Karlsruhe Institute of Technology

ETP-KA/2024-9

Search for resonant di-Higgs production in $bb + \tau\tau$ final states in pp collisions at $\sqrt{s} = 13\,\text{TeV}$

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN (Dr. rer. nat.)

von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)
angenommene

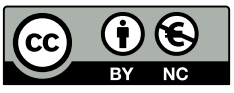
DISSERTATION

von M.Sc. Nikita Shadskiy

aus Moskau

Tag der mündlichen Prüfung: 10. Januar 2025

Referent:	Prof. Dr. Ulrich Husemann	Institut für Experimentelle Teilchenphysik
Korreferent:	Prof. Dr. Günter Quast	Institut für Experimentelle Teilchenphysik



Erklärung der selbstständigen Anfertigung der Dissertationsschrift
Hiermit erkläre ich, dass ich die Dissertation mit dem Titel

Search for resonant di-Higgs production in $bb + \tau\tau$ final states in pp collisions at $\sqrt{s} = 13$ TeV

selbstständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht habe, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde.

Ich versichere außerdem, dass ich die Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.

Karlsruhe, 4. Dezember 2024

.....
(**Nikita Shadskiy**)

Contents

1 Introduction

In the field of particle physics, the theoretical and experimental communities go hand in hand. Theorists try to find mathematical explanations of experimental observations and experimentalists try to prove or disprove existing theories that describe the physics surrounding us by conducting experiments. Currently, the most profound existing theory that describes all known fundamental particles and three out of four fundamental forces is the Standard Model (SM) of particle physics. It was developed throughout the 20th century and organizes particles into well-defined groups based on their properties and interactions, derived from underlying symmetries. Its predictions have been extensively verified, including the existence of particles like the W and Z bosons, gluon, charm quark or top quark. All of them were experimentally discovered after their theoretical formulation in the SM.

A major milestone in the success of the SM was the discovery of the Higgs boson. It was first predicted in 1964 [1–3] and finally observed in 2012 at the Large Hadron Collider (LHC) at CERN [4, 5]. This groundbreaking discovery provided experimental confirmation of the Higgs mechanism, which explains how particles acquire mass and opens the door to further studies of the Higgs sector. This sector is of particular interest because it can serve as a window into new physics beyond the SM, potentially revealing insights into unanswered questions such as the nature of dark matter or how to incorporate gravity in the same framework with the other fundamental forces.

One such beyond the SM theory is supersymmetry. Supersymmetry is a theoretical extension of the SM that introduces a symmetry linking fermions and bosons. It addresses unresolved issues in the SM and predicts new particles that can be searched for experimentally. Despite over a decade of data collection at the LHC and previous colliders before, no supersymmetric particles have been observed, suggesting that if supersymmetry exists, it likely differs from its minimal implementation (MSSM) [6]. This has shifted experimental focus to more complex extensions like the next-to-minimal supersymmetric model (NMSSM) [7, 8]. These models introduce additional parameters and degrees of freedom that can be part of future discoveries.

In this thesis, an analysis is presented that searches for the decay of a heavy scalar Higgs boson X into the discovered 125 GeV Higgs boson H_{SM} and another light scalar Higgs boson Y. Such a process is predicted e.g. by the extended Higgs sector of the NMSSM. With respect to a previous search [9], this search significantly extends the signal phase space targeted and signal processes considered, including resolved and boosted topologies of the final state particles and different decays of the Higgs bosons. The presence of Higgs bosons is inferred indirectly through the decay products, focusing on the decays into b quark and tau lepton pairs. The analysis uses data collected by the Compact Muon Solenoid (CMS) experiment at the LHC that have been recorded during the 2018 data taking period corresponding to an integrated luminosity of $59.8\,\text{fb}^{-1}$.

This analysis faces the challenge of handling the various combinations of bb and $\tau\tau$ final states, whose kinematic properties vary significantly for different (m_X, m_Y) hypotheses. Each combination of resolved and boosted topologies requires a dedicated reconstruction and identification strategy. Further, accurately estimating background processes with event signatures similar to the signal processes adds more complexity. In this analysis, data-driven methods and simulation are used. To optimize the separation of signal and background processes and to improve the sensitivity of the search, parametric neural networks are utilized.

In the following, the theoretical foundation for the search is provided in chapter 2, focusing on the Higgs mechanism within the SM and its supersymmetric extensions. In chapter 3, an overview of the CMS detector and the reconstruction algorithms is given. Chapter 4 covers the general setup of the analysis, including the exploration of the signal phase space, event selection and background estimation methods. After that, in chapter 5 the strategy for the event classification is outlined, along with the statistical analysis and the results of this search. Finally, chapter 6 concludes the results of this thesis.

2 Standard Model of Particle Physics and Extensions

The Standard Model (SM) of particle physics provides a theoretical framework to describe the fundamental particles of nature and their interactions. It accurately predicts how these particles behave, how they are created and how annihilated, based on the symmetries that govern the universe. The main aspects of the SM will be discussed in sections 2.1 and 2.2. For more information readers are referred to [10].

Although the SM has a very successful history of predictions that lead to discoveries of new particles, it cannot describe everything. Open questions still remain, such as how gravity can be included or what dark matter is made of. These mysteries are a few of many that motivate the exploration of theories beyond the SM, one of which will be discussed in section 2.3.

2.1 Standard Model of particle physics

A complete overview of the SM particles is shown in figure 2.1. The SM contains two groups of elementary particles. The first group are fermions, which have half-integer spins and make up the matter in our universe. The second group are bosons, which have integer spins. They act as carriers of the fundamental forces that define the interactions between matter particles. Whether this division is a fundamental law of nature or simply a coincidence remains an open question.

The fundamental fermions are a group of twelve particles, each with a spin of $\frac{1}{2}$. One subgroup of fermions are leptons, which include the electron, muon, tau lepton and for each of them a corresponding neutrino. The second subgroup consists of six quarks called up, down, charm, strange, top, and bottom quarks, also referred to as flavors. The key distinction between leptons and quarks lies in their interaction with the fundamental forces. Quarks interact via the strong force, whereas leptons interact only via the electromagnetic and weak forces.

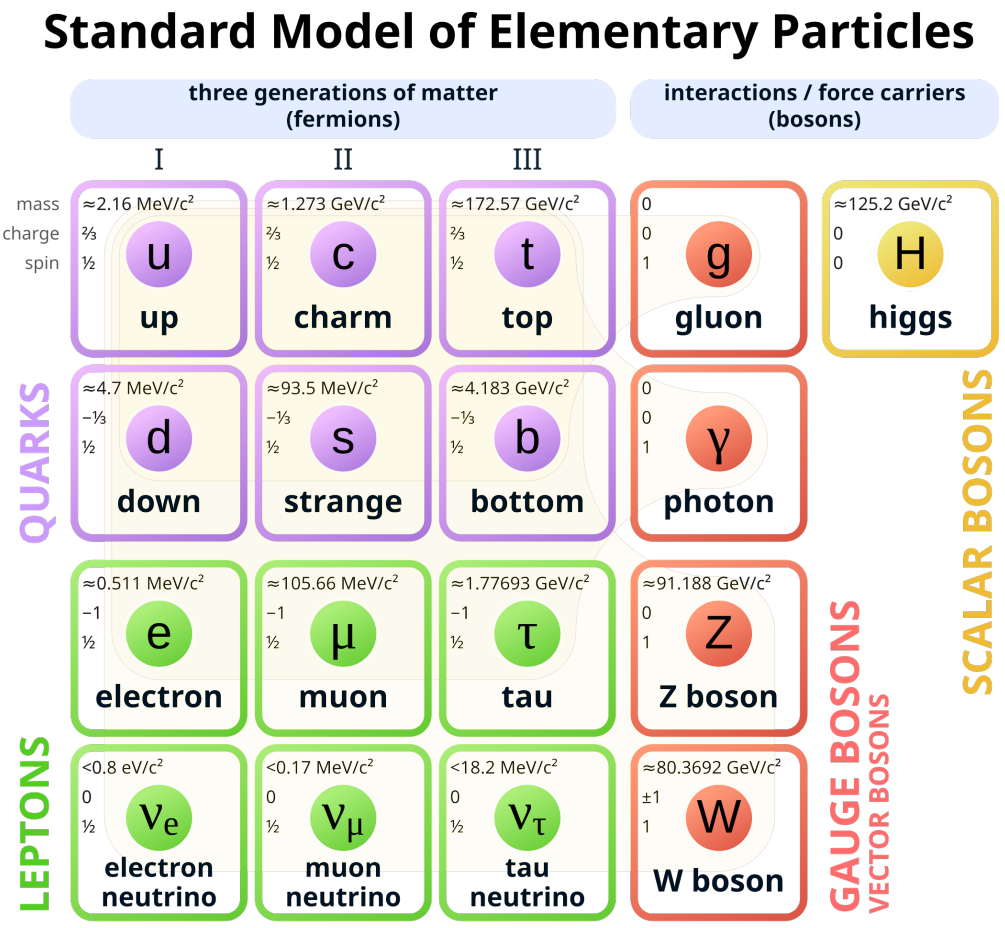


Figure 2.1: Particle content of the SM [11]. Quarks are shown in purple and leptons in green. They make up the matter in the universe. The gauge bosons in red are carriers of the fundamental forces. The Higgs boson, contrary to quarks, leptons and gauge bosons, arises from the Higgs mechanism, which is responsible for giving mass to other particles.

The everyday matter around us consists of only three fermions. The up and down quarks are building blocks of atomic nuclei, which form protons and neutrons together with electrons. Other fermions are rather unstable and rapidly decay into lighter particles. In the case of neutrinos, they cannot form bound states, which is necessary to form matter as we know.

In the SM, each fermion has a corresponding antiparticle. The additive quantum numbers of these antiparticles are inverted in comparison to their fermion partners. In this thesis, the particles and antiparticles will often be referred to with the same notation, for example, a "bottom quark" can be a bottom quark and anti-quark.

The spin one bosons in the SM include the gluons, photon and the W and Z bosons. These particles act as carriers of the fundamental forces described by the SM. In addition, the Higgs

boson is a unique particle because it has a spin of zero and does not function as a force mediator or a building block of matter. Instead, it arises from the Brout-Englert-Higgs mechanism [1–3], which is a crucial feature of the SM that explains how particles get their mass. This, shortly called, Higgs mechanism is of significant relevance to this thesis and therefore will be explained in a dedicated section 2.2.

The SM is built on the mathematical framework of a quantum field theory (QFT), which combines quantum mechanics and special relativity using the principles of classical field theories. Its foundation is based on the underlying symmetries of nature that the SM postulates, which leads to the definition of a Lagrangian function that describes the system and remains invariant under transformations of the underlying symmetries.

The Poincaré symmetry is the external symmetry of the SM, which reflects the principles of special relativity. This symmetry ensures that the laws of physics stay consistent under transformations like translations or rotations in space-time and Lorentz boosts of a system. By requiring the Poincaré symmetry, the SM ensures that its equations are Lorentz-covariant, meaning they hold true in any inertial space. Additional local gauge symmetries are related to the fundamental forces and their mediator particles.

The electromagnetic force acts on particles with electric charge and has an infinite range and therefore has noticeable effects at large scales. The mediator of the electromagnetic force is the photon. Before the quantum field theory (QFT) was developed, classical physics explained the electromagnetic force through Maxwell's equations. The SM builds on this understanding and extends it to include quantum effects that the classical approach cannot fully describe.

The weak force is comparable in strength to the electromagnetic force, but its effects are harder to observe. The mediators of the weak force are the W and Z bosons, which are very massive. This circumstance limits the weak force to a very short interaction range of about 100 times smaller than a proton. The weak force and the electromagnetic force are linked in the SM through the electroweak theory, combining both into a single fundamental force. In the SM, the electroweak force [12–14] is mathematically described by two symmetries related to two electroweak charges. The first symmetry ($SU(2)_L$) is defined for the weak isospin I, while the second symmetry ($U(1)_Y$) is related to the weak hypercharge Y. These symmetries are characterized by their generators, which are mathematical operators associated with transformations within their respective symmetry groups. For $SU(2)_L$, the generators correspond to the three components of weak isospin (I_1, I_2, I_3), while $U(1)_Y$ has a single generator corresponding to the hypercharge Y. These generators dictate how particles transform under symmetry operations and determine their associated charges.

The strong force is defined by an $SU(3)_C$ symmetry, which operates in the space of color charges. This force is carried by eight gluons that exchange color charges. The color charge is a unique property of quarks and gluons. Contrary to the other forces, the strength of the strong force weakens as the distance between two color-charged particles decreases. At extremely short distances, these particles can behave almost like free particles. This phenomenon is known

as asymptotic freedom. On the other hand, as the distance between color-charged particles increases, the energy stored in the potential field between them grows linearly. When this energy becomes large enough, it can create new particles from the quantum vacuum. This process ensures that only color-neutral combinations of particles like protons and neutrons can be observed directly. As a result, the color charge of quarks and gluons remains a hidden property that cannot be isolated and observed directly.

In the context of the mentioned gauge symmetries of the fundamental forces, the gauge refers to mathematical degrees of freedom that do not change the underlying laws of physics. For instance, the phase of a fermion field can be changed globally without altering the Lagrangian of the SM. When this is the case, the theory is said to be invariant under global gauge transformations. However, extending this invariance to local gauge symmetries breaks the symmetry of the Lagrangian. This can be resolved by introducing additional degrees of freedom to the theory. These degrees of freedom are gauge fields representing a mediator particle and each generator of the introduced local gauge symmetry needs its own gauge field and interactions with other particles. To preserve local symmetry, these mediator particles or gauge bosons must be massless.

In quantum electrodynamics (QED), the requirement of a local gauge symmetry for the $U(1)$ symmetry group predicts the existence of a massless photon. Similarly, in quantum chromodynamics (QCD), the local $SU(3)_C$ gauge symmetry leads to the prediction of eight massless gluons. These symmetries give the SM the ability to provide very precise predictions about the fundamental forces. However, a complication arises for the interactions of the electroweak force, which are described by the $SU(2)_L \times U(1)_Y$ gauge symmetry. The W and Z bosons that carry the weak force are experimentally known to have mass, but adding mass terms to the SM Lagrangian would break the required local gauge symmetry. The solution was found in the 1960s with the proposal of spontaneous symmetry breaking through the Higgs mechanism [1–3]. This mechanism provides a way to give mass to the W and Z bosons while preserving local gauge symmetry.

2.2 Higgs mechanism

The Higgs mechanism is introduced to solve the problem of the massive gauge boson of the electroweak force. The electroweak sector of the SM Lagrangian is defined based on the $SU(2)_L \times U(1)_Y$ gauge symmetry, which predicts the existence of four gauge bosons. Three of these gauge bosons are W_1, W_2 and W_3 and are a result of the $SU(2)_L$ symmetry, while the fourth gauge boson B corresponds to the $U(1)_Y$ symmetry.

These four gauge bosons do not represent the physical particles that we observe. To obtain the physical fields of the W^+, W^- and Z bosons as well as the photon γ, the gauge boson fields (W_1, W_2, W_3 and B) in the space defined by $SU(2)_L \times U(1)_Y$ are rotated. The charged bosons W^+, W^- emerge from a linear combination of W_1 and W_2,

$$\mathrm{W}^{\pm} = \frac{1}{2} \cdot (W_1 \mp \mathrm{i} W_2), \tag{2.1}$$

and the neutral bosons Z and γ are formed from a linear combination of W_3 and B by using the weak mixing angle θ_W for the rotation,

$$\mathrm{Z} = \cos(\theta_W) \cdot W_3 - \sin(\theta_W) \cdot B, \tag{2.2}$$

$$\gamma = \sin(\theta_W) \cdot W_3 + \cos(\theta_W) \cdot B. \tag{2.3}$$

This framework not only explains the weak interaction but also incorporates the electromagnetic force as part of the electroweak theory. A notable feature of the weak interaction is its violation of the parity symmetry, which means that it is not invariant under space coordinate transformation $\vec{x} \rightarrow -\vec{x}$. Specifically, this means that the W^- bosons interact only with left-handed fermions, which are particles with left-handed chirality where their spin appears opposite to their momentum, while the W^+ bosons interact only with right-handed fermions, characterized by right-handed chirality where spin aligns with their momentum. This distinction between left- and right-handed particles highlights a fundamental asymmetry, making the weak force unique among the fundamental forces.

This theory has a shortcoming which is that adding mass terms directly to the Lagrangian for the gauge bosons or fermions would break gauge invariance, making the theory incomplete. To address this, an additional Lagrangian term involving a new field ϕ is added to the electroweak Lagrangian, which is defined by

$$\mathcal{L}_{\mathrm{Higgs}} = \partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi). \tag{2.4}$$

The potential of this field is

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2 \tag{2.5}$$

and the field itself is a scalar doublet $\phi = (\phi^+, \phi^0)^T$ in the weak isospin space, consisting of two complex components. Since ϕ is complex, it has four degrees of freedom. The charged component ϕ^+ as well as the neutral component ϕ^0 have hypercharge $Y = 1$ and therefore an electric charge $Q = +1$ and $Q = 0$, respectively. These charges are determined by the relation $Q = I_3 + \frac{Y}{2}$, where I_3 is the third component of the weak isospin.

At the minimum of the potential $V(\phi)$, the energy ground state is defined. This minimum is reached at the vacuum expectation value

$$v = \sqrt{\frac{-\mu^2}{2 \cdot \lambda}}. \tag{2.6}$$

For $\lambda > 0$ and $\mu^2 < 0$, the vacuum expectation value is real and non-zero. Such a configuration breaks the symmetry of the field while still maintaining the overall structure of the theory. A visualization of the Higgs potential in figure 2.2 illustrates this symmetry breaking, where the energy ground state is no longer at the symmetric origin of the potential.

Considering the radial symmetry of the system, the field ϕ can be expressed as an expansion around its vacuum expectation value v. This expansion introduces the Higgs field H and the

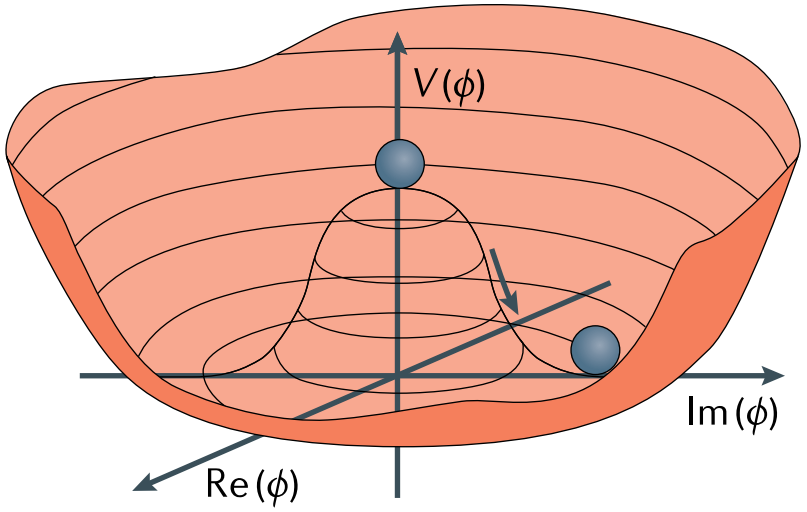


Figure 2.2: Illustration of two components of the Higgs potential $V(\phi)$ for $\mu^2 < 0$. Choosing any of the points at the bottom of the potential as the minimum spontaneously breaks the rotational symmetry. Taken from [15].

ground state of ϕ can be defined as

$$\phi_0 = \begin{pmatrix} 0 \\ v + \frac{H}{\sqrt{2}} \end{pmatrix} \tag{2.7}$$

where the charged component is zero. The expansion is specifically chosen to only affect the lower component of the scalar doublet ϕ, ensuring that the symmetry of the $U(1)$ group remains intact. This preservation of symmetry guarantees the existence of a massless photon.

With this theoretical setup, a mass term naturally arises from the interaction between the gauge bosons and the vacuum expectation value v. Usually, such a mass term would break gauge invariance, but in the Higgs mechanism the introduction of the Higgs field and its specific couplings to the gauge bosons preserves the overall gauge invariance. The introduced degrees of freedom of the field ϕ are eliminated due to the unitary gauge choice in equation 2.7, however, they are effectively re-introduced as the longitudinal polarization components of the massive gauge bosons. This process explains how the gauge bosons acquire mass while maintaining the underlying symmetry of the theory.

The scalar field ϕ_0 initially has four degrees of freedom. Three of these are used to give mass to the W^+, W^- and Z bosons. This leaves one degree of freedom remaining, which corresponds to the radial excitation of the Higgs field H. The Higgs field itself also gains mass due to its interaction with the vacuum expectation value. This mechanism of electroweak symmetry breaking, achieved through the Higgs mechanism, predicts the existence of a massive, neutral scalar particle, the Higgs boson. This prediction was confirmed with the discovery of the Higgs boson at the Large Hadron Collider (LHC) in 2012 [4, 5] and subsequent measurements of its properties.

The mechanism for giving fermions mass is different from that of the gauge bosons. Gauge bosons cannot have mass at all without the Higgs mechanism. In contrast, fermions can theoretically have mass, but adding mass terms to the theory would break the $SU(2)_L$ symmetry. This happens because the weak force interacts differently with fermions depending on their helicity. Left-handed fermions are $SU(2)_L$ doublets, while right-handed fermions are $SU(2)_L$ singlets. To maintain the symmetry, the issue of fermion masses requires a different approach. Again, the Higgs field provides a solution through what is known as a Yukawa coupling [16]. For example, the mass of the electron can be generated by adding a Yukawa interaction term to the Lagrangian, which is defined as

$$\mathcal{L}_{\text{Yukawa}} = -y_e \cdot (\bar{\psi}_L \phi_0 \psi_R + \bar{\psi}_R \phi_0 \psi_L) \tag{2.8}$$

where y_e is the Yukawa coupling constant of the electron and ψ_L and ψ_R represent the left-handed and right-handed electron fields. The mass of the electron can then be determined by the vacuum expectation value v and the Yukawa coupling y_e to the Higgs field as

$$m_e = \frac{v \cdot y_e}{\sqrt{2}}. \tag{2.9}$$

Similar terms can be added to the Lagrangian to account for the masses of other fermions, with each particle having its own Yukawa coupling to the Higgs field.

Another impact of the introduced Higgs mechanism is a distinction between the flavor eigenstates and the mass eigenstates of quarks, resulting in a mixing of the quark flavors. The Yukawa coupling for quarks is a matrix that defines a mass matrix. The mixing is introduced by diagonalizing the mass matrices of the quarks. This mixing allows quarks to transition between different quark generations and is described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix [17, 18]. The CKM matrix is a unitary 3×3 matrix that determines the probabilities of a quark transitioning to any other quark across the three generations. The mixing follows a clear hierarchy. Transitions within the same generation are much more likely than transitions between different generations. For example, the top quark almost always decays into a bottom quark because transitions from the third generation (top and bottom quarks) to other generations are highly suppressed. On the other hand, the bottom quark cannot decay into the much heavier top quark. As a result, the bottom quark has a relatively long lifetime compared to other quarks.

2.3 Supersymmetric extension of the SM Higgs sector

The SM can accurately explains many observed phenomena at particle colliders. However, it cannot describe everything. There are phenomena, which will be mentioned in the following, for which the SM still needs extensions to describe them correctly. One such extension of the SM involves supersymmetry (SUSY), a theoretical framework that establishes a symmetry between fermions and bosons. Supersymmetric theories are able to resolved some of the open questions of the SM, however, with no experimental evidence so far.

