSSM is one of the goals of this thesis, as detailed in chapter 4. This search is done in several MSSM scenarios, aiming for topologies where the differences with respect to the SM are enhanced.

1.3 Phenomenology of MSSM Higgs bosons production and decay into $\tau\tau$ in pp collisions at the LHC

In order to challenge not only the SM but also any BSM theories like the MSSM, many experiments have been designed. Those experiments aim to make indirect measurements of parameters, or directly observe new particles through their decay products. The Large Hadron Collider (LHC) was designed with both in mind. Indeed the LHC accelerates beams of protons to achieve the necessary energy in the centre of mass to generate processes of interest. When two protons collide, most interactions are scatterings due to the repulsion of their electric charge. But in some cases the interaction is inelastic, resulting in the production of new particles. Hadrons are composite particles formed by three valence quarks that are bound together by a continuous exchange of gluons. Protons are hadrons where the quark composition is *uud*. Within the proton, gluons are continuously transformed into pairs of quark-antiquark, which form what is called the sea of quarks.

The main collision, called event, comes from the hard scattering of the protons at high energy. The elements that are effectively taking part in the collision, either quarks (valence or from the sea) or gluons, are the constitutive elements of the protons, called partons. As they are bound within the proton, they carry part of its energy in a dynamic way. Since the partons carry a fraction of the total energy carried by the proton, the energy available at the collision creates a spectrum which is bounded from above by the energy given to both protons. Several different probability distributions, called Parton Distribution Functions (PDFs) are used to estimate this energy distribution. Different PDF schemes, such as CTEQ [41], MSTW [42] and NNPDF [43] have been developed and tested at different energy regimes.

The cross-section of the pp collisions, $\sigma(pp \to X)$ is given by the QCD factorization theorem [44]:

$$\sigma(pp \to X) = \sum_{i,j} \int\int f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2) \hat{\sigma}_{ij\to X}(s x_1 x_2, \mu_R^2, \mu_F^2) dx_1 dx_2 \qquad (1.82)$$

where x_1 and x_2 variables are the fraction of the total momentum that the partons i, j carry, providing an effective center-of-mass energy of $\hat{s} = s x_1 x_2$, and the variables μ_R^2 and μ_F^2 are the renormalization and factorization scales respectively, which are obtained by truncating the strong coupling constant. Finally, the variables f_i and f_j are the parton densities, obtained from PDFs.

The partial cross-section $\hat{\sigma}_{ij\to X}$ can be computed using the perturbative method, up to Leading Order (LO), or adding further corrections to next orders, (NLO, NNLO,...). However, the physical process does not end here: the partons involved in the collision can radiate soft-gluons (parton shower) which later will hadronise, forming a cascade of particles. An accurate theoretical modeling of these effects is not possible and thus, their simulation is constrained using experimental data.

On top of all this happening at the collision, several other interactions can occur simultaneously to the main collision event, due to the interactions of either other

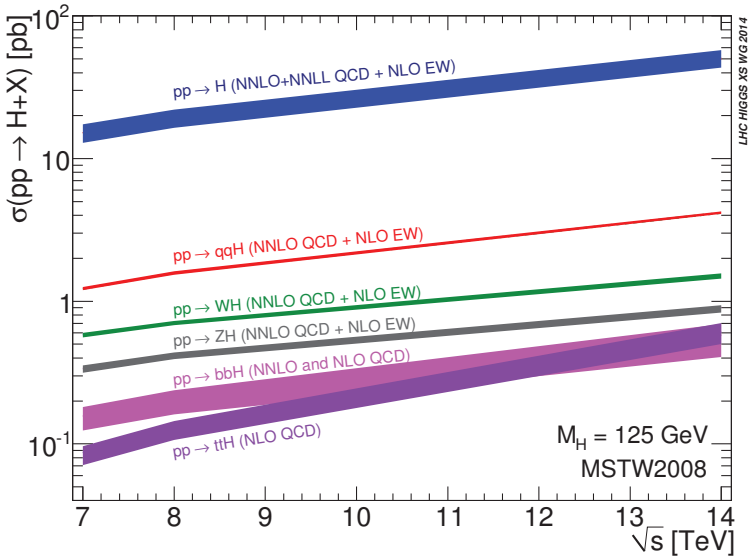


Figure 1.4: Cross-sections for different SM Higgs boson production processes as a function of the center of mass energy.

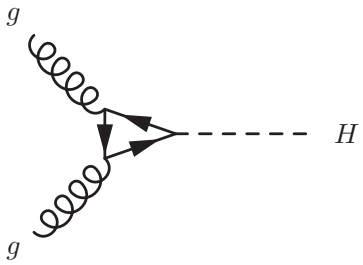


Figure 1.5: Feynman diagram for the gluon fusion process (ggH) at lowest order.

protons or of the partons not involved in the main event. These are called pile-up and underlying events, respectively.

1.3.1 Higgs boson production

At the LHC, neutral Higgs bosons can be produced in several different processes, most of which have been measured with great accuracy for the SM Higgs boson [45]. The cross-section of the SM Higgs boson production modes as a function of center-of-mass energy is shown in figure 1.4.

The dominant Higgs production process at the LHC is the gluon fusion, labeled ggH (Figure 1.5). This interaction is mediated by a loop of quarks. Since the Yukawa coupling of the Higgs boson depends on the mass of the fermion, the top quark dominates this process. The ggH process is the most abundant production mode in proton colliders, with up to 85% of SM Higgs bosons production.

The second most abundant production mode at the LHC is the vector boson fusion [46], abbreviated as VBF (Figure 1.6). This process consists of two quarks directly or indirectly producing two vector bosons ($W^{\pm}$ or Z) that fuse into a Higgs boson. Despite having a cross-section ten times lower than ggH at LHC, this pro-

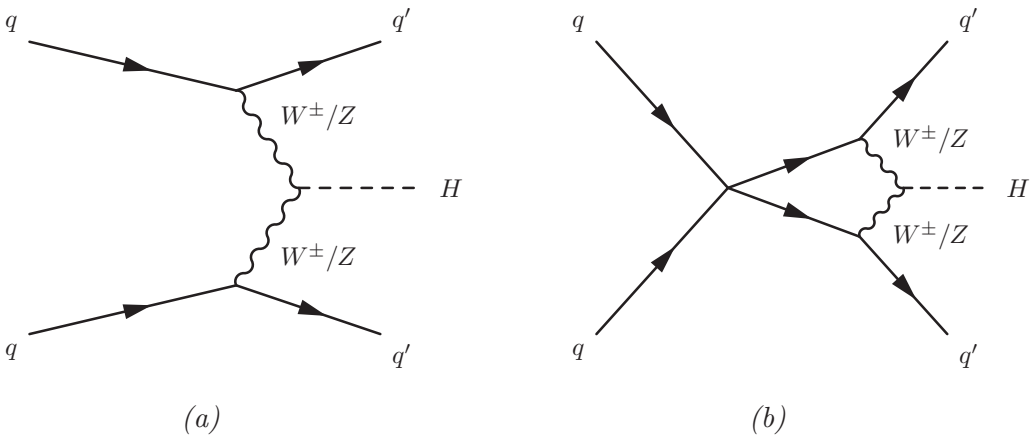


Figure 1.6: Feynman diagrams for the Vector Boson Fusion process of a Higgs boson with two jets at leading order for the t (a) and u (b) channels.

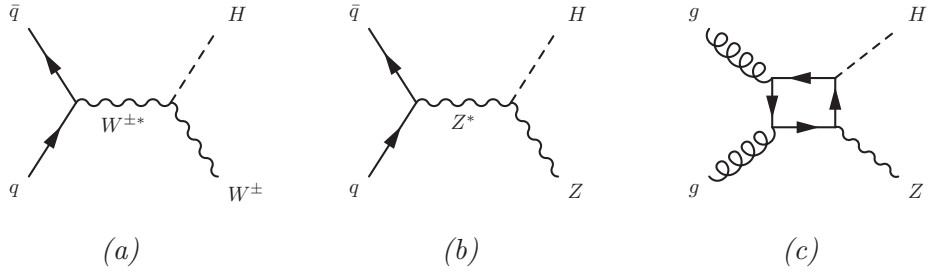
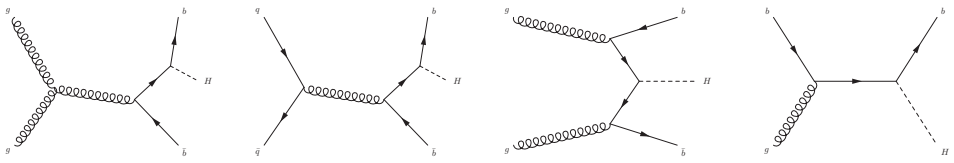


Figure 1.7: Feynman diagrams for the vector boson associated production process (VH) at leading order for (a) the W boson and (b) the Z boson. Last diagram (c) corresponds to a gluon fusion via top quark loop which contributes to the ZH mode.

cess is particularly recognizable thanks to the behaviour of the two outgoing quarks, which hadronise giving two observable energetic jets back-to-back in the Higgs reference frame. The current computation of the VBF cross-section includes NNLO QCD corrections and NLO EW corrections [47].

Another process is the production of a Higgs boson in association with a vector boson, VH (Figure 1.7), also called Higgs-strahlung. In this process, the quark-antiquark pair collisions give rise to an energetic vector boson ($W^{\pm}$ or Z), which then radiates a Higgs boson. These cross-sections are computed up to NNLO for the QCD corrections plus NLO EW corrections [47].

Last but not least, another production process is the associated production with heavy fermions, namely top quarks ($t\bar{t}H$) and bottom quarks ($b\bar{b}H$), the latter shown in Figure 1.8. Even though these modes are not significant in SM Higgs boson production at the LHC, the MSSM takes advantage of the enhanced coupling of the b-quark to the Higgs bosons for large tanβ values. In such cases the $b\bar{b}H$ mode becomes a significant or even dominant source of Higgs bosons. The cross-section for the $b\bar{b}H$ modes are computed at NLO for the 4-flavour scheme, in which it is assumed that only the 4 lightest quarks (u, d, s, c) are available in the proton, and NNLO for the 5-flavour scheme [47].



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

Figure 1.8: Feynman diagrams for the b-associated production process at leading order in the four-flavour scheme (a,b,c) and the five-flavour scheme (d).

1.3.2 Higgs bosons decay

Higgs bosons have a very short lifetime. For example, the SM Higgs boson has a lifetime of $\sim 10^{-22}$ s [48]. Hence direct observations are not even considered and the searches have to focus on the signatures of its decays. Each of the decay modes has a different topology as well as a different branching ratio (BR).

The MSSM introduces a second Higgs doublet as stated in section 1.2, leading to five observable Higgs bosons, namely h, H, A and $H^{\pm}$. The chosen interpretation of the discovered Higgs boson at $\sim 125\,\text{GeV}$ is to assume that it holds the role of the lightest boson h. This thesis will rely on this interpretation to look for other neutral Higgs bosons with masses larger than $m_h = 125 \pm 3\,\text{GeV}$. As seen previously, the correction to the masses at tree level can be written to depend uniquely on the m_A and $\tan\beta$ parameters, while the corrections to the Yukawa couplings of the heavy neutral Higgs bosons to down type fermions can be written in terms of the angles α and β as

$$\lambda(H \to bb, \tau\tau) \propto \frac{\cos\alpha}{\cos\beta} = \frac{\cos\alpha}{\sin\beta} \times \tan\beta \tag{1.83}$$

$$\lambda(A \to bb, \tau\tau) \propto \tan\beta\,. \tag{1.84}$$

Therefore, the coupling of both heavy neutral Higgs bosons are proportional to the value of the free parameter $\tan\beta$. For large values of $\tan\beta$, the coupling of these Higgs bosons to d-type fermions, such as the b quark and the τ lepton is then enhanced with respect to the SM. Large $tan\beta$ values would lead to an enhancement of the $H \to \tau\tau$ and $H \to bb$ branching ratios, and especially of the $\mathrm{b\bar{b}H}$ production mode cross-section shown in Fig. 1.8.

A large $\tan\beta$ would also affect the $\mathrm{g\bar{g}H}$ mode, as the Higgs boson production in this mode happens through a quark loop. Contrarily to the SM, where the top loop dominates because of its mass, in the MSSM, a large value $\tan\beta$ would enhance b-quark loops, which could even dominate over t-loops.

While the $H, A \to bb$ channel is expected to have a greater branching ratio, it suffers from significant backgrounds at the LHC. On the other hand, the $H, A \to \tau\tau$ is a very sensitive fermionic channel because its final state provides a clear signature (high energetic τ leptons). This thesis searches for a massive neutral Higgs boson decaying to a pair of τ leptons, denoted as $H, A \to \tau\tau$ channel, whose Feynman diagram is shown in Figure 1.9.

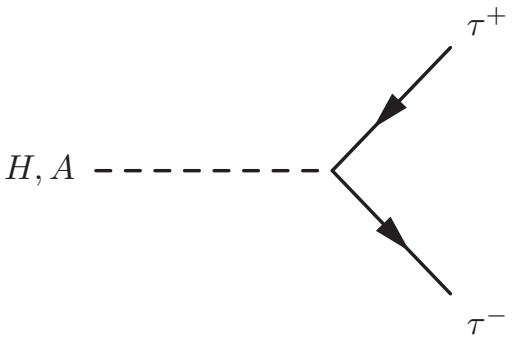


Figure 1.9: Feynman diagrams for the $H, A \to \tau\tau$ decay at tree level.

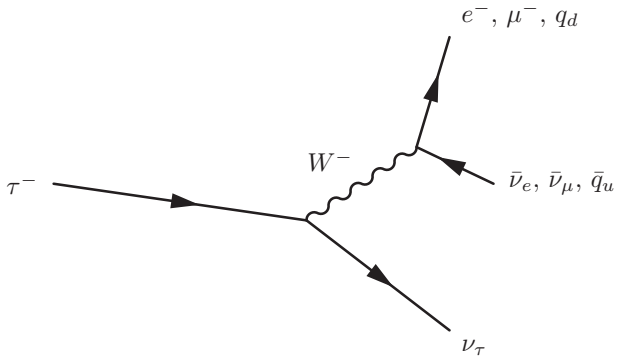


Figure 1.10: Feynman diagram for the decay of a τ^- particle, mediated by a W^- boson at tree level.

1.3.3 The τ lepton

The τ lepton is an unstable particle with a mean life of $\sim 10^{-13}$ s [48] and therefore, a decay length of $87.03\,\mu$m. Its decaying vertex is therefore usually close to the production vertex. The τ decays via the electroweak interaction, involving a virtual W, as illustrated in figure 1.10.

The decay products of the virtual W can lead to two different types of final states, the leptonic final state, and the hadronic final state. Leptonic τ decays produce an electron or a muon, and two neutrinos. Hadronic τ decays, denoted τ_{h}, lead to a single neutrino and a quark-antiquark pair, which leads to observed final states of mainly either 1 or 3 charged hadrons, and potentially several π^0 which in turn decay to photons. The main τ decays and their branching fractions are detailed in Table 1.5.

While these different decays can lead to many different final states of $H, A \to \tau\tau$ events, this thesis will focus on the fully hadronic channel, denoted $\tau_h\tau_h$.

Table 1.5: Branching fractions of the main (negative) τ decay modes. The generic symbol h^- represents a charged hadron, pion or kaon. In some cases, the decay products arise from an intermediate mesonic resonance.

Decay mode	Meson resonance	Branching fraction [%]
$\tau^- \to e^- \bar{\nu}_e \nu_\tau$		17.8
$\tau^- \to \mu^- \bar{\nu}_\mu \nu_\tau$		17.4
$\tau^- \to h^- \nu_\tau$		11.5
$\tau^- \to h^- \pi^0 \nu_\tau$	$\rho(770)$	26.0
$\tau^- \to h^- \pi^0 \pi^0 \nu_\tau$	$a_1(1260)$	10.8
$\tau^- \to h^- h^+ h^- \nu_\tau$	$a_1(1260)$	9.8
$\tau^- \to h^- h^+ h^- \pi^0 \nu_\tau$		4.8
Other modes with hadrons		1.8
All modes containing hadrons		64.8

2

The CMS experiment

Most ordinary matter is made up of stable particles. Indeed, rare particles are not easily found since they decay into ordinary stable matter before they can be observed. The best way to observe unstable or rare particles is at the place and time of their production. Physicists have therefore turned their detectors towards the skies, as the interactions between cosmic rays and the atmosphere can create many of the desired new particles. Even though this approach has lead to great discoveries, and to a better understanding of particle physics, ways to produce such processes in a controlled environment can be designed and have many advantages. One such production environment is the Large Hadron Collider (LHC), which collides protons together with the goal of creating new particles. The LHC apparatus is presented in section 2.1. One of the advantages of this experimental context is the possibility of having a global understanding of collisions through the detection of most of the outgoing products. The Compact Muon Solenoid (CMS) experiment consists in a detector being placed around one of the collision points of the LHC. This detector is presented in section 2.2. In order to conclude whether nature agrees with theory, real data will be compared to simulations of the collisions and the detection. This simulation process is described in section 2.3. To interpret the signals gathered by the detector, a reconstruction algorithm is applied to data, and is also applied to simulated events for comparison, as detailed in section 2.4.

2.1 The Large Hadron Collider

The LHC is the biggest and most powerful collider in the world. It was built in a 27 km long underground circlular cave, at a depth of 100 m below the surface. It is situated at the French-Swiss border at the CERN (European Organisation for Nuclear Research) facility.

Two rings accelerate protons in opposite directions. The acceleration gives up to 7 TeV of energy to each hadron in order to create a collision, totalling 14 TeV in the centre of mass [49–51]. This acceleration is possible through the use of 16 radio frequency cavities, and the trajectories are kept along the circular cavern by about 9500 magnets. These magnets are cooled down to 1.8 K with superfluid helium, and through superconductivity are able to deliver a nominal magnetic field of 8.33 T.

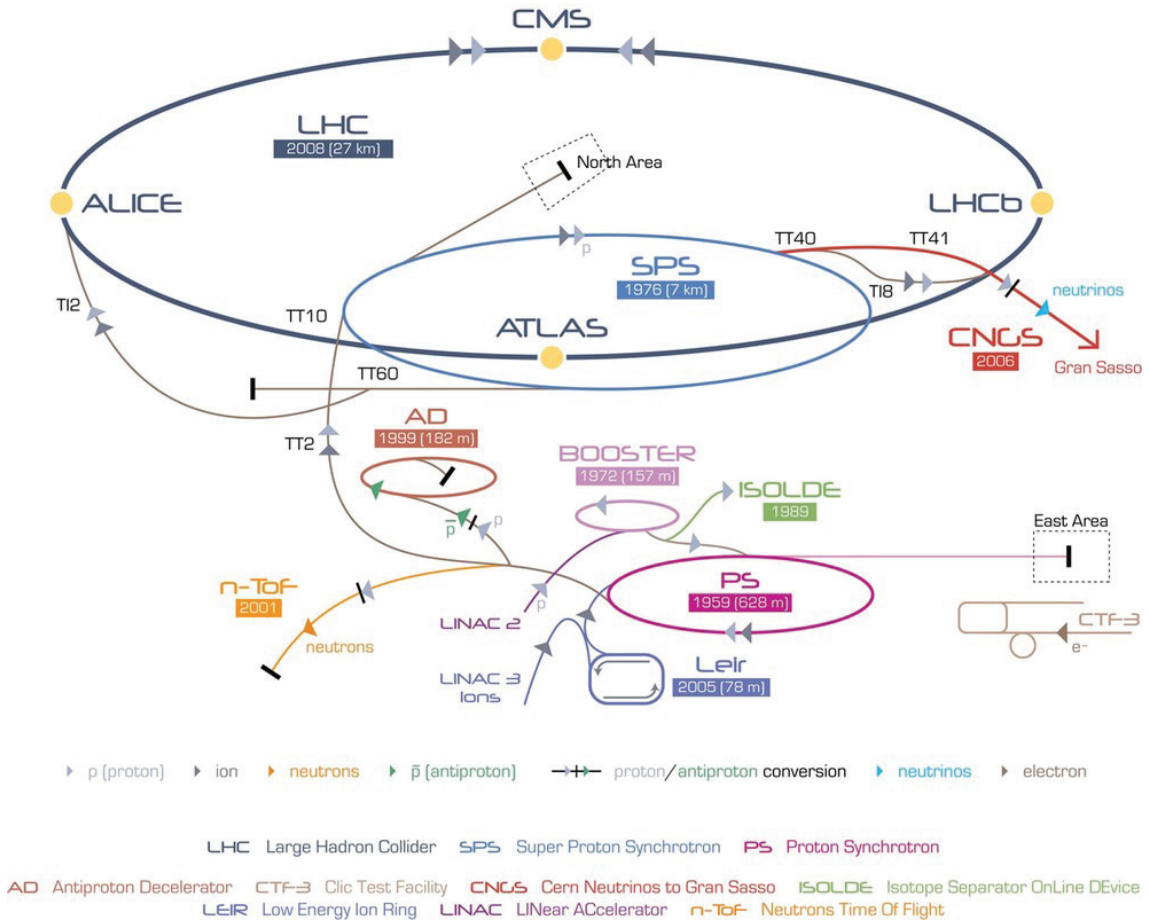


Figure 2.1: Diagram of the full accelerator complex at CERN, including the LINAC2, BOOSTER (PSB), PS, SPS and LHC accelerators.

2.1.1 Proton acceleration

Before they reach nominal energy in the LHC rings, bunches of protons are gradually accelerated by smaller accelerators as illustrated in Figure 2.1. The main accelerators are listed below.

LINAC 2 is the start of the whole acceleration process. Hydrogen is ionized by an electric field to provide protons, which are accelerated to an energy of 50 MeV.

The Proton Synchrotron Booster (PSB) is a circular accelerator where the beams of protons reach an energy of 1.4 GeV.

The Proton Synchrotron (PS) is another circular accelerator, accelerating protons to an energy of 25 GeV.

The Super Proton Synchrotron (SPS) is the last accelerator before the LHC. Protons are accelerated to 450 GeV before being injected in the LHC to reach 7 TeV.

2.1.2 Luminosity

The instantaneous luminosity is a key variable in a collider experiment. Expressed in units of $\mathrm{cm}^{-2}\mathrm{s}^{-1}$, it is proportional to the number of collisions per second and per square centimeter. It can be expressed as

$$\mathcal{L}_{\text{inst}} = \frac{\gamma f n_p N_p^2}{4\pi\epsilon_n\beta^*} = \frac{f n_p N_p^2}{4\pi\sigma_x\sigma_y} \tag{2.1}$$

where γ is the Lorentz boost, f is the revolution frequency of the bunches, n_p is the number of bunches, N_p is the number of protons per bunch, ϵ_n is the transverse emittance which is a measure of the parallelism of the beam, β^* is the amplitude function that measures the distance between the interaction point and the place where the beam gets twice as wide, and $\sigma_{x,y}$ the width of the overlapping beams in the (x,y) at the interaction point.

The integrated luminosity over a period of data-taking is defined by $\mathcal{L} = \int \mathcal{L}_{\text{inst}} dt$. This variable represents the amount of data that is or can potentially be collected by the experiment. The number of events produced by collisions for a given process is

$$N = \mathcal{L}\sigma\,. \tag{2.2}$$

In this equation, σ is the cross-section of a given process. This equation shows that to study rare particles or rare decays, it is beneficial to combine both an important instantaneous luminosity and a long data-taking period.

2.1.3 Pileup

When bunches of protons meet, several pp interactions can happen. The main interaction is called the hard process, and ideally should be the only source of particles that will be detected. But other collisions, whether elastic or inelastic, introduce unwanted noise, and can be difficult to separate from the hard process elements. This effect is called pileup (PU). The number of PU events per collision depends on the LHC configuration. Two types of PU can be distinguished, the in-time PU from other collisions occurring at the same time as the hard process and the out-of-time PU, which originates from leftover activity in the detectors from previously occurring collisions.

2.1.4 The experiments

The LHC is circular, thus allowing for several interaction points to be set up along the tunnel. Indeed, 4 major experiments have been set up along the LHC, namely ALICE, ATLAS, CMS and LHCb.

A Large Ion Collider Experiment (ALICE) This experiment is mainly aimed at the study of nuclear matter deconfinement, creating quark and gluon plasma. Its data gathering is focused around heavy ion collisions (Pb-Pb or p-Pb), but still uses pp collisions, for example as a way to calibrate the detector.

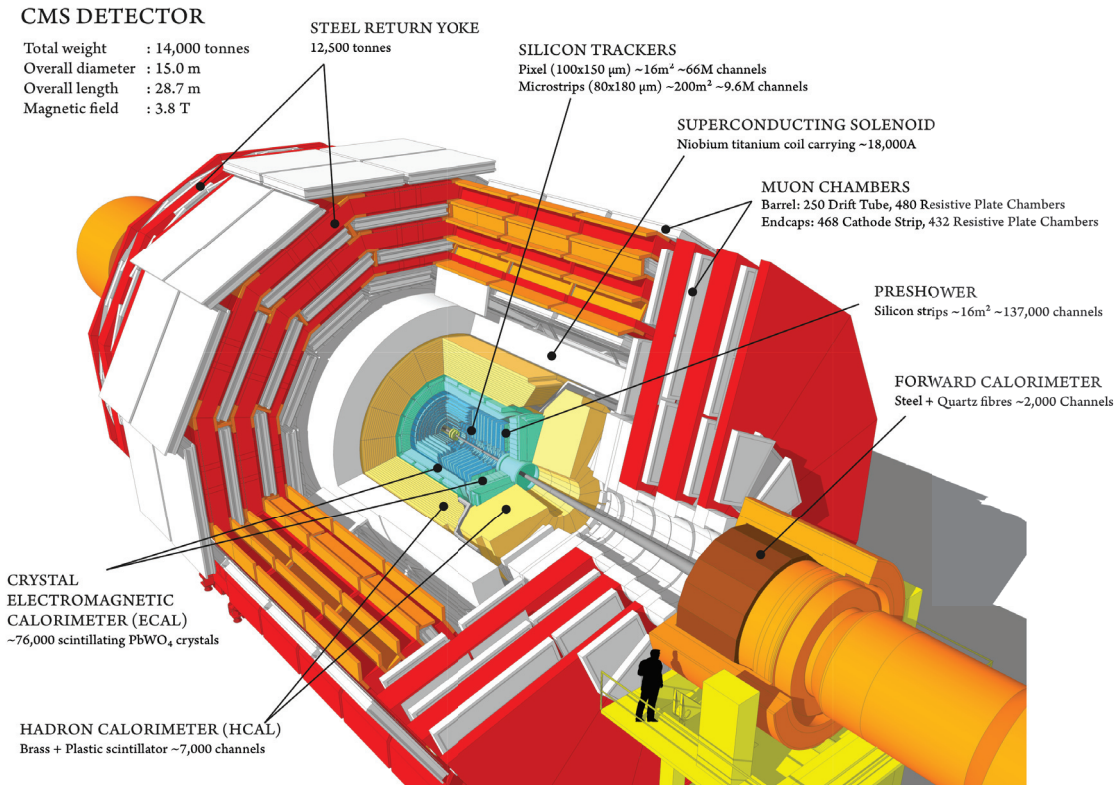


Figure 2.2: Illustration of the CMS detector and its concentric layers of subdetectors along with the superconducting solenoid and the steel return yoke.

Large Hadron Collider Beauty This experiment is dedicated to SM precision measurements as well as CP (charge-parity symmetry) violation studies. This is done through the extensive study of the b quark.

A Toroidal LHC ApparatuS (ATLAS) and Compact Muon Solenoid (CMS) These are the generalist experiments of the LHC. Indeed, ATLAS and CMS physics aims range from the Higgs boson discovery, which was successfully achieved in 2012, to precision measurements of the standard model parameters, while also allowing BSM searches, like the search for dark matter candidates, or extra Higgs bosons. The CMS experiment is detailed in the following section 2.2.