For example, including gravity as the last missing force cannot be explained by the SM. Using the supersymmetry, a local symmetry can be introduced that incorporates the general relativity into the framework of the SM [19]. Another open question in the SM is the existence of dark matter, which is supported by astrophysical observation. The SM does not predict any particle candidates that could fully explain the dark matter content of the universe. The introduction of supersymmetry goes along with naturally introducing new particles that can serve as dark matter candidates [20]. There are more problems of the SM that supersymmetry can solve, such as force unification [21, 22] or the hierarchy problem [23, 24]. More details can be found in the corresponding literature.

The Minimal Supersymmetric Standard Model (MSSM) [6] is one of the simplest extensions of the SM designed to incorporate supersymmetry. It introduces the smallest number of new particles required to create a supersymmetric framework. In this model every SM particle is paired with a supersymmetric partner. For fermions, bosonic superpartners are added and for bosons, fermionic superpartners. This further requires a second Higgs doublet to provide masses for all types of fermions, which are defined as $H_u = (H_u^+, H_u^0)^T$, $H_d = (H_d^0, H_d^-)^T$. The two Higgs doublets lead to four more degrees of freedom and five new Higgs bosons are predicted. Two of them are charged bosons $H^{\pm}$, a third is a pseudoscalar boson A and the last two are neutral scalar bosons. One of the neutral scalar bosons is expected to be relatively heavy and the other is attributed to the discovered SM Higgs boson H_{SM}.

Colliders like the LHC are designed to search for the particles introduced by supersymmetric theories like the MSSM. Although the Higgs boson H_{SM} was successfully observed and its properties are being studied, so far no evidence of supersymmetric particles has been found. This raises questions about whether supersymmetry exists in its simplest form (MSSM) or requires a more complex extension.

Because the MSSM does not parametrize all supersymmetric extensions that are possible in the SM, it has limitations. For example, it predicts that without applying any quantum corrections, the lightest neutral scalar Higgs boson has a mass below the Z boson mass ($m_Z = 91\,\text{GeV}$). However, this contradicts the observed Higgs boson mass of $125\,\text{GeV}$. Large quantum corrections must be introduced to adjust the MSSM prediction. This is possible but against the design of supersymmetry, the aim of which is to describe the underlying physics without such corrections. Besides the Higgs boson mass, there are other parameters in the MSSM that need such unnatural fine-tuning like, for example, the μ parameter of the superpotential which needs to be at the electroweak scale. Therefore, an extension of the MSSM in the Higgs sector is introduced as a consequence to resolve these issues. In the next-to-minimal supersymmetric SM (NMSSM) [7, 8, 25], a complex Higgs singlet is added to the two Higgs doublets already introduced.

In the following the NMSSM will be discussed focusing only on the Higgs sector of the superpotential because this is the relevant part of the theory for this thesis. In the NMSSM three scalar Higgs field are defined as $H_u = (H_u^+, H_u^0)^T$, $H_d = (H_d^0, H_d^-)^T$ and S. The complex scalar doublets H_u and H_d are the same as in the MSSM. In addition, the complex scalar singlet

S is introduced. The scalar part of the superpotential of the NMSSM is

$$W_{\text{NMSSM}} = \tilde{u}_R^* y_u (\tilde{Q}^T \epsilon H_u) - \tilde{d}_R^* y_d (\tilde{Q}^T \epsilon H_d) - \tilde{e}_R^* y_e (\tilde{L}^T \epsilon H_d) + \lambda S (H_u^T \epsilon H_d) + \frac{1}{3}\kappa S^3 \quad (2.10)$$

with dimensionless Yukawa couplings y_u, y_d, y_e, λ and κ and the scalar components of the supermultiplets of the NMSSM, which are summarized in table 2.1. The matrix ϵ is defined as

$$\epsilon = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \quad (2.11)$$

In the NMSSM, the structure of the superpotential closely resembles that of the MSSM, but with a significant modification. The μ term from the MSSM, written as $\mu(H_u^T \epsilon H_d)$, is replaced by two new terms proportional to the parameters λ and κ. These two parameters are free parameters within the NMSSM and define the behavior of the Higgs sector in the NMSSM.

For the NMSSM, to account for the observed Higgs boson mass without extensive fine-tuning, like in the MSSM, λ needs to be sufficiently large ($\lambda > 0.5$). On the other hand, λ is also constrained by an upper threshold of $\lambda < 0.8$ to ensure that the model remains perturbative up to the scale of the grand unified theory ($\approx 10^{16}$ GeV). This perturbative requirement ensures the mathematical consistency of the theory and the reliability of its predictions at high energies. In summary, the parameters λ and κ have enough freedom in the NMSSM to be able to resolve the problems of the MSSM.

Table 2.1: The particle content of the NMSSM is described in terms of supermultiplets, which group together particles and their supersymmetric partners. These supermultiplets are categorized based on their charges with respect to the $SU(3)_C \times SU(2)_L \times U(1)_Y$ symmetry. The superpartners of the SM components are marked with a $\sim$.

Chiral supermultiplet		Spin 0	Spin $\frac{1}{2}$	$SU(3)_C$	$SU(2)_L$	$U(1)_Y$
quark / squark	$\hat{Q}$	$\tilde{Q} = (\tilde{u}_L, \tilde{d}_L)^T$	$Q = (u_L, d_L)^T$	3	2	$\frac{1}{6}$
	$\hat{u}$	$\tilde{u}_R^*$	$u_R^\dagger$	3	1	$-\frac{2}{3}$
	$\hat{d}$	$\tilde{d}_R^*$	$d_R^\dagger$	3	1	$\frac{1}{3}$
lepton / slepton	$\hat{L}$	$\tilde{L} = (\tilde{\nu}_e, \tilde{e}_L)^T$	$L = (\nu_L, e_L)^T$	1	2	$-\frac{1}{2}$
	$\hat{e}$	$\tilde{e}_R^*$	$e_R^\dagger$	1	1	1
Higgs / Higgsino	$\hat{H}_u$	$H_u = (H_u^+, H_u^0)^T$	$\tilde{H}_u = (\tilde{H}_u^+, \tilde{H}_u^0)^T$	1	2	$\frac{1}{2}$
	$\hat{H}_d$	$H_d = (H_d^0, H_d^-)^T$	$\tilde{H}_d = (\tilde{H}_d^0, \tilde{H}_d^-)^T$	1	2	$-\frac{1}{2}$
	$\hat{S}$	S	$\hat{S}$	1	1	0
Gauge multiplets		Spin $\frac{1}{2}$	Spin 1	$SU(3)_C$	$SU(2)_L$	$U(1)_Y$
gluon / gluino		$\tilde{g}$	g	8	1	0
W boson / Wino		$\tilde{W}^\pm, \tilde{W}^0$	$W^\pm, W^0$	1	3	0
B boson / Bino		$\tilde{B}^0$	B^0	1	1	0

The scalar Higgs potential has a specific form in the NMSSM, consisting of two terms and can be written as

$$V(\Phi) = V_F + V_D \tag{2.12}$$

where V_F and V_D are derived from the superpotential W_{NMSSM} with

$$V_F = \sum_i \left| \frac{\delta W_{\text{NMSSM}}}{\delta \Phi_i} \right|^2 = |\lambda|^2 |S|^2 (H_u^\dagger H_u - H_d^\dagger H_d) + |\lambda (H_u^T \epsilon H_d) + \kappa S^2|^2 \tag{2.13}$$

and

$$V_D = \frac{1}{2} \sum_{i,j} g_a^2 (\Phi_i^\dagger \mathbf{T}^a \Phi_i)(\Phi_j^\dagger \mathbf{T}^a \Phi_j) = \frac{1}{2} g_2^2 |H_u^\dagger H_d|^2 + \frac{1}{8} (g_1^2 + g_2^2)(H_u^\dagger H_u - H_d^\dagger H_d)^2. \tag{2.14}$$

In these equations Φ represents the three Higgs fields as $\Phi = (H_u, H_d, S)$. Further, the generators of the gauge groups $U(1)_Y$ and $SU(2)_L$ are written as $\mathbf{T}^a$ with the two corresponding gauge couplings g_a.

Until now, no superpartners of SM particles have been observed, which means that supersymmetry must be broken at low energy scales, as long as it exists. Breaking supersymmetry requires adding extra terms to the potential. The exact mechanism of supersymmetry breaking is still unknown, therefore, all possible terms that preserve matter parity and avoid reintroducing issues due to quadratic divergences are added in the NMSSM. The soft supersymmetry breaking terms in the scalar Higgs potential are

$$V_{\text{soft}} = m_{H_u}^2 H_u^\dagger H_u + m_{H_d}^2 H_d^\dagger H_d + m_S^2 |S|^2 + \left(\lambda A_\lambda (H_u^T \epsilon H_d) S + \frac{1}{3} \kappa A_\kappa S^3 + \text{c.c.} \right) \tag{2.15}$$

where three mass terms are introduced for the three scalar Higgs fields together with the trilinear supersymmetry breaking parameters A_λ and A_κ. The complete scalar Higgs potential in the NMSSM is obtained by combining the three potential terms,

$$V(\Phi) = V_F + V_D + V_{\text{soft}}. \tag{2.16}$$

In summary, the Higgs potential of the NMSSM depends on seven free parameters λ, κ, A_λ, A_κ, $m_{H_u}^2$, $m_{H_d}^2$ and m_S^2. The three mass parameters should not be confused with the actual physical masses of the Higgs bosons. The physical masses are derived by diagonalizing the mixing matrices obtained during electroweak symmetry breaking.

The process of electroweak symmetry breaking in the NMSSM is similar to the mechanism in the SM. The complex scalar fields H_u, H_d and S are expanded around their vacuum expectation values v_u, v_d and v_s. The vacuum expectation values are assumed to be positive and real and

the ground states are

$$H_u = e^{i\phi_u} \begin{pmatrix} H_u^+ \\ \frac{1}{\sqrt{2}}(v_u + h_u + i a_u) \end{pmatrix}, \tag{2.17}$$

$$H_d = \begin{pmatrix} \frac{1}{\sqrt{2}}(v_d + h_d + i a_d) \\ H_d^- \end{pmatrix}, \tag{2.18}$$

$$S = \frac{1}{\sqrt{2}} e^{i\phi_s}(v_s + h_s + i a_s). \tag{2.19}$$

In these definitions, h_u, h_d and h_s are neutral CP-even states, a_u, a_d and a_s are neutral CP-odd states and H_u^+, H_d^- are charged states. Further, the phases ϕ_u and ϕ_s are used for gauge transformations to ensure the positive and real definition of the vacuum expectation values.

Similar to the SM, a change of basis allows the identification of massless Goldstone bosons, which are absorbed by the gauge bosons to provide their longitudinal degrees of freedom. The Higgs fields in the NMSSM contain ten degrees of freedom, four from each Higgs doublet and two from the Higgs singlet. As in the SM, three degrees of freedom are used to give mass to the gauge bosons. This leaves seven degrees of freedom for additional physical Higgs bosons. Two bosons are charged (H^+, H^-), two are neutral and CP-odd pseudoscalars (A_1, A_2) and three are neutral CP-even scalars (H_{SM}, Y, X). The notation of the last three bosons is not common in the theory community, but will be used in this thesis for simplification since these bosons are the target of the search.

3 CMS experiment at the LHC

Theories like the Standard Model (SM) and other theories going beyond the SM (BSM) that explain known and unknown phenomena of particle physics must be tested through experiments. The collaboration between developing these theories and testing them experimentally is a key aspect of particle physics. Part of the experimental side is the Compact Muon Solenoid (CMS) experiment located close to Geneva, Switzerland, at the Large Hadron Collider (LHC). This chapter will introduce the setup for the LHC machine that is needed to produce highly energetic particles from proton-proton (pp) collisions in section 3.1. Then the setup of the CMS detector that is needed to measure signals of the produced particles is explained in section 3.2 and afterwards the chapter focuses on the reconstruction (section 3.3) and identification (section 3.4) of these particles.

3.1 Large Hadron Collider (LHC)

The Large Hadron Collider (LHC) [26] is located at the European Organization for Nuclear Research (CERN) and is a circular particle accelerator with a 27 km circumference. It is designed to accelerate and collide hadrons, such as protons or heavy ions. The analysis in this thesis focuses on pp collisions, therefore, only they will be mentioned in the following. The LHC is the most powerful accelerator built so far and enables high energy particle collisions needed to study rare and massive particles like the Higgs boson, the top quark or not yet discovered particles. The entire CERN complex including the LHC is shown in figure 3.1.

The acceleration process of the colliding particles consists of many steps before the LHC ring. First, protons are extracted from hydrogen atoms by ionizing them with strong electric fields. The protons are first accelerated to 50 MeV in a linear accelerator (LINAC4). Then they are forwarded to a circular synchrotron (BOOSTER) where they are further accelerated to 1.4 GeV and grouped into bunches of approximately 10^{11} protons. The proton bunches are then subsequently passed to the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS) where they are accelerated in two steps to 450 GeV.

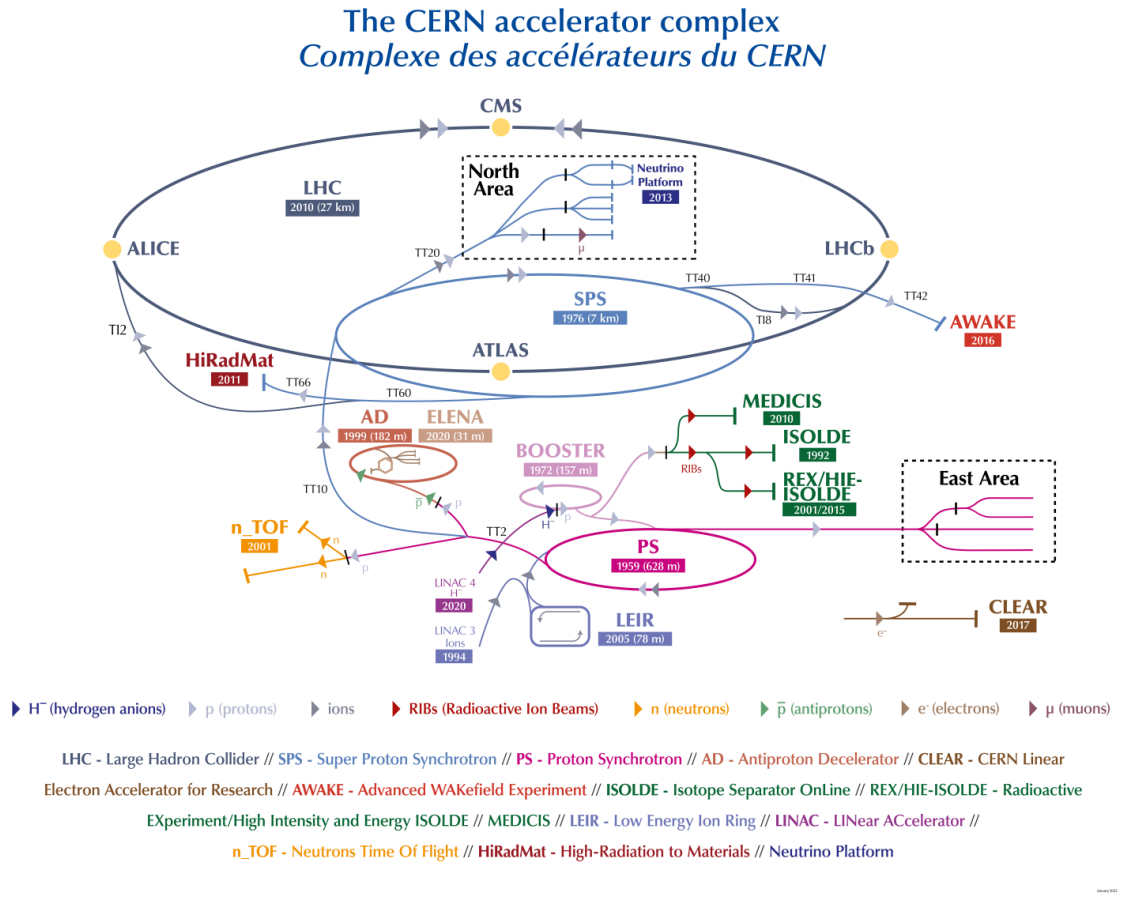


Figure 3.1: Accelerator complex at CERN [27]. Protons are generated by ionizing hydrogen gas. Before they are injected into the LHC, they are pre-accelerated with smaller accelerators starting with LINAC4 and then BOOSTER, Proton Synchrotron (PS) and Super Proton Synchrotron (SPS).

After all that, the protons are finally injected into the LHC, which uses powerful superconducting dipole magnets to bend the particles on a circular trajectory. Additionally, quadrupole and higher order magnets are used to focus the proton beam. To accelerate the protons even further, superconducting radio frequency (RF) cavities are located along the beam line. During the 2018 data taking period on which this thesis is focusing, the protons were accelerated to an energy of 6.5 TeV.

The LHC contains two separate beam pipes where proton bunches circulate in opposite directions. At four special points along the beam line it is possible to collide the two beams. These points are called interaction points and around them big experiments are located to measure the results of these collisions. The focus of this thesis will be on the CMS experiment [28] as one of two large 4π multi-purpose experiments at LHC. The other rivaling experiment is ATLAS [29].

3.1.1 Luminosity

The interesting processes produced from pp collisions often have small production cross sections (production probabilities), therefore, a high number of collisions is needed to be able to analyze them. To achieve this, proton bunches with approximately 10^{11} protons are collided at the LHC every $25\,\text{ns}$. The collision rate can be defined based on the total cross section for pp collisions σ_{pp} and the instantaneous luminosity L of the LHC,

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \sigma_{\text{pp}} \cdot L. \tag{3.1}$$

The instantaneous luminosity is a quantity that defines the potential of an accelerator and can, for head on collisions, be calculated from beam parameters as

$$L = \frac{N_b^2 \cdot n_b \cdot f_{\text{rev}}}{4 \cdot \pi \cdot \sigma_x \cdot \sigma_y} \tag{3.2}$$

where N_b is the number of protons in one bunch, n_b is the number of colliding bunches in the beam, f_{rev} is the revolution frequency with which the bunches are collided at the LHC and σ_x/σ_y are the widths of the bunches in the x-y plane.

To quantify the amount of data taken at the LHC over a longer period of time, the integrated luminosity is calculated,

$$L_{\text{int}} = \int \mathrm{d}t \cdot L. \tag{3.3}$$

With the information about the integrated luminosity, cross section measurements of interesting processes can be performed by counting how many of such processes happened in the taken data.

The analysis in this thesis will focus on data taken at the CMS experiment during the 2018 period. The integrated luminosity delivered by the LHC during that time is $59.8\,\text{fb}^{-1}$.

3.2 Compact Muon Solenoid (CMS) detector

The CMS experiment [28] at the LHC is one of the most advanced particle detectors ever built, enabling exploration of the fundamentals of particle physics. This experiment is designed to investigate a wide range of high energy processes, including the measurement of the SM Higgs boson and potential discoveries of new particles predicted by BSM theories.

A cutaway sketch of the CMS detector is shown in figure 3.2. The detector components are positioned around one of the interaction points along the LHC ring covering almost the full 4π solid angle. The overall length of the CMS detector is approximately 29 meters, it has a diameter of around 15 meters and its weight exceeds 14 kilotons. All of this is the result of the design and engineering of state-of-the-art technologies put together with the purpose of measuring high energetic particles. The subdetector components of CMS will be discussed in more detail in the following sections.

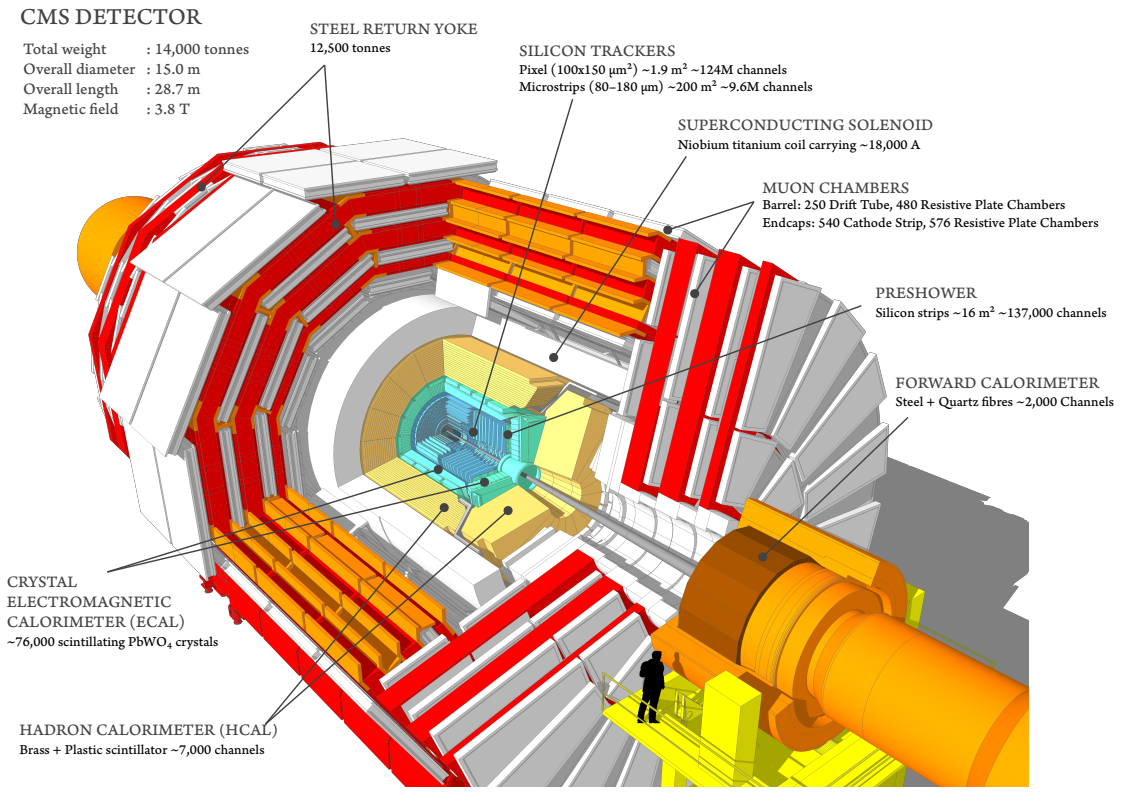


Figure 3.2: Cutaway sketch of the CMS detector [30]. The detector cylindrically surrounds the beam pipe. The pp collisions happen in the center of the detector system and the resulting particles are measured by the different subdetectors.

Any analysis that is performed on data recorded with the CMS detector uses the same Cartesian coordinate system. The origin is set to the interaction point where the pp collisions take place. The z-axis aligns with the LHC beam line and points in clockwise direction. The x-axis points inwards to the center of the LHC ring and the y-axis points vertically upwards.

Since the CMS detector is cylindrical many CMS analyses used polar instead of Cartesian coordinates. The azimuthal angle $\phi \in [-\pi, \pi]$ is defined in the x-y plane and the polar angle $\theta \in [0, 2\pi]$ in the radius-z plane. Normally, the pseudorapidity

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right) \tag{3.4}$$

is used instead of the polar angle θ as an approximation for rapidity which is valid as long as the energy of the produced particles in a collision is much larger than their mass. The ranges of η and θ can be compared and the locations of the individual detector components can be seen in figure 3.3, where a quadrant slice of the CMS detector is shown.

Further, some important observables that are used in CMS analyses need to be defined. In a pp collision the momentum of the colliding particles is unknown because the protons themselves do

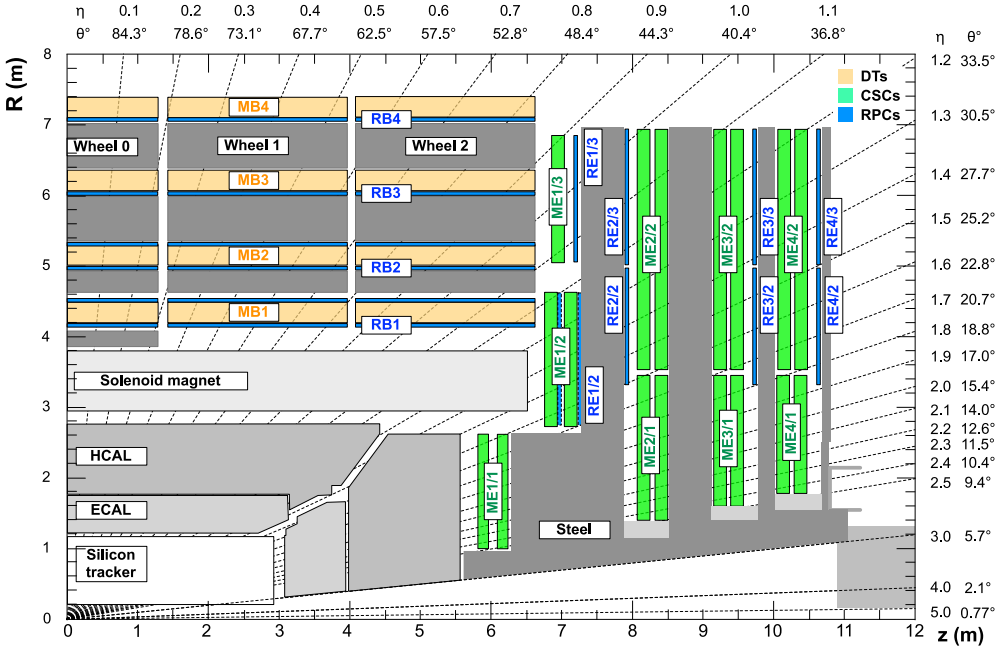


Figure 3.3: Upper right quadrant in the radius-z plane of the CMS detector [31]. The setup of all subdetectors is shown with their coverage in pseudorapidity η and polar angle θ.

not collide but partons within the protons. However, it is known that a collision happens along the z-axis and all of the momentum before the collision is directed only along the z-axis and is zero along the x- and y-axis. Due to the law of momentum conservation this is also valid after a collision. One of the main observables is the transverse momentum that is defined as

$$\vec{p}_{\mathrm{T}} = (p_x, p_y, 0), \tag{3.5}$$

$$p_{\mathrm{T}} = |\vec{p}_{\mathrm{T}}| = \sqrt{p_x^2 + p_y^2}. \tag{3.6}$$

Because of the mentioned conservation law the sum of all measured particle p_{T}'s is expected to be zero. Reasons why this sum might deviate from zero are mismeasurements of the particle momenta and energies or particles which are not measured because they do not interact with the detector. The most prominent candidates for that are neutrinos, but there could also be not yet known particles which result in a missing transverse momentum. This quantity is defined as

$$\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} = -\sum_i \vec{p}_{\mathrm{T},i} \tag{3.7}$$

where the index i represents all reconstructed particles.

Using the mentioned variables, a commonly used definition of a Lorentz vector for a reconstructed particle is $(p_{\mathrm{T}}, \eta, \phi, \mathrm{mass})$ and the spatial distance between two reconstructed particles can be calculated with

$$\Delta R = \sqrt{(\eta_1 - \eta_2)^2 + (\phi_1 - \phi_2)^2}. \tag{3.8}$$

3.2.1 Silicon trackers

The silicon trackers of the CMS detector are used to precisely reconstruct the paths of charged particles, allowing to measure their momenta and reconstruct interaction vertices. The tracker system consists of two parts, the inner tracker with the pixel detector and the outer tracker with the strip detector [32]. A sketch of the tracker system is shown in figure 3.4.

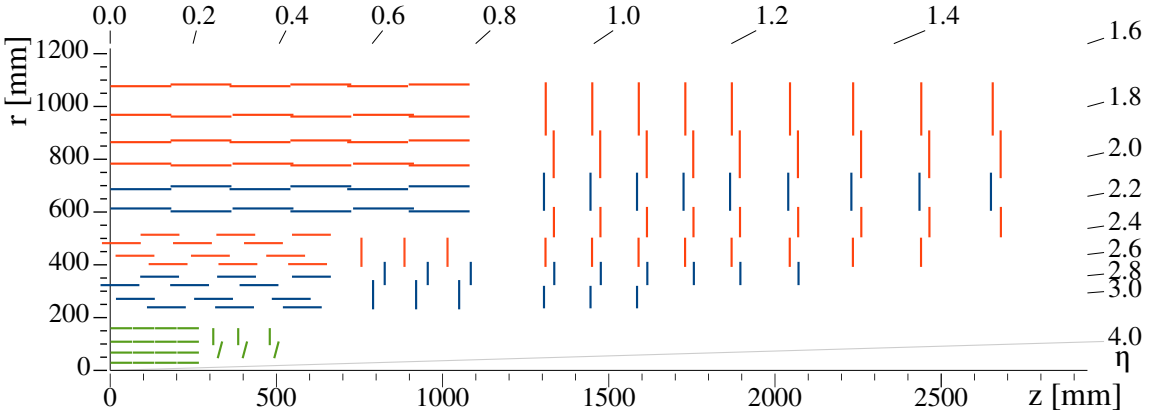


Figure 3.4: Sketch of a quadrant of the CMS silicon tracker layers in the radius-z plane [33]. The layers of the pixel tracker system are shown in green and are the innermost part of the detector closest to the beam line along the z-axis. In blue and red the strip detector is visualized for the barrel and endcap regions.

The pixel detector is positioned just a few centimeters away from the interaction point. It consists of silicon pixels with dimensions of $100\,\mu\text{m} \times 150\,\mu\text{m}$ designed to provide a three dimensional reconstruction with high resolution. The pixels measure electrical signals induced by charged particles passing through the pixel. Due to effects such as charge sharing between neighboring pixels, the resolution of reconstructed hit can be improved to around $10\,\mu\text{m}$ in the radius-z direction and around $20 - 40\,\mu\text{m}$ in the z direction. The pixel detector has four barrel layers and three endcap disks covering a range up to $|\eta| = 2.5$, which will be referred to as central region. As will be discussed later in sections 3.3.1 and 3.4.5, the pixel detector is the most important component allowing to precise identify the primary interaction vertex and any secondary vertices that may occur when for example B hadrons decay.

The strip detector is located after the pixel detector. The strips are also made of silicon and have a pitch of around $100 - 200\mu\text{m}$. Although they provide one less dimension for the reconstruction this is mitigated by setting up stereo layers where the strips are rotated by a small angle to each other, thereby recovering the three dimensional reconstruction. The strip detector consists of two barrels with four layers in the inner barrel and six layers in the outer barrel and two endcaps with three inner disk layers and nine outer disk layers. The hit resolution is approximately $15 - 45\,\mu\text{m}$.

3.2.2 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) of the CMS detector [34, 35] is designed as a homogeneous calorimeter and has the purpose of measuring particles that mainly interact electromagnetically, such as electrons and photons. The ECAL consists of over 75 thousand lead tungstate ($PbWO_4$) crystals which are chosen for the excellent properties of lead tungstate as both a scintillating and showering material. Lead tungstate is tolerant to radiation and has a high density ($8.28\,\frac{g}{cm^3}$) resulting in a relatively short radiation length of 0.89 cm, which is relevant for the depth of electromagnetic showers, and a Molière radius of $r_M = 2.19\,\text{cm}$ defining the typical width of the showers. Further, the response time of lead tungstate is fast enough to collect more than 99% of the light within 100 ns, making it suitable for the collision rates at the LHC. Each crystal has a length of 23 cm and a cell size of $2.2\,\text{cm} \times 2.2\,\text{cm}$. Therefore, each crystal can hold about 26 radiation lengths and is able to almost fully absorb the energy of electrons and photons without extra layers of absorber material between the crystals.

Like the tracker system, the ECAL is divided into a barrel region with $|\eta| < 1.479$ and an endcap region that extends the range to $|\eta| = 3$, which is also outlined in figure 3.5. Between the barrel and endcap regions there is a small gap without detector material which means that for $1.479 < |\eta| < 1.653$ almost no particles are reconstructed. Further, the endcap crystals are 1 cm shorter than the barrel crystals due to geometric constraints of the detector.

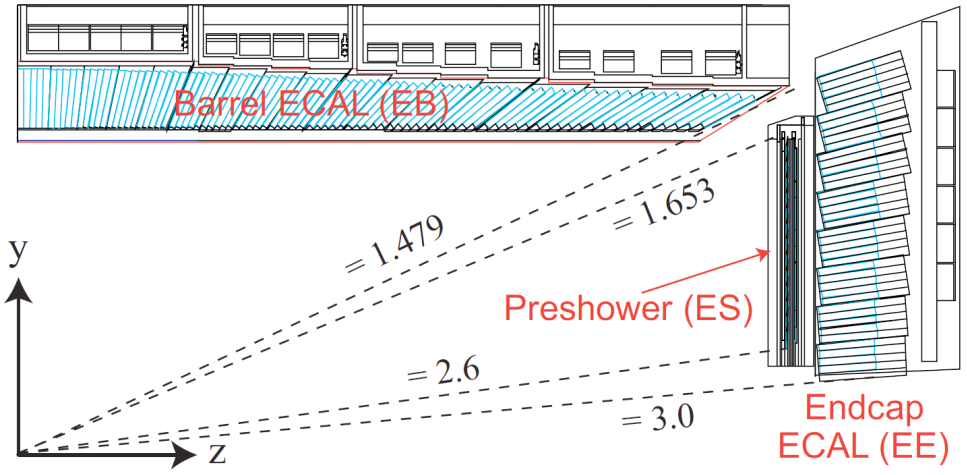


Figure 3.5: Sketch of a quadrant of the CMS electromagnetic calorimeter in the radius-z plane [35]. The ECAL consists of three parts, the barrel, endcap and preshower layers before the endcap.

The main drawback of $PbWO_4$ is its rather low light yield, which means that photodetectors with internal amplification are needed. In the barrel region silicon avalanche photodiodes (APDs) are used, while in the endcap region vacuum phototriodes (VPTs) are utilized.

In front of the endcaps, a preshower detector is located and consists of two lead absorbers with scintillating layers in between. The preshower detector is used to improve the identification of neutral pions π^0 decaying into two photons and to separate them from prompt photons.

The performance of the ECAL is mainly defined by the energy resolution of individual particles, which is measured from electron pairs decaying from Z bosons. The relatice energy resolution has the following dependencies on the energy in units of GeV

$$\frac{\sigma_E}{E} = \frac{a}{E} \oplus \frac{b}{\sqrt{E}} \oplus \text{const.} \tag{3.9}$$

The constant term is around 0.3% and is mainly caused by energy leakage or changes in the detector response from radiation damage over time. This constant term becomes dominant at higher energies. The term proportional to $\frac{1}{\sqrt{E}}$ is related to statistical fluctuations in the showers and the detection process and improves the resolution for higher numbers of particles and consequently higher energies in electromagnetic cascades produced by electrons or photons. The value of b is around 2.8%. The $\frac{1}{E}$ proportionality is mainly related to noise in the read-out of the electronics and becomes more important for low energies. This term scales with a value of $a \approx 12\%$.

3.2.3 Hadron calorimeter

The hadron calorimeter (HCAL) of the CMS detector [36, 37] is located beyond the ECAL and measures the energy of particles that are not fully absorbed by the ECAL. These are mainly hadrons like protons, neutrons, pions and kaons. The HCAL is the most enclosed part of the CMS detector, covering a wide range up to $|\eta| = 5$ and capturing as many particles as possible produced from the pp collision. The only particles that are able to pass the HCAL are muons, which only loosely interact with the detector material, and neutrinos, which do not (to a negligible degree) interact with the material.

The HCAL needs to be deep and dense to effectively stop particles. On the other hand, the size of HCAL is restricted by the design of the CMS detector to fit both calorimeters inside the superconducting solenoid. Unlike the ECAL, the HCAL is a sampling calorimeter with alternating layers of brass as absorber material and active scintillating material. It consists of a barrel region that goes up to $|\eta| < 1.5$, an endcap region for $1.5 < |\eta| < 3.0$ and a forward calorimeter covering $3 < |\eta| < 5$, which is sketched in figure 3.6. The total thickness of the HCAL depends on the η region and varies from around 5.8 nuclear interaction lengths λ at $\eta = 0$ and $10 \cdot \lambda$ for $|\eta| > 1.3$. The interaction length of brass is $\lambda = 16.42\,\text{cm}$.

Due to the compact design choice of CMS, it can happen that hadrons with high energy can sometimes produce showers that extend beyond the HCAL. Therefore, an outer HCAL component is added outside the solenoid to capture these residuals, consisting of two calorimeter layers.

The energy resolution of the HCAL is worse than the resolution of the ECAL because of the alternation of inactive absorber material in the calorimeter, a smaller number of interaction

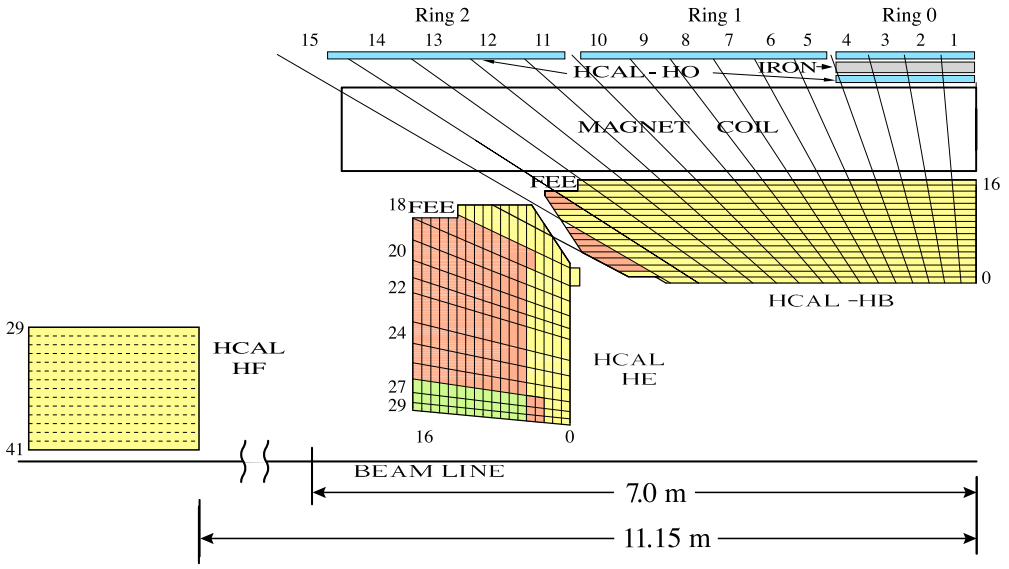


Figure 3.6: Sketch of a quadrant of the CMS hadron calorimeter in the radius-z plane [37]. The HCAL is split into four different parts, the barrel, endcap and forward calorimeters as well as a outer calorimeter component.

lengths and larger energy variations and compositions in hadronic showers compared to electromagnetic showers. This can partially be mitigated by reconstruction algorithms that take into account multiple subdetectors for the reconstruction. The approach is described in more detail in section 3.3.3.

3.2.4 Superconducting solenoid

A key part of the CMS detector is the superconducting solenoid magnet [38], which is arranged around the tracker system and both calorimeters. It creates a strong magnetic field that bends charged particles through the Lorentz force on a curved trajectory. This allows the measurement of the momentum of charged particles. The radius r of the curvature is proportional to the transverse momentum p_T of the particles, which is perpendicular to the direction of the magnetic field $B_\perp$.

$$r = \frac{p_\mathrm{T}}{q \cdot B_\perp} \tag{3.10}$$

The magnetic field runs along the z-axis causing particles to bend in ϕ direction. Furthermore, the charge sign q of the particles can be identified based on the direction of the bending.

The magnet is made from superconducting niobium-titanium coils. It is cooled down to around 4.65 K where the material enters a superconducting state and has zero resistance. This allows to produce a homogeneous magnetic field inside the solenoid of 3.8 T. To capture the magnetic flux outside the solenoid, it is surrounded by a 12 kiloton steel yoke. A measurement of the magnetic flux density produced by the magnet, both within the solenoid and in the return yokes, is illustrated in figure 3.7.

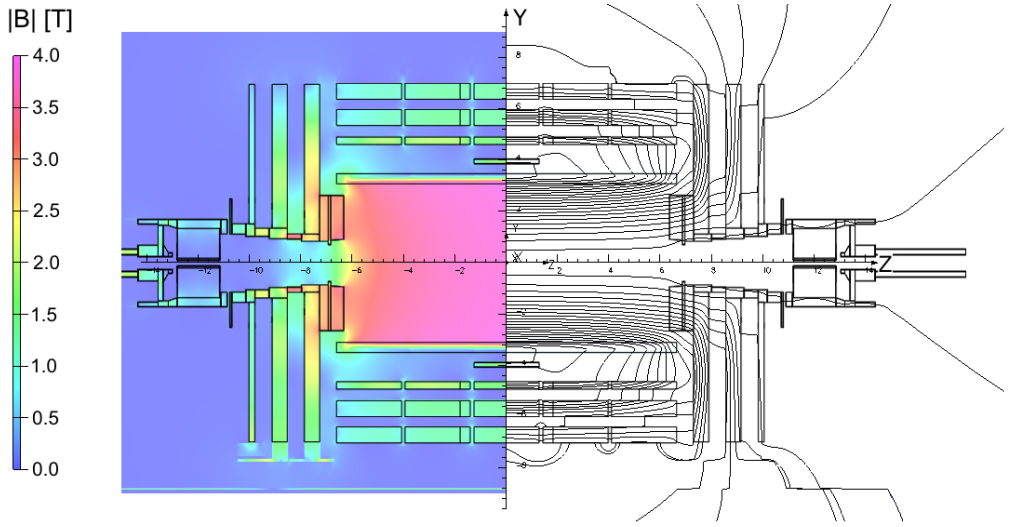


Figure 3.7: Measurement of the magnetic field in the radius-z plane [38] of the CMS detector. The left part shows value of the magnetic flux density and the right part the magnetic field lines.

3.2.5 Muon chambers

The muon system of the CMS detector [39, 40] is located outside the solenoid and covers a range up to $|\eta| < 2.4$. Its sole purpose is to measure the momentum of muons because they are the only SM particles, besides neutrinos, that are not stopped by the previously mentioned subdetectors. The muon system consists of gaseous detectors placed between layers of the steel yoke, as shown in figure 3.3, to detect muons at multiple points along their paths through the muon system. There are three different types of muon detectors in use, namely drift tubes (DTs), cathode strip chambers (CSCs) and resistive plate chambers (RPCs). The chambers are arranged in a barrel region that goes up to $|\eta| < 0.9$, an overlap region in the range of $0.9 < |\eta| < 1.2$ and an endcap region ($1.2 < |\eta| < 2.4$) to ensure maximum coverage and some overlap between the different muon detector types.

The drift tubes (DTs) cover a region of $|\eta| < 1.2$ and are located in the barrel part. They are filled with a gas mixture of Ar and CO_2 and have a stretched wire located within it. When charged particles pass through the tubes, they ionize the gas and the freed electrons drift towards the positively charged wire. The hit resolution is approximately $0.1\,\text{mm}$.

The cathode strip chambers (CSCs) cover a region of $0.9 < |\eta| < 2.4$ and are located in the endcap part. Each CSC consists of multiple layers of parallel wires and cathode strips segmented perpendicularly to the wires. When charged particles pass through the chambers, the gas in the chambers ionizes and the freed charged particles drift to both the wires and the cathode strips. Due to the design of the wires and strips, a very precise time and position measurement is possible. The hit resolution for the CSCs is approximately $75\,\mu\text{m}$.

The resistive plate chambers (RPCs) are present in both the barrel and endcap, and cover a range of $|\eta| < 1.9$. One RPC has two flat parallel plates made from synthetic material with high resistance. These plates are separated by a few millimeters in a gas tight volume and are coated with conductive graphite paint of opposite charge to act as electrodes. The RPCs are not as good regarding the hit resolutions but they have the best time resolution of around 3 ns out of the three used muon detectors. Therefore, the RPCs are able to provide hit information for the trigger system (see section 3.2.6) to make fast decisions if an event should be saved.

3.2.6 Trigger system

Proton bunches collide in the interaction point of the CMS detector at rates of around 40 million times per second. In each of these bunch crossings, multiple pp interactions can happen at the same time. For the 2018 data taking era, on average, 55 pileup interactions happened per bunch crossing. This results in around 1 MB of information for each event that must be read out. With a resulting data rate of about 40 TB/s, it is clear that recording all the data is not possible. Therefore, a two stage trigger system is incorporated to reduce the amount of data and only record events of interest.

The first stage is the Level-1 (L1) trigger [41], which is a hardware-based system within the CMS detector components and decides within 4 µs whether an event should be recorded. The decision is made based on relatively primitive information taken from the calorimeters and the muon system. The L1 system analyzes energy clusters that can be reconstructed as jets, charged leptons or photons. Further, possible reconstructed tracks in the muon system are checked and in the end a decision is made based on some predefined selection criteria, e.g. a lower p_T threshold for muons. If an event meets these criteria, it is fully read out. The L1 trigger reduces the event rate to around 100 kHz.

The second stage is the high-level trigger (HLT) [42], which is a software-based system that uses a computer farm to process events. During this stage, an event reconstruction is performed similar to the main reconstruction methods discussed in section 3.3. Since decisions if an event should be recorded long-term have to be made in real time, a more streamlined reconstruction process is used. The HLT employs jet clustering algorithms and identification algorithms for charged leptons and photons using predefined paths that sequence the reconstruction steps efficiently. If at least on HLT trigger path is passed, the event is stored to be used later in offline data analyses discussed in the following sections. This procedure reduces the data rate to about 1 kHz. Some HLT paths still have a higher data rate, which is why they are prescaled. This means that only each n-the event is stored for the specific HLT path.

3.3 Event reconstruction at CMS

In a collision event at CMS, thousands of different particles are produced and while they scatter in all directions or decay into other particles the detector is used to measure their energies and momenta. The difficult part is to reconstruct and identify characteristic signatures in this

busy environment corresponding to specific types of particles like electrons, muons or hadron originating from quarks or gluons. In this section the general event reconstruction as well as the object identification algorithms relevant to this thesis are discussed.

3.3.1 Track and vertex reconstruction

The event reconstruction starts in the tracker system of CMS [32] since it is the most precise measuring device in CMS and also closest to the interaction point. The first step is a local reconstruction of hit information from charged particles. Measured signals in the pixel and strip channels are clustered together and used to estimate the position of the hits.

These hits are further used to reconstruct tracks, which represent the trajectory of charged particles in the detector, and to estimate their momentum and position. Even in a single event, the number of hits is quite high, which makes the reconstruction of tracks computationally challenging. The track finding algorithm in CMS is based on a Kalman filter [43]. To make the track finding as efficient as possible, an iterative approach is used starting from easily identifiable tracks, for example due to a relatively large p_T or being close to the interaction point. The hits related to these tracks are then removed for the next track finding iteration. The reduced set of hits makes it easier to find less evident tracks. This procedure is repeated a few times until in the last iteration only some very displaced (from the interaction point) or low p_T tracks are left.

The next step after the track reconstruction is the reconstruction of interaction vertices (proton-proton interaction points). First, tracks are selected based on some quality criteria. For a track to be selected it has to be prompt, which means it has a low impact parameter (transverse distance) relative to the center of the beam crossing, the number of associated strip and pixel hits is at least five with more than two pixel hits and the χ^2 from the fit of the track trajectory is small.

Next, these tracks are clustered into groups that most probably originate from the same interaction vertex. The clustering of tracks is performed based on the z-coordinate along the beam axis using a *deterministic annealing* algorithm [44]. This algorithm determines the number of vertices and assigns tracks to them based on a probability approach.