2.2 The CMS experiment

The CMS detector is situated in a cavern at LHC point 5, near Meyrin in France. It is a roughly cylindrical-shaped detector of 28.7 m in length and 15 m in diameter, with a weight of 14 000 tons. It is composed of concentric layers, each layer being a different subdetector, as illustrated in Figure 2.2. Each detector has a different role, which will be detailed in this section. The geometry of the overall detector and of the collisions has pushed to define measurements in terms of a cylindrical frame of reference. The x-axis is defined as pointing toward the centre of the LHC, the y-axis as pointing effectively upward, and the z-axis as being directed along the beam axis,

together forming a Cartesian coordinate system. The ϕ angle is defined in the (x,y) plane from the x-axis. The θ angle is defined from the z-axis in the (y,z) plane. But the θ angle is rarely used and usually replaced by the pseudo-rapidity η defined as

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] . \tag{2.3}$$

The pseudo-rapidity gives an indication of the polar angle of a particle, and is equal to zero in the transverse plane. It is used instead of the polar angle because, in pp collisions, the particle production density along η is mostly constant. Since the LHC is effectively colliding partons within the incoming protons, it is impossible to know the momentum or energy that is available in the partonic collision. But from the LHC geometry, the momentum of the initial partons in the (x,y) plane, called the transverse plane, is by definition null. From momentum conservation the sum of the momentum of all the outgoing particles in the transverse plane should be null as well. It is therefore convenient to define the projected momentum and energy in the transverse plane as

$$p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2} = \frac{|\vec{p}|}{\cosh\eta} \tag{2.4}$$

and

$$E_{\mathrm{T}} = E\sin\theta = \frac{E}{\cosh\eta} . \tag{2.5}$$

One of the objectives of CMS is to measure with great precision the momentum and energies of outgoing particles, even the most boosted ones. To do so, a 3.8 T magnetic field is created by the solenoid layer. This magnetic field is oriented along the z-axis, to curve the trajectories of charged particles around this axis, i.e. in the transverse plane. The goal of the layer-based design is to be able to identify, and measure the characteristics of outgoing particles, such as their momentum and energies.

2.2.1 The silicon inner tracker

The full-silicon inner tracking system [52, 53] is the closest layer to the interaction point. It is a cylinder-shaped subdetector with an outer radius of 1.10 m and a length of 5.6 m. This detector layer is further subdivided in silicon layers. The barrel (each of the two endcaps) comprises four (two) layers of pixel detectors, surrounded by ten (twelve) layers of micro-strip detectors. The 16,588 silicon sensor modules are finely segmented into 66 million 150 $\times$ 100 μm pixels and 9.6 million strips, with a width ranging from 80 to 200 μm, and length from 8 to 25 cm. Its role is to detect charged particles passing through its different layers in order to reconstruct their trajectories, their charge, their momentum and approximate their vertex of origin, based on the knowledge of the intensity of the magnetic field. As displayed in Figure 2.3, these layers and the pertaining services (cables, support) represent a substantial amount of material in front of the calorimeters, up to 0.5 interaction lengths or 1.8 radiation lengths. The large number of emerging secondary particles turns out to be a major source of complication for reconstruction, but can be mitigated through the use of the combination of information with other subdetectors.

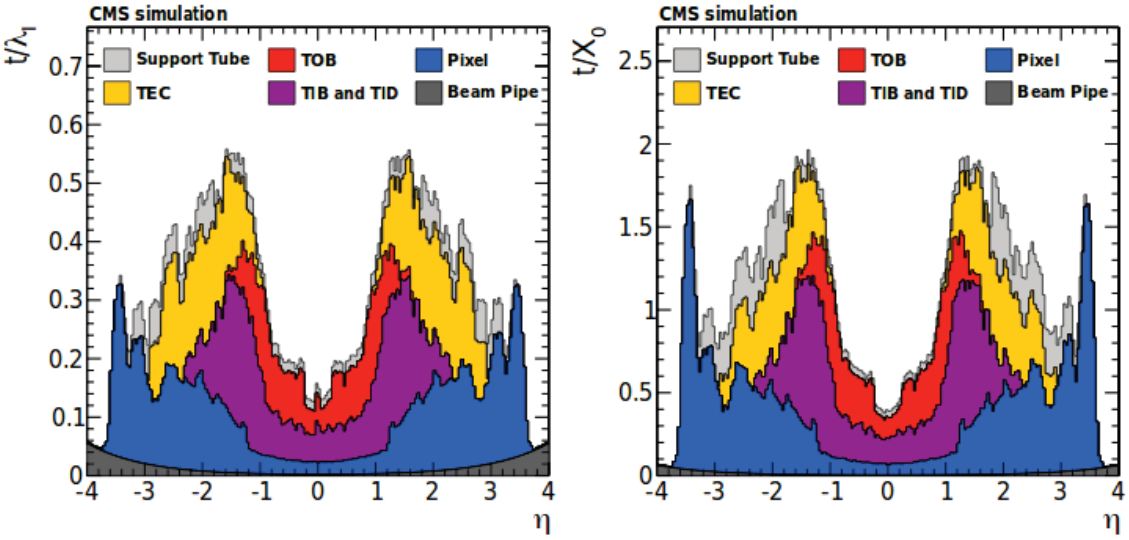


Figure 2.3: Total thickness of the inner tracker material expressed in units of interaction lengths λ_I (left) and radiation lengths X_0 (right), as a function of the pseudorapidity η. The acronyms TIB, TID, TOB and TEC stand for "tracker inner barrel", "tracker inner disks", "tracker outer barrel", and "tracker endcaps" respectively. The two figures are taken from Ref. [2]

The tracker measures the p_T of charged hadrons at normal incidence with a resolution of 1% for $p_T < 20\,\text{GeV}$. The relative resolution then degrades with increasing p_T to reach the calorimeter energy resolution for track momenta of several hundred GeV.

2.2.2 The electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) [54, 55] is a hermetic homogeneous calorimeter made of lead tungstate ($PbWO_4$) crystals. Its role is to stop and measure the energy of electrons and photons. The barrel covers $|\eta| < 1.479$ and the two endcap disks $1.479 < |\eta| < 3.0$. The barrel (endcap) crystal length of 23 (22) cm corresponds to 25.8 (24.7) radiation lengths, sufficient to contain more than 98% of the energy of electrons and photons up to 1 TeV. The crystal material also amounts to about one interaction length, causing about two thirds of the hadrons to start showering in the ECAL before entering the next subdetector.

The crystal transverse size matches the small Molière radius of $PbWO_4$, 2.2 cm. This fine transverse granularity makes it possible to fully resolve hadron and photon energy deposits as close as 5 cm from one another. The front face of the barrel crystals has an area of $2.2 \times 2.2\ \text{cm}^2$, equivalent to 0.0174×0.0174 in the (η,ϕ) plane at normal incidence. In the endcaps, the crystals are arranged instead in a rectangular grid, with a front-face area of $2.9 \times 2.9\ \text{cm}^2$. The intrinsic energy resolution of the ECAL barrel was measured with an ECAL supermodule directly exposed to an electron beam, without any attempt to reproduce the material of the tracker in front of the ECAL [56]. The relative energy resolution is parametrized as a function of the electron energy as

$$\frac{\sigma}{E} = \frac{2.8\%}{\sqrt{E/\mathrm{GeV}}} + \frac{12\%}{E/\mathrm{GeV}} + 0.3\%\,. \tag{2.6}$$

2.2.3 The hadronic calorimeter

The hadronic calorimeter (HCAL) [57] is a hermetic sampling calorimeter consisting of several layers of brass absorber and plastic scintillator tiles. Its role is to stop and measure the energy of outgoing hadrons. It surrounds the ECAL, with a barrel ($|\eta| < 1.3$) and two endcap disks ($1.3 < |\eta| < 3.0$). In the barrel, the HCAL absorber thickness amounts to almost six interaction lengths at normal incidence, and increase to over ten interaction lengths at larger pseudorapidities. It is complemented by a tail catcher (HO), installed outside the solenoid coil. The HO material (1.4 interaction lengths at normal incidence) is used as an additional absorber. At small pseudorapidities ($|\eta| < 0.25$), the total absorber thickness is enhanced to a total of three interaction lengths by a 20 cm-thick layer of steel. The total depth of the calorimeter system (including ECAL) is thus extended to a minimum of twelve interaction lengths in the barrel. In the endcaps, the thickness amounts to about ten interaction lengths.

The HCAL is read out in individual towers with a cross-section $\delta\eta \times \delta\phi = 0.087 \times 0.087$ for $\eta < 1.6$ and 0.17×0.17 at larger pseudorapidities. The combined (ECAL+HCAL) calorimeter energy resolution was measured in a pion test beam to be

$$\frac{\sigma}{E} = \frac{110\%}{\sqrt{E}} + 9\%\,, \tag{2.7}$$

where E is expressed in GeV.

2.2.4 The muon detectors

Outside the solenoid coil, the magnetic flux is returned through a yoke consisting of three layers of steel interleaved with four muon detector planes [2, 58]. Drift tube (DT) chambers and cathode strip chambers (CSC) detect muons in the regions $|\eta| < 1.2$ and $0.9 < |\eta| < 2.4$, respectively, and are complemented by a system of resistive plate chambers (RPC) covering the range $|\eta| < 1.6$. The goal of this subdetector is to detect muons passing through, as muons are the only isolated charged particles expected to go through all the previous layers. Indeed, the reconstruction of muon trajectories, described in section 2.4, involves a global trajectory fit across the muon detectors and the inner tracker.

2.2.5 The trigger system

The nominal time between each collision is 25 ns, meaning a frequency of 40 MHz. With an estimated 1 Mo per event, this would lead to a bandwidth of 40 To per second, which is far too much to be handled in real time. In order to limit the bandwidth, a trigger system has been set up, getting rid of fairly common events, which have already been well studied, and storing only events with characteristics

that are considered to be relevant. This trigger system has two distinct levels called the level-1 (L1) and high-level (HLT).

Level-1 trigger (L1) This trigger level has the goal of reducing the 40 MHz bandwidth into a 100 kHz bandwidth. It does so by relying on ultra-fast electronic hardware, using inputs straight from the muon chambers and calorimeters. The decision to keep or reject an event is then taken in less than 3.2 μs.

High level trigger (HLT) The goal of this level is to reduce the bandwidth to about $1\,kHz$ in order to be able to store the data. This triggering is done on a computer farm installed in a room adjacent to the detector. This allows for more complex triggering algorithms to be used, and even to perform physics object reconstruction. Several trigger algorithms are implemented, usually based on certain physics object requirements.

2.3 Simulation

In order to compare experimental results with theoretical predictions, both the collision process and the interaction of the outgoing particles with the subdetectors need to be simulated. The hard process is first simulated with a Monte-Carlo event generator, which simulates the parton interaction and produces the list of outgoing particles. Then the flight and decay of each particle through the different subdetectors is extrapolated with the kinematics, while the interactions and subdetector outputs are simulated as well.

2.3.1 Physics event generation

When two protons collide, only the constituents, namely the partons, will actually collide. Each parton carries a fraction of the total momentum of the incoming protons. The partonic collision is a hard process described analytically using perturbative theory. A generator will then determine the transition matrix between initial and final state, following the Feynman rules. The leading order calculation is available for most processes, the following orders are only available for specific subsets of processes. Leading order processes are computed using software like MadGraph [59], while next-to-leading order computations are done using software like Powheg [60] or MC@NLO [61]. These generators will then provide the quadri-moment of the outgoing particles produced at the collision.

Particles holding a colour charge that are produced in the collision will radiate gluons and photons. These gluons can themselves radiate other gluons, creating a cascade. Once the created gluons reach the threshold dictated by the generator, gluons then hadronize. Since they are subjected to the strong interaction, coloured particles that are created will also hadronize, leading to colour-neutral hadrons. This hadronization happens roughly 5×10^{-24} s after the coloured particles are produced. But, as stated before, there is no theory describing hadronization, only phenomenological models. The energy scale at which these effects happen makes

QCD non-perturbative. These phenomenological models can be separated into two distinct classes: Lund string models [62] and cluster hadronization models [63]. Lund string models use colour links between partons: when the partons in a pair get further from each other, the amount of energy rises, until it is high enough to create a quark-antiquark pair. This breaks the link and creates two new links between the old particles and the new ones. The cluster hadronization models use quantum numbers conservation between partons and hadrons. First, parton clusters are created from the involved partons, neutral from a QCD point of view. Then, a cluster is identified as a hadron if its mass is close to a known hadron. If not, it is considered as a resonance of two lighter hadrons. The created particles can be stable or not. If they are not stable, these hadrons decay following the experimentally measured branching ratios. However, dedicated software based on these models, for example PYTHIA [64] and Herwig [65] allows to simulate the whole hadronization chain.

2.3.2 Subdetectors and interactions

Once the outgoing particles have been obtained, the response of the different subdetectors to the particles that are stable enough to reach them is simulated. This level of simulation includes the flight of the particles through the subdetectors, the electromagnetic and hadronic cascades, which are the interaction between particles and the matter of the subdetectors, the electrical response of the subdetectors, and the decay of the particles whose mean lifetime leads to their decay within the detector geometry. Two approaches are available, a detailed simulation based on GEANT4 [66], which is very computationally heavy, and a fast simulation, allowing for less precise but faster simulation, which will not be used here.

The full simulation relies on a very detailed description of the subdetectors. Indeed, even the cables and hardware material are included in this 3D simulation. This description is then used by the GEANT4 software, which simulates the propagations, interactions, scatterings and detection through each subdetectors.

2.4 Event reconstruction

At this point, the signals delivered by the different subdetectors, whether in simulation or data, is hard to interpret. In order to make sense of the detection on a particle physics level, an event reconstruction algorithm is applied. Its goal is to provide the list of outgoing particles from the detection data. Event reconstruction is done identically for both simulated events and real collision data. In the CMS experiment, particles often interact with several subdetectors, as illustrated in Figure 2.4. The reconstruction is therefore done by an algorithm specifically developed to optimally combine the information gathered by all subdetectors. This algorithm is called the particle-flow algorithm (PF).

The PF algorithm can be split into two distinct steps: detection elements are first reconstructed and then used to identify and reconstruct particles.

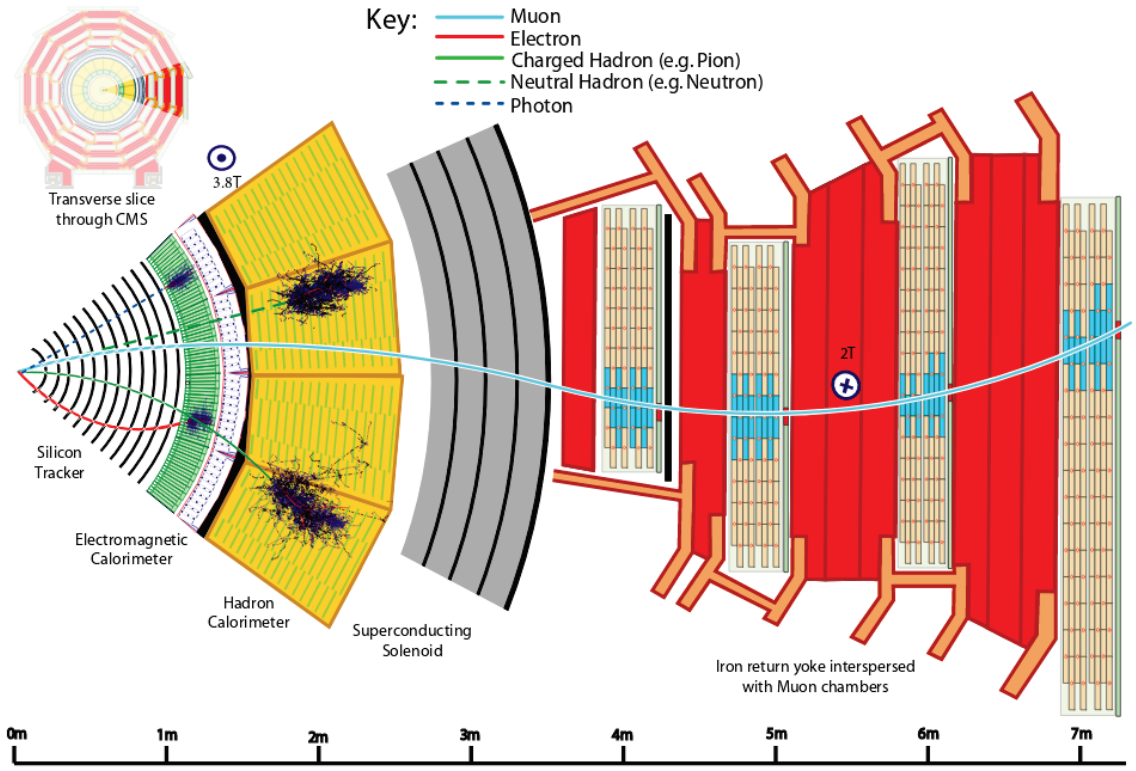


Figure 2.4: A sketch of the specific particle interactions in a transverse slice of the CMS detector, from the beam interaction region to the muon detector. The muon and the charged pion are positively charged, and the electron is negatively charged.

2.4.1 Particle-flow elements

Charged-particle tracks and vertices

An iterative process is first used to reconstruct tracks [67]. This is done using a combinatorial track finder based on Kalman Filtering (KF) that reconstructs tracks in three stages: initial seed generation with a few hits compatible with a charged-particle trajectory; trajectory building to gather hits from all tracker layers along this charged-particle trajectory; and final fitting to determine the charged-particle properties: origin, transverse momentum, charge and direction. This track reconstruction is then performed iteratively, the same operation being repeated several times with progressively more complex and time-consuming seeding, filtering, and tracking algorithms. The reduction of the misreconstruction is accomplished with quality criteria on the track seeds, on the track vertices, and on the track compatibility with the reconstructed primary vertices, adapted to the track p_{T}, $|\eta|$, and number of hits. The hits associated with the selected tracks are masked in order to reduce the probability of random hit-to-seed association in the next iteration. The remaining hits may thus be used in the next iteration to form new seeds and tracks with relaxed quality criteria, increasing in turn the total tracking efficiency without degrading the purity.

Nuclear interactions in the tracker material may lead either to a kink in the original hadron trajectory, or to the production of a number of secondary particles. A dedicated algorithm was thus developed to identify tracks linked to a common

secondary displaced vertex within the tracker volume [68, 69].

Tracking for electrons

Electron reconstruction, originally aimed at characterising energetic, well-isolated electrons, is naturally based on the ECAL measurements. More specifically, the traditional electron seeding strategy (hereafter called the ECAL-based approach) [70] makes use of energetic ECAL clusters ($E_{\mathrm{T}} > 4$GeV). The cluster energy and position are used to infer the position of the hits expected in the innermost tracker layers under the assumptions that the cluster is produced either by an electron or by a positron. Due to the significant tracker material thickness, most of the electrons emit a sizable fraction of their energy in the form of bremsstrahlung photons before reaching the ECAL. The performance of the method therefore depends on the ability to gather all the radiated energy, and only that energy. The energy of the electron and of possible bremsstrahlung photons is collected by grouping the ECAL clusters into a supercluster. The ECAL clusters are reconstructed in a small window in η and an extended window in ϕ around the electron direction (to account for the azimuthal bending of the electron in the magnetic field). However, for non-isolated electrons, such as those originating from QCD jets, the overlapping contributions from other particle deposits can lead to large inefficiencies. In addition, the backward propagation from the supercluster to the interaction region is likely to be compatible with many hits from other charged particles in the innermost tracker layers, causing a substantial misreconstruction rate.

A tracker-based approach, based on the iterative tracking, is designed to have a large efficiency for electrons that can be easily missed by the ECAL-based approach. All the tracks from the iterative tracking are therefore used as potential seeds for electrons, if their p_{T} exceeds 2 GeV. The large probability for electrons to radiate in the tracker material is exploited to disentangle electrons from charged hadrons. When the energy radiated by the electron is small, the corresponding track can be reconstructed across the whole tracker with a well-behaved χ^2 and be safely propagated to the ECAL inner surface, where it can be matched with the closest ECAL cluster as will be detailed later. For these tracks to form an electron seed, the ratio of the cluster energy to the track momentum is required to be compatible with unity. In the case of soft photon emission, the pattern recognition may still succeed in collecting most hits along the electron trajectory, but the track fit generally leads to a large χ^2 value. When energetic photons are radiated, the pattern recognition may be unable to accommodate the change in electron momentum, causing the track to be reconstructed with a small number of hits. A preselection based on the number of hits and the fit χ^2 is therefore applied and the selected tracks are fitted again with a Gaussian-sum filter (GSF) [71]. The GSF fitting is more adapted to electrons than the KF used in the iterative tracking, as it allows for sudden and substantial energy losses along the trajectory. A final requirement is applied to the score of a boosted-decision-tree (BDT) classifier that combines the discriminating power of the number of hits, the χ^2 of the GSF track fit and its ratio to that of the KF track fit, the energy lost along the GSF track, and the distance between the extrapolation of the track to the ECAL inner surface and the closest ECAL cluster. The tracker-based seeding is also effective at selecting electrons and positrons from

conversions in the tracker material, for both prompt and bremsstrahlung photons. The recovery of the converted photons of the latter category and their association to their parent electrons is instrumental in minimising energy double counting in the course of the PF reconstruction.

Tracking for muons

The muon spectrometer allows muons to be identified with high efficiency over the full muon detector acceptance. A high purity is granted by the upstream calorimeters, meant to absorb other SM particles (except neutrinos). The inner tracker provides a precise measurement of the momentum of these muons. The high-level muon physics objects are reconstructed in a multifaceted way, with the final collection being composed of three different muon types:

- standalone muon. Hits within each DT or CSC detector are clustered to form track segments, used as seeds for the pattern recognition in the muon spectrometer, to gather all DT, CSC, and RPC hits along the muon trajectory. The result of the final fitting is called a standalone-muon track.

- global muon. Each standalone-muon track is matched to a track in the inner tracker (hereafter referred to as an inner track) if the parameters of the two tracks propagated onto a common surface are compatible. The hits from the inner track and from the standalone-muon track are combined and fit to form a global-muon track.

- tracker muon. Each inner track with p_{T} larger than 0.5 GeV and a total momentum in excess of 2.5 GeV is extrapolated to the muon system. If at least one muon segment matches the extrapolated track, the inner track qualifies as a tracker muon track.

Charged hadrons may be misreconstructed as muons if some of the hadron shower remnants reach the muon system (punch-through). Different identification criteria can be applied to the muon tracks in order to obtain the desired balance between identification efficiency and purity.

Calorimeter clusters

The purpose of the clustering algorithm in the calorimeters is fourfold: detect and measure the energy and direction of stable neutral particles such as photons and neutral hadrons; separate these neutral particles from charged hadron energy deposits; reconstruct and identify electrons and all accompanying bremsstrahlung photons; and help the energy measurement of charged hadrons for which the track parameters were not determined accurately, which is the case for low-quality and high-p_{T} tracks. The clustering is performed separately in each concerned subdetector. The values of all parameters of the clustering algorithm result from optimizations based on the simulation of single photons, π^0, K^0_{L} and jets.

First, cluster seeds are identified as cells with an energy larger than a given seed threshold, and larger than the energy of the neighbouring cells. Second, topological

clusters are grown from the seeds by aggregating neighbouring cells above twice the noise threshold.

An expectation-maximisation algorithm based on a Gaussian-mixture model is then used to reconstruct the clusters within a topological cluster. The energy and position of the seeds are used as initial values for the parameters of the corresponding Gaussian functions and the expectation maximisation cycle is repeated until convergence.

Calorimeter cluster calibration

In the PF reconstruction algorithm, photons and neutral hadrons are reconstructed from calorimeter clusters. Calorimeter clusters separated from the extrapolated position of any charged-particle track in the calorimeters constitute a clear signature of neutral particles. On the other hand, neutral-particle energy deposits overlapping with charged-particle clusters can only be detected as calorimeter energy excesses with respect to the sum of the associated charged-particle momenta. An accurate calibration of the calorimeter response to photons and hadrons is instrumental in maximising the probability to identify these neutral particles while minimising the rate of misreconstructed energy excesses, and to get the right energy scale for all neutral particles. A first estimate of the absolute calibration of the ECAL response to electrons and photons, as well as of the cell-to-cell relative calibration, has been determined with test beam data, radioactive sources, and cosmic ray measurements, all of which were collected prior to the start of collision data taking. The ECAL calibration was then refined with collision data collected at $\sqrt{s} = 7$ and 8 TeV [72]. Hadrons generally deposit energy in both ECAL and HCAL. The initial calibration of the HCAL was realized with test beam data for 50 GeV charged pions not interacting in the ECAL, but the calorimeter response depends on the fraction of the shower energy deposited in the ECAL, and is not linear with energy. The ECAL and HCAL cluster energies therefore need to be substantially recalibrated to get an estimate of the true hadron energy. The calibrated calorimetric energy associated with a hadron is expressed as

$$E_{\text{calib}} = a + b(E)f(\eta)E_{\text{ECAL}} + c(E)g(\eta)E_{\text{HCAL}}\,, \tag{2.8}$$

where E_{ECAL} and E_{HCAL} are the calibrated energies measured in the ECAL and HCAL, and where E and η are the true energy and pseudorapidity of the hadron. To avoid the need for an accurate estimate of the true hadron energy E (which might not be available in real data), the constant a is chosen to minimise the dependence on E of the coefficients b and c. Isolated charged hadrons selected from early data recorded at $\sqrt{s} = 0.9$, 2.2, and 7 TeV have been used to check that the calibration coefficients determined from the simulation are adequate for real data.

2.4.2 Particle identification and reconstruction

Link algorithm

A given particle is, in general, expected to give rise to several PF elements in the various CMS subdetectors. The reconstruction of a particle therefore first proceeds

with a link algorithm that connects the PF elements from different subdetectors. The probability for the algorithm to link elements from one particle only is limited by the granularity of the various subdetectors and by the number of particles to resolve per unit of solid angle. The probability to link all elements of a given particle is mostly limited by the amount of material encountered upstream of the calorimeters and the muon detector, which may lead to trajectory kinks and to the creation of secondary particles. The link algorithm can test any pair of elements in the event. If two elements are found to be linked, the algorithm defines a distance between these two elements aimed at quantifying the quality of the link. The link algorithm then produces PF blocks of elements associated either by a direct link or by an indirect link through common elements. Tracks are linked with clusters if their extrapolated trajectory matches the position of considered clusters. Only the closest track-to-cluster links are kept. Calorimeter cluster-to-cluster links are sought between HCAL clusters and ECAL clusters by checking whether one cluster's position is within another's envelope. Charged-particle tracks may also be linked together through a common secondary vertex, for nuclear-interaction reconstruction. Finally, a link between a track in the central tracker and information in the muon detector is established to form global and tracker muons.

Muons

First, muon identification proceeds by a set of selections based on the global and tracker muon properties. Isolated global muons are selected by considering additional inner tracks and calorimeter energy deposits close to its trajectory to carry less than 10% of the muon p_{T}, which is sufficient to adequately reject hadrons that would be misidentified as muons. For non-isolated global muons, the tight-muon selection [73] is applied. The PF elements that make up these identified muons are masked against further processing in the corresponding PF block, i.e. they are not used as building elements for other particles.

Electrons and isolated photons

Electron reconstruction is based on combined information from the inner tracker and the calorimeters. Due to the large amount of material in the tracker, electrons often emit bremsstrahlung photons and photons often convert to $\mathrm{e}^{+}\mathrm{e}^{-}$ pairs, which in turn emit bremsstrahlung photons, etc. For this reason, the basic properties and the technical issues to be solved for the tracking and the energy deposition patterns of electrons and photons are similar. Isolated photon reconstruction is therefore conducted together with electron reconstruction. In a given PF block, an electron candidate is seeded from a GSF track, provided that the corresponding ECAL cluster is not linked to three or more additional tracks. A photon candidate is seeded from an ECAL supercluster with E_{T} larger than 10 GeV, with no link to a GSF track. For ECAL-based electron candidates and for photon candidates, the sum of the energies measured in the HCAL cells with a distance to the supercluster position smaller than 0.15 in the (η,ϕ) plane must not exceed 10% of the supercluster energy. The total energy of the collected ECAL clusters is corrected for the energy missed in the association process, with analytical functions of E and η. The final energy

assignment for electrons is obtained from a combination of the corrected ECAL energy with the momentum of the GSF track and the electron direction is chosen to be that of the GSF track. Electron candidates must satisfy additional identification criteria in the form of a BDT score and associated working points. This BDT takes up to fourteen variables of the electron candidate and is trained separately for ECAL barrel and endcaps, and for isolated and non-isolated electrons. Photon candidates are retained if they are isolated from other tracks and calorimeter clusters in the event, and if the ECAL cell energy distribution and the ratio between the HCAL and ECAL energies are compatible with those expected from a photon shower.