For each identified vertex with at least two associated tracks, the *adaptive vertex fitter* [45] is used to calculate vertex parameters like position, covariance matrix and number of degrees of freedom, representing the quality of the fit. The efficiency of the vertex reconstruction is almost 100% if more than two tracks are used and the resolution of the vertex also improves with an increasing number of tracks.

In a single event, normally several vertices are reconstructed due to the fact that not only one interaction happens in a collision. To identify the vertex from the hard interaction, the magnitudes of the transverse momenta of the tracks of each vertex are summed quadratically $\sum p_T^2$. The vertex with the highest sum is defined as the primary vertex. All other vertices are considered *pileup*.

3.3.2 Energy cluster reconstruction

After the tracker system, particles pass through the ECAL and HCAL and deposit energy in the cells of the calorimeters. These energy deposits are reconstructed as clusters, for ECAL and HCAL separately [46]. To reconstruct a cluster, seed cells are identified with an energy deposit higher than a given seed threshold (several 100 MeV) and a higher energy than neighboring cells. Starting from a seed cell, neighboring cells are added to the cluster, as long as their energy deposit is higher than a given noise threshold, forming a topological cluster. Since clusters often overlap, a Gaussian-mixture model is used to reconstruct the clusters assuming that each seed cell corresponds to one normally distributed energy deposit.

Due to the applied energy thresholds for the clustering, the measured energy is expected to be smaller than the true energy of the particles. Therefore, an energy calibration is performed. First, the ECAL is calibrated dependent on the pseudorapidity using simulated single photons. Especially for low energy photons the energy is underestimated and the correction goes up to 20%. The calibration is validated using the abundant production of neutral pions in the collision data. Because neutral pions decay almost always into a pair of photons, the invariant mass of the photons can be fitted and compared to the known pion mass of 135 MeV.

The HCAL is calibrated after the ECAL because hadrons deposit energy in both ECAL and HCAL and the response to hadrons in the ECAL is substantially different compared to photons or electrons. Therefore, the HCAL calibration depends not only on the pseudorapidity but also on the deposited energy fraction between ECAL and HCAL and is performed using simulated neutral K_L^0. Also for the HCAL, the energy is especially underestimated for low energies and the correction goes up to 40%. The calibration is validated using isolated charged hadrons from collision data.

3.3.3 Particle flow algorithm

In a proton-proton collision, all sorts of particles are created. The different subdetector systems of the CMS detector are developed to target the measurement of trajectories and energies of different groups of those particles, like charged or neutral hadrons, photons or leptons. However, these particles do not interact with only one subdetector system. Depending on the particle, the particle itself or its decay products leave traces in multiple detector layers. In CMS the particle flow (PF) algorithm [46] uses this circumstance to improve the reconstruction of these particles by combining the information from different subdetectors.

The PF algorithm takes the so-called PF elements from the various detector layers and tries to link them. PF elements are, for example, the already introduced reconstructed tracks (see section 3.3.1) or the reconstructed energy clusters in the ECAL and HCAL (see section 3.3.2).

First, tracks are extrapolated to the ECAL based on their outermost hit in the tracker system. If energy clusters are found within an angular acceptance of the track extrapolation, these clusters are linked to the track. If multiple tracks are linked to the same cluster or multiple clusters are linked to the same track, only the link with the smallest link distance is kept. The

link distance is defined as the distance between an extrapolated track position and the position of a cluster in the η-ϕ plane.

Further links are set due to electron bremsstrahlung. For that, tangents are extrapolated for each intersection of a track and a tracker layer. Bremsstrahlung photons either produce clusters in the ECAL or convert to e^+e^- pairs, which again are reconstructed as tracks. A link is made if a tangent is compatible with an ECAL cluster or with the sum of the momenta of two other tracks.

A similar procedure is performed between ECAL and HCAL clusters where a cluster in the ECAL has to be within the envelope of a HCAL cluster to be linked. How all these links are used to reconstruct particles is described in the next sections.

After all PF candidates are identified, the particle collection is still contaminated by particles from pileup. Such pileup particles affect the reconstruction of jets (see section 3.3.7), missing transverse momentum or the isolation of leptons. To mitigate this problem, reconstructed charged hadrons that can be associated to a pileup vertex are removed from the particle collection. For other particles like photons or neutral hadrons, which do not interact with the tracker, this is not possible.

3.3.4 Electron reconstruction

The electron reconstruction is part of the PF algorithm [46]. While passing through the detector and interacting with the material, electrons emit photons due to bremsstrahlung and, in turn, photons convert to e^+e^- pairs. This results in cascades of electromagnetic decays called electromagnetic showers that are mainly measured in the ECAL. Due to this close entanglement between electrons and photons, they are reconstructed together [47].

The reconstruction starts with the tracks. Due to the bremsstrahlung photons emitted from electrons while they pass through the tracker material, the trajectory of the electrons changes. To adapt to this circumstance, an additional Gaussian-sum filter (GSF) [48] is applied to the tracks. The difference from normal track fitting is that the GSF is better adapted to sudden and substantial energy loss along the track trajectory.

While originating from only one electron or photon, several electrons/photons reach the ECAL because of the electromagnetic showering. To capture the energy from the original object, the ECAL clusters that are geometrically close to a seed cluster are combined into a supercluster. Further, no HCAL clusters should be present within close distance to the ECAL supercluster direction, and if there is one, its energy should not exceed 10% of the energy deposited in the ECAL supercluster.

The GSF tracks, superclusters, but also normal tracks are forwarded to the PF algorithm that links them together define PF candidates. The normal tracks are added to the candidate if they can be identified as electrons/positrons coming from a photon conversion. The difference between photon and electron PF candidates is that an ECAL supercluster is not linked to a

GSF track, since photons are neutral particles and do not leave tracks. All tracks and clusters used for the electron and photon reconstruction are removed from the collection in further processing.

Although the ECAL energy is already calibrated as discussed in section 3.3.2 the reconstruction of electrons and photons using superclusters introduces new sources of errors like energy loss through shower leakage or dead crystals in the ECAL. There is also energy lost on the way through the tracker material, leading to a smaller reconstructed energy compared to the initial energy.

To account for this, the electron/photon energy is corrected using a multivariate regression with boosted decision trees (BDTs). The regression target is the ratio of the true energy and the reconstructed energy of an electron or photon and is applicable individually to each PF candidate. The correction consists of three sequential steps. First, the supercluster energy itself is corrected. Second, the resolution of the supercluster energy is corrected to better match the real detector conditions. And third, the combined energy estimate of supercluster and tracks is corrected, for electrons only.

The BDTs are trained on simulated electron and photon pairs. This means that differences between simulation and real data still remain after the correction, especially the resolution is usually better in simulation. To account for that, an additional correction is derived by fitting a Breit-Wigner function to the invariant mass distribution of di-electron pairs targeting the Z boson peak, in data and simulation separately. The difference of both fits is used to obtain the scale offset. To get the correction for the resolution the simulated invariant mass distribution is directly fit to data with different variations of Gaussian smearing (0.1 to 1.5%) applied to the simulation. The best fit result is used as the energy resolution correction.

3.3.5 Muon reconstruction

Muons are a special reconstruction case because their reconstruction is not directly part of the PF algorithm. This has to do with the very high purity and efficiency of the muon reconstruction using the dedicated muon detector system of CMS. The efficiency is given by the wide coverage of the muon system and the purity is a result of the absorption of almost all other particles before they reach the muon system. The only exceptions are neutrinos or some *punch-through* hadron showers.

There are three types of reconstructed muons [40]. The first type depends only on the muon detector system to reconstruct a muon. Several hit segments have to be present in the DT, CSC or RPC detector parts along the muon trajectory. These hits are then fitted to obtain a *standalone muon* track. Such muons have a worse momentum resolution compared to the other muon types and are often faked by cosmic muons.

The second type is a *global muon*. For that a *standalone muon* track has to match to a track in the tracker system. The hits from the tracker and muon system are combined and fitted

to obtain a *global muon* track. Due to the combination, especially for high p_T muons, the momentum resolution increases significantly compared to a standalone muon.

The third type is a *tracker muon*. This reconstruction type mainly relies on the tracker system. Tracks in the tracker system are extrapolated to the muon system and, if at least one segment or hit can be matched to the extrapolated track, it is counted as a *tracker muon*. This muon type has a higher efficiency for low p_T muons but is also more affected by muon misidentification from hadron shower remnants in the innermost layers of the muon detector system.

About 99% of the reconstructed muons are global or tracker muons. After the reconstruction, the muons are given to the PF algorithm that applies additional quality criteria to the muons to select the muon PF candidates. The tracks involved are removed to simplify the PF reconstruction of other particles.

3.3.6 Hadron reconstruction

After muons, electrons and photons are reconstructed and the corresponding tracks, clusters and segments are removed from the PF algorithm, the leftover particles to be reconstructed are charged and neutral hadrons. Within the tracker acceptance of $|\eta| < 2.5$ hadrons leave only a small fraction of their energy in the ECAL (about 3%). Therefore, only HCAL clusters are considered to reconstruct the hadron energy. Outside of the tracker acceptance this changes but this is not of relevance for this analysis because no objects are selected with $|\eta| > 2.5$.

HCAL clusters without any links to tracks are assigned to neutral hadrons. For charged hadrons the remaining HCAL clusters are linked to at least one track in the tracker system that is not linked to any other HCAL cluster. Then, for these charged hadron candidates the sum of the track momenta is compared to the calibrated energy in the calorimeters. Depending on the compatibility a redefinition of the particle hypothesis is done. If the calibrated energy and the track momenta are compatible, the candidate stays a charged hadron and the momenta are redefined based on a combined fit of tracker and calorimeter information. If the calibrated energy is significantly larger than the sum of track momenta, a split is done into multiple candidates. The tracks are used for new charged hadron candidates and depending on the energy allocation in the ECAL/HCAL additional photon or neutral hadron candidates are added. In rare cases, the calibrated energy is significantly smaller than the sum of track momenta. In this case a global muon candidate that failed the quality criteria before but would match the tracks and calorimeter clusters is looked for. After a re-evaluation of the global muon momentum with the additional information, a few more muons can be added to the muon PF candidates.

3.3.7 Jet reconstruction

During a pp-collision several quarks and gluons are produced or scattered. After losing enough energy they start to hadronize and build hadrons like pions or kaons. The initial quarks and gluons which originate from a hard process often have a high momentum, which leads to the

effect that all hadrons originating from these quarks or gluons are collimated in the same direction and can be clustered together within a narrow cone, starting from the initial particle. These reconstructed cones are referred to as jets.

In CMS the mainly used jet clustering is based on the anti-k_t algorithm [49] but in this thesis also other jet clustering algorithms are used depending on the object of interest. All of these algorithms use sequential recombination and can be written down in a generalized way. For the recombination, PF candidates are used. Starting from a hard particle i (high p_T), close by particles j are checked and recombined to a new pseudo-jet object based on a distance measure

$$d_{ij} = \min(p_{\mathrm{T},i}^{2n}, p_{\mathrm{T},j}^{2n}) \cdot \frac{\Delta_{ij}^2}{R^2} \quad \text{with} \tag{3.11}$$

$$\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2. \tag{3.12}$$

This recombination is repeated sequentially with the pseudo-jet as the new i as long as d_{ij} is smaller than

$$d_{iB} = p_{\mathrm{T},i}^{2n}. \tag{3.13}$$

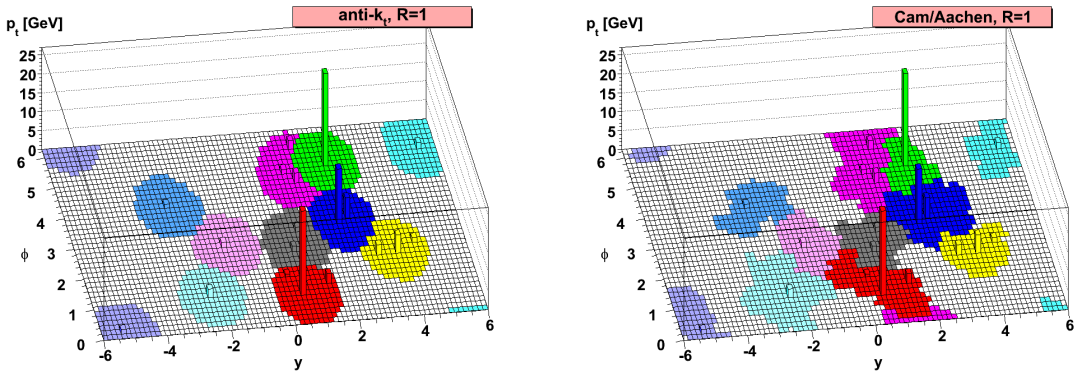
In this equation, B is the beam, which means the recombination stops if the distance of the pseudo-jet is closer to the beam than the next PF candidate. After the clustering is finished the pseudo-jet is defined as a proper reconstructed jet.

Further, the constant radius parameter R is of interest. It defines the size of the jet cone and is set to 0.4 for normal jets in CMS. Additionally, for this thesis also wide cone jets are of interest with $R = 0.8$. A bigger radius parameter increases the probability of catching more than one hard particle within one jet. This can be useful when looking for boosted resonances (more on this in sections 3.4.4 and 3.4.6).

As mentioned before, the above equations are a generalized way of describing several jet clustering algorithms. To get to a specific algorithm, the parameter n has to be defined. For this thesis, two values of n are relevant.

First, for the standard anti-k_t algorithm, $n = -1$ is used. By this, the distances d_{ij} and d_{iB} contain not only a spatial dependence but also depend on the transverse momentum of the particles. The inverse p_T dependence prefers to accumulate soft particles around one hard particle, which leads to a very conical representation for a jet. An illustration of the anti-k_t algorithm applied to a simulated event is shown in figure 3.8a. Only for the hardest jets a totally conical structure is visible which means soft particles cluster preferably to harder jet clusters. For nearby overlapping jet clusters, the cone structure is deformed.

Second, for boosted hadronic tau leptons, which are described in section 3.4.4, the Cambridge-Aachen (CA) algorithm [50] is used. To obtain the definition of the CA algorithm, n is set to 0, which means that the jet clustering depends only on the spatial distance between the particles. The effect of this is visible in the illustration of the CA algorithm applied to a simulated event in figure 3.8b. The jet clusters do not have a conical structure like with the anti-k_t algorithm



(a) Anti-k_t algorithm (b) Cambridge-Aachen algorithm

Figure 3.8: A single simulated parton-level event with randomly added soft particles is reconstructed with two different clustering algorithms [49]. For both algorithms the radius parameter is set to $R = 1$. In (a) the anti-k_t algorithm is used and shows a very conical clustering with the hard particle located in the middle. In (b) the CA algorithm is used. Because only the spatial distance is relevant for the clustering a bigger fluctuation is visible from a cone shape and the hard particle is not always located centrally in the jet.

since no preference is given for soft or hard particles. Although the CA algorithm is worse in resolving jets than the anti-k_t algorithm, it has the advantage that the de-clustering of jets is much easier. The de-clustering of wide cone CA jets ($R = 0.8$) is one of the key ingredients for the identification of boosted hadronic tau leptons (see section 3.4.4).

For the jet clustering algorithms the pileup mitigation strategy mentioned in section 3.3.3 is used. However, for the calculation of the missing transverse momentum $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ an alternative approach is used that includes all particles and not only charged hadrons. The pileup per particle identification (PUPPI) algorithm [51] assigns a weight with values between zero and one for each particle. Particles with values close to zero are expected to be from pileup while particles with values close to one are associated to the primary vertex. For charged hadron an exact value of zero or one is assigned based on the track association to a pileup or primary vertex, respectively. The weights are used during the calculation of $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ and are applied to the particle four momenta to basically remove pileup particles from the reconstruction.

The weight of a particle is estimated via a discriminating distribution called α. This distribution distinguishes between pileup and primary vertex particles based on several kinematic features like e.g. p_{T}. For charged particles it is clear whether they are from the primary vertex (PV) or from pileup. This information is used to identify regions in α that are closer to pileup or closer to the PV. The discriminating distribution α is similar for charged and neutral particles and the weights for neutral particles are estimated by comparing it with the known distribution from charged particles.

Although an energy calibration has already been done on the level of calorimeter clusters, another energy calibration is performed specifically for reconstructed jets [52]. The jet energy scale and resolution corrections are measured depending on the pseudorapidity η and transverse momentum p_T of jets using several methods. These methods exploit well known measurements of $Z \to \mu\mu/ee$ or a photon γ in association with a single jet. The muons, electrons and photons can be measured very precisely and the p_T of the reconstructed Z boson or photon is used to calibrate the p_T of the recoiling jet. Additionally, to better account for gluon induced jets, QCD di-jet events are used for the energy calibration where one of the jets is already corrected from the Z/γ+jet calibration.

Regarding wide cone jets ($R = 0.8$), a special mass reconstruction of the jets is relevant for this analysis and is performed to better identify and reconstruct two-prong jet substructures. The *SoftDrop* algorithm [53] is used to groom jets and remove soft radiation originating from initial state radiation (ISR), underlying event (UE) or pileup. The algorithm iteratively de-clusters a jet down to its first two jet constituents. In each step the two subcomponents i and j with their transverse momenta $p_{T,i}$ and $p_{T,j}$ as well as their distance $R_{i,j}$ are checked and the softer one is removed if the following condition is not fulfilled:

$$\frac{\min(p_{T,i}, p_{T,j})}{p_{T,i} + p_{T,j}} > z_{\text{cut}} \cdot \left(\frac{R_{i,j}}{R}\right)^{\beta} \tag{3.14}$$

If the condition is fulfilled, the algorithm stops and the two subcomponents are defined as two subjets representing the two-prong substructure of the wide cone jet and used to reconstruct the jet mass. In the following, this mass will be referred to as softdrop mass m_{SD}. For CMS analysis the soft cut-off threshold z_{cut} is set to 0.1 and the angularity exponent β to 0.

3.4 Object identification at CMS

After the reconstruction of particles or clusters of particles in an event, additional algorithms are applied to even stricter identify them. While electrons and muons are identified as their own object and have no subgroups, jets as an object can represent multiple particles like gluons, quarks of all the different flavors or tau leptons which decayed hadronically. The following sections describe the object identification algorithms relevant to this thesis.

3.4.1 Electron identification

Electron candidates reconstructed with the PF algorithm include not only prompt electrons coming directly from a hard process but also electrons from photon conversion or misidentified charged hadrons. To reduce the contribution from such misidentified electron candidates, some quality criteria are applied [54].

The isolation criterion has the goal of removing jets misidentified as electrons and genuine electrons produced as a decay product from other particles, e.g. within a b quark induced jet. In both cases there is a significant energy contribution present around the electron trajectory.

An isolation quantity is defined including the p_T of all PF candidates around the electron within a radius of $\Delta R < 0.3$.

$$\text{Iso} = \sum p_T^{\text{charged}} + \max\left[0, \sum p_T^{\text{neutral had.}} + \sum p_T^{\gamma} - p_T^{\text{PU}}\right] \tag{3.15}$$

The charged PF candidates include only those associated with the primary vertex. For neutral hadrons and photons the vertex association is not possible. Instead, a pileup related correction p_T^{PU} is subtracted from the p_T sum of considered neutral particles. The pileup correction depends on the energy density ρ. The density is measured for each event as the median of the energy distribution per area in (η, ϕ) and, in turn, depends almost linearly on the number of reconstructed vertices in the event. With the energy density ρ the pileup correction is calculated as

$$p_T^{\text{PU}} = \rho \cdot A_{eff}\,. \tag{3.16}$$

The effective area A_{eff} is defined by the isolation cone around the electron and the area that it covers in the η-ϕ plane. To have a consistent definition of the isolation for all electrons, the isolation relative to the p_T of the electron in question,

$$\text{Iso}_{\text{rel}} = \frac{\text{Iso}}{p_T^e}, \tag{3.17}$$

is used to identify isolated electrons. In this thesis different kinds of electrons and isolation cuts are relevant depending on the targeted final state topology of the signal processes. For example, for boosted tau lepton pairs it is required that the electron is not isolated. A detailed description of the event and electron selection is discussed in section 4.3.1 and 4.3.2.

Since the selection based on the relative isolation is quite analysis specific, additionally a more general selection of electrons is carried out using a boosted decision tree (BDT) algorithm [54]. The BDT is trained to separate prompt electrons from non-prompt or misidentified electrons. The inputs to the BDT are variables related to the reconstruction of the electrons. This includes, for example, the matching quality of tracks to superclusters, more specific information about the supercluster shape or the relative contribution of bremsstrahlung to the supercluster energy.

Each electron candidate gets a score assigned by the BDT with a value between minus one and one, where one corresponds to a prompt electron and minus one to a misidentified electron. For the analysis in this thesis electrons are selected based on the BDT score equivalent to a 90% selection efficiency and a misidentification rate of around 3%.

3.4.2 Muon identification

The identification of muons works similar to the electron identification [40]. The isolation of a muon is calculated with the same equation 3.15 as already introduced for electrons. The only difference is in the definition of how the pileup correction p_T^{PU} is derived and the isolation radius is set to $\Delta R < 0.4$.

To obtain the pileup correction for each event, the p_T of all charged hadrons originating from pileup vertices is computed. Further, the assumption is made based on simulation results that in a pp collision the number of produced charged hadrons is approximately the same as the number of produced neutral hadrons and photons. Using this approximation the pileup correction results to

$$p_T^{PU} = 0.5 \cdot \sum p_T^{\text{charged, PU}}. \tag{3.18}$$

Similar to the electron identification, in this thesis the muon isolation selection depends on the targeted final state topology of the signal processes and is further discussed in section 4.3.1 and 4.3.2.

The muon identification algorithm is applied independent of the isolation and uses a set of variables that are combined to define different selection working points, allowing an analysis specific balancing between purity and efficiency. These variables are related to different parts of the muon reconstruction like the fit of the tracks, the number of hits in the tracker or muon system or the matching quality of tracks and segments for global muons.

In this analysis the medium working point (WP) is used, which targets prompt muons and muons from heavy flavor decays. To pass the medium WP the muon has to be classified as either a tracker or a global muon by the PF algorithm. Further, a muon track must pass through at least 80% of the tracker layers. Next, for global muons the fit with both tracks and segments has to fulfill some quality criteria related to the degrees of freedom (dof) and χ^2 of the fit. The criteria are $\chi^2 < 12$ and $\chi^2/\text{dof} < 3$. In addition, a kink-finding algorithm is run to check the quality of the tracks. The algorithm splits a muon track at different points along the trajectory in the tracker system into two tracks and compares them. The χ^2 of this comparison must be smaller than 20 for the initial track to be considered as a single track. All these checks lead to an overall muon identification efficiency of 99.5% for simulated W and Z boson decays into muons.

More working point definitions can be found in [40], however, they are not discussed further in this thesis because they are not used.

3.4.3 Hadronic tau lepton identification

Tau leptons have a relatively short mean lifetime of around $2.9 \cdot 10^{-13}$ s and decay before they reach the CMS detector. Additionally, the tau lepton is the only lepton that can decay into hadrons due to its higher mass compared to the other leptons and light hadrons. The branching fraction for a tau lepton to decay hadronically is around 65% and the decay products are reconstructed as jets in the CMS detector. This also means that jets from tau leptons blend in with the abundant jet production from hadronized quarks and gluons, making it difficult to differentiate them from each other. Such hadronically decaying tau leptons will be further denoted as τ_h.

To identify τ_h, the hadron-plus-strip (HPS) algorithm [55] is used. This algorithm is applied to each reconstructed jet. The possible final states of a hadronic tau decay can be found in

table 3.1. The neutral pions (π^0) almost always decay into a pair of photons which, in turn, with a high probability convert into e^+e^- pairs. To reconstruct the energy of the π^0, photon and electron candidates within a jet are clustered within a certain $\Delta\eta \times \Delta\phi$ region. The exact region values depend on the p_T of the candidates involved, but the upper and lower limits are $\Delta\eta \in [0.05, 0.15]$ and $\Delta\phi \in [0.05, 0.3]$. The resulting object is called *strip*. The charged hadrons within a jet must be associated with the primary vertex and have a p_T of at least 0.5 GeV.

Table 3.1: List of all possible weak decays of a tau lepton with the corresponding branching fractions $\mathcal{B}$ in percent [56]. Charged hadrons are denoted as $h^{\pm}$. For simplicity only the τ^- decays are mentioned but the same branching fractions are also valid for the charge conjugated decays. The decay mode values are a CMS specific classification of the hadronic tau lepton decays.

	Decay	Decay mode	$\mathcal{B}$ in %
Leptonic			35.2
	$\tau^- \to e^- \bar{\nu}_e \nu_\tau$		17.8
	$\tau^- \to \mu^- \bar{\nu}_\mu \nu_\tau$		17.4
Hadronic			64.8
	$\tau^- \to h^- \nu_\tau$	0	11.5
	$\tau^- \to h^- \pi^0 \nu_\tau$	1	25.9
	$\tau^- \to h^- \pi^0 \pi^0 \nu_\tau$		9.5
	$\tau^- \to h^- h^+ h^- \nu_\tau$	10	9.8
	$\tau^- \to h^- h^+ h^- \pi^0 \nu_\tau$	11	4.8
	other		3.3

The identified charged hadrons and strips are then classified into the decay modes mentioned in table 3.1. In this analysis four decay modes are considered, with either one or three charged hadrons and with or without a π^0. The HPS algorithm only considers combinations of hadrons and strips with charge ± 1 and if the hadronic system fits within a narrow signal cone range between 0.05 and 0.1 defined by $R_{sig} = (3\,\text{GeV})/p_{T,system}$. If more than one τ_h candidate is identified within a jet, only the one with the highest p_T is kept.

The HPS algorithm already rejects a significant part of jets originating from quarks or gluons. Nevertheless, the collection of τ_h candidates is still contaminated not only by misidentified jets from quarks and gluons but also by misidentified electrons and muons reconstructed as jets. To increase the sensitivity even further in identifying τ_h, a convolutional neural network based classifier called *DeepTau* [57, 58] is used. DeepTau is trained on low-level detector information from the tracker system, the calorimeters and the muon system together with higher-level information about the reconstructed τ_h candidates. Especially the low-level features significantly improve the classification compared to the previously used τ_h identification algorithms. The

training is performed using simulated data from different SM processes like Drell-Yan, $t\bar{t}$ and W+jets production and has the task to separate genuine τ_h candidates from quark/gluon, electron or muon induced τ_h candidates. The probability-like outputs p_{particle} of DeepTau are combined into three discriminants

$$D^{\alpha}_{\tau_h} = \frac{p_{\tau_h}}{p_{\tau_h} + p_{\alpha}} \tag{3.19}$$

where α corresponds to quarks/gluons, electrons or muons. For the discrimination against quark or gluon induced jets a *medium* working point (WP) is chosen. This WP corresponds to a misidentification rate of about 1%. The WPs for the discrimination against electrons and muons will be discussed in section 4.3.1 because they differ depending on the final state of the analysis.

3.4.4 Boosted hadronic tau lepton identification

Part of the signal processes in this analysis is a boson that decays into a pair of tau leptons. In some cases this boson can be quite heavy or have large p_T. Either way, the pair of tau leptons gets a significant boost and the tau leptons decay further in almost the same direction. This means that during the reconstruction both tau leptons would not be resolvable and clustered within a single jet. Unfortunately, the HPS algorithm used to identify τ_h candidates only reconstructs one τ_h candidate per jet.

To still be able to reconstruct hadronic tau leptons, a different reconstruction procedure is utilized [55]. As introduced in section 3.3.7, wide cone jets (with $R = 0.8$) reconstructed with the Cambridge-Aachen algorithm (CA8) are used. The first step is to undo the last step of the jet clustering algorithm and get the two subjets of the CA8 jets. These two subjets are expected to originate from the two tau leptons. This also means that both subjets should have similar kinematic properties. Each of them must have a p_T greater than $10\,\text{GeV}$ and the reconstructed mass of the heavier subjet must be less than two thirds of the full CA8 jet mass. These criteria reduce misidentification of jets produced by multijet QCD events or similar processes where a single highly energetic quark or gluon is clustered as a CA8 jet.

At this point there is no real differentiation between hadronic and leptonic tau decays because an electron or muon from a leptonic tau decay could also be reconstructed as a subjet due to the enriched hadronic activity around it. If a CA8 jet is found in which both subjets fulfill the above mentioned criteria, a similar procedure is applied as discussed before for the common hadronic tau leptons. Both subjets are individually passed to the HPS algorithm and if they pass, they are defined as boosted τ_h candidates.

Similar to the common τ_h candidates, the boosted τ_h candidates can be misidentified jets from quarks/gluons, electrons or muons. To increase the selection efficiency of genuine boosted τ_h candidates, three dedicated discrimination algorithms are used for each of the misidentified objects.

The separation from quark or gluon induced jets is carried out by a BDT. The BDT is trained on a set of variables related to the τ_h candidate, such as the decay mode, isolation, p_T and

spatial distances of the charged hadrons and strips within the cone of the reconstructed τ_h candidate. The working points of the BDT discriminant are defined based on the isolation efficiency of the τ_h candidates. In this analysis the *loose* WP is used, which corresponds to an efficiency of 60%. The reason for not using a tighter WP is that the τ_h candidate should not be too isolated since the final state of the analysis requires a second τ_h candidate or a non-isolated electron or muon to be close.

The discrimination against electron and muon induced τ_h candidates is only applied in some specific boosted final states of this analysis and the WPs are discussed in section 4.3.2. To separate genuine τ_h candidates from electrons, a second BDT is specifically trained for this task. The separation from muons is simpler and is based on some reconstruction criteria. To pass the muon WP used in this analysis, not more than one segment should be present in the muon system within a ΔR cone of 0.3 around the τ_h candidate. More details about the electron BDT and other muon working points are discussed in [55].

3.4.5 Identification of b quark induced jets

Among all quarks and gluons, b quarks produce a rather unique jet signature in the detector. The b quark decay is CKM-suppressed and results in a relatively long lifetime ($\sim 10^{-12}$ s) of hadronized b quarks (B hadrons). This means the B hadrons start to decay slightly displaced from the primary vertex (PV). A visualization of this behavior is shown in figure 3.9. These so-called secondary vertices (SVs) can be reconstructed because of the great resolution of the CMS tracker system.

The SVs are reconstructed using the inclusive secondary vertex finding algorithm [59]. This algorithm runs independent of the jet clustering over all tracks associated with the PV. First, seed tracks with a three dimensional impact parameter d_{xyz} of at least 50 µm and a two dimensional impact parameter significance of $S_{2D} = d_{xy}/\sigma(d_{xy}) > 1.2$, where σ is the uncertainty, are identified. The impact parameter and flight distance are measures of the spatial distance of the closest point of a track to the reconstructed PV and are defined depending on the dimensionality in the x-y-z plane, x-y plane or along the z-axis. After identifying the seed tracks, other nearby tracks are clustered to a seed track if they are closer to the seed track than to the PV. The clustered tracks are fitted to identify the SV position similar to the PV fit (see section 3.3.1). The SVs with more than 70% overlap with each other are merged together and only those with a two(three) dimensional impact parameter significance of greater than 2.5(0.5) are kept.

Finding a SV in a jet significantly increases the identification efficiency for b-jets (b quark induced jets) up to around 75%. However, a SV is not the only discriminating feature that can be used to separate b-jets from light flavor (u, d, s, c quarks or gluons) induced jets. For example, in 20% of the cases soft leptons from semi-leptonic decays are present in b-jets. For light flavor jets this is much less common. Soft leptons in this case are low-energetic, non-isolated electrons or muons.

DeepJet [60] is a deep neural network classifier that combines this information to obtain an even better b-jet identification efficiency. The task of DeepJet is to assign the most probable jet

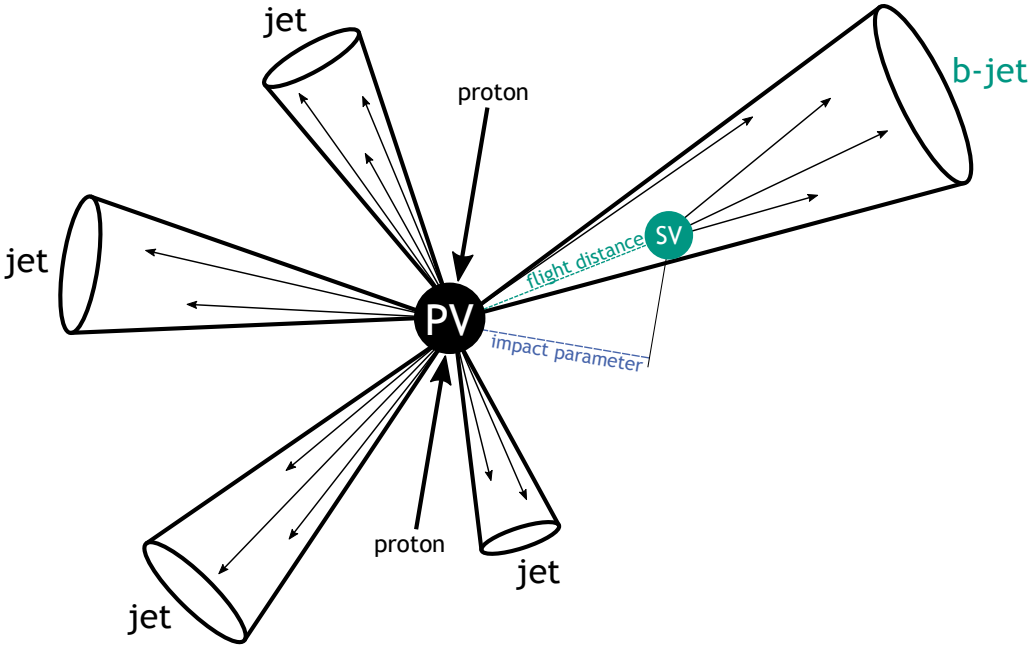


Figure 3.9: Visualization of a b quark induced jet (b-jet) with a secondary vertex (SV). Due to their long lifetime, the flight distance of B hadrons is large enough to reconstruct a SV displaced from the primary vertex (PV). The two dimensional impact parameter is a relevant quantity to define the quality of a reconstructed SV. Illustration adapted from [59].

flavor to each jet in a collision event. To achieve that, it uses information from the secondary vertex associated with a jet, from charged and neutral PF candidates also associated with this jet and global high-level jet information like its kinematics. The network architecture is a combination of convolutional, recurrent and dense layers. First, three independent branches are defined for features from charged PF candidates, neutral PF candidates and features from the SV. Each branch starts with convolutional layers followed by recurrent layers. After that the three branches are combined with the global jet features through a dense layer to produce a probability-like output that sums up to one for all classes. DeepJet is trained to classify the following jet classes:

- *bb* for two B hadrons in a jet
- *b* for one B hadron in a jet
- b_{lep} for a B hadron decaying semi-leptonically in a jet
- *c* for c quark induced jets
- *l* for light flavor induced jets (u, d, s quarks)
- *g* for gluon induced jets

The discriminant for the identification of b-jets is defined by the sum of the predictions of all B hadron related classes (bb, b, b_{lep}). In this analysis two different working points (WPs)

of this DeepJet discriminant are used. The WPs are defined based on the misidentification rate of light flavor induced jets. The *loose* WP has a misidentification rate of 10% and the *medium* WP of 1%. The b-jet identification efficiency for these WPs is around 93% and 83%, respectively. These values are computed based on jets from $t\bar{t}$ events with $p_T > 30$ GeV [60]. In the following, jets passing one of the mentioned WPs will be referred to as b-tagged jets or b-jets.

An energy correction for jets was discussed before in section 3.3.7, but as already mentioned, b-jets differ from jets induced by other quarks or gluons. Especially a higher presence of leptons during the decay is accompanied by neutrinos which leads to a lower energy estimate compared to other jets. Therefore, an energy calibration is performed specifically for b-jets [61].

A neural network is trained on simulated top quark events with the target to learn the correction of the reconstructed b-jet p_T from the top quark decay to the true p_T of the b quark. The neural network utilizes general jet information as well as information about the PF candidates of the jet to estimate the correction factor $p_{T,gen}/p_{T,reco}$. The results of the energy regression are validated with data selected for the production of $Z \to \mu\mu/ee$ in association with at least one b-tagged jet.

In addition, the effect of the b-jet energy regression is validated for one of the signal processes used in this analysis. The signal processes are discussed in more detail later in section 4.2 but a part of the signal is a hypothetical neutral boson which decays into a pair of b quarks. The invariant mass of reconstructed b-jet pairs is shown in figure 3.10. The difference between applied and not applied energy regression is small but clearly visible. Due to the energy regression, the mass peak is corrected closer to the initial boson mass of 100 GeV. Based on the standard deviation divided by the mean of the distribution σ/μ as a figure of merit, the mass reconstruction improves by around 8% for this specific boson mass. The range of improvement can change depending on the initial boson mass but is always present. Therefore, the b-jet energy regression will be applied to all b-tagged jet entering this analysis.

An additional output of the neural network used for the b-jet energy regression is an estimate of the b-jet energy resolution for each jet. This information is needed at a later point in the analysis where a kinematic fit is performed and is discussed in section 4.8.1.

3.4.6 Identification of resonant b quark pair induced jets

Similar to the boosted tau leptons, part of the signal processes in this analysis is a boson that decays into a pair of b quarks. These b quarks can also be boosted in one direction depending on the initial mass of the boson and its p_T. If the boost is large enough, the jets from the two b quarks cannot be resolved individually in the detector anymore. To still be able to probe this boosted phase space in the analysis, a dedicated identification algorithm called *ParticleNet* [62, 63] is used to identify resonantly produced b quark pairs resulting in a wide cone jet. These jets are reconstructed with the anti-k_t algorithm with a radius parameter of 0.8 (AK8), as introduced in section 3.3.7.

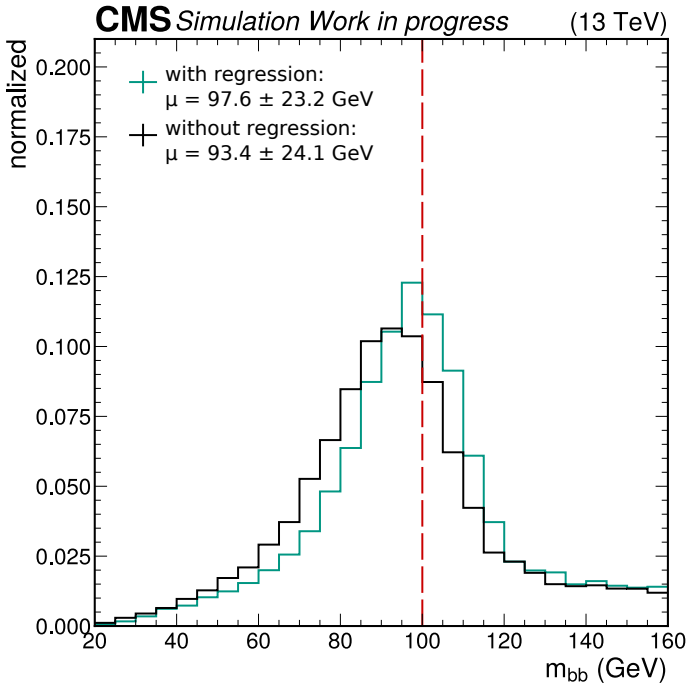


Figure 3.10: Invariant mass of a reconstructed b-jet pair with and without b-jet energy regression applied. The distributions are calculated for the $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ signal process produced for the mass hypothesis $m_X = 500\,\text{GeV}$ and $m_Y = 100\,\text{GeV}$.

ParticleNet is a graph based neural network [64]. Commonly used neural networks (NNs) like feed-forward NNs have a fixed number of input features and the order in which they are given to the NN is also set beforehand. A graph based structure resolves or mitigates these issue. The number of nodes in a graph can be arbitrary and each node can be connected to any other node making the graph permutation invariant. The constant part of a graph is the features or properties vector that each node must have. Such a structure fits well with the general event signatures that are present in the CMS detector. In a pp collision, particles are created in different numbers and decay in different direction through the detector. These particles can originate from the same initial particles or in general be close to each other in the detector, which means that a connection can be established between these particles. Also a feature vector can be defined which is valid for each particle. The feature vector can include the four-vector information like p_T, η or ϕ but also other information like particle type or charge.

For ParticleNet [62, 63] the same idea is applied to jets. Jets are a collection of particles and depending on the initial particle a jet is produced from, the structure within the jet can be very characteristic. In this thesis a mass decorrelated version of ParticleNet(MD) is used. The reason is the previously mentioned boson from the signal processes for which the mass is unknown. ParticleNetMD is trained to identify boosted particles $\mathcal{X}$ with a two prong hadronic decay $\mathcal{X} \to b\bar{b}$, $\mathcal{X} \to c\bar{c}$ or $\mathcal{X} \to q\bar{q}$ (with $q \in u, d, s$). In addition, QCD multijets are used in

the training to consider background jet structures. To ensure the mass independence, simulated signal samples are used with a flat distribution of the spin-0 $\mathcal{X}$ boson mass $m_{\mathcal{X}}$ between 15 and 250 GeV. Both signal and background samples are additionally reweighted to have a flat p_{T} and m_{SD} distribution. The final discriminant used in this analysis is then defined based on the probability-like predictions of ParticleNetMD

$$D_{\mathcal{X}(\mathrm{b\bar{b}})} = \frac{p_{\mathcal{X}(\mathrm{b\bar{b}})}}{p_{\mathcal{X}(\mathrm{b\bar{b}})} + p_{\mathrm{QCD}}}. \tag{3.20}$$

The performance of ParticleNetMD is shown in figure 3.11 and compared to a previously used algorithm in CMS (DeepAK8 [65]) as well as the mass dependent version of ParticleNet. The background versus signal efficiency is measured for the SM case where a boosted Higgs boson decays into a pair of b quarks. The selection of AK8 jets is optimized accordingly by limiting the jet p_{T} to high values between 500 and 1000 GeV and the softdrop mass to be around the Higgs boson mass between 90 and 140 GeV.

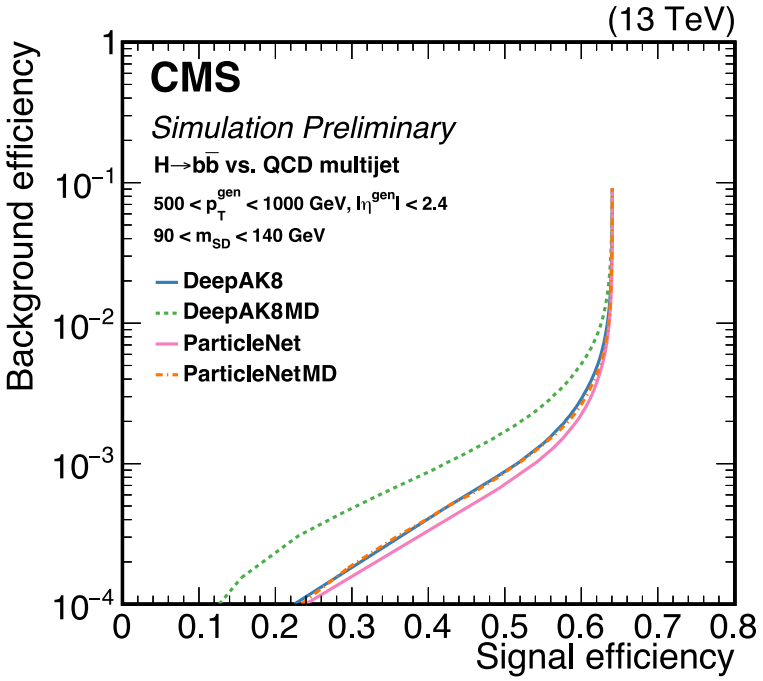


Figure 3.11: Performance of algorithms used for the identification of boosted Higgs bosons decaying into pairs of b quarks reconstructed within AK8 jets is visualized. The plot shows the background efficiency for QCD multijets versus the Higgs boson signal efficiency for mass dependent and mass decorrelated versions of ParticleNet and DeepAK8. Plot taken from [63].

First of all, a clear improvement in performance is visible compared to the previously used DeepAK8 algorithm. The improvement of the MD version is even more significant and basically

on par with the DeepAK8 version specifically trained for the Higgs boson mass. Both algorithms, DeepAK8 and ParticleNet, use more or less the same input information in the training like kinematic features of PF candidates as well as secondary vertices connected to an AK8 jet. Therefore, the performance improvement is a result of the superior graph based architecture of ParticleNet. On the other hand, the architecture of DeepAK8 is based on a convolutional neural network (CNN) in one dimension, which is the list of particles in a jet. The mass decorrelation for DeepAK8MD is achieved by using an adversarial branch in the DeepAK8 architecture that predicts the mass and penalizes the training if the mass is predicted too accurately. ParticleNetMD shows that directly training on mass independent data leads to better results.

The difference in performance between ParticleNet and ParticleNetMD is not as significant as is for the DeepAK8 algorithms, so for simplicity only the mass decorrelated version of ParticleNet is used in this analysis although one of the signal processes has a SM Higgs boson decaying into a b quark pair.

In the following, AK8 jets passing a certain ParticleNetMD WP will be referred to as bb-tagged AK8 jets. As a working point for identifying bb-tagged AK8 jets a value of $D_{\mathcal{X}(\mathrm{b\bar{b}})} > 0.6$ is used. This value is chosen to select as many signal jets as possible, while still rejecting the majority of jets coming from background processes. The efficiency of this WP for selecting $\mathrm{H_{SM}} \to \mathrm{b\bar{b}}$ jets is around 94% [66].

4 Search for di-Higgs events

After the discovery of the Higgs boson and the measurement of its mass, one of the most interesting research topics in the field of particle physics and specifically Higgs physics is the measurement of the Higgs boson self-coupling. The currently available data from the ATLAS and CMS experiments is only sensitive enough to set upper limits on the self-coupling [67, 68]. But in the upcoming years it is expected that the trilinear Higgs self-coupling can be directly measured with the data taken at the High-Luminosity LHC (HL-LHC) [69, 70].

All of these measurements are likely to confirm what the SM already predicts. However, the SM is not the end because it cannot describe everything. Open questions within particle physics lead to the assumption that there must be something beyond the SM. The Higgs sector is a promising field to explore BSM theories as introduced in section 2.3. The parameter space of BSM theories like the NMSSM is large enough to be accessible with the data that is already available from the ATLAS and CMS experiments.

This thesis focuses on a search for di-Higgs events in the CMS data taken in 2018. The di-Higgs production is assumed to be resonant and involves three different Higgs bosons, more specifically neutral scalar Higgs bosons with three different masses. A heavy Higgs boson X is produced via gluon fusion and then decays into two lighter Higgs bosons Y and H, one of which is the Higgs boson known from the SM. For the resonant case the kinematic requirement

$$m_X > m_Y + m_{H_{SM}} = m_Y + 125\,\text{GeV} \tag{4.1}$$

has to be fulfilled. The measurable final state is defined by the decay of Y and H_{SM} into a bb and $\tau\tau$ pair. For the search both possible decay combinations $Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ and $Y(\tau^-\tau^+)H_{SM}(b\bar{b})$ are taken into account as signal processes. The search for the $Y(\tau^-\tau^+)H_{SM}(b\bar{b})$ is performed for the first time. To better visualize the decay chain the Feynman diagrams for both signal processes are drawn in figure 4.1.