Hadrons and non-isolated photons

Once muons, electrons, and isolated photons are identified and removed from the PF blocks,the remaining particles to be identified are charged hadrons, neutral hadrons, non-isolated photons, and more rarely additional muons. The ECAL and HCAL clusters that are not linked to any track give rise to photons and neutral hadrons. Each of the remaining HCAL clusters of the PF block is linked to one or several tracks (not linked to any other HCAL cluster) and these tracks may in turn be linked to some of the remaining ECAL clusters (each linked to only one of the tracks). If the calibrated calorimetric energy is in excess of the sum of the track momenta by an amount larger than the expected calorimetric energy resolution for hadrons, the excess may be interpreted as the presence of photons and neutral hadrons. If the calibrated calorimetric energy is compatible with the sum of the track momenta, no neutral particle is identified. The charged-hadron momenta are redefined by a χ^2 fit of the measurements in the tracker and the calorimeters, which reduces to a weighted average if only one track is linked to the HCAL cluster. In rare cases, the calibrated calorimetric energy is significantly smaller than the sum of the track momenta. When the difference is larger than three standard deviations, a relaxed search for muons, which deposit little energy in the calorimeters, is performed.

2.4.3 High level objects

Jets

As explained in section 1.1.3, the quarks can only be observed as compound states with no colour-charge. When a quark is produced in one of the collision events, the strong force generates pairs of particles, and this leads to colour neutralness of the final objects. The shower of particles created in this hadronization process is called a jet.

QCD jets are reconstructed in the event from the list of previously created particles, using clustering algorithms. Several lists of reconstructed jets are defined from the method used to cluster the particles into jets. Sequential recombination jet algorithms is a class of clustering algorithms that is widely used in collider experiments. From the list of particles, they identify the pair of particles that are closest in a chosen distance metric, recombine them, and then repeat the procedure over and over, until some stopping criterion is reached, usually the maximum size of the cone associated with the constructed jet. The definition of the metric used to express

distance between particles and the maximum size of the cone are the two defining parameters of such an algorithm, as will be detailed in chapter 4. For example, in the CMS experiment, one of the most used definitions is the anti-k$_t$ metric with a distance parameter of 0.4. Jets are usually clustered on the subset of particles that are considered as coming from the primary vertex, to avoid contamination from the pileup events.

Several corrections are applied sequentially to the four-momenta of the jets. First, the expected pileup energy contribution is subtracted. This energy is estimated from the average pileup energy density in the event and the jet area. The next level of correction is done in simulation, by comparing the reconstructed p_{T} to the particle-level one. Following levels of correction are made to account for small differences between simulation and data, and can be specific to the analysis, and therefore the relevant parts for this thesis are presented in chapter 4.

Missing transverse momentum ($E_{\mathrm{T}}^{\mathbf{miss}}$)

Momentum is conserved in each direction of space in the collision. Since collisions happen between partons that have close to zero momentum in the transverse directions, the sum of the transverse momentum of all the products of the collision should be zero. However, due to the purely weak interaction of neutrinos, they cannot be detected by the CMS subdetectors. The missing transverse momentum, labeled $E_{\mathrm{T}}^{\mathrm{miss}}$, is defined as the module of the vectorial sum of the transverse momentum of all reconstructed objects involved in the event, which allows for an estimation of the presence of neutrinos. A good reconstruction of the $E_{\mathrm{T}}^{\mathrm{miss}}$ is crucial for the H$\rightarrow \tau\tau$ analyses, due to the presence of several neutrinos (two, three or four, depending on the final state) in the expected signal events. The $E_{\mathrm{T}}^{\mathrm{miss}}$ is usually adapted to the corrections applied to the other reconstructed objects in the event, as well as being itself corrected to account for difference in simulation and data. Since these corrections are analysis-based, they will be presented in chapter 4.

Hadronic τ decays (τ_h)

Contrarily to a τ lepton decaying leptonically, which is generally considered in analyses as a lepton and some $E_{\mathrm{T}}^{\mathrm{miss}}$ in the final state, hadronic tau decay products (τ_h) are reconstructed as single high-level objects. The τ_{h} decay products are expected to give rise to a reconstructed jet, therefore the τ_{h} identification starts from a collection of seeding jets. The standard τ_{h} identification algorithm, as well as a new identification procedure based on recursive neural networks created as part of this thesis, will be presented in chapter 3.

3

A recursive neural network for hadronic tau decay identification

As seen before in section 1.3.3, when a τ is produced in a collision, it can decay into several different final states. Hadronic final states, denoted τ_h, represent about 65% of tau decays. These hadronic final states are characterized by one or three charged hadrons with or without π^0. But similar particles can be reconstructed from other processes and decays, such as events involving QCD jets. Since these QCD jets greatly outnumber τ_h in the final states of proton-proton collisions at the LHC, τ_h identification algorithms have been designed in CMS to reject QCD jets as much as possible while keeping a τ_h identification efficiency somewhere between 35% and 70%, depending on the purity needed by the analyses. Identification algorithms classify all reconstructed jets as either background which means QCD jets, or signal which means τ_h decay products. The standard τ_h identification algorithm used in CMS, which will be detailed in section 3.1, has reached excellent performance, thanks to the use of particle-flow reconstruction, but do not make use of its full potential.

Deep learning algorithms have shown an ability to use available information as efficiently as possible in order to accomplish the task for which they are trained. In the field of particle physics, for example, their use in heavy flavour jet-tagging [74] has shown significant improvements over previously used techniques. These networks show best results when their design, or architecture, simplifies the task at hand. New architectures specifically intended for high energy proton-proton collisions have shown promising results. One such architecture is the Recursive Neural Network (RecNN) [75] used to identify boosted jets originating from hadronically decaying W bosons. Deep learning techniques will be introduced by level of complexity leading to the RecNN in section 3.2. This architecture has been adapted to τ_h identification, as it is a similar task. The adaptations of this architecture as well as some improvements that have been implemented will be presented in section 3.2.3, along with the reached performance.

3.1 The standard CMS hadronic τ decays identification

3.1.1 Decay mode finding

First, the particles of the reconstructed jet are fed as input to the hadrons-plus-strip (HPS) algorithm [76] to reconstruct and identify τ_h candidates. A first selection is the requirement that reconstructed jets have $p_T > 14\,\text{GeV}$ and $|\eta| < 2.5$. The constituent particles are combined into τ_h candidates compatible with one of the main τ decay modes, $\tau^- \to h^-\nu_\tau$, $\tau^- \to h^-\pi^0\nu_\tau$, $\tau^- \to h^-\pi^0\pi^0\nu_\tau$, $\tau^- \to h^-h^+h^-\nu_\tau$, and charge conjugates. The decay mode $\tau^- \to h^-h^+h^-\pi^0\nu_\tau$ is not considered, owing to its relatively small branching fraction and high contamination from quark and gluon jets.

The π^0 produced in some decay modes have a mean life of about $9 \times 10^{-17}\,\text{s}$ and about 99% of their decays lead to two photons. Because of the large amount of material in the tracker, illustrated in Figure 2.3, photons often convert before reaching the ECAL. The resulting electrons and positrons can be identified as such by the PF algorithm or, in the case their track is not reconstructed, as photons displaced along the ϕ direction because their trajectory is bent by the magnetic field. These neutral pions are therefore obtained by adding iteratively reconstructed photons and electrons located in a strip of size 0.05×0.20 in the (η,ϕ) plane:

- Every reconstructed electron and photon of $p_T > 0.5\,\text{GeV}$ in the strip is added iteratively from the highest to the lowest p_T.
- At each step, the position of the center of the strip is re-computed as a p_T-weighted average of the position of all constituents.
- Any electron or photon not included in an existing strip is used as seed to another new strip.

τ_h candidates are then formed by creating all combinations of either one or three charged-particles and up to two strips in the jet. Each τ_h candidate is then required to have a mass compatible with its decay mode and to have unit charge. Collimation of the products is ensured by requiring all charged hadrons and neutral pions to be within a circle of radius $\Delta R = (2.8\,\text{GeV})/p_T$ in the (η,ϕ), which is called the signal cone. The size of the signal cone is, however, not allowed to increase above 0.1 at low p_T, nor to decrease below 0.05 at high p_T. It decreases with p_T to account for the boost of the τ decay products. Finally, the highest p_T selected τ_h candidate in a given jet is retained. The four-momentum of the τ_h candidate is determined by adding up the four-momenta of its constituent particles.

3.1.2 Isolation

τ_h candidates reconstructed from QCD jets are likely to be surrounded by other particles coming from the jet. Isolation is therefore a powerful way to reject background QCD jets.

Two approaches are available : cut-based and multi-variate.

Cut-based Isolation is computed from the p_{T} sum of charged particles and photons with $p_{\mathrm{T}} > 0.5$ GeV within an isolation cone of $dR = 0.5$ centered around the τ_{h} direction, excluding particles used to form the τ_{h} candidate. In order to mitigate pileup contribution, tracks associated to the considered charged particles are required to be compatible with the τ_{h} production vertex within a distance $\Delta z < 0.2\,\mathrm{cm}$ and $\Delta r < 0.03\,\mathrm{cm}$ where r is the distance in the (x,y) plane. The contribution from pileup is estimated as $\Delta\beta$ calculated from the contribution of charged displaced particles and removed as

$$I_\tau = \sum_{charged,\Delta z<0.2cm} p_{\mathrm{T}} + max\Big\{0, \sum_{\gamma} p_{\mathrm{T}} - \Delta\beta\Big\}\,, \tag{3.1}$$

$$\Delta\beta = 0.46 \sum_{charged,\Delta z>0.2cm} p_{\mathrm{T}}\,. \tag{3.2}$$

Working points are defined by the thresholds on the value taken by the isolation defined in equation 3.1. Usual working points such as loose, medium, tight correspond to thresholds of 2.0, 1.0 and 0.8 GeV respectively.

Multi-variate This method is based on decision trees. These are machine learning techniques that rely on finding the best successive cuts on the available variables to separate signal and background in a training set. Boosting is a method of combining many weakly classifying trees into a strong classifier. A BDT is trained on an appropriate choice of isolation variables to give best separation between QCD jets and τ_{h}. The variables are:

- charged- and neutral-particle isolation sums defined as in equation3.1;
- the reconstructed decay mode;
- the transverse impact parameter d_0 of the leading tack of the τ_{h} candidate and its significance d_0/σ_{d_0};
- the distance between the τ decay and production vertices, $|\vec{r}_{SV} - \vec{r}_{PV}|$, and its significance $|\vec{r}_{SV} - \vec{r}_{PV}|/\sigma_{|\vec{r}_{SV}-\vec{r}_{PV}|}$, along with a flag indicating whether a decay vertex has successfully been reconstructed for a given τ_{h} candidate. The positions of the vertices, $\vec{r}_{SV}$ and $\vec{r}_{PV}$, are reconstructed using the adaptive vertex-fitter algorithm [77].

More details on the variables can be found in [76].

3.1.3 Anti-leptons discriminants

Electrons and muons can easily be misidentified as τ_{h}, particularly in the $h^{\pm}$ decay mode. Electrons radiating a bremsstrahlung photon that subsequently converts may also get reconstructed in the $h^{\pm}\pi^0$ decay mode. Therefore, discriminants have been developed to separate such lepton decays from real τ_{h} decay products.

Electrons are rejected by a BDT using observables that quantify the distribution in energy depositions in the ECAL and HCAL, in combination with observables

sensitive to the amount of bremsstrahlung emitted along the leading track. The BDT also uses observables related to the overall particle multiplicity, to distinguish electromagnetic from hadronic showers. All these variables are listed in [76].

τ_h candidates are rejected if no track segments are found in at least two muon stations within a cone of size $\Delta R = 0.3$ around its direction. τ_h candidates are also rejected when the sum of the energies in the ECAL and HCAL corresponds to less than 20% of the momentum of their leading track.

3.1.4 Simulation of QCD jets and hadronic τ decays

In order to compare both classification methods, their performance is expressed in terms of signal efficiency and background rejection. To quantify these, and to provide a training set for our deep learning algorithm, datasets of QCD jets and hadronic tau decays have been selected from the simulated CMS datasets that were introduced in 2.3.1. Instead of simulating isolated QCD jets and τ_h separately, entire collision events are used as they provide a realistic environment similar to the one in which the classifier will be used.

Selected processes leading to τ_h and QCD jets in the final state are detailed in table 3.1. QCD jets are obtained from QCD multijet events in which true τ_h are extremely rare, while τ_h are taken from events with true taus in the final state (MSSM $H \to \tau\tau$, DY $Z \to \tau\tau$).

The proton-proton collision events are generated with PYTHIA 8 [64], and are then processed by the CMS GEANT4 simulation, as detailed in section 2.3.1. The generation-level information is kept and will be referred to as gen level. All the information coming out of the detector simulation is fed to the CMS reconstruction algorithms, where the particle flow algorithm provides the list of reconstructed stable particles. Higher-level objects such as jets are reconstructed by combining these particles using the clustering algorithms introduced in 2.4.3.

As stated before, the identification of τ_h and QCD jets is performed jet by jet. A reconstructed jet is defined as signal if it is selected from the genuine tau samples, and as background if selected from the QCD multijet samples. To ensure purity, some extra cuts are applied, these cuts can be found in table 3.1.

The matching between reconstructed jets and gen level τ_h is done by ensuring their respective directions are aligned. This is done by requiring the distance separating their orientation in the (η, ϕ) plane to be $\Delta R < 0.1$.

Table 3.1: Provenance and cuts applied to reconstructed jets defining signal and background.

	τ_h (signal)	QCD jets (background)
Hard processes	SUSY ggH $\to \tau\tau$	QCD multijets samples ordered by p_T of jet (in GeV) : 15-30, 30-50, 50-80, 80-120, 120-170, 170-300
	SUSY bbH $\to \tau\tau$	
	DY $\to \tau\tau$	
phase-space cuts	$20 < p_T < 100$ GeV and $\lvert\eta\rvert < 0.8$	
specific extra cuts	matched with gen-level τ_h	any

Parts of these sets will also be used in the training of the deep learning algorithm. But biases coming from the difference between signal and background distribution

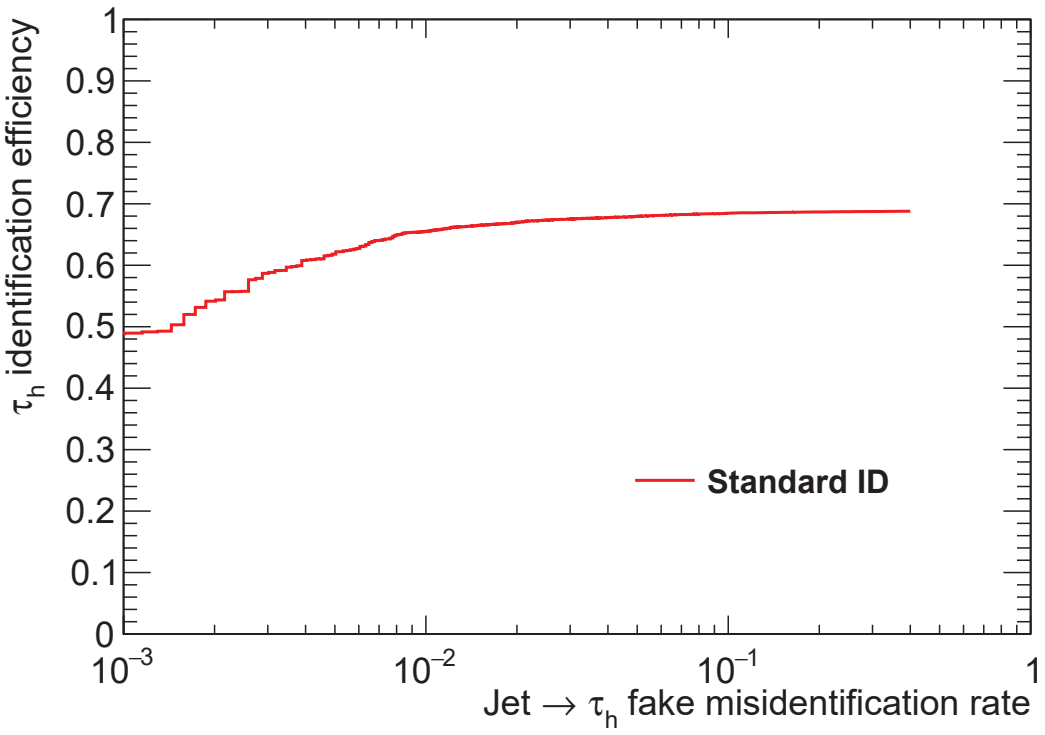


Figure 3.1: τ_h identification ROC curve for the standard method. The x axis is set to a logarithmic scale.

can cause mistraining. For example, if both background and signal samples had different p_T distribution, the network could wrongly assume that the p_T of a reconstructed jet is correlated to its nature as a τ_h or a QCD jet, which should be avoided. To prevent this bias, the signal and background jets are split by p_T bins, and the selection of signal and background jets is done in each p_T bin.

3.1.5 Performance

Performance is evaluated in terms of signal efficiency and background rejection. The signal efficiency is defined as the ratio of the number of well-tagged τ_h over the total number of τ_h in the selected population. The background rejection is similarly defined as the ratio of the number of well-tagged QCD jets over the total number of QCD jets in the population.

While the goal of a classifier is to tag objects as signal or background, most will instead provide a score between 0 and 1. A score close to 0 is to be interpreted as very likely to be background, and a score close to 1 as very likely to be signal. This continuous score can then be translated into a discrete tag by the choice of a working point (WP) value. For a given WP, each reconstructed jet that scores below the WP value is then tagged as QCD jet, and each reconstructed jet that scores higher is tagged as τ_h. The values of the signal efficiency and background rejection can be measured for a continuous scan of WP values. The created curve in the signal efficiency vs background rejection space is called a Receiver Operating Characteristic (ROC) curve. The standard identification ROC curve is shown in

Figure 3.1 A numerical figure of merit often used to quantify the overall performance of a classifier is the area under the ROC curve, called ROC AUC, as it is maximum when signal efficiency and background rejection is perfect. In our case, ROC AUC might not be the best figure of merit, as it does not take into account which regions are best covered by the technique. Indeed, in our case the standard identification technique is based on applying hard cuts such as decay mode finding and anti-lepton discriminant before scanning the score of the isolation BDT. This is why the ROC of the standard technique reaches a plateau, as the plateau corresponds to maximum efficiency allowed by the previous cuts.

QCD jets are also overwhelmingly more present in collisions than τ_{h}, which means useful working points are in the region of high background rejection.

3.1.6 Intrinsic limitations

The cut-based method relies on a single variable, namely isolation, to classify. The BDT-based method is an improvement on the cut-based method as it combines isolation with other variables susceptible to bring more information relevant to the classification process. But the BDT still takes a limited amount of information encoded into a strict number of variables. The construction of such variables does not take into account all the information gathered in the detection and reconstruction phases. A possible improvement should therefore be expected from using all the available information rather than a chosen subset. Deep learning techniques are conceptually adequate for such a task as they take all available information in and the processing of such information is then completely derived from training.

3.2 From a single neuron to recurrent networks

Indeed, neural networks have revolutionised fields such as big data, image recognition, and even pseudo-data generation. Neural networks are based on processing units called neurons. The combination of such neurons into networks allows a huge amount of information-processing possibilities. Those neural networks are then trained to give the desired output by a trial and error process on a set of examples, called the training set. The possibilities gained by the complexity of a neural network comes with a need for a bigger training set, which can easily be obtained by simulation.

The organisation of the neurons in the network is called architecture. Neural networks have shown their best achievements when their architecture is specifically designed for the task at hand.

3.2.1 Basics : neurons, dense networks, deep learning

Neuron

A neuron is defined by an activation function f, a set of scalar weights w_i, and a bias b. It takes as input in a number of variables denoted X_i. The layout of a neuron taking two inputs is illustrated in Figure 3.2. First, the weighted output z is defined

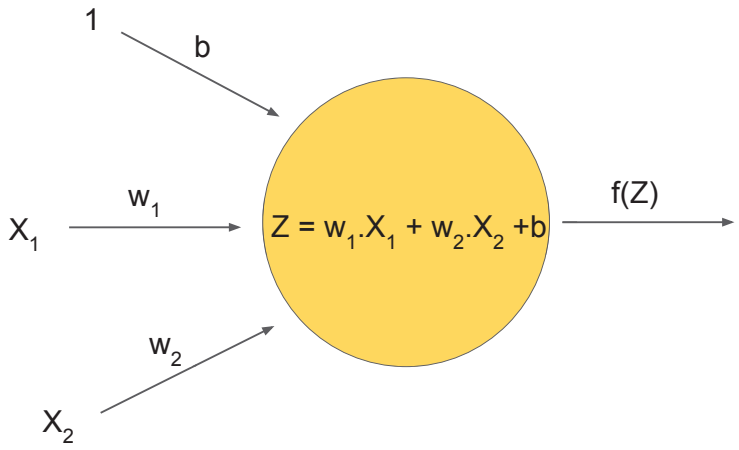


Figure 3.2: Diagram of a single neuron, in the case of two input variables X_1 and X_2. The output of the neuron is the value of the activation function f chosen for the neuron.

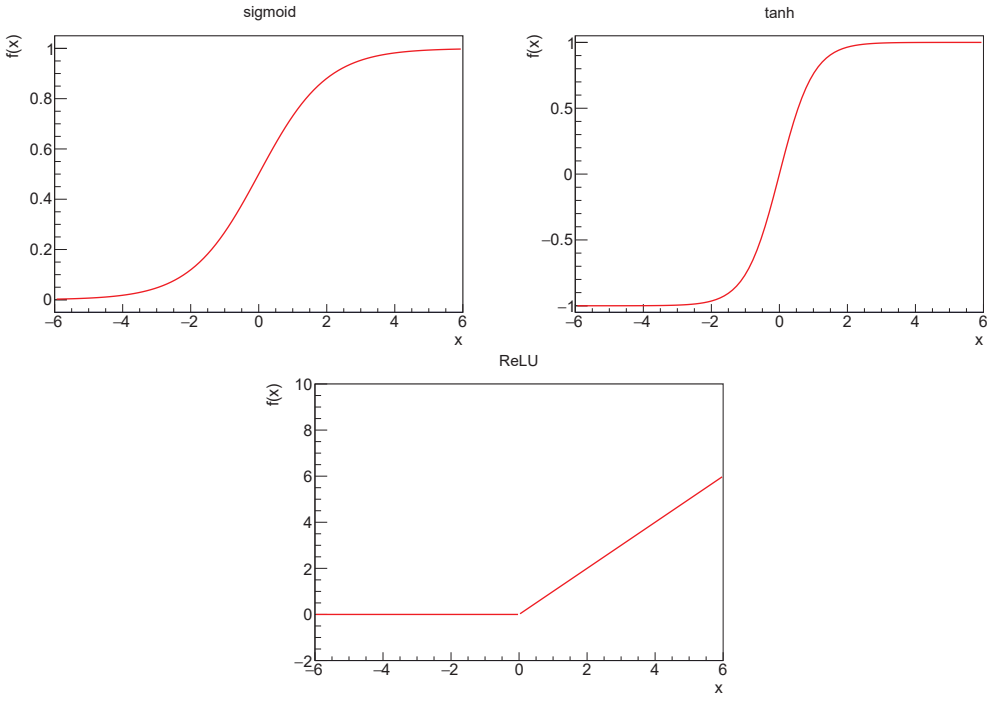


Figure 3.3: Visualization of the used activation functions.

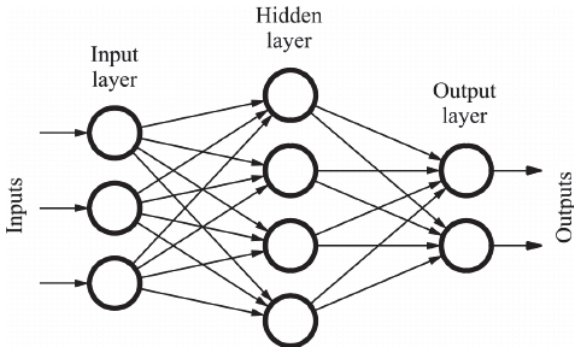


Figure 3.4: Diagram of an example of a feed-forward densely connected network with 3 input variables, 4 neurons in the hidden layer and 2 output neurons.

as $z = \sum w_i \times X_i + b$. The output of the neuron is then $f(z)$. While the activation function can be any nonlinear function, the different activation functions used here are the sigmoid function

$$f(x) = \frac{1}{1 + e^{-x}}\,, \tag{3.3}$$

the tanh function

$$f(x) = \frac{2}{1 + e^{-2x}} - 1\,, \tag{3.4}$$

and the ReLU function

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}\,. \tag{3.5}$$

These functions are displayed in Figure 3.3.

Densely connected network

A network is created from neurons by connecting their inputs and outputs. A simple case is the feed-forward densely connected network, such as the one shown in Figure 3.4. In this architecture, neurons are organised in layers. The output of each neuron in a given layer is sent to all neurons in the next layer. The inputs of the first layer are the input variables. The information then propagates through the evaluation of the neurons in each layer. The output of the last layer is then the output of the network for the given set of inputs, in other words the example data. An architecture is called feed-forward when there is no cycle in the propagation of the evaluation from inputs to outputs. Such an architecture creates the possibility of theoretically approximating any function of the inputs, as stated by the universal approximation theorem [78], as long as the activation functions used in the neurons are nonlinear.

Training: loss function and backpropagation

In order to find the set of weights and biases that allows the network to perform the desired task, the network is trained. As previously mentioned, the training phase relies on a set of examples of inputs associated with their target output value. The training is done iteratively. The first step is comparing the output of the network for a set of input variables to the desired target output. This comparison is quantified

through the use of a metric called the loss function. The loss function is required to be a differentiable function and is designed to reach a minimum when the output is equal to the target value. Training the network to perform the task is therefore equivalent to finding the network parameters (weights and biases) that minimise the value of the loss function for the whole training set, provided the training set is a representative sample of the global population.

But the space of configurations for the weights and biases of the network has a huge dimensionality. The iterative process of training the network case by case can then lead to stagnation if examples are evaluated and parameters adapted for each example of the training set. To avoid this stagnation, the parameters are changed to minimise the average of the loss function over a number of examples, called a mini-batch. The number of examples in each mini-batch is referred to as the mini-batch size.

The way the parameters are changed to minimise the loss function depends on which optimizer algorithm is used. Most optimizers rely on backpropagation, which means the variation that a parameter should undergo is computed by propagating the change of the loss function backwards through all the layers of neurons between the output of the network and the considered neuron [79].

A classical problem that can occur in the training of neural networks is the existence of local minima of the loss function. Indeed, local minima can lead to a sub-optimal training, as it prevents the network from reaching a potentially lower global minimum. To mitigate such effects, diminishing learning rates as well as momentum-based optimizers are used. The learning rate is a simple scalar that multiplies the changes in parameters for a given change in loss. By starting at a high value of this learning rate, it is possible to avoid local minima that are too small. After several iterations of training, this rate can be lowered to help reaching the lowest point of the minimum. Momentum-based optimizers try to avoid minimums by multiplying the learning rate by a factor proportional to the gain of the last step. Indeed, the more a training step helped to minimise the loss function, the bigger the next step, avoiding local minima on the way to a global minimum.

Backpropagation comes with another important problem called vanishing gradients. This is due to the change of output of a neuron being relatively small compared to change of its input. Indeed, an activation function such as the sigmoid, illustrated in Figure 3.3, can lead to states where the derivative is close to 0. But to compute the update to a given parameter for a given neuron, backpropagation uses the derivatives of the activation functions of the neurons in the layers between the neuron and the output as factors. Thus, the more layers there are between a neuron and the final layer, the more likely it is that its parameters will update based on a derivative that is close to 0. In other words, this means that a change in the parameters of a neuron in an early layer will have a relatively small effect on the loss compared to a similar change in the last layers. This leads to a slower training of the early layers compared to the training of the last layers. Therefore, a very deep network will explore the space of its parameters very slowly compared to a shallow network. This can even lead to a stagnation of the overall performance, which means the network does not gain anything from training. This is mitigated by the use of an activation function such as the ReLU. This activation function mitigates

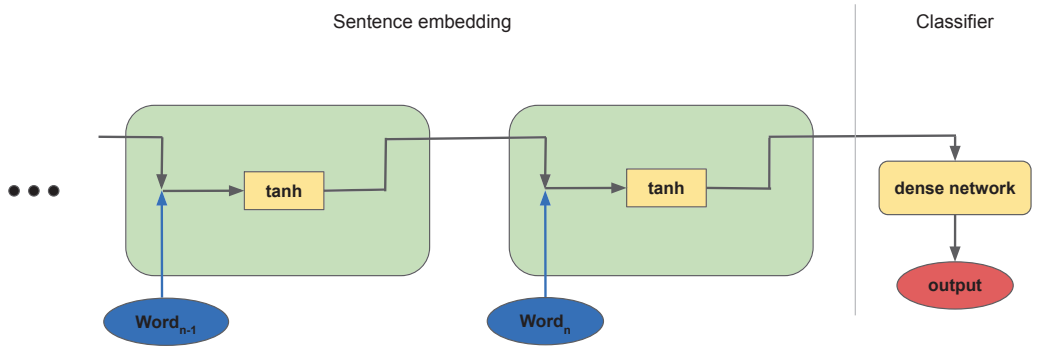


Figure 3.5: Diagram of the iterative evaluation of a recurrent neural network in the language processing case. Only the last two iterations corresponding to the two last words of a sentence are represented. The green boxes represent several iterations of the same unit. The yellow rectangular boxes represent a single neuron layer and the rounded yellow box represents a dense network, which can be made of several layers.

the small derivative issue by having a much larger domain where the derivative is not close to zero. The use of cross-entropy as a loss function also allows to mitigate the vanishing gradients problem. Cross-entropy is defined as

$$f(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \,. \tag{3.6}$$

where y is the target output and $\hat{y}$ the output of the network. Indeed, the derivative of sigmoid activation functions is cancelled out in the computation of the derivative of the cross-entropy function with respect to a parameter of the network [80]. One can also mitigate this problem further by designing an architectural workaround. For instance, the network can be designed to be less deep in practice, as is the case for recurrent neural networks, which are introduced in the following section.

3.2.2 Recurrent neural networks

The dense neural network that has been introduced is not well suited to our case. In order to use all the information available in jets, all the characteristics of the particles of each jet must be fed into the network. But each jet has a different number of particles, and a dense network cannot accommodate a varying number of inputs. Recurrent neural networks (RNN) are designed to do just that. They have been particularly useful in language processing tasks. Indeed, language processing has a varying number of input requirements, as the number of words changes in different sentences. RNNs will be introduced in this section in the case of an example task of classifying sentences as positive or negative.

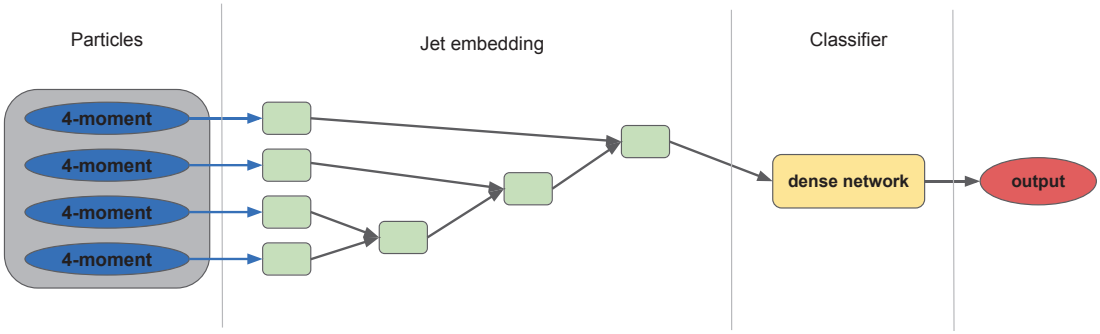


Figure 3.6: Diagram of the overall structure of the RecNN, including the jet embedding structure. The order in which the nodes are merged is determined by the chosen jet clustering metric. Each green box represents a node. Each node is evaluated with the same set of layer parameters. The black arrows represent the flow of both the 4-momentum and the embedding array. These arrows therefore represent both black and blue arrows from the node diagrams.

Every RNN is composed of two distinct parts, as illustrated in Figure 3.5. The first part has the goal of embedding the information gathered from the inputs into a fixed-size array of values, called the embedding array. The second part is a dense neural network that takes the elements of this array as input, and its output is considered the output of the RNN as a whole. The embedding part consists in applying a unit iteratively on each input element, in this case each word. The output of the previous iteration is also taken as a secondary input. The layout of a unit in terms of network layers and information flow then defines a specific RNN architecture.

While this architecture has the advantage of being able to accommodate inputs of varying size, it also benefits from the fact that the same embedding unit is applied at each iteration, with the same parameters. This mitigates the vanishing gradient problem, as all parameters of the network are evaluated close to the final layer. Conceptually, this allows the use of a small amount of layers by taking advantage of the symmetries of the inputs.

3.2.3 Recursive neural networks

Just as RNNs were designed to accommodate the needs of language processing, another architecture type has been designed to fit the needs of jet classification. This architecture type, called the recursive neural network (RecNN), was developed following the idea behind jet clustering. This architecture was first originally applied

to the problem of boosted W-jet tagging [75].

3.2.4 Recursive architecture

As in a RNN, the network architecture is made of two parts, a jet embedding part and a final classifier consisting of a dense neural network. Contrarily to the linear structure of the sentence embedding, the embedding part is organised as a binary tree that represents the iterations of the jet clustering algorithm. The iterations of base units in a RecNN are called nodes. There are two types of nodes, namely the leaf nodes and the inner nodes. The leaf nodes are designed to each take the 4-momentum of an input particle, as illustrated in Figure 3.6. Inner nodes take the output of two other nodes, leaf or inner, as input. First, a criterion is evaluated on every pair of nodes, and the pair that best fulfills this criterion is selected. This criterion can be chosen among the following:

- randomised : two nodes are selected at random;
- p_{T}-ordered : nodes holding the two pseudo-jets with the highest p_{T};
- reversed p_{T}-ordered : nodes holding the two pseudo-jets with the lowest p_{T};
- k_t : nodes holding the closest pseudo-jets following the k_t clustering metric;
- Cambridge : nodes holding the closest pseudo-jets following the Cambridge clustering metric;
- anti-k_t : nodes holding the closest pseudo-jets following the anti-k_t metric.

The distance of two particles i and j in a clustering metric is defined as

$$d_{ij} = \min(p_{Ti}^{2k}, p_{Tj}^{2k}) \frac{\Delta_{ij}}{\mathrm{R}} \,. \tag{3.7}$$

In this expression Δ_{ij} corresponds to the distance of the pseudo-jets in the (η,ϕ) plane, R is a distance parameter, and p_{Ti} corresponds to the modulus of the transverse momentum of particle i. The k parameter is equal to 1 for the kt metric, 0 for the Cambridge metric, and -1 for the anti-kt metric.

Then, the output of the two chosen nodes are fed into an inner node. The 4-momentum associated with this new node is the sum of the 4-momenta of the two previous nodes, effectively creating a new pseudo-jet. The list of nodes is updated by removing the two merged nodes and adding the new inner node. The criterion is computed again for all pair combinations and the process is repeated until only one node remains. The output of the final node is finally fed to the classifier part to provide a final output.

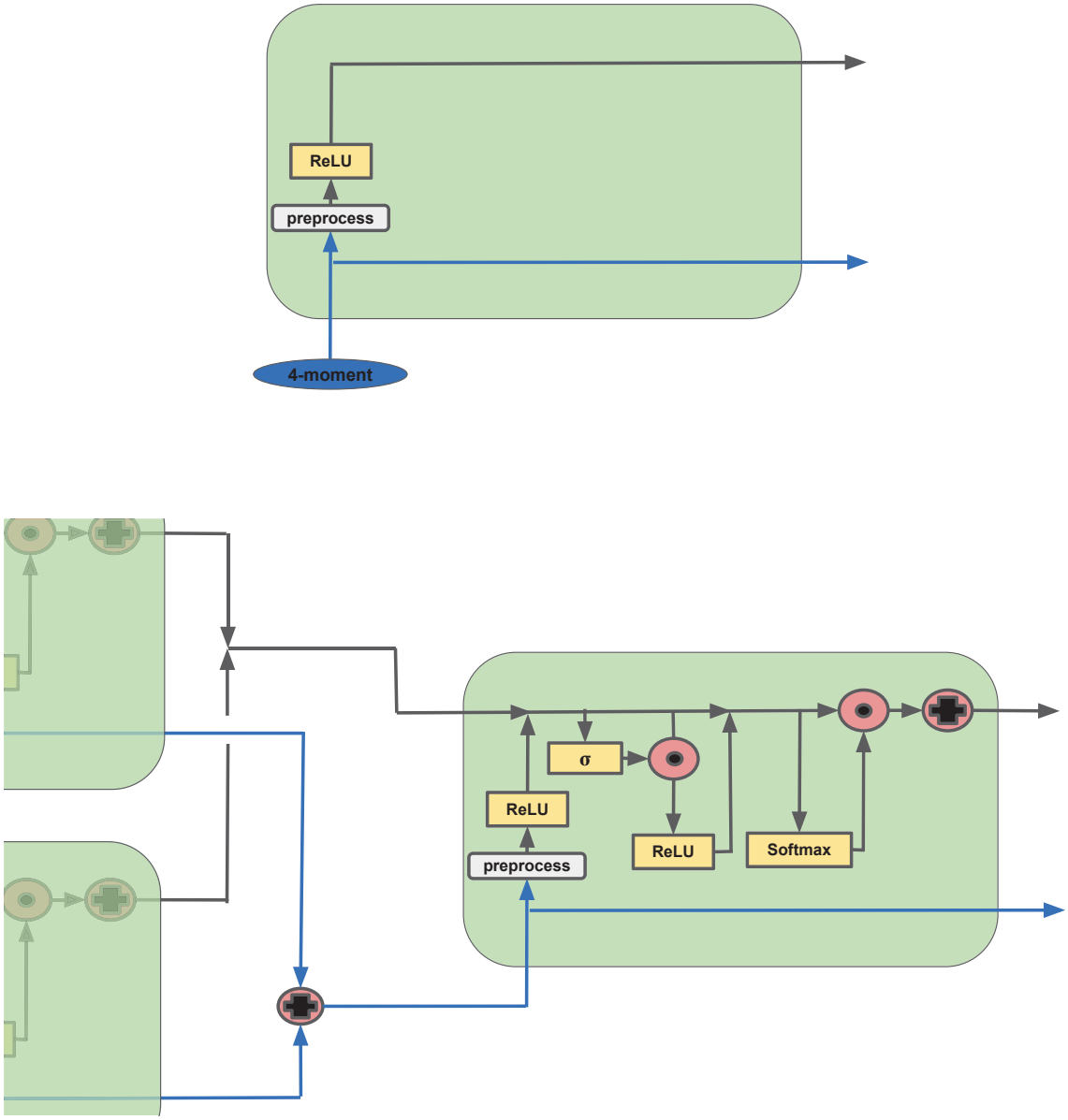


Figure 3.7: Diagrams of the nodes that can be found in a RecNN architecture. The top diagram represents a leaf node, which directly takes a particle 4-momentum as only input. The bottom diagram illustrates the inner structure of a node as well as how its input is taken from the previous nodes. The green boxes represent several iterations of the same unit. The yellow rectangular boxes represent a layer of neurons, and their names correspond to the activation function of the neurons in that layer. The blue arrows represent the path of 4-momentum merging at each iteration. The black arrows represent the path of the embedding arrays. The plus and dot signs in the red circles represent element-wise sum and element-wise product, respectively. The white box represents the preprocessing step that involves the transformation into cylindrical coordinates, as well as the scaling of each variable.

Node composition: gating

Similarly to a RNN, the role of an inner(leaf) node is to modify(create) an embedding array. An inner node layout therefore takes two inputs, namely the 4-momentum of a pseudo-jet and an embedding array. While the 4-momentum is then directly given as output for the next node, the embedding array will be modified through the use of one or several neuron layers. Thus, the information effectively takes two parallel paths. Indeed, a first path is defined by the propagation of the 4-momenta, while a second path is defined by the merging and modifications of embedding representation arrays.

While several node layouts are tested in [75], the gated layout has the best results and is the basis of our version. Inspired by the long short-term memory (LSTM) architecture [81] and gated recurrent unit (GRU) architectures [82], the gating conceptually adds the possibility for the network to select and mix available information more easily. While a mathematical formulation of this layout is available in the appendix of this article [75], the following explanation relies on the illustration in Figure 3.7.

First, each leaf node takes the 4-momentum of a particle and directly propagates it as its secondary output. A copy of this 4-momentum is preprocessed in two distinct steps. In the first step, the 4-momentum expression is changed from the easily addable Cartesian coordinates, to the more interpretable spherical coordinates. A few more variables are also computed at this step, like the modulus of the momentum p and the mass m. Each of those variables are then scaled by the robustScaler method [83] trained on the whole dataset. The goal of this scaling is to make the distributions of all variables comparable in terms of median and quantiles. Conceptually, this avoids the following layers to be forced to learn the scales of each variable.

The output of this preprocessing step is given as input to a first ReLU layer. The output array of this layer, designed to be of the same shape as each of the two input embedding arrays, is concatenated with these arrays. A second ReLU layer is evaluated by combining the information in this new embedding array using a reset gate. Conceptually, the goal of this reset gate is to actively select which parts of the embedding array are to be emphasised, and which parts are to be forgotten. This reset gate is implemented by evaluating a sigmoid layer, which takes the embedding array as input and creates an output array of the same shape, with values between 0 and 1. An element-wise product between the output array and a copy of the embedding array is then computed before feeding the produced array into the second ReLU layer. The output of this second ReLU layer, designed to have the same shape as the input embedding arrays, is then also concatenated to the new embedding array.

At this point the embedding array consists of the concatenation of four different arrays of the same shape: two embedding array outputs from the previous nodes, a local evaluation of the pseudo-jet, and an array produced from the combination of all the previous ones. In order to combine all those information into an array of the desired shape, a softmax gate is applied. This softmax gate is designed to weight each variable of the embedded array, before the four concatenated arrays are added element-wise. These weights are determined by a softmax neuron layer. The output

weights w_i of this softmax layer are determined from the activation of each output neuron Z_i by

$$w_i = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}} . \tag{3.8}$$

The product of this softmax gate is then the embedding array output of this node. If the node is the final node, this output is then directly fed into the dense neural network classifier part of the architecture.

Jet centering

The position of the jet in the detector impacts all the variables, while having little to do with the identification process. To avoid adding the task of learning this fact, the reconstructed jets are centered around the highest p_{T} particle. After centering particles appear to the network as orientated around the center (0,0) of the (η,ϕ) plane. The jet is then re-clustered into three subjets. Jets are finally rotated and if needed mirrored so that the positions of the re-clustered subjets are similar in every jet.

3.3 Recursive neural network for hadronic tau decay identification

3.3.1 Implementation

While most neural network architectures can easily be implemented using widely available libraries like Keras [84], the RecNN architecture cannot be implemented from these libraries, mainly because of its changing structure that adapts to each jet. The RecNN architecture was therefore implemented by the authors of Ref. [75] with classes from the scikit-learn library [85] as a base, which has been the basis for our implementation.

The first difference brought by our implementation is purely an optimisation of the code. While the code was already designed to run in parallel on several cores at the evaluation and training phases, both the preprocessing and centering steps were not. By adapting those steps and running the computation in parallel the time needed for this part was reduced from several hours to about 15 min using 20 cores. Time was also gained by changing the format under which the arrays were saved on disk.

In order to compare the standard and RecNN methods and to study score distributions on population subsets, the code was adapted to be able to track jets through the evaluation process. Indeed, in the original code the formatting of particles into the arrays used as input to the RecNN meant the loss of its link with any other information, such as the score of the jet with the standard technique, or the gen-level information. This tracking also allowed to build a display of jets with their associated gen-level information, allowing a case-by-case study.

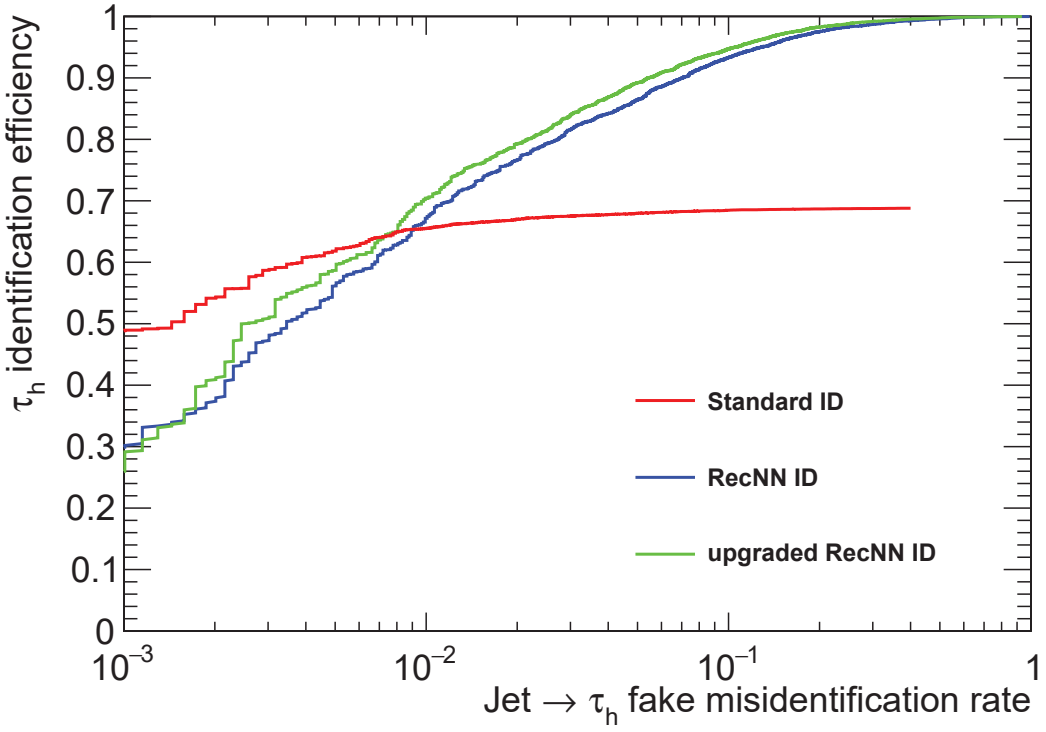


Figure 3.8: τ_h identification ROC curve for the standard method, the RecNN method, and the upgraded RecNN method. The x axis is set to a logarithmic scale.

3.3.2 Upgrade

In the original implementation, particles were described only by their 4-momentum. But the nature of those particles as well as other available information was not used. In order to add this information, the 4-momentum was changed to an array holding extra information. Just like the 4-momentum, the extra information must consist of variables that can be added when merging two nodes.

The first variables added were the energy and p_T contributions to the pseudo-jet from each particle type, namely photons, electrons, muons, neutral hadrons, charged hadrons from the primary vertex, and charged hadrons from pileup. Charged hadrons are considered as coming from the primary vertex if the distance between the primary vertex and the closest point of their reconstructed trajectory is found to be less than 0.2 cm away along the z-axis.

As the number of particles of each type is relevant to identify the different τ_h decay modes, the total number of each particle type in a given pseudo-jet was also added.

3.3.3 Performance

The ROC curves of both the standard method, the base RecNN implementation and the upgraded RecNN approach are presented in Figure 3.8. While the area under the curve is strictly better in the RecNN approaches compared to the standard method, the efficiency of the standard method is still better at low jet misidentification rate.

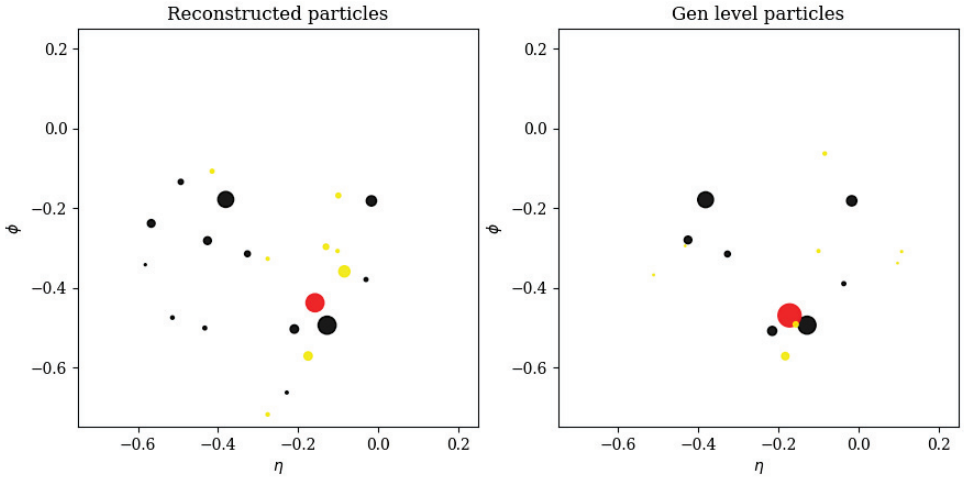


Figure 3.9: Scatter plots of the particle constituents in a QCD jet misidentified by the RecNN network. Left: reconstructed jet. Right: gen-level jet. The size of the points are proportional to the p_{T} of the particles and their colour depends on their type. Black points represent charged hadrons, yellow represent photons and red represent neutral hadrons.

The gain from the upgrade of the RecNN, while significant, is less than expected.

3.3.4 Possible optimisations

Although the results have shown some potential, this study has not reached the level of optimisation that could help the RecNN to outperform the standard technique systematically. Indeed, several upgrades that could benefit the RecNN approach have been considered, but were not fully implemented in time.

Even while limiting the computational times as much as possible, the latest versions of the RecNN network have proven to take a long time to train, about two days on a dedicated machine with 40 cores. All the clustering orderings have shown similar results in several stages of optimisation. While all should eventually be optimised and tested, the presented results have been produced with the anti-k_t ordering.

One such upgrade idea came from the study of QCD jets misidentified by the RecNN. Indeed, many examples such as the QCD jet mistagged as a τ_{h} by the RecNN network displayed in Figure 3.9 show cases where the RecNN should be able to classify such QCD jets as background from the number of reconstructed charged hadrons. These observations imply that the RecNN approach does not use the number of particles as efficiently as it should. An upgrade would be to directly add the number of particles per type at the classifier-level, rather than at the jet embedding level. This could help the network to reject trivial cases, while being able to specialise the jet-embedding part for the less straight-forward cases.

In our training sample, a majority of QCD jets can be trivially rejected from the number or composition of reconstructed particles. This implies that most training cases are not useful for the network. Conceptually, this means that the only useful training is done on a subset of examples, while the other training cases only contribute to an effect called overtraining. This effect appears when the networks

learns characteristics that are specific to the training sample, but those characteristics are not generalisable. To avoid this overtraining, the training phase is stopped when performances measured on an independent sample start getting worse as the training continues. A possible change to be tested could be to select more QCD jets that are similar to τ_h products in the training sample, and select less trivial cases.

While all these optimisations were attempted, they were not implemented in time. The RecNN approach has already shown good performances, and while the implemented upgrades have improved these performances, further improvements are expected to be within reach, as the case-by-case study shows. The potential optimisations that have been discussed could help reach a better performance.

4

Search for a MSSM heavy Higgs boson

This chapter describes the search for a neutral MSSM Higgs boson decaying to a pair of τ. The studied final state is defined by the presence of a pair of reconstructed τ decaying hadronically, and this channel is thus named $\tau_\mathrm{h}\,\tau_\mathrm{h}$. The Standard Model Higgs boson decay into a pair of τ leptons has been discovered by both the CMS and ATLAS collaborations [86, 87]. In the context of the MSSM, three neutral Higgs bosons are predicted: the CP-even states h and H and the CP-odd state A, which may all decay to τ pairs. A model-independent search for a single Higgs boson, denoted ϕ, is performed in a mass range of 110 to 2900 GeV. The analysis is sensitive to production via gluon-gluon fusion and production in association with b-quarks. The cross-section of the latter increases for larger values of tanβ due to the enhanced down-type fermion Yukawa couplings. The search is also performed in the $m_A - \tan\beta$ parameter space of the m_h^{max} scenario [88] for the same mass range.

Similar searches for MSSM neutral Higgs bosons have previously been performed by the collaborations at LEP [89], the Tevatron [90], and at LHC by the CMS and ATLAS collaborations [91, 92] with no excess observed above the background expectation.

The results presented here follow those published by CMS in 2018 using 2016 data, but make use of the new 2017 data for the first time while being restricted to the $\tau_\mathrm{h}\,\tau_\mathrm{h}$ channel. The 2017 data corresponds to an integrated luminosity of 41.5 fb^{-1}. Event categorisation is used to enhance sensitivity to particular production modes. The production of neutrinos in the τ decays makes it difficult to reconstruct the invariant mass of the candidate Higgs boson. Statistical inference is therefore performed on the distribution of the $m_\mathrm{T}^\mathrm{tot}$ variable, designed to account for $E_\mathrm{T}^\mathrm{miss}$ in the mass estimation, thus improving signal to background separation.

Section 4.1 outlines the datasets of recorded collisions, defined by which trigger these collisions fired, and the simulation used to estimate the contribution of some of the background processes. The event selection, which is done in a framework partially developed for this purpose is then detailed in section 4.2. The estimation of the contribution of each background process, using data-driven methods where possible, is detailed in section 4.3 and followed by a summary of the experimental and theoretical uncertainties affecting the signal and background estimations in

section 4.5. Validation checks of the used techniques were performed on the resulting distributions, as presented in section 4.6. The statistical procedure used to quantify the presence of signal in the data is given in section 4.7 and is followed by the results of the search in section 4.8.

4.1 Data samples and simulation

4.1.1 Trigger

Collision datasets are defined by the trigger pattern that fired their recording. At both the level one (L1) trigger and the high level trigger (HLT), the trigger patterns are defined by requirements of reconstructed objects. The definitions of these trigger objects vary from level to level, and the complexity of the reconstruction techniques increases at each level, and the rate at which events must be treated lowers. Therefore, the object properties determined in the trigger reconstruction, such as p_{T} and isolation, are only approximate to those in the full reconstruction.

Events are first selected at the L1 trigger level by an algorithm requiring either one L1 τ_{h} of $p_{\mathrm{T}} > 70\,\mathrm{GeV}$ and $|\eta| < 2.1$, or two L1 τ_{h} of $p_{\mathrm{T}} > 28\,\mathrm{GeV}$ and $|\eta| < 2.1$. At the HLT level, the di-τ_{h} triggers require two HLT τ_{h} objects to be isolated, to be identified, not to overlap, and to each have $p_{\mathrm{T}} > 35\,\mathrm{GeV}$ and $|\eta| < 2.1$. In the analysis, the full reconstruction τ_{h} candidates are required to have $p_{\mathrm{T}} > 40\,\mathrm{GeV}$. Such requirements are referred to as offline selections. Since the reconstruction at trigger level is not as complete as the full reconstruction, events that pass the offline selections can be rejected at the HLT level. The trigger efficiency with respect to the offline selections typically plateaus at 85% for large τ_{h} p_{T} [2]. At the p_{T} threshold applied in the analysis, it is around 60%.