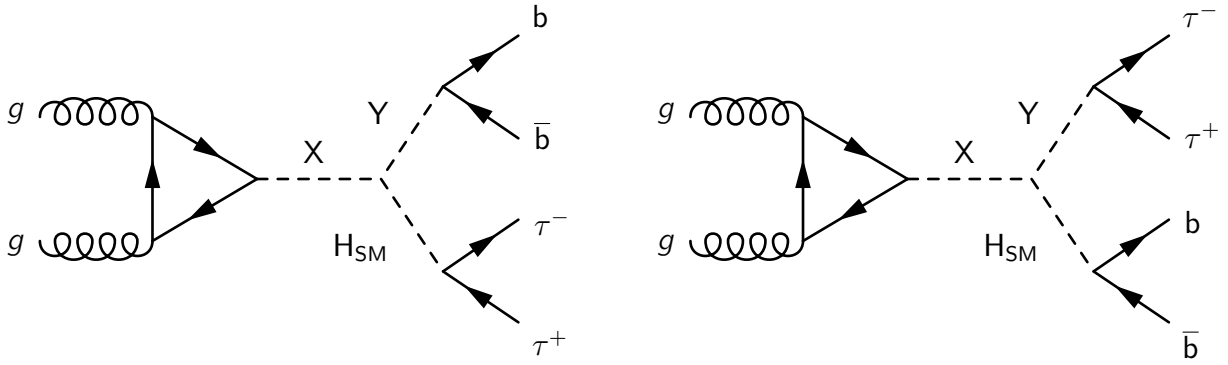
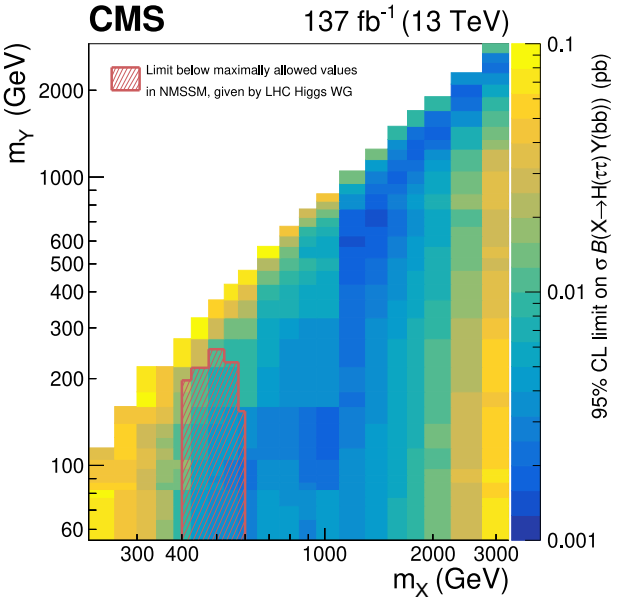


Figure 4.1: Feynman diagrams of the two signal processes which are the target of the search conducted in this thesis. A heavy Higgs boson X produced via gluon fusion decays into two lighter Higgs bosons Y and H_{SM}, which then further decay into either a bb or a $\tau\tau$ pair.

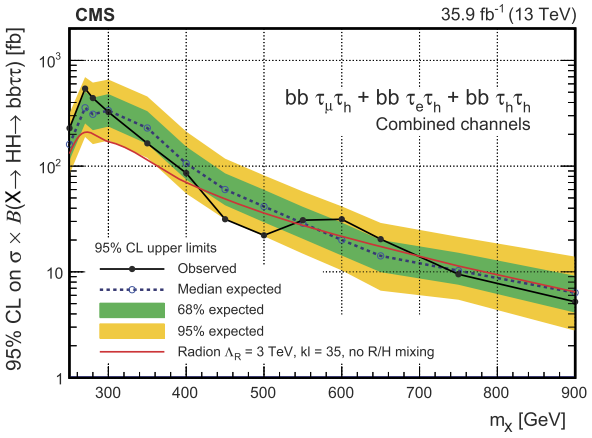
The masses of the X and Y bosons are free parameters of the search, therefore, a two dimensional grid scan of possible mass pair hypotheses, fulfilling equation 4.1, is performed. The details about the event simulation are discussed in section 4.7.2. The analyzed mass ranges are $240\,\text{GeV} \leq m_X \leq 4000\,\text{GeV}$ and $60\,\text{GeV} \leq m_Y \leq 2800\,\text{GeV}$ with 574 mass pair hypotheses in total, while only for 92 of them an upper limit measurement is performed. Depending on the mass pair hypothesis, different final state topologies may emerge from the decay products of the Y and H_{SM} bosons, for example, being highly boosted and therefore need dedicated methods of identification. In the following sections the analysis strategy is discussed. This includes the motivation and introduction of the object and event selection, how the estimation of background processes is done and which additional reconstruction algorithms are used specifically targeting the signature of the signal processes.

4.1 Review of state-of-the-art results

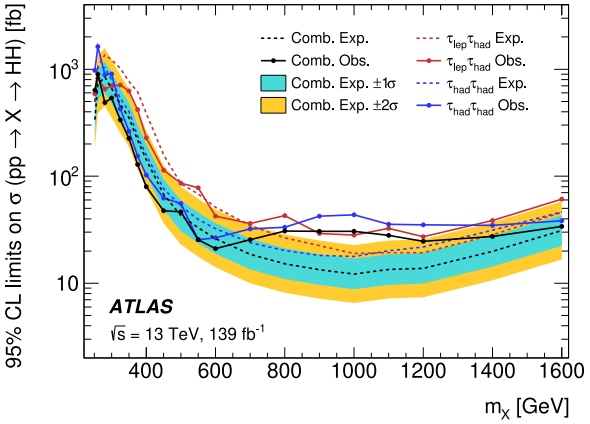
Before introducing the analysis of this thesis, an overview is given about the current status of research in the field of resonant di-Higgs searches. The published result closest to the analysis of this thesis is its predecessor analysis from CMS [9]. The focus of the previous analysis was on the resolved final states of the bb and $\tau\tau$ pairs and the $X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ signal process was considered. The dataset analyzed corresponds to an integrated luminosity of $137.2\,\text{fb}^{-1}$. The included mass range of the X boson mass was $240\,\text{GeV} \leq m_X \leq 3000\,\text{GeV}$ and of the Y boson $60\,\text{GeV} \leq m_Y \leq 2800\,\text{GeV}$. In the analysis no excess was found and exclusion limits on the maximally allowed cross section (σ) times branching fraction ($\mathcal{B}$) in the context of the NMSSM were set for $400\,\text{GeV} \leq m_X \leq 600\,\text{GeV}$ and $60\,\text{GeV} \leq m_Y \leq 200\,\text{GeV}$. In the most sensitive mass region ($m_X = 450\,\text{GeV}$, $60\,\text{GeV} \leq m_Y \leq 80\,\text{GeV}$) the observed limits are five times smaller than the maximal theoretical prediction of the $\sigma \times \mathcal{B}$. The results are shown in figure 4.2a.



(a) $X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ CMS search [9]



(b) $X \to H_{SM}(b\bar{b})H_{SM}(\tau^-\tau^+)$ CMS search [71]



(c) $X \to H_{SM}(b\bar{b})H_{SM}(\tau^-\tau^+)$ ATLAS search [72]

Figure 4.2: Published upper limit results for resonant di-Higgs production with bb + $\tau\tau$ final states from the two multipurpose experiments at the LHC, CMS and ATLAS.

A second CMS measurement [71] was performed with a dataset from 2016, corresponding to an integrated luminosity of $35.9\,\text{fb}^{-1}$. The analysis focused on the symmetric resonant di-Higgs production $X \to H_{SM}(b\bar{b})H_{SM}(\tau^-\tau^+)$ with resolved final states of the bb and $\tau\tau$ pairs. The included mass range of the X boson is $250\,\text{GeV} \leq m_X \leq 900\,\text{GeV}$. The results were interpreted in the context of models with warped extra dimensions [73, 74] and no exclusion limits were set for the selected model parameters, which were the radion mass scale $\Lambda_R = 3\,\text{TeV}$, the size of extra dimensions $kl = 35$ and the assumption that the radion does not mix with the Higgs boson. Due to the narrow width approximation [75], which can be assumed when the detector resolution is worse than the width of a particle, the radion can also be interpreted in other

models as a heavy resonance. In the case of NMSSM, the radion would correspond to the X boson. The results of this search are shown in figure 4.2b.

Further, there is a similar measurement from the ATLAS collaboration [72], which also focuses the symmetric resonant di-Higgs production $X \to H_{SM}(b\bar{b})H_{SM}(\tau^-\tau^+)$ with resolved final states of the bb and $\tau\tau$ pairs. The results are shown in figure 4.2c. The measurement was performed on a dataset that corresponds to an integrated luminosity of $139\,\text{fb}^{-1}$. The range of scanned masses of the X boson is $251\,\text{GeV} \leq m_X \leq 1600\,\text{GeV}$. In the publication no specific comparison with theoretical predictions is made, however, an excess was observed at an X boson mass of $1000\,\text{GeV}$ with a local (global) significance of 3.1 (2) standard deviations. In future searches, it needs to be confirmed if this excess is a sign of BSM physics or just a statistical fluctuation.

In the results section 5.3.2, these three measurements are compared with the results of this thesis.

4.2 Signal phase space

In comparison to this thesis the previous search [9] focused only on the $Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ final state and the selection of events focused on resolved final states for both bb and $\tau\tau$ pairs. A resolved final state means that all four final state objects (bb, $\tau\tau$) can be reconstructed individually in the detector. Since the default jet size parameter at CMS is set to a radius of 0.4, objects can be reconstructed well as long as the spatial distance $\Delta R(\text{obj}_1, \text{obj}_2)$ in the η-ϕ plane between them is approximately larger than 0.4. This is not a discrete transition but the probability to resolve two objects drops steadily for $\Delta R(\text{obj}_1, \text{obj}_2) < 0.4$. Studies on this are presented in section 4.3.

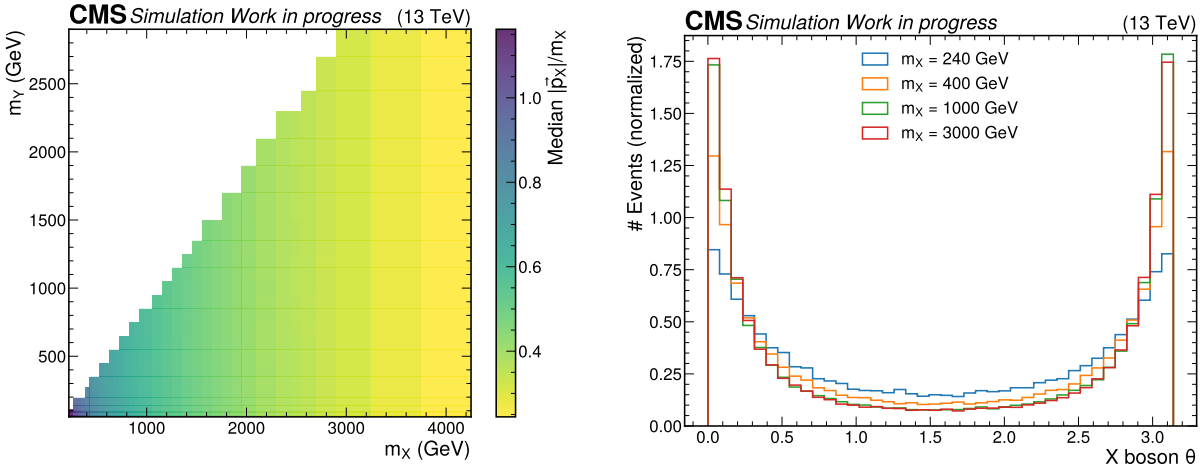
Only selecting resolved objects as was done in the previous search [9] cuts into the acceptance of expected signal events. The spatial distribution of the decay particles in the detector strongly depends on the mass pair hypotheses used for the X and Y bosons during the signal event generation. The full simulation process of the signal events is discussed in a later section 4.7. Since simulation is used to generate signal events, information about the true b quarks and tau leptons is available before any further decay and before the detector response is simulated. In the following a generator study, based on some previous studies [76, 77], is conducted to better understand how the individual particles in the signal processes are produced and decay geometrically. This is important to define a phase space for the object and event selection.

The first particle of interest is the X boson. This boson is relatively heavy even for the lowest mass hypothesis ($240\,\text{GeV}$). This means that the largest part of the energy with which the X boson is produced is used for its mass. This is visualized in figure 4.3a where for all mass pair hypotheses of the analysis the median of the distribution of the X boson momentum normalized to its mass is calculated. Only for the smallest X boson masses, this ratio

$$\gamma = \frac{|\vec{p}_X|}{m_X} \tag{4.2}$$

is larger than one. From this it can be assumed that for higher X boson masses the boson is produced more at rest. Since the X boson comes before the Y boson in the decay chain, the ratio γ does not depend on the Y mass.

Additionally, looking at the θ distribution of the produced X boson in figure 4.3b, it is clear from the two peak structure that the X boson is preferably produced in forward and backward directions of the detector. This means the gluons involved in the gluon fusion producing the X boson point to an imbalance of their momenta towards one of the gluons. A possible reason for this behavior is that the probability for a very high energetic gluon is higher than having two moderately high energetic gluons, at least for the energies needed to produce the X boson. It is also visible that this can change for smaller X boson masses where the θ distribution is less peaking. These results are valid for both signal processes because up to this point in the decay chain they are identical.



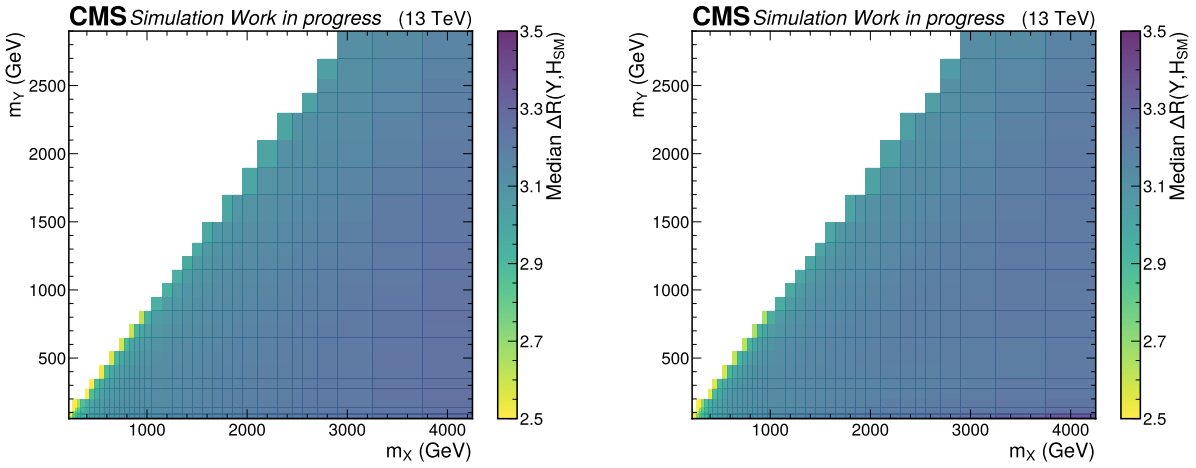
(a) Median of $\frac{|\vec{p}_X|}{m_X}$

(b) θ of the X boson

Figure 4.3: Kinematic information of the X boson. In (a) the median of the distribution of the X boson momentum normalized to its mass is plotted in a two dimensional histogram including all mass pair hypotheses. In (b) the θ distribution of the X boson is plotted for some selected masses of the X boson to visualize the kinematic differences.

Next, it is interesting to identify how the Y and H_{SM} bosons are produced from the decay of the X boson. In figure 4.4, it is shown that for both signal processes the spatial distance ΔR between the Y and H_{SM} bosons is basically independent from m_X and m_Y. The median of ΔR is calculated using all simulated events for each individual mass pair hypothesis of m_X and m_Y. The result is almost always close to the value of π, which indicates that the Y and H_{SM} bosons are produced back-to-back. Only in the kinematic case where the mass of the X boson is very close to the sum of the Y and H_{SM} boson masses and m_X is in general relatively small (below 1000 GeV) the median ΔR gets smaller. This is consistent with the ratio γ getting larger since more momentum results in a larger Lorentz boost. But still, the smallest found value is bigger

than 1.3 which makes it quite clear that the Y and H_{SM} bosons are well separated from each other for all mass pair hypotheses. Since in the simulation the decay chain up to the Y and H_{SM} bosons is the same for both signal processes, the result shown in figure 4.4a and 4.4b are basically identical and the small differences are related to statistical fluctuations.

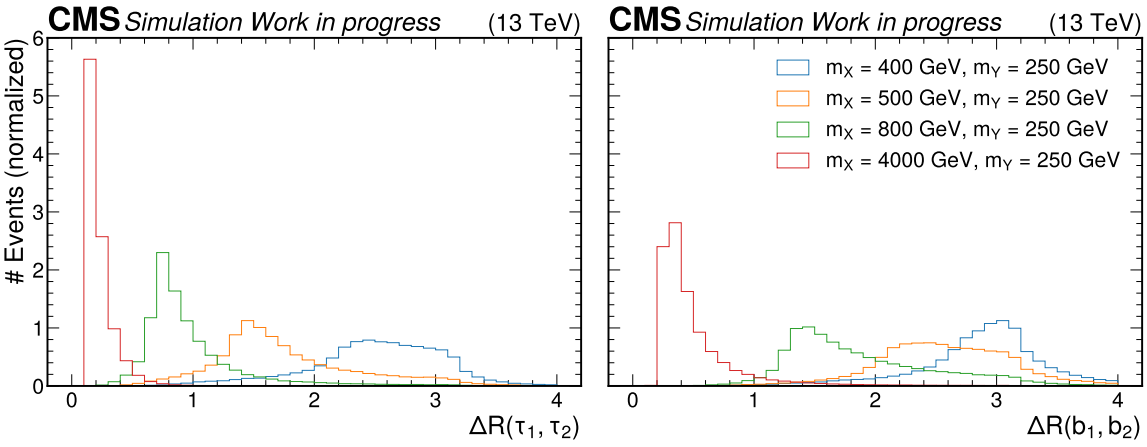


(a) $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ (b) $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$

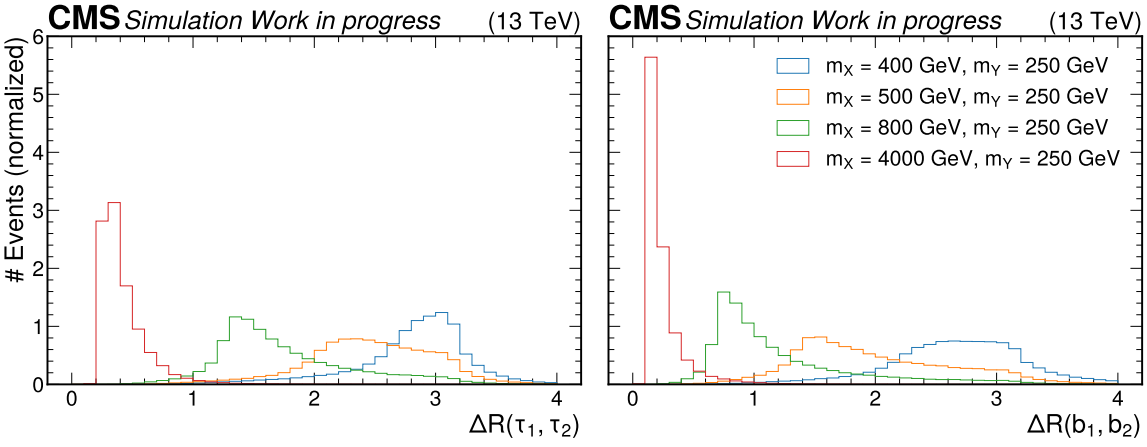
Figure 4.4: The median spatial distance ΔR between the Y and H_{SM} bosons for the two signal processes. All simulated m_X and m_Y hypotheses are used and plotted in a two dimensional histogram with the median spatial distance ΔR as color encoded bin yields.

Differences between both signal processes start to appear when looking at the decay products of the Y and H_{SM} bosons. The kinematic properties and spatial decay patterns of the bb and $\tau\tau$ pairs are the most crucial part of the object selection because these are the particles that produce the measurable signatures in the detector. In contrast to the ΔR between the Y and H_{SM} bosons, the spatial distance between the two b quarks and the two tau leptons strongly depends on the parameters chosen for m_X and m_Y. This effect can be seen in figure 4.5 where the mass of the Y boson is fixed to 250 GeV and m_X is varied. Increasing X boson masses lead to increasing momenta of the Y and H_{SM} bosons when they are produced. These momenta affect the direction of the decay products of the Y and H_{SM} bosons. A higher momentum leads to a stronger Lorentz boost and the b quark or tau lepton pairs decay spatially closer to each other. This effect is indicated by the peak of the ΔR distribution moving to smaller values for higher values of m_X.

Further, the asymmetry between the Y boson mass and the mass of H_{SM} has an effect on the collinearity of the final state pairs (bb and $\tau\tau$). The decay products of the boson with the lower mass get a stronger boost and therefore are closer to each other. In the case of the shown mass pair hypotheses in figures 4.5a and 4.5b, the H_{SM} mass is smaller, resulting in overall smaller ΔR values for the $\tau\tau$ pair for the $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ signal process and the bb pair



(a) $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$

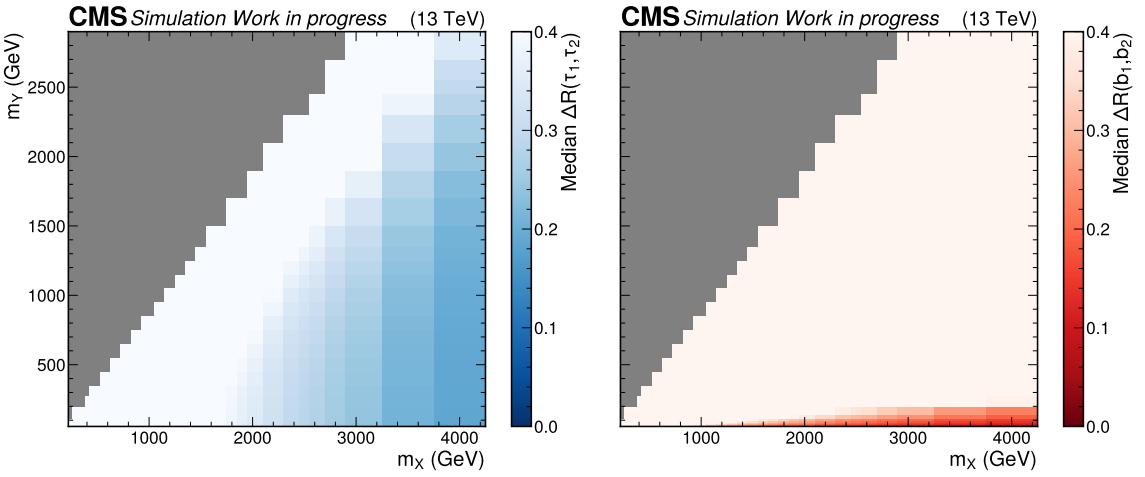


(b) $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$

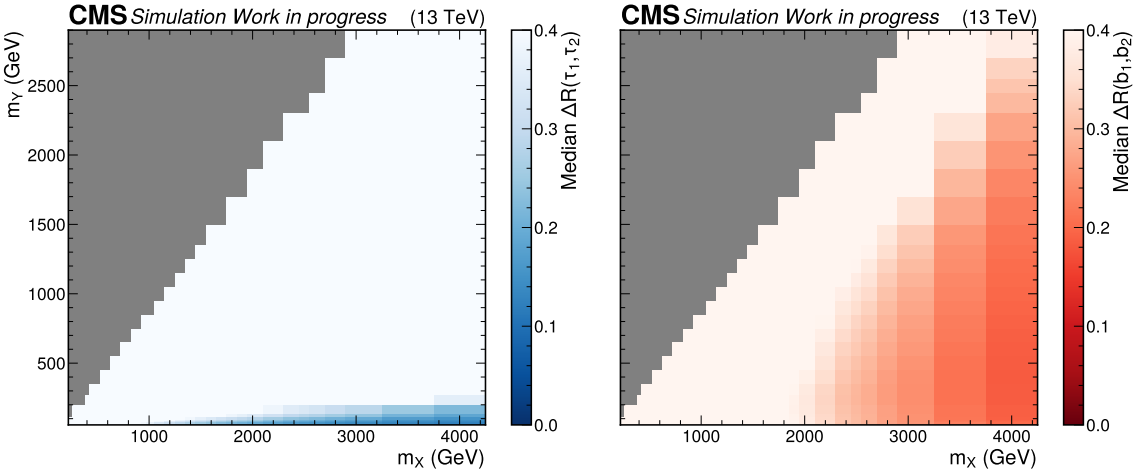
Figure 4.5: Distributions of the spatial distance ΔR for the final state particle pairs. In (a) and (b) the ΔR distributions for the two different signal processes are shown. The two left plots refer to the $\tau\tau$ pair indicated by τ_1 and τ_2 and the two right plot refer to the bb pair indicated by b_1 and b_2. Some selected values of m_X and m_Y are used to demonstrate the increasing collinearity of the bb and $\tau\tau$ pairs with increasing X boson mass.

for the $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$ signal process compared to their bb and $\tau\tau$ counterparts, respectively, from the Y boson decay.