4.1.2 Trigger optimisation

In order to maximise the trigger efficiency for our analysis, asymmetric p_{T} thresholds have been considered at the HLT level. The main goal of this study was to find a set of p_{T} thresholds that would improve efficiency without increasing the trigger rate. In order to estimate the efficiency gain, a new trigger pattern similar to the classical double τ_{h} trigger but without any p_{T} requirements was implemented. It was then applied to simulated datasets of $H \to \tau\tau$ and DY $Z \to \tau\tau$ events. The p_{T} requirements on the trigger-level objects were then applied after the full reconstruction on the HLT level τ_{h} trigger objects to simulate the use of new p_{T} thresholds, and the efficiency was calculated as

$$\epsilon = \frac{n_{\mathrm{trigger}}}{n_{\mathrm{selection}}}\,. \tag{4.1}$$

In this expression, $n_{selection}$ is the number of events that pass the analysis selection. This selection requires that two offline reconstructed τ_{h} are found in the event, and that both τ_{h} pass the tight WP of the BDT-based identification criterion, which was defined in chapter 3. The analysis selection also requires the kinematic cuts of $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ and $|\eta| < 2.1$. In the same expression, n_{trigger} is the number of these

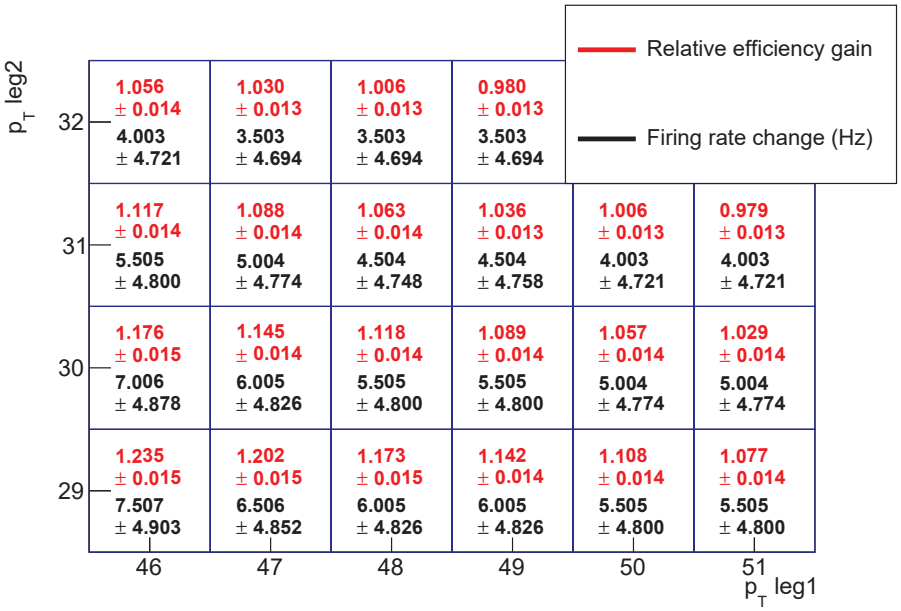


Figure 4.1: Values of the relative efficiency gain and the firing rate increase for different values of p_{T} thresholds. The displayed uncertainties are only statistical. The complete studied range was 35 to 60 GeV for the leading HLT τ_{h} object and 20 to 40 GeV for the sub-leading, but only the region that was found to be most interesting in terms of efficiency gain and rate increase is displayed here.

events that pass also the trigger p_{T} requirements. This efficiency is then computed for different values of p_{T} requirements for each of the two τ_{h} trigger objects.

To estimate the firing rate in real collisions, the trigger pattern without p_{T} requirement was then applied to randomly selected collision events, to avoid introducing bias from the use of a trigger. The rates were then computed from the random selection rate as

$$f = \frac{n_{\mathrm{pass}}}{n_{\mathrm{total}}} \times f_{\mathrm{unbiased}} \,. \tag{4.2}$$

In this expression, n_{pass} is the number of events that pass the trigger threshold, n_{total} is the total number of events of the unbiased sample, and $\mathrm{f_{unbiased}}$ is the rate of random selection used to produce the unbiased sample.

Some of the results are shown in figure 4.1. The statistical uncertainties on the measured rates are very large due to the very low number of events passing the offline selection in the randomly selected datasets. This study showed a gain of up to 6% in efficiency through the use of asymmetric trigger thresholds, while the rate was kept unchanged within the statistical uncertainties. But before more unbiased datasets were used to lower the uncertainties, a $Z \to \tau\tau$ polarisation analysis showed that the use of such a trigger pattern could reduce their acceptance of about 20%. The use of an asymmetric τ_{h} trigger p_{T} threshold was therefore not implemented by the CMS collaboration.

4.1.3 Simulation

In the analysis, several Monte Carlo (MC) generators are employed to produce simulated samples of signal and background events. The MADGRAPH [59] matrix element generator is used for Z+jets, W+jets, $t\bar{t}$+jets and diboson production. The POWHEG [60] generator is used for single top-quark production. The SM gluon-gluon fusion and VBF production modes of the Higgs boson, treated as a background in this analysis, are also simulated with POWHEG at NLO precision. Both ggH and bbH MSSM signal production modes are provided by PYTHIA [64]. All samples utilise PYTHIA for parton showering and hadronisation, and TAUOLA [93] for tau decays.

4.2 Analysis sequence

This section describes the event-based analysis sequence. The goals of this sequence are applying an event selection, applying corrections to the reconstructed physics objects and deriving quantities such as weights and physical variables like $m_{\mathrm{T}}^{\mathrm{tot}}$, the final discriminating variable. The interpretation will be made from the distribution of this variable, defined as

$$m_{\mathrm{T}}^{\mathrm{tot}} = \sqrt{m_{\mathrm{T}}^2(\tau_{\mathrm{h}}^{(1)}, E_{\mathrm{T}}^{\mathrm{miss}}) + m_{\mathrm{T}}^2(\tau_{\mathrm{h}}^{(2)}, E_{\mathrm{T}}^{\mathrm{miss}}) + m_{\mathrm{T}}^2(\tau_{\mathrm{h}}^{(1)}, \tau_{\mathrm{h}}^{(2)})} \tag{4.3}$$

where

$$m_{\mathrm{T}}(x, y) = \sqrt{2 \times p_{\mathrm{T}}^x \times p_{\mathrm{T}}^y \times (1 - \cos(\Delta\phi_{x,y}))}\,. \tag{4.4}$$

In these expressions, $\tau_{\mathrm{h}}^{(1)}$ and $\tau_{\mathrm{h}}^{(2)}$ are the offline τ_{h} candidates forming the selected pair. The τ_{h} pair selection process is detailed later in this section. The variables p_{T}^x and p_{T}^y are the p_{T} of objects x and y respectively, and $\Delta\phi_{x,y}$ is the angle between the projections of the 4-momentum of objects x and y in the transverse plane.

To maximise the sensitivity, the $m_{\mathrm{T}}^{\mathrm{tot}}$ distributions are derived in two categories, motivated by the fact that the b-associated production of Higgs bosons is favoured at high $\tan\beta$ values. The events are then categorised from the presence of a b-tagged jet, leading to two distinct categories, one called the b-tag category where at least one b-tagged jet is found in the event, and the other called the no b-tag category defined by the absence of b-tagged jet in the event.

To create the distributions, data and simulation events are first processed using a framework called Heppy. Heppy is a python event-processing framework for high energy physics based on ROOT. While it can take different ROOT-based types of inputs, the inputs used in this analysis follow the MINIAOD format of the CMS collaboration. This format was designed to hold the event-based information needed by most analyses. Therefore, the input files hold a lot of information, i.e. the lists of reconstructed physics objects, making them fairly sizeable. Events are first selected based on the criteria defined in this section, while also trimmed of the information that is not useful to the analysis, leading to a new lightweight format.

Heppy is a modular framework, which means that all the processing is done in a feed-forward workflow with each step being encoded into a module called an

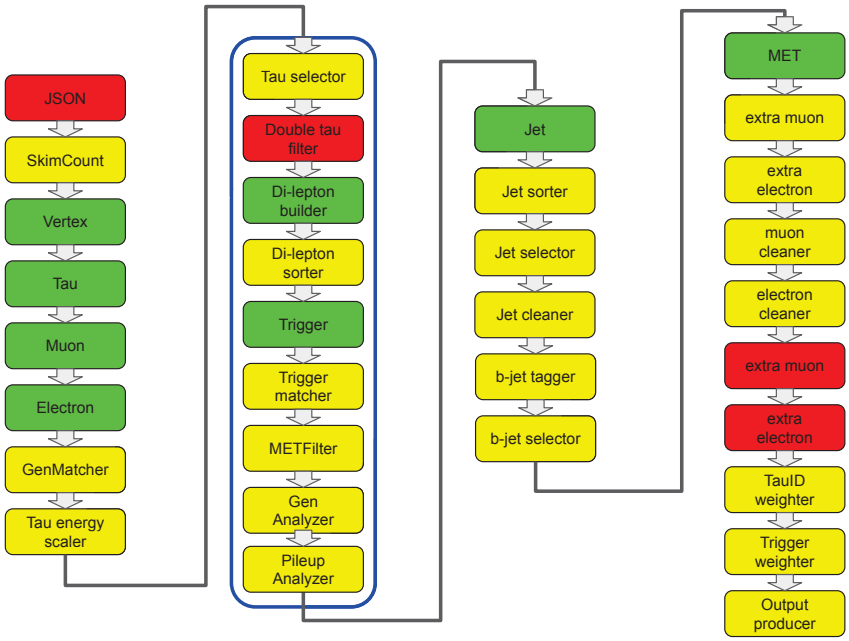


Figure 4.2: Diagram of the selection flow implemented in Heppy. Every box represents a module, called an analyzer. Green analyzers create collections in the event instance by wrapping objects from the input in dedicated python classes. Yellow analyzers modify or compute a variable of the event. Red analyzers reject events that do not match given selection criteria. Finally, the blue highlighted section is the only part of the sequence that is specific to the $\tau_h\tau_h$ channel. The rest of the sequence is used in other channels, like the semileptonic channels $e\tau_h$ and $\mu\tau_h$, and has been successfully tested.

analyzer. Physics objects retrieved from the ROOT files are wrapped in python classes, allowing the definition of useful methods. Most analyzers of this workflow manipulate the physics objects to either create useful new variables, create lists of objects passing a criterion or reject events based on specific criteria. New analyzers have been created in order to provide as much modularity and clarity as possible so that other channels and any equivalent stage of future analyses will be able to easily and promptly be implemented.

The following paragraphs are focused on the description of the sequence for the τ_h τ_h channel. This sequence is applied to both simulation and data, some analyzer being specific to either simulation or data. A diagram of the workflow is shown in figure 4.2. In the order of use in the analysis flow, the role of the analyzers are the following:

- JSON: Only active when running on real data. Rejects the events that have not been validated by the CMS collaboration.

- SkimCount: Only active when running on simulation. Counts the number of generated events before selection. This is used later to renormalise the number of generated events to match the data integrated luminosity.

- Vertex: creates a collection of the vertices that passed quality criteria [67]. These quality criteria are applied in order to select genuine pp interactions and reject beam-induced backgrounds.

- Tau/Muon/Electron: creates collections of the respective objects, and adds useful methods and attributes to these objects.

- Gen matcher: Only active when running on simulation. Matches reconstructed τ_h with closest generator-level particles, and classifies them following the scheme described in table 4.1.

- Tau energy scaler: only active when running on simulation, scales the energy of τ_h, depending on their gen level match. Stores each change for later propagation to the E_T^{miss}.

- Tau selector: first analyzer of the channel-specific sequence. Selects τ_h which have:

  - $p_T > 40\,\text{GeV}$ and $|\eta| < 2.1$;
  - passed the decay-mode finding discriminator detailed in section 3.1;
  - $d_z < 0.2\,\text{cm}$, where d_z is the longitudinal distance between the point of closest approach of the leading charged track and the selected primary vertex;
  - passed the very loose working point of the anti-electron discriminator and the loose working point of the anti-muon discriminator.
  - passed the loosest WP of τ_h BDT-based identification criterion. While the signal region is defined by both selected τ_h passing the tight WP, events with τ_h passing the loosest and not the tight WP identification criterion are used for the fake factor method as will be described in section 4.3.2, and should therefore also be selected.

- Double tau filter: Rejects the event if less than two τ_h fulfilling the requirements of the Tau selector have been found.

- Di-lepton builder: creates all possible combinations of two τ_h that have passed the selections, provided the τ_h candidates:

  - are separated by $\Delta R > 0.5$.
  - have opposite-sign electric charges.

- Di-lepton sorter: After these requirements, several pairs of τ_h candidates can remain. In this case, the pair with the τ_h of highest p_T, called leading τ_h, is chosen. If two pairs have the same p_T for their leading τ_h, the pair with the most-isolated leading τ_h is chosen. In case of more than one pair with the same leading τ_h p_T and isolation, the next tested criterion is the p_T of the other τ_h and the last criterion is the isolation of this other τ_h.

- Trigger: Retrieves the trigger information.

- Trigger matcher: Checks if the selected τ_h pair matches with any of the L1 trigger patterns, and any of the HLT trigger patterns.

- MET Filter: Retrieves several flags that are provided by the CMS collaboration to reject events in order to mitigate several $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction issues.

- Gen analyzer: Only active when running on simulation. Retrieves generator-level information in order to compute several weights, i.e the top quark and Drell-Yan p_{T} reweighting that are detailed in the next section.

- Pileup analyzer: Only active when running on simulation, retrieves pileup information and computes pileup weights, as detailed in the next section.

- Jet: Creates collection of jets, and adds useful methods and attributes. Also applies the jet energy corrections as detailed in next section, while also storing the information for propagation to the $E_{\mathrm{T}}^{\mathrm{miss}}$.

- Jet sorter: Sorts the jet collection by p_{T}.

- Jet selector: Jets are required to have $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 4.7$, and to pass identification criteria to reject fake jets originating from detector noise or pileup.

- Jet cleaner: Discards jets overlapping with one of the two selected leptons, i.e. the distance between jets and any lepton must be $\Delta R > 0.5$.

- b-jet tagger: Applies a tag to each jet that defines whether it is considered as a jet originating from a b-quark (b-tagged). The medium WP of the deepCSV method [94] is used. Also applies b-tagging corrections as detailed in the next section.

- b-jet selector: Creates a b-tagged jet collection from all jets passing the b-tagging requirements defined in the b-jet tagger, and of $|\eta| < 2.5$.

- MET: Retrieves $E_{\mathrm{T}}^{\mathrm{miss}}$ of the event. Applies all needed corrections detailed in the next section and adjusts the $E_{\mathrm{T}}^{\mathrm{miss}}$ to compensate for all corrections applied to other physics objects.

- extra muon (electron) cleaners: Rejects events if any muon (electron) passes a set of quality criteria.

- TauID/trigger weighters: compute and apply the respective correction weights detailed in the next section.

- Output producer: Gathers all desired information and stores it in a flexible ROOT format allowing production of the distributions for statistical inference.

The output of this stage is used to perform a synchronisation with other CMS institutes working on the same analysis. For example a prototype of this sequence was used to synchronise on the previous MSSM search for a heavy Higgs bosons [92] and helped to figure out tweaks and upgrades in implementations of the same analysis by other groups. For this 2017 data analysis, a new synchronisation has been successfully performed.

Table 4.1: MC simulation generator matching.

Value	Type	Generator level object properties
1	Prompt electron	$\|\text{pdgID}\| = 11$, $p_\text{T} > 8\,\text{GeV}$, status flag IsPrompt
2	Prompt muon	$\|\text{pdgID}\| = 13$, $p_\text{T} > 8\,\text{GeV}$, status flag IsPrompt
3	$\tau \to e$	$\|\text{pdgID}\| = 11$, $p_\text{T} > 8\,\text{GeV}$, status flag IsDirectPromptTauDecayProduct
4	$\tau \to \mu$	$\|\text{pdgID}\| = 13$, $p_\text{T} > 8\,\text{GeV}$, status flag IsDirectPromptTauDecayProduct
5	$\tau \to \tau_\text{h}$	Gen-tau jet
6	Jet or pu fake	Anything that does not fall into any of the above categories

4.3 Background estimation methods

The most important backgrounds are estimated using two data-driven techniques. Data-driven techniques are preferred over simulation as they improve the estimation of the background and reduce the associated systematic uncertainties. The contribution of DY $Z \to \tau\tau$ as well as other backgrounds that lead to two genuine τ_h in the final states is estimated with the embedding method, detailed in section 4.3.1. The contribution of jets misidentified as τ_h is estimated using the fake factor method, described in section 4.3.2. For all background processes except QCD multijet events, appropriate samples of simulation are also used, either as part of one of the data-driven techniques, or to estimate the backgrounds not covered by these techniques. The corrections applied to the simulation are detailed in section 4.4.

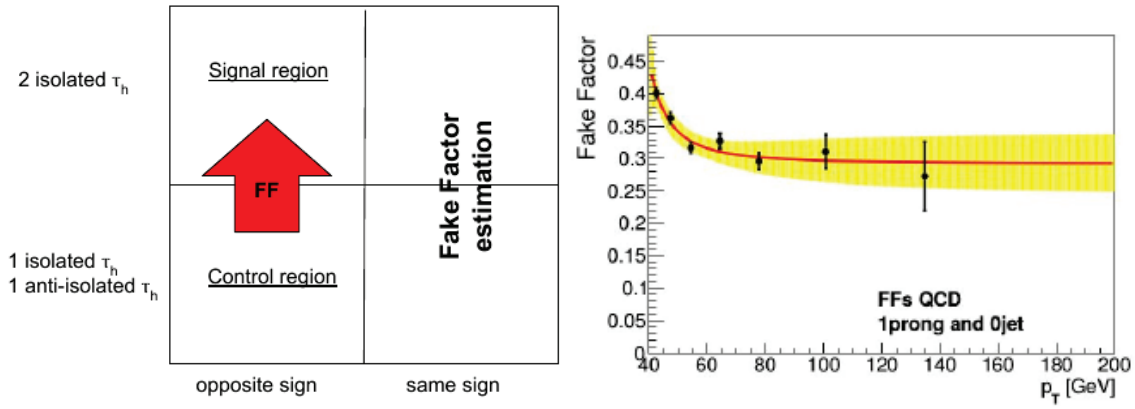
To distinguish between the contributions that are covered by the data-driven techniques and the contributions that are directly estimated from simulation, simulated events are classified depending on a matching between the generator level physics objects and the reconstructed τ_h that have been selected. This matching process is referred to as gen matching. The exact definitions used to distinguish the different matched types can be found in table 4.1. In the table, the status flags are information added by the Monte-Carlo generators to describe the provenance of the physics object. The flag IsPrompt means the object comes directly from the hard scattering process, and the flag IsDirectPromptTauDecayProduct means the physics object is the product of the decay of a τ coming from the hard scattering process. Gen-tau jet refers to a cluster of the generator level decay products of a τ_h. Each simulated background sample is split into three contibutions labeled T, J and L. The T contribution corresponds to events with gen match equal to 3, 4, or 5 for both tau candidates. The T contribution is covered by the embedding technique. The J contribution corresponds to events where gen match is equal to 6 for at least one of the hadronic tau candidates. The total J contribution is covered by the fake factor technique. The L contribution corresponds to all remaining events, and is covered by simulated samples.

4.3.1 Embedding

The embedding technique estimates from data the contribution of the standard model background processes that lead to two τ_h in the final state, with minimal input from simulation. This technique relies on a recorded sample of di-muon events. The two muons are removed from the event and replaced with simulated τ_h with the same kinematic properties. A set of hybrid data-simulation events is obtained, where most of the event comes from data, and where simulation is only used to model the decay of the τ leptons. Challenges in describing the underlying event or the production of associated jets in the simulation are thus avoided. A detailed description of the embedding technique can be found in Ref. [95].

The embedded samples make it possible to avoid using simulated samples for $Z \to \tau\tau$ and the parts of $t\bar{t}$, di-boson and electroweak events where both tau candidates are matched to genuine taus at generator level.

4.3.2 Fake factor method



(a) *(b)*

Figure 4.3: Illustration of the fake factor method. Diagram a illustrates the way fake factors are measured, while b illustrates some of the values and the fitted function used in the case of τ_h decaying to 1 prong and 0 jet in the event.

The fake factor method (FF) is used to predict all background sources where at least one of the reconstructed τ_h is actually a misidentified gluon- or quark-initiated jet. The contribution of such cases is estimated in the well-isolated signal region, defined by the presence of two offline τ_h passing all the identification criteria described in chapter 3 including the tight WP of the BDT-based isolation criterion as well the kinematic cuts of $p_T > 40\,\text{GeV}$ and $|\eta| < 2.1$. This contribution is estimated from collision events in the anti-isolated signal region. The anti-isolated signal region is defined by recorded events where one of the τ_h is loosely isolated but not as isolated as the signal region, which means the considered τ_h passes the loose working point of the τ_h identification criteria but not the tight one. As illustrated in 4.3, a weight, called fake factor, is then applied to each event of the anti-isolated region to estimate the fake jet contribution in the signal region.

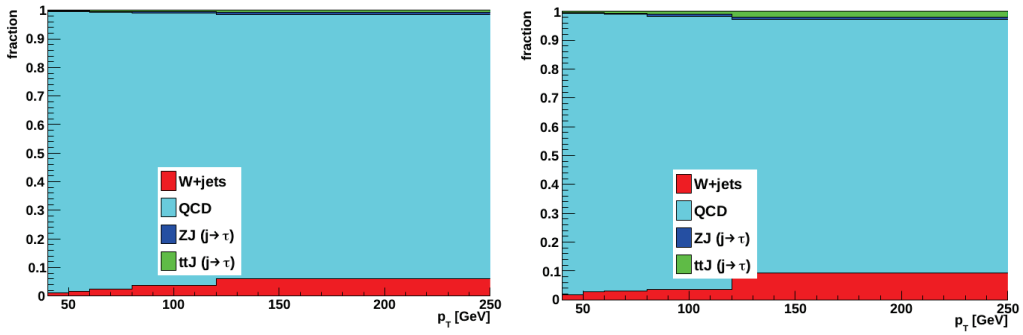


Figure 4.4: The fractions of events from different processes in the anti-isolated signal region as a function of the p_{T} of the anti-isolated τ_{h} for 1-prong candidates (left) and 3-prong candidates (right).

For all channels involving a τ_{h}, the fake factors are measured in different regions for each considered background process, namely QCD multi-jet, W+jets and $t\bar{t}$. Each region is selected for their purity in the considered background. The fake factor applied to a given event in the anti-isolated region is a weighted average of the values measured for the different processes. The weight is given by the expected fraction of events of a given process in the anti-isolated region, and is binned in m_{vis} and number of jets. For the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel, the contribution of each background is measured, as illustrated in figure 4.4. Since QCD jets are the overwhelming contribution, the fake factors derived in the QCD control region are also used to estimate any other backgrounds, contrarily to the semi-leptonic channels.

In the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel, the QCD fake factor measurement region corresponds to events where the two reconstructed τ_{h} have the same electric charge sign. The value of the fake factor (FF) is then measured in this region as

$$FF = \frac{\text{number of } \tau_{\mathrm{h}} \text{ passing tight } \tau_{\mathrm{h}} \text{ isolation discriminant}}{\text{number of } \tau_{\mathrm{h}} \text{ passing loose but not tight } \tau_{\mathrm{h}} \text{ isolation discriminant}} . \quad (4.5)$$

The fake factors are measured as a function of the jet multiplicity (0 or ≥ 1) and the p_{T} of the anti-isolated τ_{h}. These measured fake factors are then extrapolated to the opposite-sign region using a correction function. This function is derived in the region where the other τ_{h} is also anti-isolated. The fake factors are then interpolated by a p_{T}-dependent fit.

For the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel, the fake factor is applied to all events in the anti-isolated region twice, once considering a τ_{h} as the fake, then considering the other as a fake, while multiplying both weights by 0.5. Since genuine τ_{h} events in data are also present in the anti-isolated region, the expected contribution from events with genuine τ_{h} is subtracted in the isolated region by applying the fake factors to simulated events with genuine τ_{h} in the anti-isolated region.

4.4 Correction of the Monte Carlo simulation

In order to mitigate the differences between data and simulation, several measurements have provided corrections that are applied on the simulated samples used in the analysis. These corrections are:

Pileup reweighting MC simulated samples are generated with a given instantaneous luminosity which does not match the instantaneous luminosity of the data, which is often recorded after or during the production of the MC samples. In order to better fit the recorded pileup distribution in data a pileup reweighting is applied to MC events.

EE noise jets removal Due to noise in the ECAL endcaps in 2017, all jets in a pseudo-rapidity gap of $2.65 < |\eta| < 3.139$ and of less than 50 GeV are removed in both data and MC. The $E_{\mathrm{T}}^{\mathrm{miss}}$ is modified accordingly.

τ_{h} triggering efficiency Trigger scale factors are also measured using a tag-and-probe method. The tag is a τ_{h} that passes the ID and isolation requirements applied in the analysis and is matched to a trigger τ_{h} object that passes the requirements of the considered trigger pattern. The probe is any τ_{h} that passes the ID and isolation requirements applied in the analysis, except the tag τ_{h}. The efficiency ϵ of the trigger is then the ratio of the number of events where the probe is matched to a τ_{h} trigger object to the total number of events where a probe is found. This efficiency is then computed in both data and simulation. The scale factor is then

$$SF = \frac{\epsilon(\text{data})}{\epsilon(\text{simulation})}\,. \tag{4.6}$$

The scale factors for the double-τ_{h} trigger are measured in bins of the tau p_{T}, η, and ϕ. The total scale factor in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ is then the product of the scale factor associated to each τ_{h}. The scale factors for the embedded samples use the same ϵ(data) measured for the simulation scale factors but use the efficiency measured for the embedded taus as denominator.

τ_{h} identification efficiency A data/simulation scale factor is measured in $\mathrm{Z} \to \mu\tau_{\mathrm{h}}$ events also using a tag-and-probe approach. Events with a muon and a τ_{h} are selected in both data and simulation and the efficiencies are extracted from a fit to the di-lepton invariant mass in the mass window around the Z mass. An additional correction is applied for embedded taus to correct for biases due to higher tracking efficiencies in embedded events than data.

τ_{h} energy scale Similarly to the previous corrections, the correction on the energy of reconstructed τ_{h} was measured using a tag-and-probe approach in $\mathrm{Z} \to \mu\tau_{\mathrm{h}}$ events. This time a profile-likelihood fit is performed on the invariant mass of the muon and τ_{h} pair between simulation with energy scale applied and data. The best fitted value of the energy scale for each decay mode then gives an energy correction factor that is applied to any simulated τ_{h} of the given decay mode.

Efficiency of leptons misreconstructed as τ_{h} Electrons(muons) can be misreconstructed as τ_{h} decay products, as their track can be misinterpreted as arising from the presence of a charged hadron. As described in chapter 3, anti-lepton discriminants are applied. A scale factor is applied to correct for the data/simulation

discrepancies coming from its use. To measure these scale factors, a tag-and-probe approach is used on $Z \to e^+e^-$ ($Z \to \mu^+\mu^-$) events. The tag is then a well-identified and isolated electron (muon) and the probe is a well-identified τ_h. The efficiency is the number of τ_h probes passing the anti-lepton discriminant over the number of all found τ_h probes. The scale factor is then derived from a profile-likelihood fit performed on the visible mass distributions of events where the probes pass the anti-electron (anti-muon) discriminant and events where the probes fail.

Energy scale of lepton misreconstructed as τ_h A correction to the values of the τ_h energies for τ_h candidates originating from electron (muon) is applied for the 1-prong and 1-prong+1-π^0 decay modes [96]. This correction is derived with the same approach as for the measurements of the efficiency of leptons misreconstructed as τ_h.

Jet energy On top of the corrections detailed in section 2.4.3, corrections derived by the CMS collaboration to mitigate data/MC discrepancies are also applied. All the correction measurements are detailed in Ref. [97].

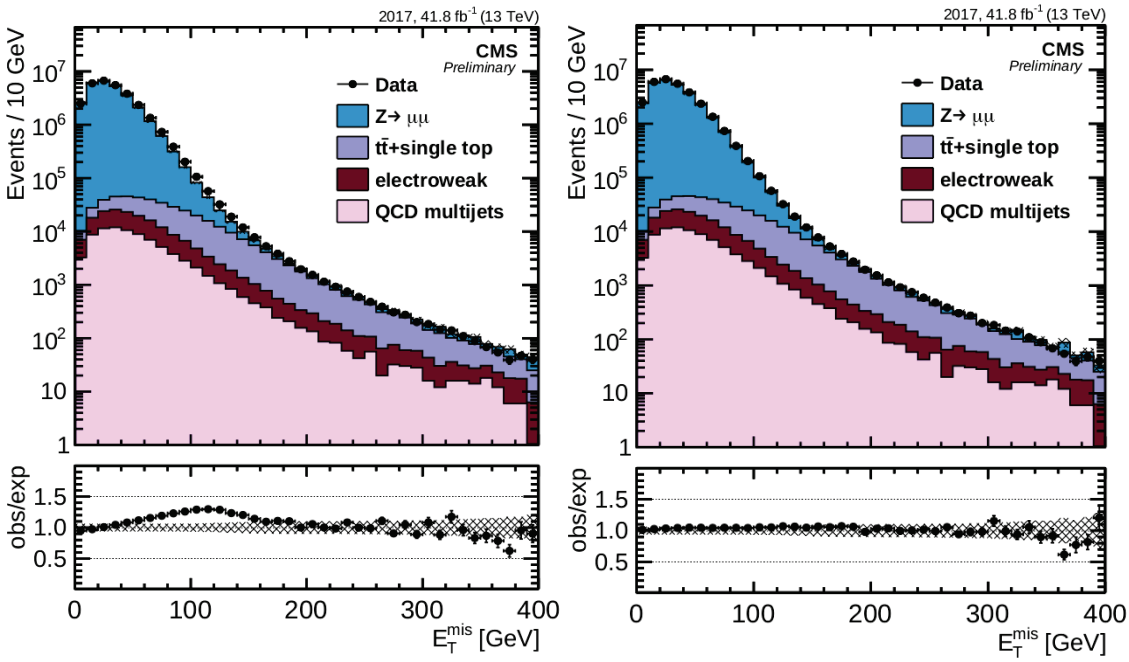
b-tagging efficiency The efficiency for tagging the b jets and the mistagging rate for light-flavour jets have been measured in both data and simulation. The efficiency and mistagging rate in the simulation are corrected through the application of efficiency and mistagging scale factors. The values of these factors and a description of the methods used to determine them can be found in Ref. [94]. The simulation is corrected by un-tagging a fraction of jets that pass the requirements of b-tagging, and tagging a fraction of jets that do not pass the b-tagging requirements. The tagging of a jet is called promotion, and the untagging of a jet is called demotion. The promotion or demotion probabilities for each jet are defined as

$$P(\text{demote}) = 1 - SF\,, \quad \text{when } SF < 1$$

$$P(\text{promote}) = \frac{(SF-1)}{\frac{1}{\epsilon}-1}\,, \quad \text{when } SF > 1\,.$$

In this expression, the scale factors SF are p_T, η and jet-flavour dependent. They are the ratio of data over simulation efficiencies, and their measurement are provided by the CMS collaboration. The tagging efficiency ϵ is determined in the simulated samples used in the analysis.

Recoil corrections Recoil corrections are applied to correct for the mismodeling of E_T^{miss} in the simulated Drell-Yan, W+Jets and Higgs production. The corrections are derived in $Z \to \mu\mu$ events, where the leptonic recoil does not contain neutrinos and the four-vector of the Z boson can be measured precisely. The effect of the recoil corrections on the E_T^{miss} distribution of $Z \to \mu\mu$ events, as measured by another CMS institute, is shown in figure 4.5.



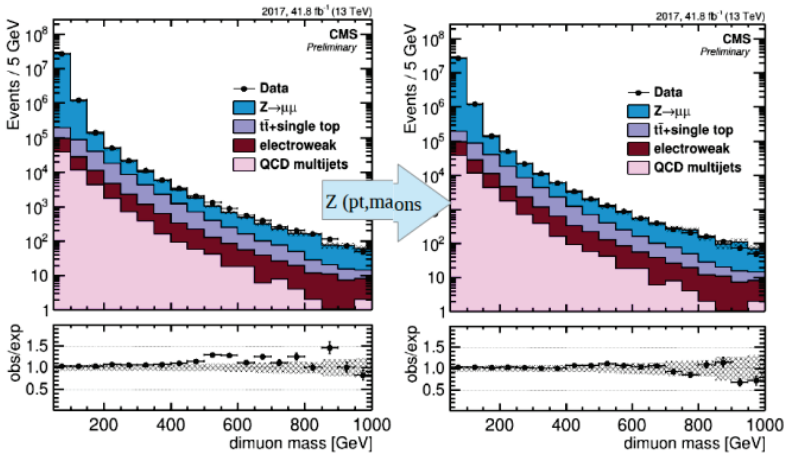
(a) Without recoil corrections *(b) With recoil corrections*

Figure 4.5: Effect of applying recoil corrections to the E_{T}^{miss} distribution in the $\mathrm{Z} \to \mu\mu$ selection.

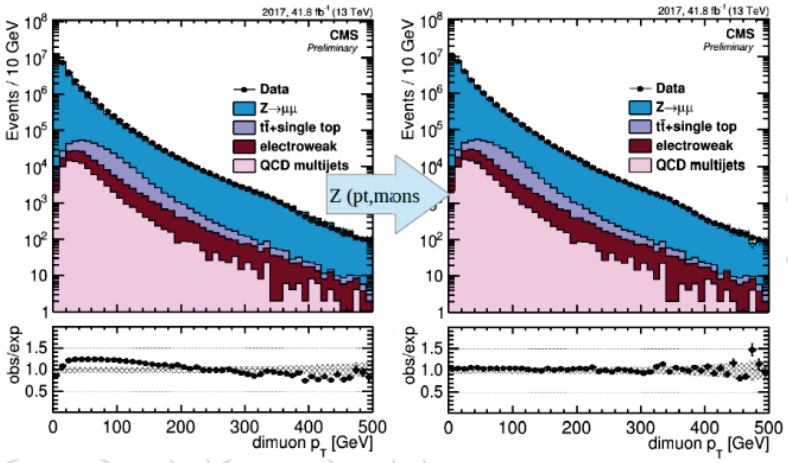
DY mass and transverse momentum reweighting A reweighting is applied to Drell-Yan MC samples to correct the gen level di-lepton p_{T} and mass distributions in LO madgraph samples. These corrections are measured in a $\mathrm{Z} \to \mu\mu$ control region. The weights are computed in such a way as to make the two-dimensional distributions of the Z p_{T} and the Z boson reconstructed mass match between data and simulation. The weights are then corrected not to introduce a general yield variation of the Drell-Yan background, but just to have a shape effect on the considered distributions. This correction was derived by the DESY group, and the $M_{\mu\mu}$ and $p_{\mathrm{T}\mu\mu}$ distributions of the $\mathrm{Z} \to \mu\mu$ events before and after reweighting are shown in figure 4.6.

Top quark transverse momentum reweighting The modeling of the $t\bar{t}$ background is improved by reweighting the p_{T} spectrum of the top quarks. The correction follows the strategy developed for Run I [98], as it provides the best description of this background.

τ_{h} tracking efficiency in embedded sample In embedded events, tracking is simulated in an empty detector environment. This causes difference with respect to tracking in complete event simulation. Correction scale factors are derived by first applying the embedding technique to a simulated sample of $\mathrm{Z} \to \mu\mu$ events instead of data, and then comparing with $\mathrm{Z} \to \tau_{\mathrm{h}}\tau_{\mathrm{h}}$ simulated events.



(a) $\mu\mu$ mass, before reweighting (b) $\mu\mu$ mass, after reweighting



(c) $p_T(\mu\mu)$, before reweighting (d) $p_T(\mu\mu)$, after reweighting

Figure 4.6: Di-muon mass and p_{T} distributions in $\mathrm{Z} \to \mu\mu$ data before and after the DY p_{T} reweighting.

4.5 Systematic uncertainties

This section describes the sources of uncertainty that affect the signal and background predictions of the $m_{\mathrm{T}}^{\mathrm{tot}}$ distribution in each category. The experimental uncertainties typically concern either physics object reconstruction and identification or the methods used to estimate the backgrounds described in the previous section, and are inherent to their respective measurement. The object selection is more important for the signal prediction, whereas the estimation methods have a large effect on the background estimation. Theoretical uncertainties affect the predictions of both signal and simulated background but are larger for the signal. Uncertainties can affect either the yield of the distributions only, or affect both their shape and yield. Each source of uncertainty will be represented by a nuisance parameter in the global fit described in the next section.

4.5.1 Normalisation uncertainties

The yield from each process in the $m_{\mathrm{T}}^{\mathrm{tot}}$ distribution is affected by a normalisation uncertainty, described by a nuisance parameter following a log-normal distribution.

Luminosity A 2.3% luminosity uncertainty is applied to the yield of contributions that are purely estimated from simulation [99].

τ_{h} identification efficiency A 7.9% uncertainty is applied to the yield of contributions that are purely estimated from simulation, as well as for the embedded sample. The embedded samples also have an additional uncertainty applied to cover a tracking efficiency correction, with a magnitude of 2% per τ_{h}.

Trigger efficiency The uncertainty in the trigger efficiency amounts to 10%. The uncertainty is applied to all processes with a contribution predicted from simulation, and to the embedded samples. The embedded and MC uncertainties are uncorrelated. In order to account for the efficiency of the double muon trigger that was used to select input events for the embedding technique, an additional 4% uncertainty (2% per muon) is applied to the embedded samples.

Background normalization uncertainty 4%, 5%, 6% and 4% uncertainties are applied to the $Z \to ll$, di-boson/single top, $t\bar{t}$ and EWKZ processes, respectively, to account for the uncertainty in the production cross-section of these processes.

Fake factor normalization The uncertainty due to the subtraction of the genuine tau contribution is estimated by varying the subtracted number of events by $\pm 10\%$, and amounts to about 2% of the jet faking τ_{h} yield.

4.5.2 Shape uncertainties

Uncertainties that have an influence on the shape of the $m_{\mathrm{T}}^{\mathrm{tot}}$ distribution are treated as shape uncertainties. In this case, the yield is described by a nuisance parameter

following a log-normal distribution, as it is done for the normalisation uncertainties. The shape variations are taken into account by vertical interpolation between the $m_{\mathrm{T}}^{\mathrm{tot}}$ histograms corresponding to a $\pm 1\sigma$ variation, also called up or down variation respectively, of the considered source of uncertainty. The histograms corresponding to the variation of most of the uncertainties are evaluated by re-running the concerned simulated events through the analysis sequence, while applying the up or down fluctuation to the considered variable. This leads to a multiplication of the overall computing needed to perform the full analysis, and is the main reason the semi-leptonic channels are not yet included in this analysis. The following uncertainties are the ones that cause variations in the $m_{\mathrm{T}}^{\mathrm{tot}}$ distributions.

τ_{h} Energy scale Since the $m_{\mathrm{T}}^{\mathrm{tot}}$ variable depends on the p_{T} of the selected τ_{h}, a separate shape uncertainty is applied for each of the corrected decay modes in the simulated samples. In the embedded samples we have hybrid events where the simulated taus might be mixed with calorimeter deposits remaining from the removal of the original muons in the events. For this reason, the tau energy scale uncertainties for embedded samples are split into two parts, where 50% are fully correlated with the uncertainty for fully simulated samples and 50% are uncorrelated.

Energy scale of leptons misreconstructed as τ_{h} For the same reason, shape uncertainties are propagated to $m_{\mathrm{T}}^{\mathrm{tot}}$ from the p_{T} of a misidentified τ_{h} arising from the presence of leptons in simulation samples uncorrelated between decay modes.

Jet energy Since the change in jet energy is propagated to the $E_{\mathrm{T}}^{\mathrm{miss}}$, the jet energy scale impacts the shape of the $m_{\mathrm{T}}^{\mathrm{tot}}$ distribution. In general, the CMS collaboration derives uncertainties in the jet energy scale from 28 sources and combines them in a single uncertainty with one nuisance parameter. The uncertainty is split into 5 groups, instead of all 28 sources, because this would result in an unnecessary amount of parameters in the fit as well as a large technical effort. Instead, the sources are grouped according to the affected detector regions.

$E_{\mathrm{T}}^{\mathrm{miss}}$ unclustered energy uncertainty The $E_{\mathrm{T}}^{\mathrm{miss}}$ takes into account the energy that is not clustered in the reconstruction process, directly impacting the value of $m_{\mathrm{T}}^{\mathrm{tot}}$. An uncertainty is therefore applied to all simulated processes that do not have recoil correction applied.

$E_{\mathrm{T}}^{\mathrm{miss}}$ recoil correction uncertainties For all simulated processes where recoil corrections are applied, uncertainties determined during the computation of the recoil corrections are propagated to the variable.

Top p_{T} reweighting The uncertainty in the top p_{T} reweighting is estimated by not applying the correction (down fluctuation) and applying the correction twice (up fluctuation) in the $t\bar{t}$ simulated events.

DY p_{T} reweighting The uncertainty in the DY p_{T} reweighting is estimated by shifting the reweighting applied to $\mathrm{Z} \to ll$ events by 10%.

B-tagging efficiency Categories are defined from the presence or absence of b-tagged jets in the events, therefore variation in the b-tagging of jets can lead to migration of events from one category to the other. The uncertainties in the b-tagging scale factors provided by the CMS collaboration are therefore propagated to the $m_{\mathrm{T}}^{\mathrm{tot}}$ distribution in the categories defined by the presence or absence of b-tagged jets.

τ_{h} tracking efficiency for the embedded samples Since the scale factors for tracking efficiencies are dependent on the τ_{h} p_{T}, variations in their values can cause a shape effect on the $m_{\mathrm{T}}^{\mathrm{tot}}$ distributions. An uncertainty in the tracking efficiency of hadronic taus in the embedded samples is propagated uncorrelated between the 1 and 3 prong decay modes.

Fake-factor uncertainties The fake factors depend on the p_{T} of the τ_{h}, and therefore a variation in those fake factors can lead to variation in the shape of the $m_{\mathrm{T}}^{\mathrm{tot}}$ distribution. Uncertainties in the fake factor background estimation method result from several sources:

- Statistical uncertainty in the fake factor measurement in the control regions.
- Systematic uncertainties related to the QCD multi-jet fake factor corrections are propagated.
- Systematic uncertainties in the fraction of W/Z+jets events and $t\bar{t}$ events with one misidentified τ_{h} in the anti-isolated region, adding two nuisance parameters. These are evaluated by varying the fractions of these two backgrounds within uncertainties (including cross-section and experimental uncertainties), while readjusting the fractions of the other processes to keep the sum at 100%.

Bin-by-bin uncertainties To account for statistical shape uncertainties in the backgrounds due to the use of Monte-Carlo and embedded samples or templates derived from data events with a limited number of events, we introduce shape variations to the background templates in all categories following the Barlow-Beeston approach, where the statistical uncertainties in each bin are used to define alternative shapes.

4.6 Validation

4.6.1 Non-discriminant variable distributions

In order to validate our estimation methods, distributions of variables that are virtually not sensitive to the presence of signal have been produced and can be found in appendix B.

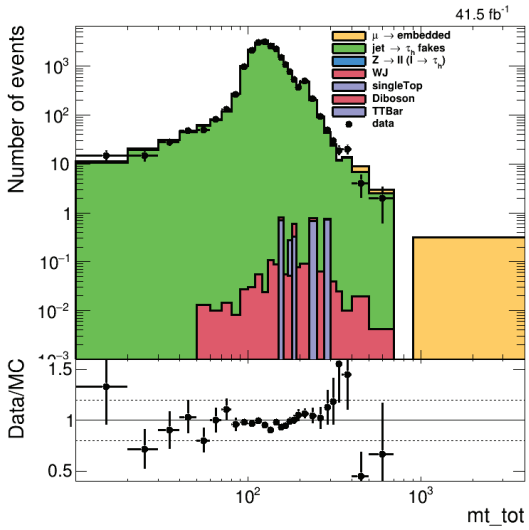


Figure 4.7: Distribution of the discriminating variable in the inclusive same sign region. This same sign region only differs from the signal region by the requirement that both τ_h have the same sign instead of opposite signs.

4.6.2 Fake factor method validation

In the fully hadronic channel, one of the main background contribution originates from QCD processes. Since in these processes there is no constraint between the charges of both τ_h, the same sign region, defined by both reconstructed hadronic taus being of the same electric charge, can be used to validate the agreement between data and the estimation method. While there is a very small correction applied to the fake factors to account for the differences between the same-sign region (SS) and opposite-sign region (OS), the same sign distribution shown in Fig 4.7 shows a good agreement between our estimation and recorded data in the SS region.

4.6.3 Embedding technique validation

In order to validate the embedding technique, its contribution to the fake factor technique must be disentangled, especially since an over-estimation of the concerned backgrounds will lead to an under-estimation in the fake-factor covered backgrounds. To disentangle both techniques, we have replaced the fake factor technique by another estimation method called the ABCD method. In this method, the ratio of the yields of the isolated same-sign region and anti-isolated same-sign region is measured in data. Then the shape of the QCD distribution is taken from the anti-isolated region, and its yield is multiplied by the previously measured ratio. The other backgrounds that are not QCD are then covered by MC simulation.

The obtained prefit plots are shown in figure 4.8. These plots show first that the use of the ABCD QCD method leads to a poor description of the background. Then, these plots also show that the region higher than 300 GeV seem to be over-estimated by the embedding technique, as the direct simulation estimation leads to less over-estimation. Although many more aspects are to be understood, the time constraints of this work did not allow to go further in this investigation. The different groups

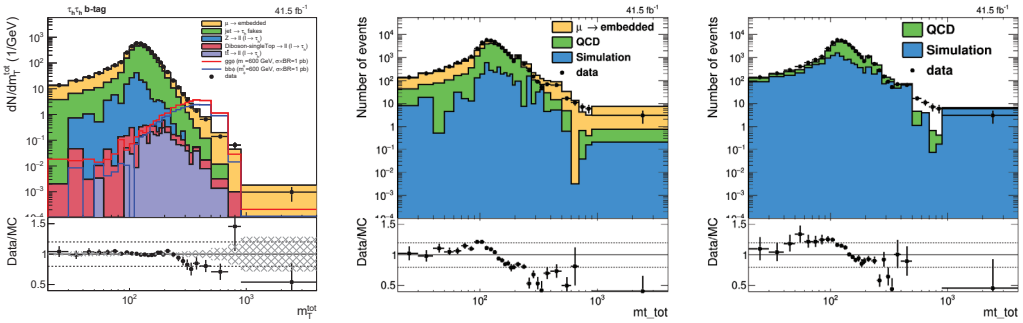


Figure 4.8: Pre-fit distributions of $m_{\mathrm{T}}^{\mathrm{tot}}$ in the inclusive category with different background estimation techniques. Left is made using both the fake factor and embedding methods, and simulation for the rest. Middle is made using the embedding, the ABCD QCD method and simulation. Right is only the QCD ABCD and simulation methods.

using the technique have since picked up on the problem revealed by this study and are actively working to finish this investigation.

4.7 Statistical interpretation

This section outlines the statistical procedure used to quantify or reject the presence of a signal in data. These methods were developed by the LHC Higgs Combination Group to provide a common strategy for both the CMS and ATLAS Collaborations and to facilitate the combination of individual search results [100].

The expected Higgs boson event yield in a given model can be denoted as s and the background yield as b. This can refer equally to a simple counting experiment, or to predicted binned distributions for use in a shape-based analysis. An additional factor μ is introduced as a signal strength modifier, which allows for the signal rate to scale as $\mu \times s$. The background-only hypothesis is therefore defined by $\mu = 0$, and any signal hypothesis by $\mu > 0$. The term "data" will refer to an observed event count or counts, which could originate from an actual experiment or from simulation. The yields s and b are, in general, functions of nuisance parameters θ representing experimental and theoretical uncertainties. The nominal values $\tilde{\theta}$ of these nuisance parameters are usually determined by external measurements, with uncertainties described by probability density functions $p(\tilde{\theta}|\theta)$. From these components the likelihood for an observed dataset, $\mathcal{L}(\text{data}|\mu,\theta)$, is defined as

$$\mathcal{L}(\text{data}|\mu,\theta) = \text{Poisson}(\text{data}|\mu \times s(\theta) + b(\theta)) \times p(\tilde{\theta}|\theta)\,, \tag{4.7}$$

where for a binned likelihood model the Poisson term is simply the product of Poisson probabilities over each bin i:

$$\text{Poisson}(\text{data}|\mu \times s(\theta) + b(\theta)) = \prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)}\,. \tag{4.8}$$

A ratio of likelihoods can be used to define a test statistic, a single number which can distinguish between two hypotheses. Such a test statistic can be used to set upper limits on the rate of signal production. Historically, a number of definitions

have been used in Higgs boson searches. The one chosen by the LHC experiments is known as the profile likelihood ratio

$$q_\mu = -2\ln\frac{\mathcal{L}(\text{data}|\mu,\hat{\theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu},\hat{\theta})}, \text{ with the constraint } 0 \leq \hat{\mu} \leq \mu\,. \tag{4.9}$$

In this expression, $\hat{\theta}_\mu$ are the values of the nuisance parameters that maximise the likelihood, given the fixed signal strength μ; and $\hat{\mu}$ and $\hat{\theta}$ are the values which give the global maximum of the likelihood. The constraint $0 \leq \hat{\mu}$ is added to prevent an unphysical negative signal strength. The constraint $\hat{\mu} \leq \mu$ is chosen to prevent the exclusion of any μ lower than the best fit $\hat{\mu}$, thus ensuring the construction of a one-sided confidence interval. Large values of q_μ indicate a value of μ that the data disagrees with, whereas values close to zero indicate good compatibility with the signal hypothesis in question. The probability of finding a value q_μ at least as large as the observed value, q_μ^{obs}, is defined as

$$\text{CL}_{s+b} = \int_{q_\mu^{\text{obs}}}^{inf} f(q_\mu|\mu,\hat{\theta}_\mu)dq_\mu\,, \tag{4.10}$$

where $f(q_\mu|\mu,\hat{\theta}_\mu)$ is the probability distribution function for q_μ. The tested value of μ is then said to be excluded at a confidence level α, where $\alpha = 1 - \text{CL}_{s+b}$. The 95% CL is typically chosen when setting upper limits. One issue with this definition is that in some cases it will lead to the exclusion of low signal strengths, where an analysis has a low sensitivity. For example, this may happen with a downward fluctuation of the data when the signal expectation is very small compared to the background expectation. To protect against this effect, an additional probability CL_b can be introduced, defined similarly to equation 4.10, but under the assumption of the background-only hypothesis, $f(q_\mu|0,\hat{\theta}_0)$. Instead, the ratio of these probabilities, denoted CL_s, where

$$\text{CL}_s = \frac{\text{CL}_{s+b}}{\text{CL}_b}\,, \tag{4.11}$$

is used to set the 95% CL exclusion limit, and this is commonly referred to as the modified frequentist approach [101].

The distributions $f(q_\mu|\mu,\hat{\theta}_\mu)$ and $f(q_\mu|0,\hat{\theta}_0)$ can be determined by generating toy MC datasets from their respective models, in which the nuisance parameters are fixed to the values found in the fits to the observed data. The value of q_μ is then determined for each toy dataset. The effect of systematic uncertainties is incorporated by sampling a set of pseudo-measurements $\tilde{\theta}$ in each toy using the chosen nuisance pdfs. It is often instructive to compare the observed exclusion limit to the expectation under the assumption of the background-only hypothesis. This can be determined by generating background-only toy datasets and determining the 95% CL limit in each. These values form a cumulative pdf from which the median exclusion and uncertainty bands can be extracted.

A profile likelihood ratio can also be used to calculate the p-value for an observed excess of events given the background-only hypothesis. For this a slightly modified

definition of the test statistic is required,

$$q_0 = -2\ln\frac{\mathcal{L}(data|0,\hat{\theta}_0)}{\mathcal{L}(data|\hat{\mu},\hat{\theta})}, \text{ with the constraint } \hat{\mu} \geq 0\,, \tag{4.12}$$

where the constraint $\hat{\mu} \geq 0$ is chosen to prevent a downward fluctuation being considered evidence against the background-only hypothesis. The p-value for the observed data is then given as

$$p_0 = \lim_{q_o^{\text{obs}}}^{\text{inf}} f(q_0|0,\hat{\theta}_0)dq_0\,, \tag{4.13}$$

where $f(q_0|0,\hat{\theta}_0)$ can be determined by generating pseudo-data from the background-only hypothesis. The p-value is typically converted to a significance, Z, by determining the number of standard deviations of a one-sided normal distribution that would yield an equal tail probability.

A major advantage of the profile likelihood test statistic is that in the limit of a large data sample, the distribution $f(q_\mu)$ follows a known formula [102]. This so-called asymptotic limit approximation removes the need for the computationally intensive step of generating and fitting toy datasets, which can take an appreciable time for models with many bins and nuisance parameters. This method relies on the properties of the Asimov dataset, a single representative dataset in which the observed rates match exactly the prediction of the model under the nominal set of nuisance parameters. Furthermore, it is possible to derive a formula for the median expected limit and uncertainty bands using only the properties of the Asimov dataset, thus completely removing the need for any toy MC [102].

4.8 Results and interpretations

The statistical interpretation in the MSSM includes the use of a simultaneous maximum-likelihood fit of the $m_{\text{T}}^{\text{tot}}$ distribution in the two categories, namely b-tag and no b-tag categories. Once the $m_{\text{T}}^{\text{tot}}$ distributions are created, a fit under the background-only hypothesis is performed on both categories. The resulting distributions are shown in figure 4.9, and the post-fit values of the nuisances parameters for background-only and for a signal plus background hypothesis, with a signal mass point of 600 GeV, are shown in figure 4.10. The shown systematics uncertainties correspond to all the previously detailed sources, except for the uncertainties concerning the jet-fakes process. Indeed, thanks to this analysis, a bug was found in the tool used to retrieve the values of the fake weights, but unfortunately the group at the origin of this tool has not yet been able to isolate the origin of the problem. The main symptom is the fact that the up and down shifted values of the fake weight give the exact same values. To avoid a mis-estimation, the jet-fakes related shape systematics were dropped, and conservative yield uncertainties of 20% were added to account for them. This temporary fix was validated by the CMS Higgs $\to \tau\tau$ group.

Upper limits in this search are determined in two different ways. The first context is for model-independent limits on the cross-section of a single neutral Higgs

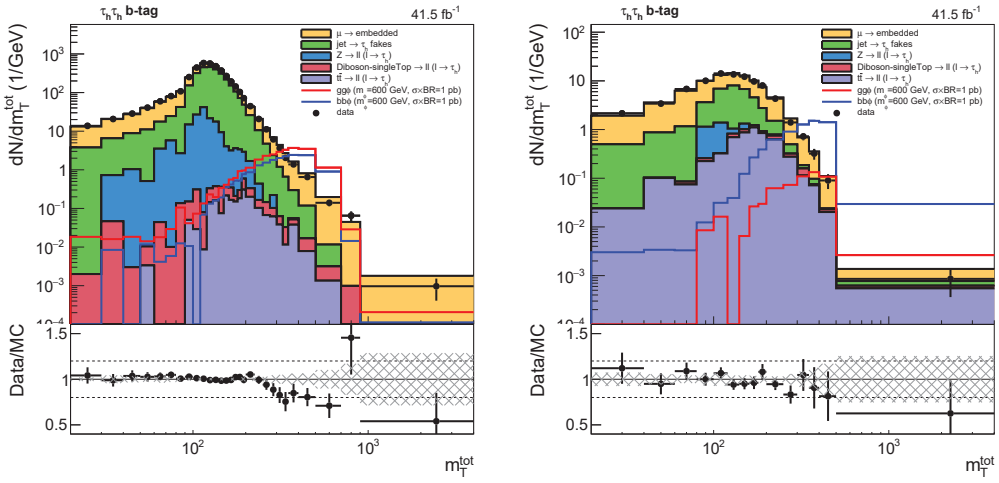


Figure 4.9: Distribution of m_T^{tot} in the no b-tag (left) and b-tag (right) categories. In these plots, the background contributions have been adjusted in a background-only fit. The signal distributions correspond to either the $gg\phi$ or the $bb\phi$ process for a mass of 600 GeV, before any fitting. Both signal processes contributions are normalised to $\sigma \times \mathrm{BR} = 1\,\mathrm{pb}$.