These findings can be extended to the full mass grid similar to $\Delta R(Y, H_{SM})$ by calculating the median of the ΔR distributions for each mass pair hypothesis of m_X and m_Y. In figure 4.6 the results of this exercise are visualized. The ΔR scale is restricted to be between 0 and 0.4 to emphasize the relevance of a dedicated object and event selection targeting boosted final states of the signal processes.



(a) $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$



(b) $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$

Figure 4.6: Two dimensional distributions of the median spatial distance ΔR for the final state particle pairs. In (a) and (b) the median ΔR distributions for the two different signal processes are shown. The two left plots refer to the $\tau\tau$ pair indicated by τ_1 and τ_2 and the two right plots refer to the bb pair indicated by b_1 and b_2. All simulated m_X and m_Y hypotheses are used and the color encoded ΔR values are restricted to a range from 0 to 0.4 to highlight mass pair hypotheses with highly boosted final state particle pairs.

Starting with the signal process $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ that was searched for in the previous analysis [9], it is clearly visible in the left plot of figure 4.6a that a resolved selection of the $\tau\tau$ pair is not suitable for a significant number of mass pair hypotheses. Going above 2000 GeV for the X boson mass, more than 50% of the $\tau\tau$ pairs are too close to each other for the tau leptons to be resolved individually. An exception is if the Y boson mass is similar

to the X mass because then the majority of the X boson energy is used to produce the Y and H_{SM} bosons and not much is left to give them a significant momentum.

On the other hand, due to the previously mentioned asymmetry effect of the Y and H_{SM} boson masses, the bb pair is much more often in a resolvable state as shown in the right plot of figure 4.6a. This is the case because m_Y is larger than $m_{H_{SM}}$ for the majority of the mass pair hypotheses. Only for very low Y boson masses ($< 200\,\text{GeV}$) is a dedicated boosted bb pair selection gaining relevance.

For the other signal process $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$ the results are basically reversed. As displayed in the plots in figure 4.6b, the $\tau\tau$ pair is expected to be much more often in a resolvable state and a dedicated boosted selection of the bb pair becomes more relevant.

Considering the results of these studies, the previous analysis [9] is missing out by not always selecting the expected phase space of the signal processes. There is a transition from resolved to boosted final states of the bb and $\tau\tau$ pairs dependent on the masses of the X and Y bosons. The goal of this analysis is to introduce boosted object selections in addition to the resolved selection and by that to improve upon the previously conducted measurement.

4.3 Physics object selection

Based on the findings of section 4.2 on the expected signal signatures, reconstructed objects, as discussed in section 3.4, like electrons, muons, τ_h and jets, have to be selected specifically targeting these event signatures. The first selection step is to identify genuine tau lepton pairs. This is described in sections 4.3.1 and 4.3.2 for resolved and boosted tau lepton pairs, respectively. Next, a jet selection is applied to identify the genuine b quark pairs. The details of this selection are summarized in sections 4.3.3 and 4.3.4 for resolved and boosted b quark pairs, respectively.

4.3.1 Selection of resolved tau lepton pair decays

The different orthogonal analysis channels are set up based on the selection of the tau lepton pairs. The selection of the resolved tau lepton pair is basically the same as in the previous analysis [9]. Since a pair of tau leptons has to be selected, final state combinations depending on the exact decay of each tau lepton into an electron, muon or τ_h are considered. Combining the individual decays and branching fractions given in table 3.1 results in six final states visualized in figure 4.7. For the analysis only the semi-leptonic $e\tau_h$ and $\mu\tau_h$ channels are considered as well as the fully hadronic $\tau_h\tau_h$ channel. First, these three channels already cover around 87% of all tau lepton pair decays. Second, the fully leptonic channels $e\mu$, ee and $\mu\mu$ suffer from a much higher background contribution from the production and decay of top quark pairs (especially $e\mu$) and Z bosons (especially ee, $\mu\mu$) compared to the semi-leptonic or fully hadronic channels. For these reasons, they do not contribute significantly to the analysis sensitivity.

To select the final state objects for the three channels in consideration, different quality criteria are applied to the electrons, muons or τ_h. These criteria include cuts e.g. on the transverse

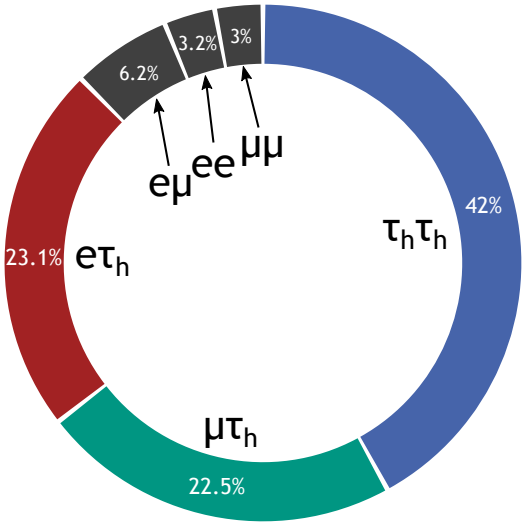


Figure 4.7: Branching fractions of all possible $\tau\tau$ decays. For the analysis only the $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$ final states are considered. These channels already cover around 87% of all possible $\tau\tau$ final states. The other channels are not part of the analysis due to their small contribution.

momentum, relative isolation of the object or identification. Further, an event needs to be triggered based on some online criteria which are usually related to leptons.

For the $e\tau_h$ ($\mu\tau_h$) channel, a single electron (muon) trigger is used, which starts with a lower p_T threshold of $32\,\text{GeV}$ ($24\,\text{GeV}$). The trigger fires depending on the p_T, isolation and electron (muon) identification which can already be applied online during data taking. These criteria impact the cuts that can be applied offline. The offline p_T threshold is increased by $1\,\text{GeV}$ compared to the lower p_T threshold of the trigger to remove events from the trigger turn-on region that are difficult to describe by simulation. All applied cuts for electrons (muons) are summarized in table 4.1. These cuts define the electron (muon) candidates considered for an $e\tau_h$ ($\mu\tau_h$) pair in an event.

The offline selection for τ_h candidates depends on the channel and the cuts applied for this analysis are summarized in table 4.2. For the $\tau_h\tau_h$ channel, a set of di-tau triggers is used which is specifically designed to trigger on the presence of two τ_h candidates are present in an event. The offline p_T cut for τ_h candidates is increased by $5\,\text{GeV}$ compared to the trigger p_T requirement due to a less steep turn-on region.

After the electron, muon and τ_h candidates are selected, pairs of electron and τ_h, muon and τ_h and two different τ_h are constructed for the $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$ channels, respectively. A good pair is required to have opposite charge and have a spatial separation of $\Delta R > 0.4$. If after applying these conditions, still multiple pairs are present, the pair where the two candidates are the most isolated and have the highest p_T are selected.

Table 4.1: List of quality criteria applied to electron and muon candidates that have to be fulfilled for the electrons/muons to be counted as good for the resolved tau lepton pair selection in the $e\tau_h$ and $\mu\tau_h$ channels, respectively.

	Transverse momentum	Pseudo-rapidity	Identification	Relative isolation	Distance from PV
Electrons	$p_T > 33\,\text{GeV}$	$\lvert\eta\rvert < 2.1$	90% eff. WP	$\text{Iso}_{\text{rel}} < 0.15$	$d_{xy} < 0.045\,\text{cm}$ $d_z < 0.2\,\text{cm}$
Muons	$p_T > 25\,\text{GeV}$	$\lvert\eta\rvert < 2.1$	medium WP	$\text{Iso}_{\text{rel}} < 0.15$	$d_{xy} < 0.045\,\text{cm}$ $d_z < 0.2\,\text{cm}$

Table 4.2: List of quality criteria applied to τ_h candidates that have to be fulfilled for the τ_h to be counted as good for resolved tau lepton pairs in the $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$ channels.

Final state	Transverse momentum	Pseudo-rapidity	DeepTau identification	Distance from PV
$e\tau_h$	$p_T > 30\,\text{GeV}$	$\lvert\eta\rvert < 2.3$	Medium WP vs. jet Tight WP vs. e VLoose WP vs. μ	$d_z < 0.2\,\text{cm}$
$\mu\tau_h$	$p_T > 30\,\text{GeV}$	$\lvert\eta\rvert < 2.3$	Medium WP vs. jet Tight WP vs. e VLoose WP vs. μ	$d_z < 0.2\,\text{cm}$
$\tau_h\tau_h$	$p_T > 40\,\text{GeV}$	$\lvert\eta\rvert < 2.1$	Medium WP vs. jet Tight WP vs. e VLoose WP vs. μ	$d_z < 0.2\,\text{cm}$

The offline object reconstruction and identification are more sophisticated than the object reconstruction that is performed online. To ensure that the offline selected electron, muon or τ_h pair are responsible for firing the single electron, muon or di-tau trigger, respectively, the online and offline objects are required to spatially match to each other. This means that their spatial distance $\Delta R(\text{obj}_{\text{off}}, \text{obj}_{\text{on}})$ has to be smaller than 0.4 and the particle ID should be the same. In case of the di-tau trigger each of the selected τ_h candidates has to match one of the trigger legs.

4.3.2 Selection of boosted tau lepton pair decays

The selection of boosted tau lepton pairs proceeds similar to the resolved tau lepton pair selection. This includes the same final states of the $\tau\tau$ pair decay that are considered as in the resolved case, which are the semi-leptonic $e\tau_h$ and $\mu\tau_h$ channels and the fully hadronic $\tau_h\tau_h$ channel.

The main difference are the triggers used to select the events. The tau lepton pair is boosted, which means that both pair objects are expected to be close to each other. Therefore, isolation of an objects is not suited as selection criterion anymore and not part of the object selection criteria. For the $e\tau_h$ ($\mu\tau_h$) channel a non-isolated single electron (muon) trigger is used as the main trigger in combination with an isolated single electron (muon) trigger. The first trigger without isolation, which is not prescaled, is used with a p_T greater than 115 GeV (50 GeV). The offline p_T threshold is increased by 5 GeV to avoid the turn-on region of the trigger. The isolated triggers are the same as the triggers already used for the resolved tau lepton pair selection and are only used for electrons (muons) in the p_T range of $33-120$ GeV ($25-55$ GeV). Depending on the spatial distance between the electron (muon) and the boosted τ_h, the electron (muon) can still be sufficiently isolated to be triggered by a trigger with isolation and since the p_T threshold of this trigger is lower than for the non-isolation electron (muon) trigger more events can be selected that way. The offline selection criteria for electron (muon) pair candidates are listed in table 4.3.

Table 4.3: List of quality criteria applied to electron and muon candidates that have to be fulfilled for the electron/muon to be counted as good for the boosted tau lepton pair selection in the $e\tau_h$ and $\mu\tau_h$ channels, respectively.

	Transverse momentum	Pseudo-rapidity	Identification	Distance from PV
Electrons	$p_T > 33$ GeV	$\|\eta\| < 2.1$	90% eff. WP	$d_{xy} < 0.045$ cm $d_z < 0.2$ cm
Muons	$p_T > 25$ GeV	$\|\eta\| < 2.1$	medium WP	$d_{xy} < 0.045$ cm $d_z < 0.2$ cm

The boosted τ_h candidates are selected based on a different tau identification algorithm (see section 3.4.4) than the resolved τ_h candidates and a summary of the criteria is given in table 4.4. The tau identification includes isolation, therefore, to avoid too isolated boosted τ_h candidates the cuts are chosen relatively loose.

For the $\tau_h\tau_h$ channel none of the available di-tau triggers can be used. Therefore, a combination of an AK8 jet trigger and a $\not{E}_T$ (missing transverse energy) trigger is used. The AK8 jet trigger primarily targets the jet used to reconstruct the boosted τ_h candidates but in case of the signal processes of this analysis the AK8 jet can also come from the jets induced by the b quarks. Additionally, a MET trigger is considered because at least two neutrinos are expected from the tau lepton decays.

After the selection of the boosted electron, muon and τ_h candidates, the pairs are constructed in a similar way as the resolved tau lepton pairs. The pairs are required to have opposite charge and the pair with the highest p_T of the candidates is selected. Contrary to the resolved tau lepton pair, the spatial distance ΔR is required to be between 0.1 and 0.8. Further a trigger object matching is done for the $e\tau_h$ and $\mu\tau_h$ channels with the reconstructed electron and

Table 4.4: List of quality criteria applied to τ_h candidates that have to be fulfilled for a τ_h to be counted as good for boosted tau lepton pairs in the $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$ channels.

Final state	Transverse momentum	Pseudo-rapidity	DeepTau identification
$e\tau_h$	$p_T > 40\,\text{GeV}$	$\lvert\eta\rvert < 2.3$	Loose WP vs. jet Loose WP vs. e
$\mu\tau_h$	$p_T > 40\,\text{GeV}$	$\lvert\eta\rvert < 2.3$	Loose WP vs. jet VLoose WP vs. e Loose WP vs. μ
$\tau_h\tau_h$	$p_T > 40\,\text{GeV}$	$\lvert\eta\rvert < 2.1$	Loose WP vs. jet

muon, respectively. For the $\tau_h\tau_h$ channel no matching is performed because the trigger object can be from different parts of the signal process.

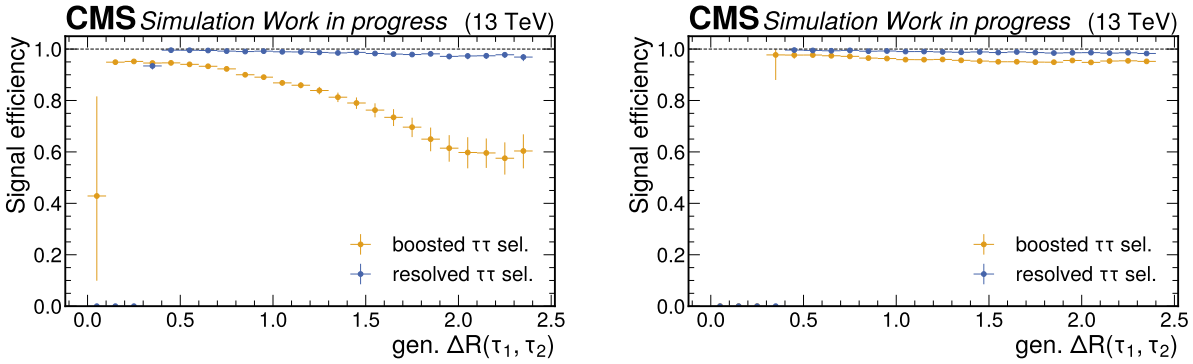
The efficiency of the boosted and resolved $\tau\tau$ pair selection is compared in figure 4.8 for the $\mu\tau_h$ channel for both signal processes to verify that both selections are targeting the expected phase space. To calculate the efficiency, generator level tau leptons are matched to the selected $\mu\tau_h$ pairs within $\Delta R < 0.2$. The shown signal events are selected from the mass pair hypotheses with $m_Y = 500\,\text{GeV}$ and a varying X boson mass from 650 to 4000 GeV to cover various topologies of the tau lepton pair in the detector.

For the efficiency measurement, the aforementioned $\Delta R < 0.8$ cut is dropped showing that the boosted τ_h reconstruction generally also works for resolved final states of the tau lepton pair. However, the resolved τ_h reconstruction yields a better efficiency, close to one and over nearly the full ΔR range. It is visible that the resolved selection efficiency drops to almost zero for $\Delta R < 0.4$ due to the resolved di-tau selection, while the boosted selection still is able to correctly identify the $\mu\tau_h$ pair as anticipated. The differences between the efficiencies of the two signal processes are a result of the mass asymmetry $m_Y > m_{H_{SM}}$, which means that the decay products of the H_{SM} boson have a stronger boost.

4.3.3 Selection of b-jet pairs

The selection of resolved b-jet pairs is based on the reconstruction and identification of b quark induced jets described in section 3.4.5. The b-jet candidates have to fulfill the selection criteria summarized in table 4.5. The p_T and η cuts are restricted by the *DeepJet* algorithm used for b-jet identification, which is highly dependent on the tracker system and therefore only applicable in the central region of the detector.

In an event several jets can pass the selection requirements. As long as a jet passes the medium WP it is defined as b-tagged. If at least two b-tagged jets are found, the two highest p_T b-tagged jets are selected for the b-jet pair. In cases where only one b-tagged jet is found, a second jet is searched for that at least passes the loose WP. If such a second jet is found, the



(a) $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ (b) $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$

Figure 4.8: Selection efficiency of the signal processes for the $\mu\tau_h$ pair selection. In (a) and (b) the two signal processes are displayed. The boosted selection is compared to the resolved selection and shows that it has a worse overall selection efficiency, except for the $\Delta R(\tau_1, \tau_2) < 0.4$ region. The efficiency of the resolved selection drops almost directly around $\Delta R(\tau_1, \tau_2) < 0.4$ due to the applied isolation criteria. The differences between the two signal processes are a result of the mass asymmetry $m_Y > m_{H_{SM}}$. In (b) the efficiency of the boosted selection also drops for $\Delta R(\tau_1, \tau_2) < 0.4$ because for the chosen value of $m_Y = 500\,\text{GeV}$ the $\tau\tau$ pairs are not significantly boosted (see figure 4.6b).

Table 4.5: List of quality criteria applied to b-jet candidates that have to be fulfilled for b-jets to be counted as good for the resolved b-jet pair. These criteria are applied in all $\tau\tau$ decay channels considered.

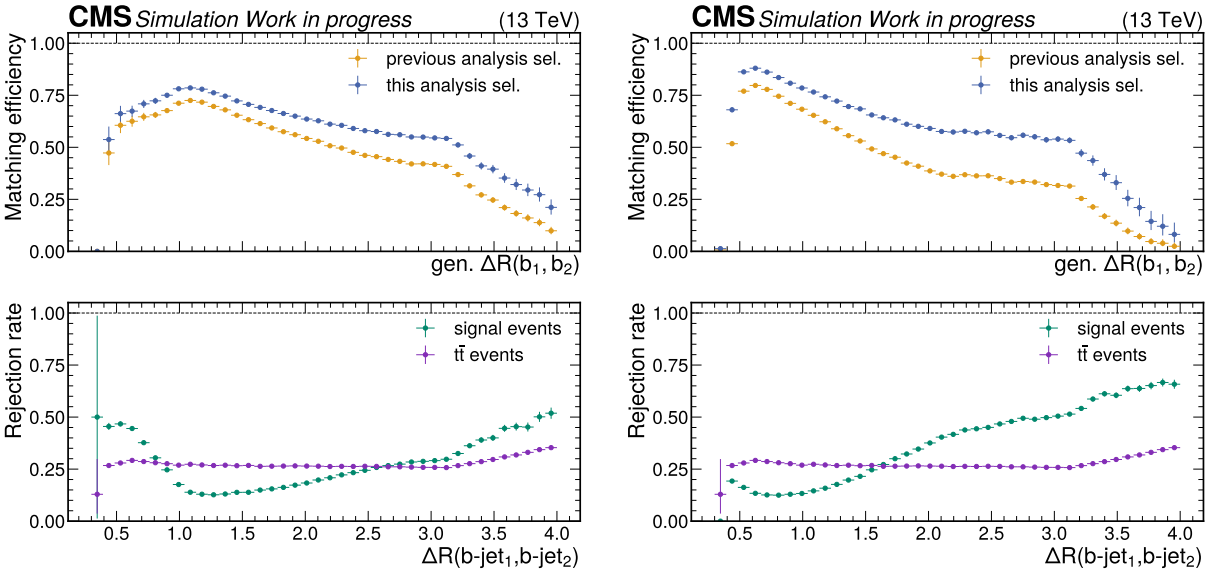
	Transverse momentum	Pseudo-rapidity	DeepJet Identification
first b-jet	$p_T > 20\,\text{GeV}$	$\lvert\eta\rvert < 2.5$	medium WP
second b-jet	$p_T > 20\,\text{GeV}$	$\lvert\eta\rvert < 2.5$	loose WP

two jets are selected as the b-jet pair. When a b-jet pair is found, it additionally has to be well separated, therefore, a spatial distance of $\Delta R(b_1, b_2) > 0.4$ is required.

Such a b-jet pair selection differs from the selection used in the previous analysis [9]. There, in case of only one b-tagged jet found, a second jet with the highest b-tagging score was used. With such a selection more events can be selected but on the other hand the second jet more often does not match the true b-jet in the signal processes. In figure 4.9 both selections are compared.

First, in the upper plots the b-jet pair selection efficiency is shown for both signal processes. The efficiency indicates for how many signal events the selected b-jet pair matches to the

generator level b quarks. For a valid match each b-jet needs to be within one of the $\Delta R < 0.2$ cones around the two b quarks. The efficiency is calculated as a function of the ΔR distance between the two generator level b quarks. The selection used in this analysis improves the b-jet pair matching efficiency significantly for both signal processes. For the mass pair hypotheses considered, on average the efficiency increases around 15% (23%) for the signal process with the Y (H_{SM}) boson decaying to bb.



(a) $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ (b) $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$

Figure 4.9: Comparison of b-jet pair selections used in this and the previous analysis [9], split for the two signal processes in (a) and (b). The two upper plots show the matching efficiency of selecting the correct b-jets induced by the b quarks of the signal processes dependent on the ΔR distance between the two generator level b quarks. The two lower plots show the event rejection rate for the signal processes and the main background process $t\bar{t}$ dependent on the ΔR distance between the two b-jets on reconstruction level.

The tighter b-jet pair selection is applied to all processes, therefore, it is important to verify that the new selection does not affect the signal to background ratio in a negative way. The lower plots in figure 4.9 show a comparison of the event rejection rate due to the tighter b-jet pair selection for the signal processes and for the main background of the analysis, which is the top quark pair production ($t\bar{t}$), dependent on the ΔR distance between the two b-jets on reconstruction level. The $t\bar{t}$ event rejection remains quite constant and is on average around 26%. On the other hand, for the signal events the rejection rate significantly depends on $\Delta R(\text{b-jet}_1, \text{b-jet}_2)$ and for some ΔR regions it is larger than for $t\bar{t}$. But after averaging out the rejection, where each bin is weighted with its event number, the overall rejection rate is around

20% (23%) for the signal process with the Y (H_{SM}) boson decaying to bb and thus smaller than for $t\bar{t}$. Based on these results, the tighter b-jet pair selection is favored in this analysis.