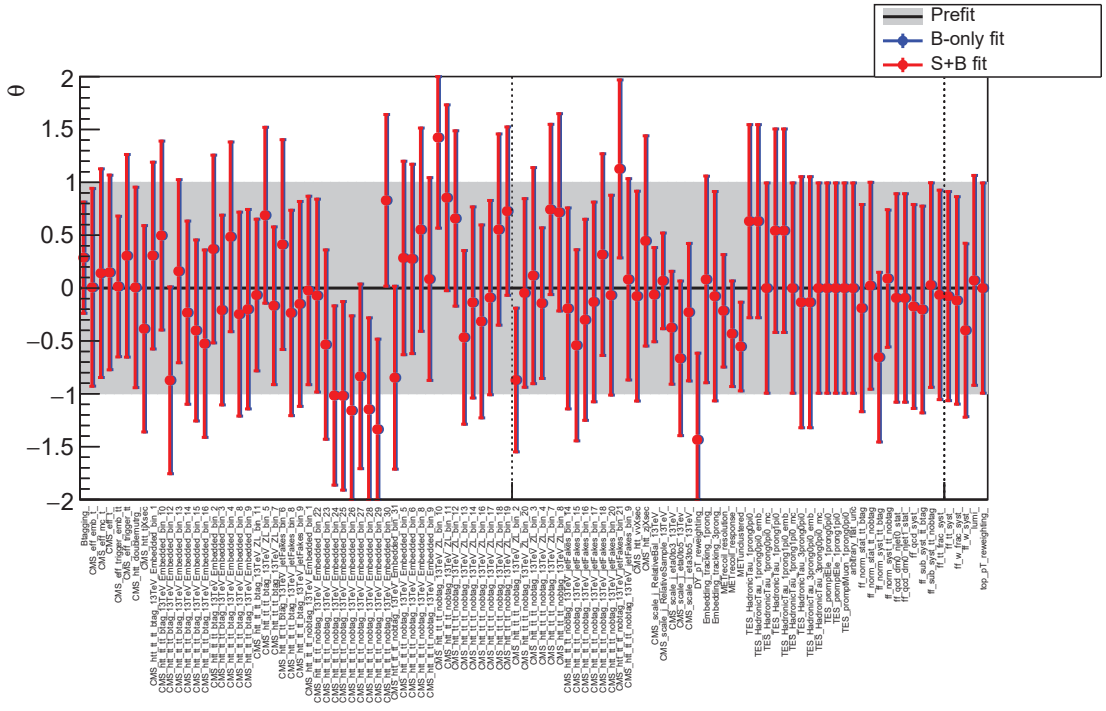


Figure 4.10: Values of the nuisance parameters from the background-only fit and a signal plus background for a resonance mass of 600 GeV.

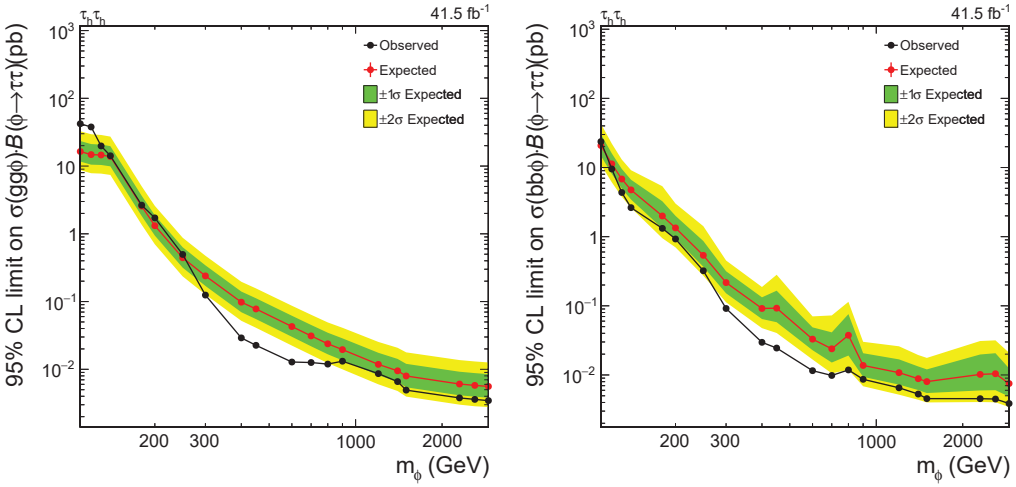


Figure 4.11: Expected and observed 95% CL upper limits for the production of a single narrow resonance, Φ, with a mass between 110 GeV and 3.2 TeV in the $\tau_h\tau_h$ final state for the production via gluon fusion $gg\Phi$ (left) and in association with b quark $bb\Phi$ (right). The green and yellow bands indicate the 68 and 95% confidence intervals for the variation of the expected exclusion limit. The black dots correspond to the observed limits.

boson, denoted Φ, produced either through the gluon-gluon fusion or b-associated production mode and decaying to $\tau_h\tau_h$. The second is in the m_h^{max} scenario, where limits on the parameter $\tan\beta$ are determined as a function of m_A. In this case, the signal model includes the three neutral Higgs bosons h, H and A with masses, cross and branching ratios computed from the chosen values of m_A and $\tan\beta$. The following paragraphs detail how those searches are performed.

Signal samples are generated for the set of m_A mass points to be tested, in the range 110 to 2900 GeV. The step size between points increases with m_A to scale with the worsening m_T^{tot} resolution. To derived model-independent expected upper limits on the production of a single neutral Higgs boson with mass m_Φ, fits under the signal plus background hypothesis will be performed. The limits on the cross-section times branching fraction, $\sigma \times B(\Phi \to \tau_h\tau_h)$, will be determined individually for gluon-gluon fusion and b-associated production. In the fit to extract gluon-gluon fusion limits the b-associated contribution will be allowed to float freely, and vice versa. This is required as neither the no b-tag or b-tag categories are completely pure in one production mode, and this avoids the need to impose any assumptions about the ratio of cross-sections between the two processes.

In order to allow the evaluation of the limit at different mass points, including between the mass points for which simulation has been processed, a horizontal morphing [103] is used. All available processed templates are used for the template morphing and the template fit is performed using the distributions produced by the morphing.

The obtained expected and observed limits are shown in figure 4.11. Several aspects are to be discussed around these. The observed limits are clearly not in agreement with the expected limits, especially in the range of 300 to 800 GeV. This disagreement can be imputed, at least in part, to the embedding technique as the validation procedures showed in section 4.6.

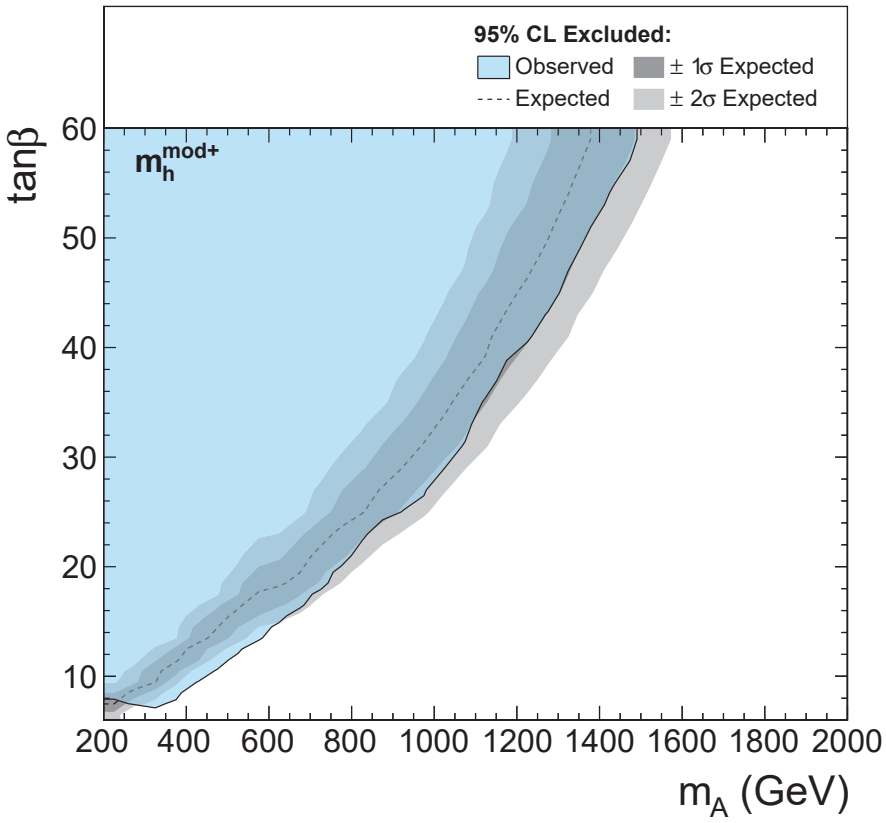


Figure 4.12: Expected and observed 95% CL exclusion contour in the MSSM $m_{h^{mod+}}$ scenario. The expected median is shown as a dashed black line. The dark and bright gray bands indicate the 68 and 95 % confidence intervals for the variation of the expected exclusion. The observed exclusion contour is indicated by the coloured blue area.

A number of additional steps are needed to determine the m_A- tanβ limits. At each m_A- tanβ hypothesis, the masses of the other two Higgs bosons are calculated using results from the Higgs Working Group [45]. In each event category, templates for the h and H are generated by the horizontal morphing [103] between templates from the two closest samples in mass. The category acceptance is similarly interpolated from the neighbouring mass points. All three templates are scaled by the appropriate cross-sections and branching ratios and combined into a single template. The 95% CL upper limit is determined for each point on the m_A- tanβ grid, with the signal strength parameter μ uniformly scaling the entire signal model. The limit in tanβ is then defined as the point on which this upper limit is found to occur at $\mu = 1.0$. Practically, this is determined by interpolation between the points on either side of this threshold. The limits are shown in figure 4.12.

Conclusion

Extensions of the Standard Model with at least two Higgs doublets, such as the MSSM, predict the existence of extra neutral Higgs bosons. The observation of such Higgs bosons would provide direct evidence for physics beyond the Standard Model, and would make the MSSM one of the most promising theories.

This thesis has presented an analysis of proton-proton collision data recorded by the CMS detector during the 2017 data-taking period. The state of search for a MSSM heavy neutral Higgs boson decaying to tau pairs has been presented. Results have been determined from distributions of the m_T^{tot} variable in the $\tau_\text{h}\tau_\text{h}$ final state. Categorisation is used to improve sensitivity to signal and to specific Higgs boson production modes. Expected upper limits at the 95% CL are determined in the $m_\text{A} - \tan\beta$ parameter space for the m_h^{max} scenario. Additionally, model-independent expected limits on the product of the cross-section and the branching fractions for a single Higgs boson, produced via either gluon-gluon fusion or in association with b-quarks, are determined for mass hypotheses in the range 90 to 3200 GeV.

A new hadronic tau decay identification technique, based on a neural network architecture called a recursive neural network has also been presented. Its performance has been compared to the standard identification technique used in the CMS collaboration. This comparison highlighted a better QCD jet rejection in the high τ_h efficiency region. Some potential improvements to this approach have also been presented.

The ambitious LHC physics programme will continue for decades. After the end of Run 3, planned for 2021–2023, the amount of collected data is expected to exceed $300\,\text{fb}^{-1}$. The next major milestone will be the installation of the high-luminosity LHC (HL-LHC), which is expected to deliver $3000\,\text{fb}^{-1}$ of data by 2035. The continuously increasing amount of data will allow extremely precise measurements of the properties of the known particles as well as ambitious searches for new physics, including extra Higgs bosons. While many BSM theories have been postulated, no experimental results have been able to confirm or even hint that any of them could accurately describe Nature better than the SM does. Only by gathering more data and continuously trying to take down any barrier that prevents us from probing higher energies, will we be able to guide theoreticians toward the formulation of a more complete theory.

Bibliography

[1] MissMJ. URL https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[2] The CMS collaboration. The performance of the CMS muon detector in proton-proton collisions at $\sqrt{s}$= 7 TeV at the LHC. *Journal of Instrumentation*, 8(11):P11002–P11002, nov 2013. doi: 10.1088/1748-0221/8/11/p11002. URL https://doi.org/10.1088%2F1748-0221%2F8%2F11%2Fp11002.

[3] Mark Thomson. *Modern particle physics*. Cambridge University Press, New York, 2013. ISBN 9781107034266. URL http://www-spires.fnal.gov/spires/find/books/www?cl=QC793.2.T46::2013.

[4] The L3 Collaboration The OPAL Collaboration The SLD Collaboration The LEP Electroweak Working Group The SLD Electroweak The ALEPH Collaboration, The DELPHI Collaboration and Heavy Flavour Groups. Precision electroweak measurements on the z resonance. *Physics Reports*, 427(5): 257 – 454, 2006. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2005.12.006. URL http://www.sciencedirect.com/science/article/pii/S0370157305005119.

[5] C. N. Yang and R. L. Mills. Conservation of isotopic spin and isotopic gauge invariance. *Phys. Rev.*, 96:191–195, Oct 1954. doi: 10.1103/PhysRev.96.191. URL https://link.aps.org/doi/10.1103/PhysRev.96.191.

[6] Enrico Fermi. Tentativo di una teoria dei raggi β. *Il Nuovo Cimento (1924-1942)*, 11(1):1, Sep 2008. ISSN 1827-6121. doi: 10.1007/BF02959820. URL https://doi.org/10.1007/BF02959820.

[7] R. P. Feynman. Mathematical formulation of the quantum theory of electromagnetic interaction. *Phys. Rev.*, 80:440–457, Nov 1950. doi: 10.1103/PhysRev.80.440. URL https://link.aps.org/doi/10.1103/PhysRev.80.440.

[8] R. P. Feynman. The theory of positrons. *Phys. Rev.*, 76:749–759, Sep 1949. doi: 10.1103/PhysRev.76.749. URL https://link.aps.org/doi/10.1103/PhysRev.76.749.

[9] R. P. Feynman. Space-time approach to quantum electrodynamics. *Phys. Rev.*, 76:769–789, Sep 1949. doi: 10.1103/PhysRev.76.769. URL https://link.aps.org/doi/10.1103/PhysRev.76.769.

[10] Julian Schwinger. Quantum electrodynamics. i. a covariant formulation. *Phys. Rev.*, 74:1439–1461, Nov 1948. doi: 10.1103/PhysRev.74.1439. URL https://link.aps.org/doi/10.1103/PhysRev.74.1439.

[11] Julian Schwinger. On quantum-electrodynamics and the magnetic moment of the electron. *Phys. Rev.*, 73:416–417, Feb 1948. doi: 10.1103/PhysRev.73.416. URL https://link.aps.org/doi/10.1103/PhysRev.73.416.

[12] S. Tomonaga. On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields*. *Progress of Theoretical Physics*, 1(2):27–42, 08 1946. ISSN 0033-068X. doi: 10.1143/PTP.1.27. URL https://doi.org/10.1143/PTP.1.27.

[13] Julian Schwinger. A theory of the fundamental interactions. *Annals of Physics*, 2(5):407 – 434, 1957. ISSN 0003-4916. doi: https://doi.org/10.1016/0003-4916(57)90015-5. URL http://www.sciencedirect.com/science/article/pii/0003491657900155.

[14] R. P. Feynman and M. Gell-Mann. Theory of the fermi interaction. *Phys. Rev.*, 109:193–198, Jan 1958. doi: 10.1103/PhysRev.109.193. URL https://link.aps.org/doi/10.1103/PhysRev.109.193.

[15] E. C. G. Sudarshan and R. E. Marshak. Chirality invariance and the universal fermi interaction. *Phys. Rev.*, 109:1860–1862, Mar 1958. doi: 10.1103/PhysRev.109.1860.2. URL https://link.aps.org/doi/10.1103/PhysRev.109.1860.2.

[16] Sheldon L. Glashow. Partial-symmetries of weak interactions. *Nuclear Physics*, 22(4):579 – 588, 1961. ISSN 0029-5582. doi: https://doi.org/10.1016/0029-5582(61)90469-2. URL http://www.sciencedirect.com/science/article/pii/0029558261904692.

[17] Jeffrey Goldstone, Abdus Salam, and Steven Weinberg. Broken symmetries. *Phys. Rev.*, 127:965–970, Aug 1962. doi: 10.1103/PhysRev.127.965. URL https://link.aps.org/doi/10.1103/PhysRev.127.965.

[18] P. W. Anderson. Plasmons, gauge invariance, and mass. *Phys. Rev.*, 130:439–442, Apr 1963. doi: 10.1103/PhysRev.130.439. URL https://link.aps.org/doi/10.1103/PhysRev.130.439.

[19] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964. doi: 10.1103/PhysRevLett.13.321. URL https://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[20] P.W. Higgs. Broken symmetries, massless particles and gauge fields. *Physics Letters*, 12(2):132 – 133, 1964. ISSN 0031-9163. doi: https://doi.org/10.1016/0031-9163(64)91136-9. URL http://www.sciencedirect.com/science/article/pii/0031916364911369.

[21] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13:508–509, Oct 1964. doi: 10.1103/PhysRevLett.13.508. URL https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[22] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. Global conservation laws and massless particles. *Phys. Rev. Lett.*, 13:585–587, Nov 1964. doi: 10.1103/PhysRevLett.13.585. URL https://link.aps.org/doi/10.1103/PhysRevLett.13.585.

[23] G. Arnison et al. Experimental observation of isolated large transverse energy electrons with associated missing energy at s=540 gev. *Physics Letters B*, 122(1):103 – 116, 1983. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(83)91177-2. URL http://www.sciencedirect.com/science/article/pii/0370269383911772.

[24] M. Banner et al. Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the cern pp collider. *Physics Letters B*, 122(5):476 – 485, 1983. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(83)91605-2. URL http://www.sciencedirect.com/science/article/pii/0370269383916052.

[25] G. Arnison et al. Experimental observation of lepton pairs of invariant mass around 95 gev/c2 at the cern sps collider. *Physics Letters B*, 126(5):398 – 410, 1983. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(83)90188-0. URL http://www.sciencedirect.com/science/article/pii/0370269383901880.

[26] P. Bagnaia et al. Evidence for z0 decay to e+e- at the cern pp collider. *Physics Letters B*, 129(1):130 – 140, 1983. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(83)90744-X. URL http://www.sciencedirect.com/science/article/pii/037026938390744X.

[27] Combined CDF and D0 Upper Limits on Standard Model Higgs-Boson Production with up to 2.4 fb^{-1} of data. 2008.

[28] The ATLAS collaboration. Measurements of the higgs boson production and decay rates and constraints on its couplings from a combined atlas and cms analysis of the lhc pp collision data at $\sqrt{s} = 7$ and 8 TeV. *Journal of High Energy Physics*, 2016(8):45, Aug 2016. ISSN 1029-8479. doi: 10.1007/JHEP08(2016)045. URL https://doi.org/10.1007/JHEP08(2016)045.

[29] Y. Fukuda and Hayakawa et al. Evidence for oscillation of atmospheric neutrinos. *Phys. Rev. Lett.*, 81:1562–1567, Aug 1998. doi: 10.1103/PhysRevLett.81.1562. URL https://link.aps.org/doi/10.1103/PhysRevLett.81.1562.

[30] Q. R. et al. Ahmad. Direct evidence for neutrino flavor transformation from neutral-current interactions in the sudbury neutrino observatory. *Phys. Rev. Lett.*, 89:011301, Jun 2002. doi: 10.1103/PhysRevLett.89.011301. URL https://link.aps.org/doi/10.1103/PhysRevLett.89.011301.

[31] J. Schechter and J. W. F. Valle. Neutrino masses in su(2) $\bigotimes$ u(1) theories. *Phys. Rev. D*, 22:2227–2235, Nov 1980. doi: 10.1103/PhysRevD.22.2227. URL https://link.aps.org/doi/10.1103/PhysRevD.22.2227.

[32] Aharon Davidson and Kameshwar C. Wali. Family mass hierarchy from universal seesaw mechanism. *Phys. Rev. Lett.*, 60:1813–1816, May 1988. doi: 10.1103/PhysRevLett.60.1813. URL https://link.aps.org/doi/10.1103/PhysRevLett.60.1813.

[33] Rabindra N. Mohapatra and Goran Senjanović. Neutrino masses and mixings in gauge models with spontaneous parity violation. *Phys. Rev. D*, 23:165–180, Jan 1981. doi: 10.1103/PhysRevD.23.165. URL https://link.aps.org/doi/10.1103/PhysRevD.23.165.

[34] Rabindra N. Mohapatra and Goran Senjanović. Neutrino mass and spontaneous parity nonconservation. *Phys. Rev. Lett.*, 44:912–915, Apr 1980. doi: 10.1103/PhysRevLett.44.912. URL https://link.aps.org/doi/10.1103/PhysRevLett.44.912.

[35] Murray Gell-Mann, Pierre Ramond, and Richard Slansky. Complex Spinors and Unified Theories. *Conf. Proc.*, C790927:315–321, 1979.

[36] Peter Minkowski. $\mu \to e\gamma$ at a rate of one out of 109 muon decays? *Physics Letters B*, 67(4):421 – 428, 1977. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(77)90435-X. URL http://www.sciencedirect.com/science/article/pii/037026937790435X.

[37] Stephen P. Martin. A Supersymmetry primer. pages 1–98, 1997. doi: 10.1142/9789812839657_0001,10.1142/9789814307505_0001. [Adv. Ser. Direct. High Energy Phys.18,1(1998)].

[38] John F Gunion, Sally Dawson, Howard E Haber, and Gordon L Kane. *The Higgs hunter's guide*, volume 80. Brookhaven Nat. Lab., Upton, NY, 1989. URL https://cds.cern.ch/record/425736. In the second printing (1990) by Perseus Books in the collection Frontiers in physics, no 80, a number of errors and omissions are corrected and the references at the end of each chapter are updated. A paperback reprint of the 1990 edition has been published in 2000.

[39] V. Barger, J. L. Hewett, and R. J. N. Phillips. New constraints on the charged higgs sector in two-higgs-doublet models. *Phys. Rev. D*, 41:3421–3441, Jun 1990. doi: 10.1103/PhysRevD.41.3421. URL https://link.aps.org/doi/10.1103/PhysRevD.41.3421.

[40] Yorikiyo Nagashima. *Beyond the standard model of elementary particle physics*. Wiley-VCH, Weinheim, USA, 2014. ISBN 9783527411771, 9783527665051. URL http://www.wiley-vch.de/publish/dt/books/ISBN3-527-41177-1.

[41] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, Pavel M. Nadolsky, and W. K. Tung. New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, 07:012, 2002. doi: 10.1088/1126-6708/2002/07/012.

[42] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt. Parton distributions for the LHC. *Eur. Phys. J.*, C63:189–285, 2009. doi: 10.1140/epjc/s10052-009-1072-5.

[43] Richard D. Ball, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Jose I. Latorre, Andrea Piccione, Juan Rojo, and Maria Ubiali. A Determination of parton distributions with faithful uncertainty estimation. *Nucl. Phys.*, B809: 1–63, 2009. doi: 10.1016/j.nuclphysb.2008.09.037,10.1016/j.nuclphysb.2009.02.027. [Erratum: Nucl. Phys.B816,293(2009)].

[44] Jonathan M. Butterworth, Guenther Dissertori, and Gavin P. Salam. Hard Processes in Proton-Proton Collisions at the Large Hadron Collider. *Ann. Rev. Nucl. Part. Sci.*, 62:387–405, 2012. doi: 10.1146/annurev-nucl-102711-094913.

[45] S et al. Dittmaier. *Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables*. CERN Yellow Reports: Monographs. CERN, Geneva, 2011. doi: 10.5170/CERN-2011-002. URL https://cds.cern.ch/record/1318996. Comments: 153 pages, 43 figures, to be submitted to CERN Report. Working Group web page: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CrossSections.

[46] Paolo Bolzoni, Fabio Maltoni, Sven-Olaf Moch, and Marco Zaro. Vector boson fusion at next-to-next-to-leading order in qcd: Standard model higgs boson and beyond. *Phys. Rev. D*, 85:035002, Feb 2012. doi: 10.1103/PhysRevD.85.035002. URL https://link.aps.org/doi/10.1103/PhysRevD.85.035002.

[47] D. et al. de Florian. *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*. CERN Yellow Reports: Monographs. Oct 2016. doi: 10.23731/CYRM-2017-002. URL http://cds.cern.ch/record/2227475. 869 pages, 295 figures, 248 tables and 1645 citations. Working Group web page: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG.

[48] C. Patrignani et al. Review of Particle Physics. *Chin. Phys.*, C40(10):100001, 2016. doi: 10.1088/1674-1137/40/10/100001.

[49] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock. *LHC Design Report*. CERN Yellow Reports: Monographs. CERN, Geneva, 2004. doi: 10.5170/CERN-2004-003-V-1. URL https://cds.cern.ch/record/782076.

[50] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock. *LHC Design Report*. CERN Yellow Reports: Monographs. CERN, Geneva, 2004. doi: 10.5170/CERN-2004-003-V-2. URL https://cds.cern.ch/record/815187.

[51] Michael Benedikt, Paul Collier, V Mertens, John Poole, and Karlheinz Schindl. *LHC Design Report*. CERN Yellow Reports: Monographs. CERN, Geneva, 2004. doi: 10.5170/CERN-2004-003-V-3. URL https://cds.cern.ch/record/823808.

[52] V Karimäki, M Mannelli, P Siegrist, H Breuker, A Caner, R Castaldi, K Freudenreich, G Hall, R Horisberger, M Huhtinen, and A Cattai. *The CMS tracker system project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997. URL https://cds.cern.ch/record/368412.

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

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

[55] Philippe Bloch, Robert Brown, Paul Lecoq, and Hans Rykaczewski. *Changes to CMS ECAL electronics: addendum to the Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 2002. URL http://cds.cern.ch/record/581342.

[56] Q Ingram. Energy resolution of the barrel of the CMS electromagnetic calorimeter. *Journal of Instrumentation*, 2(04):P04004–P04004, apr 2007. doi: 10.1088/1748-0221/2/04/p04004. URL https://doi.org/10.1088%2F1748-0221%2F2%2F04%2Fp04004.

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

[58] *The CMS muon project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997. URL https://cds.cern.ch/record/343814.

[59] Johan Alwall, Michel Herquet, Fabio Maltoni, Olivier Mattelaer, and Tim Stelzer. Madgraph 5: going beyond. *Journal of High Energy Physics*, 2011 (6):128, Jun 2011. ISSN 1029-8479. doi: 10.1007/JHEP06(2011)128. URL https://doi.org/10.1007/JHEP06(2011)128.

[60] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing nlo calculations in shower monte carlo programs: the powheg box. *Journal of High Energy Physics*, 2010(6):43, Jun 2010. ISSN 1029-8479. doi: 10.1007/JHEP06(2010)043. URL https://doi.org/10.1007/JHEP06(2010)043.

[61] Stefano Frixione and Bryan R Webber. Matching NLO QCD computations and parton shower simulations. *Journal of High Energy Physics*, 2002(06): 029–029, jun 2002. doi: 10.1088/1126-6708/2002/06/029. URL https://doi.org/10.1088%2F1126-6708%2F2002%2F06%2F029.

[62] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand. Parton fragmentation and string dynamics. , 97:31–145, July 1983. doi: 10.1016/0370-1573(83)90080-7.

[63] J.-C. Winter, F. Krauss, and G. Soff. A modified cluster-hadronisation model. *The European Physical Journal C - Particles and Fields*, 36(3):381–395, Aug 2004. ISSN 1434-6052. doi: 10.1140/epjc/s2004-01960-8. URL https://doi.org/10.1140/epjc/s2004-01960-8.

[64] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. A brief introduction to pythia 8.1. *Computer Physics Communications*, 178(11):852 – 867, 2008. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2008.01.036. URL http://www.sciencedirect.com/science/article/pii/S0010465508000441.

[65] Johannes et al. Bellm. Herwig 7.0/herwig++ 3.0 release note. *The European Physical Journal C*, 76(4):196, Apr 2016. ISSN 1434-6052. doi: 10.1140/epjc/s10052-016-4018-8. URL https://doi.org/10.1140/epjc/s10052-016-4018-8.

[66] S. Agostinelli et al. Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250 – 303, 2003. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(03)01368-8. URL http://www.sciencedirect.com/science/article/pii/S0168900203013688.

[67] The CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *Journal of Instrumentation*, 9 (10):P10009–P10009, oct 2014. doi: 10.1088/1748-0221/9/10/p10009. URL https://doi.org/10.1088%2F1748-0221%2F9%2F10%2Fp10009.

[68] Studies of Tracker Material. Technical Report CMS-PAS-TRK-10-003, 2010. URL http://cds.cern.ch/record/1279138.

[69] The CMS Collaboration. Cms tracking performance results from early lhc operation. *The European Physical Journal C*, 70(4):1165–1192, Dec 2010. ISSN 1434-6052. doi: 10.1140/epjc/s10052-010-1491-3. URL https://doi.org/10.1140/epjc/s10052-010-1491-3.

[70] Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s}$= 8 TeV. *Journal of Instrumentation*, 10 (06):P06005–P06005, jun 2015. doi: 10.1088/1748-0221/10/06/p06005. URL https://doi.org/10.1088%2F1748-0221%2F10%2F06%2Fp06005.

[71] W Adam, R Frühwirth, A Strandlie, and T Todorov. Reconstruction of electrons with the gaussian-sum filter in the CMS tracker at the LHC. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9–N20, jul 2005. doi: 10.1088/0954-3899/31/9/n01. URL https://doi.org/10.1088%2F0954-3899%2F31%2F9%2Fn01.

[72] Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s}$= 7 TeV. *Journal of Instrumentation*, 8(09):P09009–P09009, sep 2013. doi: 10.1088/1748-0221/8/09/p09009. URL https://doi.org/10.1088%2F1748-0221%2F8%2F09%2Fp09009.

[73] The CMS collaboration. Performance of CMS muon reconstruction in pp collision events at$\sqrt{s}$ = 7tev. *Journal of Instrumentation*, 7(10):P10002–P10002, oct 2012. doi: 10.1088/1748-0221/7/10/p10002. URL https://doi.org/10.1088%2F1748-0221%2F7%2F10%2Fp10002.

[74] Daniel Guest, Julian Collado, Pierre Baldi, Shih-Chieh Hsu, Gregor Urban, and Daniel Whiteson. Jet flavor classification in high-energy physics with deep neural networks. *Phys. Rev. D*, 94:112002, Dec 2016. doi: 10.1103/PhysRevD.94.112002. URL https://link.aps.org/doi/10.1103/PhysRevD.94.112002.

[75] Gilles Louppe, Kyunghyun Cho, Cyril Becot, and Kyle Cranmer. QCD-Aware Recursive Neural Networks for Jet Physics. *JHEP*, 01:057, 2019. doi: 10.1007/JHEP01(2019)057.

[76] Reconstruction and identification of lepton decays to hadrons and at CMS. *Journal of Instrumentation*, 11(01):P01019–P01019, jan 2016. doi: 10.1088/1748-0221/11/01/p01019. URL https://doi.org/10.1088%2F1748-0221%2F11%2F01%2Fp01019.

[77] Wolfgang Waltenberger, Rudolf Frühwirth, and Pascal Vanlaer. Adaptive vertex fitting. *Journal of Physics G: Nuclear and Particle Physics*, 34(12):N343–N356, nov 2007. doi: 10.1088/0954-3899/34/12/n01. URL https://doi.org/10.1088%2F0954-3899%2F34%2F12%2Fn01.

[78] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989. ISSN 1435-568X. doi: 10.1007/BF02551274. URL https://doi.org/10.1007/BF02551274.

[79] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. , 323(6088):533–536, Oct 1986. doi: 10.1038/323533a0.

[80] Nielsen M. URL http://neuralnetworksanddeeplearning.com/index.html.

[81] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[82] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014. URL http://arxiv.org/abs/1412.3555.

[83] scikit-learn developers. Robustscaler. URL https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html.

[84] François Chollet et al. Keras. https://keras.io, 2015.

[85] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[86] M. et al. Aaboud. Cross-section measurements of the higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$ with the atlas detector. *Phys. Rev. D*, 99:072001, Apr 2019. doi: 10.1103/PhysRevD.99.072001. URL https://link.aps.org/doi/10.1103/PhysRevD.99.072001.

[87] The CMS Collaboration. Observation of the higgs boson decay to a pair of leptons with the cms detector. *Physics Letters B*, 779:283 – 316, 2018. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2018.02.004. URL http://www.sciencedirect.com/science/article/pii/S0370269318301035.

[88] M. Carena, S. Heinemeyer, C.E.M. Wagner, and G. Weiglein. Suggestions for benchmark scenarios for mssm higgs boson searches at hadron colliders. *The European Physical Journal C - Particles and Fields*, 26(4):601–607, Feb 2003. ISSN 1434-6052. doi: 10.1140/epjc/s2002-01084-3. URL https://doi.org/10.1140/epjc/s2002-01084-3.

[89] S. et al. Schael. Search for neutral mssm higgs bosons at lep. *The European Physical Journal C - Particles and Fields*, 47(3):547, Jul 2006. ISSN 1434-6052. doi: 10.1140/epjc/s2006-02569-7. URL https://doi.org/10.1140/epjc/s2006-02569-7.

[90] Doug Benjamin et al. Combined CDF and D0 Upper Limits on MSSM Higgs Boson Production in tau-tau Final States with up to 2.2 fb-1. 2010.

[91] The ATLAS collaboration. Search for additional heavy neutral higgs and gauge bosons in the ditau final state produced in 36 fb1 of pp collisions at $\sqrt{s} = 13\,\text{TeV}$ with the atlas detector. *Journal of High Energy Physics*, 2018 (1):55, Jan 2018. ISSN 1029-8479. doi: 10.1007/JHEP01(2018)055. URL https://doi.org/10.1007/JHEP01(2018)055.

[92] The CMS collaboration. Search for additional neutral mssm higgs bosons in the τ τ final state in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$. *Journal of High Energy Physics*, 2018(9):7, Sep 2018. ISSN 1029-8479. doi: 10.1007/JHEP09(2018)007. URL https://doi.org/10.1007/JHEP09(2018)007.

[93] Stanisław Jadach, Johann H. Kühn, and Zbigniew Was. Tauola - a library of monte carlo programs to simulate decays of polarized leptons. *Computer Physics Communications*, 64(2):275 – 299, 1991. ISSN 0010-4655. doi: https://

doi.org/10.1016/0010-4655(91)90038-M. URL http://www.sciencedirect.com/science/article/pii/001046559190038M.

[94] A.M. Sirunyan et al. Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. *Journal of Instrumentation*, 13(05):P05011–P05011, may 2018. doi: 10.1088/1748-0221/13/05/p05011. URL https://doi.org/10.1088%2F1748-0221%2F13%2F05%2Fp05011.

[95] CMS Collaboration. An embedding technique to determine genuine $\tau\tau$ backgrounds from CMS data. 2018.

[96] The CMS collaboration. Measurement of the inclusive w and z production cross sections in pp collisions at $\sqrt{s} = 7\,\text{TeV}$ with the cms experiment. *Journal of High Energy Physics*, 2011(10):132, Oct 2011. ISSN 1029-8479. doi: 10.1007/JHEP10(2011)132. URL https://doi.org/10.1007/JHEP10(2011)132.

[97] The CMS collaboration. Determination of jet energy calibration and transverse momentum resolution in CMS. *Journal of Instrumentation*, 6(11):P11002–P11002, nov 2011. doi: 10.1088/1748-0221/6/11/p11002. URL https://doi.org/10.1088%2F1748-0221%2F6%2F11%2Fp11002.

[98] Vardan Khachatryan et al. Measurement of the differential cross section for top quark pair production in pp collisions at $\sqrt{s} = 8\,\text{TeV}$. *Eur. Phys. J.*, C75 (11):542, 2015. doi: 10.1140/epjc/s10052-015-3709-x.

[99] CMS Collaboration. CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV. 2018.

[100] Procedure for the LHC Higgs boson search combination in Summer 2011. Technical Report CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, Geneva, Aug 2011. URL http://cds.cern.ch/record/1379837.

[101] A L Read. Presentation of search results: theCLstechnique. *Journal of Physics G: Nuclear and Particle Physics*, 28(10):2693–2704, sep 2002. doi: 10.1088/0954-3899/28/10/313. URL https://doi.org/10.1088%2F0954-3899%2F28%2F10%2F313.

[102] M. Carena, S. Heinemeyer, O. Stål, C. E. M. Wagner, and G. Weiglein. Mssm higgs boson searches at the lhc: benchmark scenarios after the discovery of a higgs-like particle. *The European Physical Journal C*, 73(9):2552, Sep 2013. ISSN 1434-6052. doi: 10.1140/epjc/s10052-013-2552-1. URL https://doi.org/10.1140/epjc/s10052-013-2552-1.

[103] A.L Read. Linear interpolation of histograms. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 425(1):357 – 360, 1999. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(98)01347-3. URL http://www.sciencedirect.com/science/article/pii/S0168900298013473.

Appendix A

Mathematical extension

This appendix covers mathematical definitions not included in Chapter 1.

Gama matrices - γ^i

$$\gamma^0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad \gamma^1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}$$
$$\gamma^2 = \begin{pmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & i & 0 & 0 \\ -i & 0 & 0 & 0 \end{pmatrix} \quad \gamma^3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Chiral projector - γ^5

$$\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Pauli matrices - The generators of $SU(2)$ are the τ_i matrices defined as $\tau_i = \frac{1}{2}\sigma_i$, with σ_i as:

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_1 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Gell-Mann matrices - The generators of $SU(3)$ are the T_i matrices defined as $T_i = \frac{1}{2}\lambda_i$, with λ_i as:

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}$$

$$\lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \quad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}$$

Appendix B

Control distributions

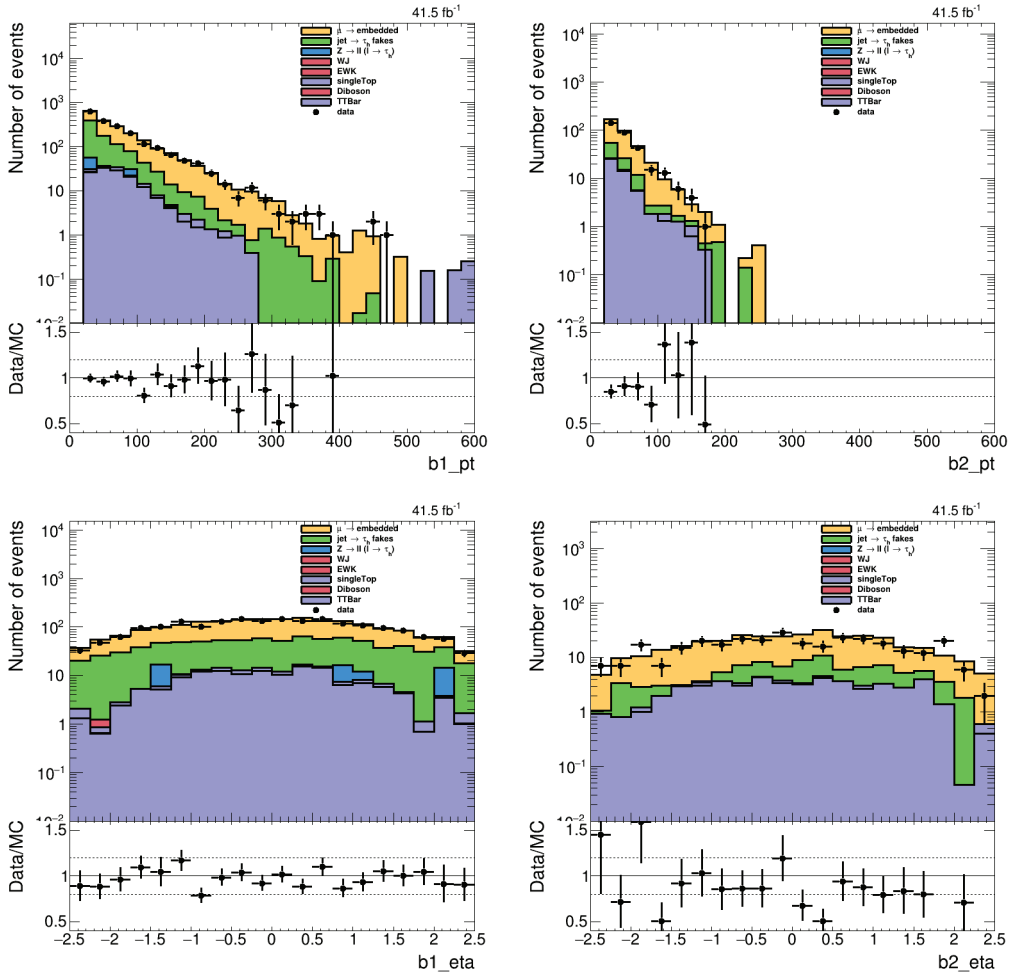


Figure B.1: Comparison of the distributions obtained from the estimation methods and observed data for the leading (b1) and sub-leading (b2) b-tagged jets kinematic variables in the b-tag category.

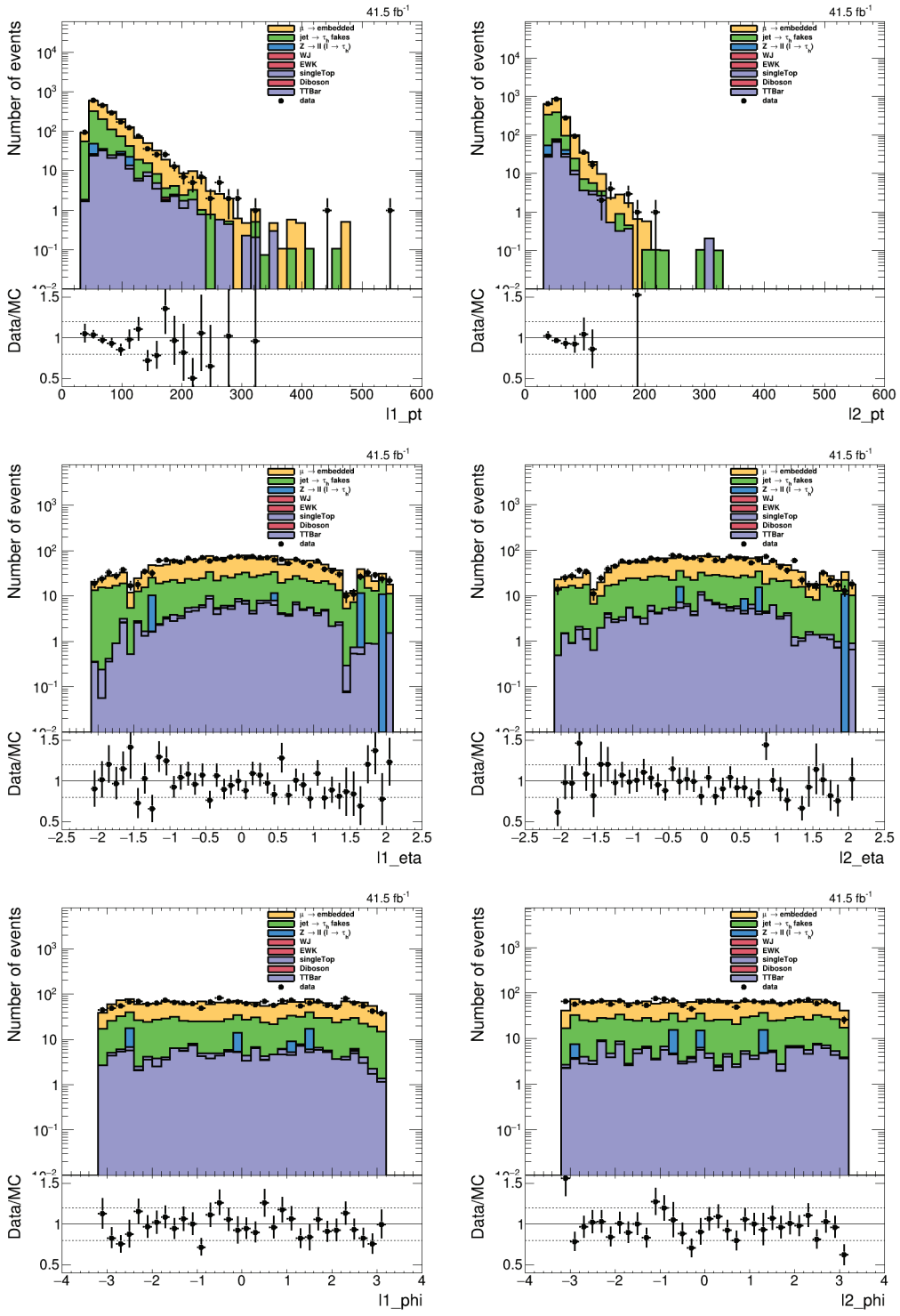


Figure B.2: Comparison of the distributions obtained from the estimation methods and observed data for the leading (l1) and sub-leading (l2) selected τ_{h} kinematic variables in the b-tag category.

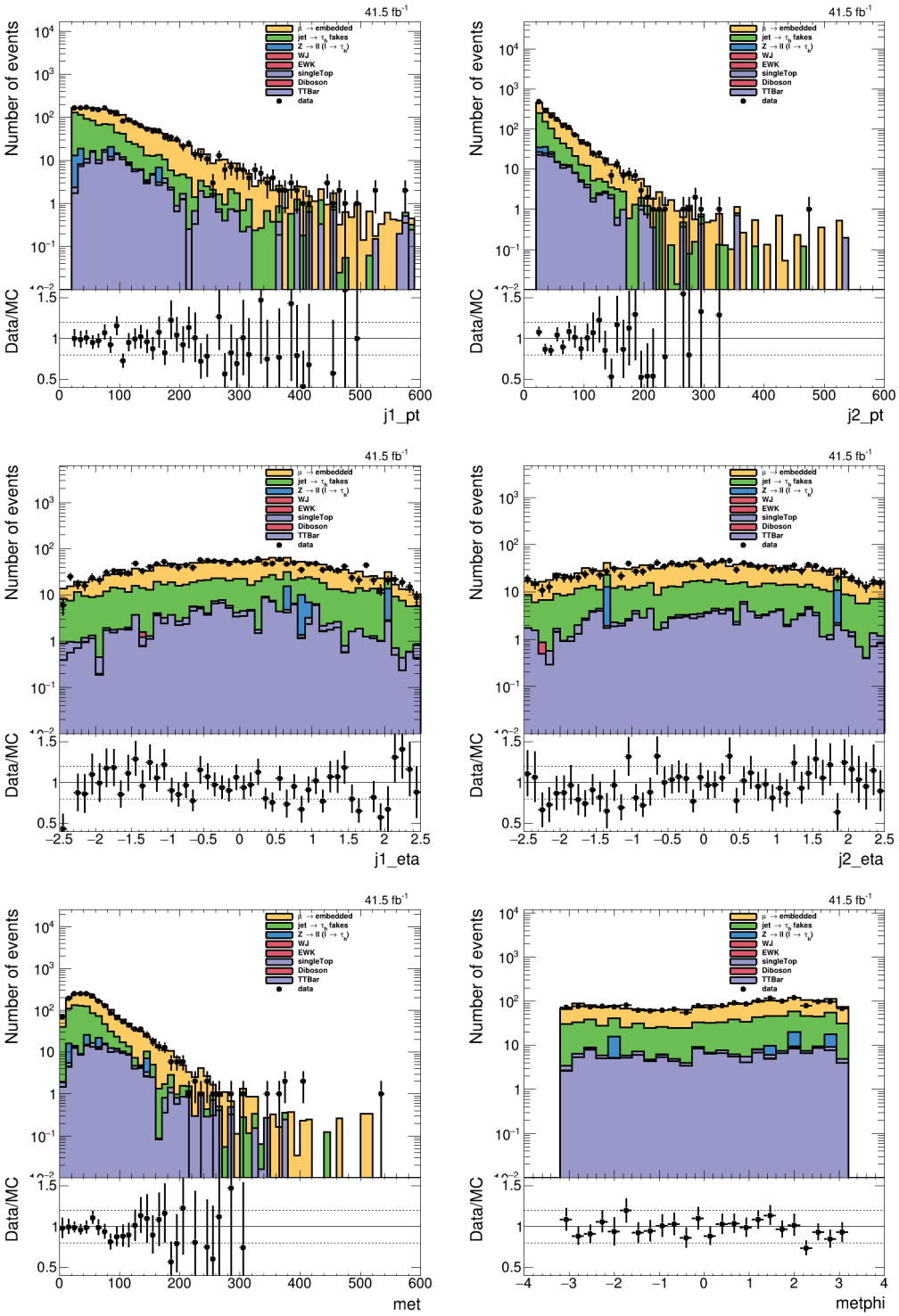


Figure B.3: Comparison of the distributions obtained from the estimation methods and observed data for the leading (j1) and sub-leading (j2) selected jets kinematic variables, as well as the E_{T}^{miss} (met) and its orientation in the transverse plane, in the b-tag category.

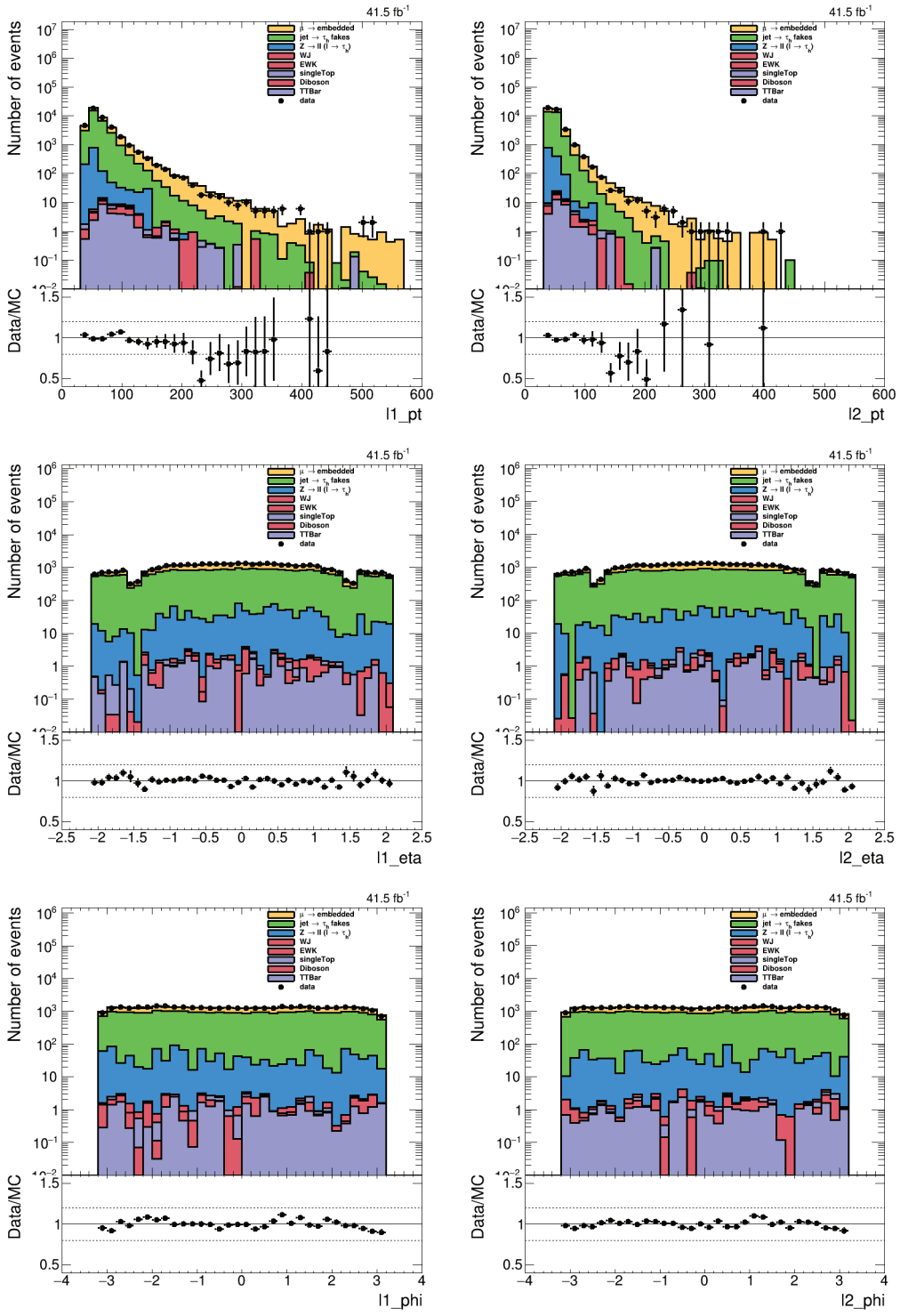


Figure B.4: Comparison of the distributions obtained from the estimation methods and observed data for the leading (l1) and sub-leading (l2) selected τ_{h} kinematic variables in the no b-tag category.

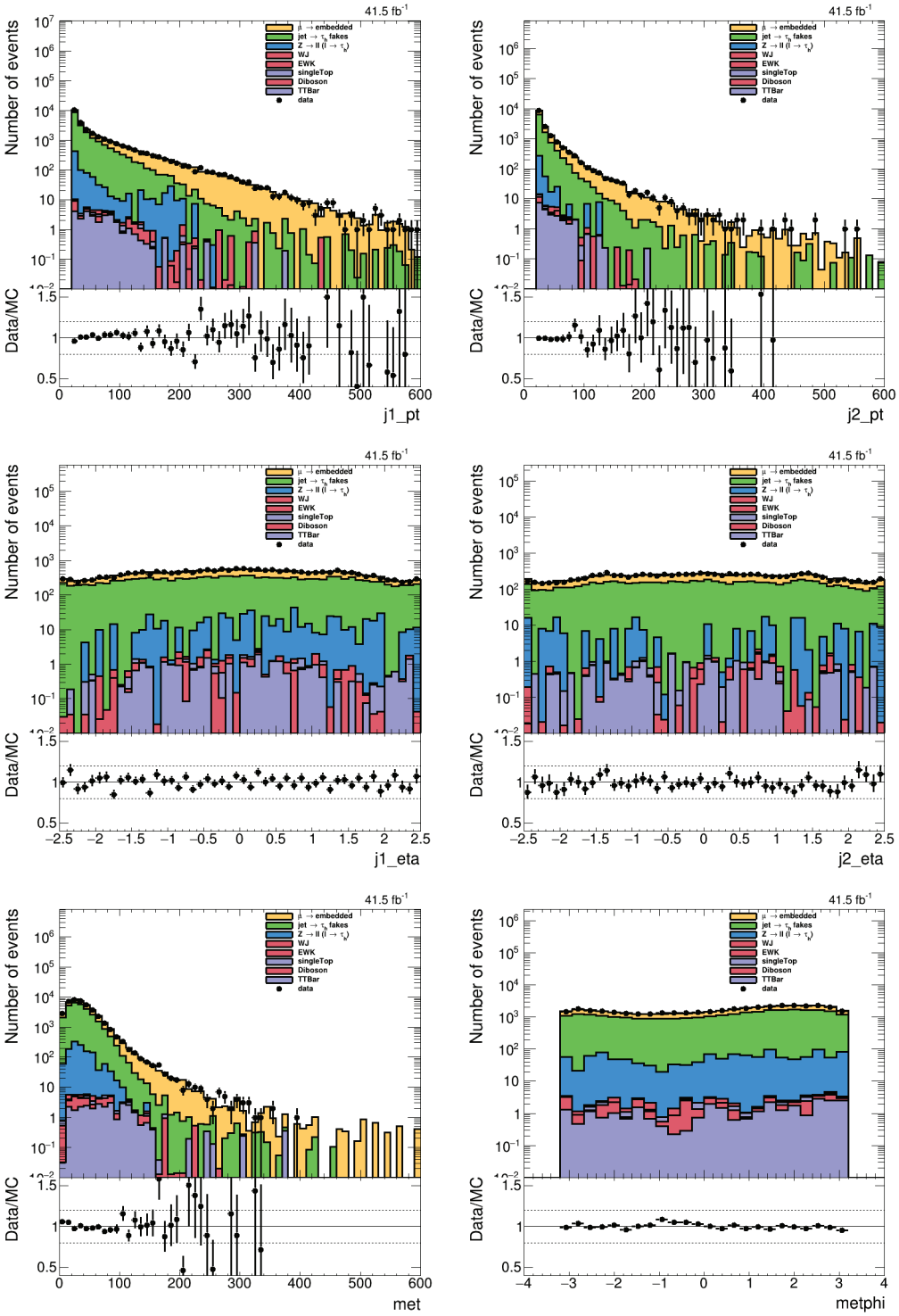


Figure B.5: Comparison of the distributions obtained from the estimation methods and observed data for the leading (j1) and sub-leading (j2) selected jets kinematic variables, as well as the E_{T}^{miss} (met) and its orientation in the transverse plane, in the no b-tag category.