4.3.4 Selection of bb-tagged AK8 jets

Boosted b-jet pairs are more likely to be reconstructed as single AK8 jets instead of individual AK4 jets due to their small spatial distance. The identification of the b quark pair induced AK8 jets are discussed in section 3.4.6. The selection criteria for an AK8 jet are listed in table 4.6 and are mainly restricted by the *ParticleNetMD* identification. This algorithm is trained on events with a certain jet selection which does not allow to go lower than 250 GeV in p_T and lower than 30 GeV in the softdrop mass.

Table 4.6: List of quality criteria applied to AK8 jets candidatees that have to be fulfilled to be counted as good for the boosted b-jet pair. These criteria are applied in all $\tau\tau$ decay channels in consideration.

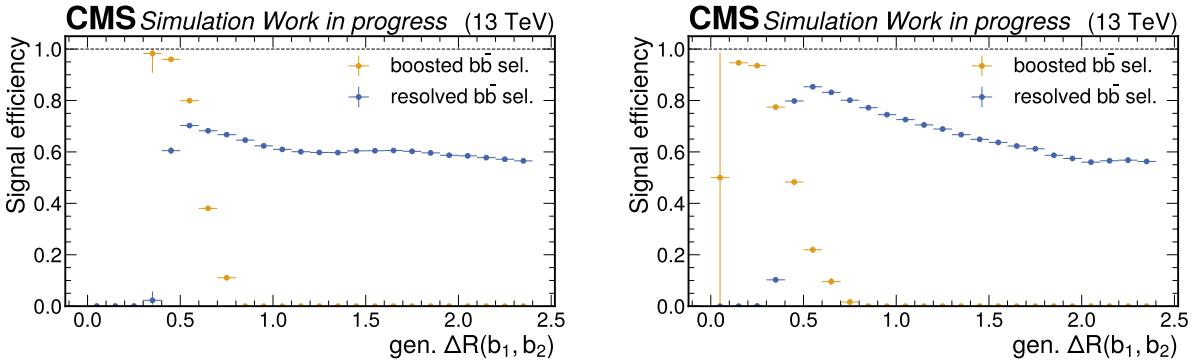
Transverse momentum	Pseudo-rapidity	ParticleNetMD Identification	Softdrop mass
$p_T > 250$ GeV	$\|\eta\| < 2.5$	$D_{\mathcal{X}(b\bar{b})} > 0.6$	$m_{SD} > 30$ GeV

To verify that the boosted bb pair selection targets the correct phase space, the efficiencies of the boosted and resolved bb pair selections are compared for both signal processes in figure 4.10. For the boosted selection the two generator level b quarks have to be within a ΔR cone smaller than 0.4 around the selected AK8 jet on reconstruction level. For the resolved selection the same generator matching is applied as already described in section 4.3.3. For the signal mass pair hypotheses m_Y is set to 500 GeV and m_X is varied between 650 and 4000 GeV to cover all possible constellations of the b quark pair in the detector.

The results show that the efficiency for the boosted bb pair selection via an AK8 jet is decreasing with increasing ΔR distance of the b quarks which means the AK8 jet cannot contain both b quarks. On the other hand, the resolved bb pair efficiency drops for small ΔR distances because the pair cannot be resolved anymore individually as AK4 jets in the detector. This finding is in general valid for both signal processes, but if the bb pair comes from the Y boson it is significantly less boosted compared to the decay from H_{SM}. This is expected due to the mass difference $m_Y > m_{H_{SM}}$ and was already described in section 4.2. In summary, the selection efficiencies show that both, boosted and resolved, selections are relevant for the analysis to cover the full signal phase space of the bb final state.

4.4 Event selection

After all $\tau\tau$ and bb pair selections are applied as described in sections 4.3.1-4.3.4, the full event selection is carried out. A schematic description of this process can be found in figure 4.11.



(a) $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ (b) $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$

Figure 4.10: Selection efficiency of the signal processes for the bb pair selection. In (a) and (b) the two signal processes are displayed. The boosted selection is compared to the resolved selection and shows the distribution from small $\Delta R(b_1, b_2)$ where the boosted selection efficiency is high to higher $\Delta R(b_1, b_2)$ where this is the case for the resolved selection. The transition occurs roughly at $\Delta R = 0.4$ depending on the mass pair hypothesis of the signal processes. In (a) the efficiency of the boosted selection also drops for $\Delta R(b_1, b_2) < 0.3$ because for the chosen value of $m_Y = 500\,\text{GeV}$ the bb pair is not significantly boosted (see figure 4.6a).

First, the events are split orthogonally into a boosted and resolved $\tau\tau$ analysis. One of the main reasons is related to the data-driven background estimation method used in this analysis (see section 4.6) which depends on the τ_h identification. Since different algorithms are used for the boosted and resolved τ_h identification, they have to be taken care of independently.

If a boosted tau lepton pair can be reconstructed in an event and the $\Delta R(\tau_1, \tau_2)$ distance is smaller than 0.8 and larger than 0.1, the event is classified into the boosted $\tau\tau$ analysis. This is done even if a resolved tau lepton pair can be reconstructed in this event because the boosted $\tau\tau$ analysis suffers from lower event yields. All other events for which a good boosted tau lepton pair could not be reconstructed are classified into the resolved $\tau\tau$ analysis as long as a well reconstructed resolved tau lepton pair is present.

Next, the events are classified into three decay channels depending on how the tau lepton pair decays. To make sure that the channels are mutually exclusive, lepton vetos are introduced for each channel. The selection requirements for electrons and muons are loosened to $p_T > 15\,\text{GeV}$ and a loose ID. These object collections are used to define the vetos. For the $e\tau_h$ channel no loose muons and no loose electrons in addition to the selected one are allowed in the event, for the $\mu\tau_h$ channel no loose electrons and no loose muons in addition to the selected one are allowed in the event and for the $\tau_h\tau_h$ channel no loose electrons or loose muons are allowed.

After that, the bb pairs are selected. For an event to be selected, it requires either a good resolved b-jet pair or a good bb-tagged AK8 jet which are well separated spatially from the

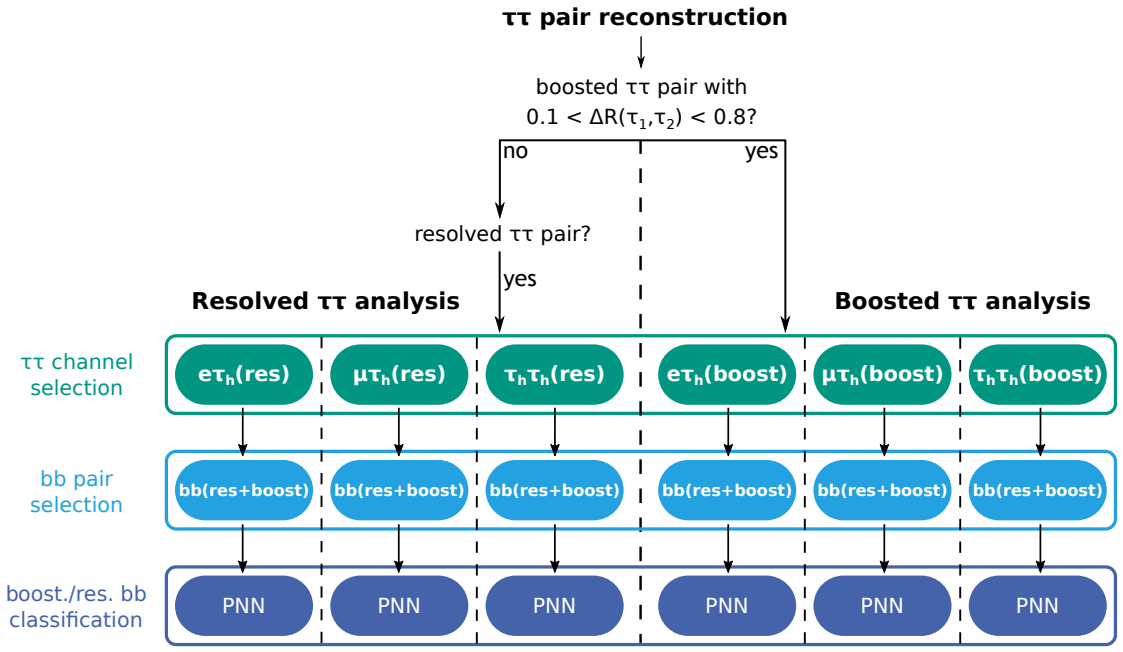


Figure 4.11: A schematic illustration of the full event selection chain. First, events are split into a boosted and resolved $\tau\tau$ analysis based on the reconstructed tau lepton pairs. Next, the events are classified into the tau lepton decay channels and the bb pair is selected. The separation between boosted and resolved bb pairs is taken care of by a parametric neural network (PNN).

selected tau lepton pair. The resolved bb reconstruction fails if at least one b-jet is closer than $\Delta R = 0.4$ to any of the tau leptons. The same happens for the boosted bb reconstruction if the bb-tagged AK8 jet is closer than $\Delta R = 0.6$ to any of the tau leptons. A larger ΔR is chosen due to the larger radius of the AK8 jet. At this point, it is allowed that both methods successfully find a good candidate for the bb pair. Information about both is passed to the classification algorithm that is used to classify events into signal and background processes. This algorithm is a parametric neural network (PNN) and is described later in section 5.1. One of its tasks is to classify signal events to have a boosted or resolved bb pair.

4.5 Background processes

The final state of the signal processes with a $\tau\tau$ and bb pair is not unique to di-Higgs events. There are other processes predicted by the SM with the same or similar final states. The data recorded at CMS mainly consist of events from background processes that are considered as irreducible backgrounds in the search for signal events. For a search it is assumed that a data excess exists on top of the background hypothesis. By comparing the background expectation with the measured data in the experiment, such an excess hypothesis can be either rejected or not rejected based on how well the data is described by the background estimation. Therefore, the backgrounds and their systematic uncertainties must be estimated as accurately as possible.

In this analysis two background estimation methods are used. One is based on simulation and the other one is a data-driven estimation method called fake factor (F_F) method to estimate the background contribution from events with jets that are falsely identified as τ_h. Both methods will be discussed in detail in section 4.7 and section 4.6, respectively. For which background processes the methods are relevant is displayed in figure 4.12. Some background processes are split because both estimation methods are used.

In the following the focus will be on describing the physics of the background processes and their significance for the analysis. The background processes considered in this analysis are

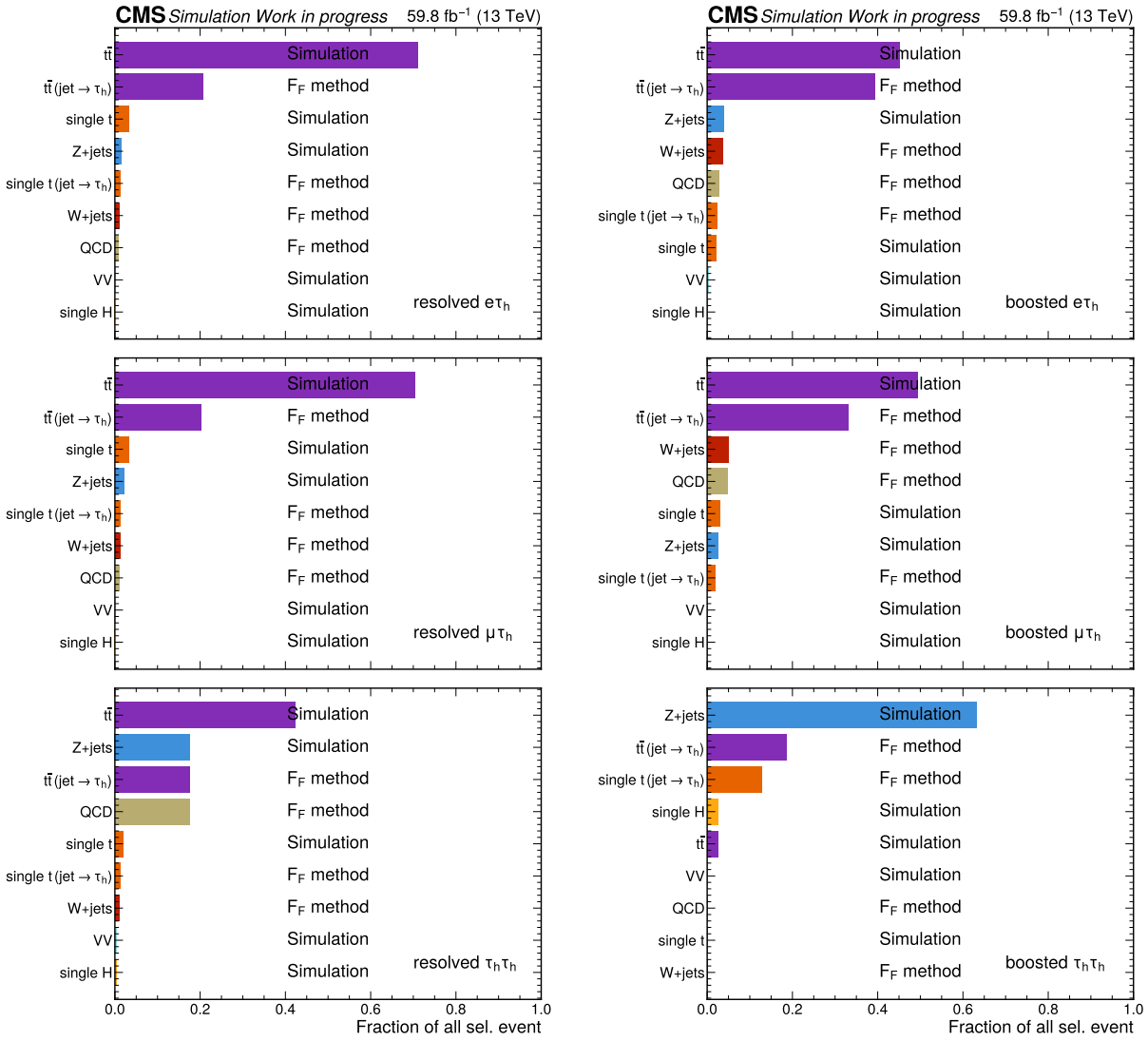
- Top quark pair production ($t\bar{t}$),
- Z boson production in association with jets (Z+jets),
- Single top quark production (single t),
- Quantum chromodynamics multijet production (QCD),
- W boson production in association with jets (W+jet),
- Diboson (WW, WZ, ZZ) production (VV),
- Single Higgs boson production (single H_{SM}).

The order of the list roughly represents the importance of the background process for this analysis. The importance changes depending on the analysis channel and is summarized in figure 4.12. The fractions for all process contributions are calculated with simulated events.

For the semi-leptonic channels $e\tau_h$ and $\mu\tau_h$ the background composition is very similar for the boosted and resolved $\tau\tau$ analyses. The $t\bar{t}$ production has the highest contribution of over 95% (90%) for the resolved (boosted) $\tau\tau$ analysis. The main reason is that besides leptonically decaying tau leptons also prompt electrons or muons can be produced in a $t\bar{t}$ decay and contaminate the selection of tau lepton pairs. Each other process contributes less than 5% to the overall yield of selected events.

For the fully hadronic $\tau_h\tau_h$ channels the background composition is different. In the resolved $\tau_h\tau_h$ channel $t\bar{t}$ still has the highest contribution but is reduced to 60% and Z+jets and QCD multijet production gain importance. In this channel, $t\bar{t}$ can be better suppressed due to the lepton (e, μ) veto. On the other hand, this veto has a smaller effect on Z+jets events due to the higher branching fraction of $Z \to \tau_h\tau_h$ (1.4%) compared to $t\bar{t} \to \tau_h\tau_h$ (0.76%) while for the semi-leptonic channels it is 1.6% versus 3.1%, respectively. The lepton veto also reduces the suppression of QCD multijet events because τ_h are easier to confuse with jets than electrons or muons.

In the boosted $\tau_h\tau_h$ channel, Z+jets production is the dominant background due to the restriction on the ΔR distance between the two τ_h to be smaller than 0.8. The Z boson is a heavy resonance which means that the tau lepton pair can have a significant boost. The single Higgs boson contributions become also more important for the same reason. Contrary to that,



(a) Resolved $\tau\tau$ analysis

(b) Boosted $\tau\tau$ analysis

Figure 4.12: Composition of background processes after the event selection for the (a) three resolved $\tau\tau$ channels and the (b) three boosted $\tau\tau$ channels. The processes are ordered by their importance for each corresponding individual analysis channel. Additionally, in the row of the each process the method is mentioned that is used to estimate this background process.

the top quarks in the $t\bar{t}$ system are produced more likely back-to-back and the tau leptons from the $t\bar{t}$ decay go in different directions. The same reasoning can be applied to QCD multijet events where the jets do not have a preferred direction and the events are suppressed by the ΔR requirement.

4.5.1 Top quark pair production

The $t\bar{t}$ production has an inclusive cross section of about 830 pb for pp collisions at $\sqrt{s} = 13$ TeV at the LHC [78]. The leading order (LO) Feynman diagrams are shown in figure 4.13.

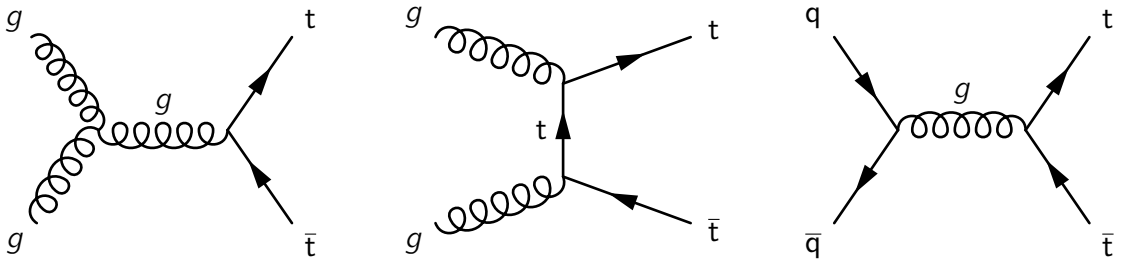


Figure 4.13: Leading order Feynman diagrams for top quark pair production. The main production channels of $t\bar{t}$ in a pp collision is via gluon fusion (left), t-channel (middle) or by quark-antiquark annihilation (right).

There are three different channels through which a $t\bar{t}$ pair can decay. A top quark almost always decays into a b quark and a W boson, other decays are CKM-suppressed. The W boson in turn can decay either into a quark-antiquark pair or a lepton and a neutrino. This leads to the three possible final state combinations which are listed in table 4.7.

Table 4.7: List of $t\bar{t}$ final states and the corresponding branching fraction calculated from [56]. The final state depends on the decay of the two W bosons into a pair of quarks ($q\bar{q}$), a lepton (e, μ) and a neutrino (ν_l) or a tau lepton (τ) and a neutrino (ν_τ).

$t\bar{t}$ channel	Final state	Branching fraction
fully hadronic	$b\bar{b} + q\bar{q} + q\bar{q}$	45.5%
semi-leptonic	$b\bar{b} + q\bar{q} + l\nu_l$	29.3%
	$b\bar{b} + q\bar{q} + \tau\nu_\tau$	14.6%
fully leptonic	$b\bar{b} + l\nu + l\nu_l$	5.3%
	$b\bar{b} + l\nu_l + \tau\nu_\tau$	3.5%
	$b\bar{b} + \tau\nu_\tau + \tau\nu_\tau$	1.8%

The semi-leptonic channels $e\tau_h$ and $\mu\tau_h$ of this analysis are highly enriched by $t\bar{t}$ events in which electrons or muons can be produced from a leptonic tau lepton decay but also promptly from the W boson decay. The possible final states of $t\bar{t}$ have a significant overlap or are even the same as the signal final state which makes it very difficult to distinguish them. Additional smaller contributions to the contaminating $t\bar{t}$ events enter the selection through jets misidentified as τ_h from one of the final state quarks of the semi-leptonic $t\bar{t}$ channel. Although this contribution is significantly suppressed by the τ_h identification algorithm at the chosen WP against jets, it still needs to be considered due to the large cross section of $t\bar{t}$ production.

For the fully hadronic $\tau_h\tau_h$ channel, the $t\bar{t}$ contribution can be better suppressed due to the lepton (e, μ) veto or the ΔR requirement for boosted $\tau_h\tau_h$ pairs. Also the contribution from misidentified τ_h can be better suppressed due to now two required τ_h. The mainly selected $t\bar{t}$ events decay into real τ_h, with a branching fraction of only 0.76% but combined with the $t\bar{t}$ cross section it results in a non-negligible background contribution.

Although the final state of the signal processes is similar to or the same as the $t\bar{t}$ final state, there are kinematic differences in the processes. These differences can be exploited to distinguish between signal and $t\bar{t}$ events. The tau lepton pair in the signal process is produced from a resonant decay of the Y or H_{SM} boson. On the other hand, in the $t\bar{t}$ process the tau leptons are produced from different top quarks and are not directly related to each other. A kinematic fit, which is introduced in more detail in section 4.8.1, of the tau lepton pair to a resonant mass can help separate signal from $t\bar{t}$ events.

4.5.2 Z boson production in association with jets

The Z boson production in association with jets has a cross section of about $2000\,\text{pb}$ for pp collisions at $\sqrt{s} = 13\,\text{TeV}$ at the LHC. The LO Feynman diagrams are shown in figure 4.14.

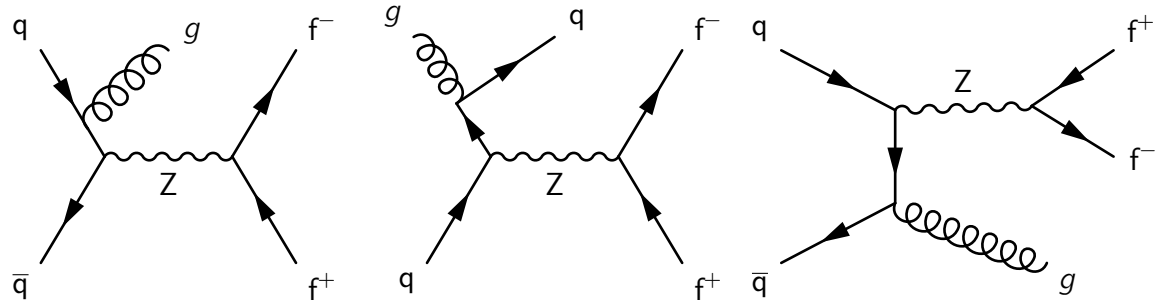


Figure 4.14: Feynman diagrams for Z boson production in association with jets. The main production channel is through quark-antiquark annihilation and a following decay of the Z boson into a pair of fermions. One or more additional jets can be present in an event due to gluon or quark radiation during the hard process.

The Z boson production as a background is more strongly suppressed than the $t\bar{t}$ background because, at leading order, the Z boson can only decay into a pair of b quarks or a pair of tau leptons and the event selection requires both. Since the cross section for the Z+jets process is even higher than for the $t\bar{t}$ process, it still has a significant contribution to the background. An even smaller background contribution from $Z \to ee/\mu\mu$ decays occurs if electrons or muons are misidentified as τ_h. Such events are suppressed by the τ_h identification algorithms against electrons/muons at the chosen WPs.

In Z+jets events that enter the analysis the tau lepton pair is produced from the Z boson decay and the two b-tagged jets or the one bb-tagged AK8 jet are either identified from real radiated b quarks or jets from other quarks or gluons misidentified as b-jets.

The Z boson is a resonance like the Y and H_{SM} bosons and has a mass of about $91\,\text{GeV}$. In a kinematic fit the reconstructed masses for these bosons can be close to each other making

it difficult to distinguish between signal events and the Z boson background. On the other hand, the reconstruction of the invariant mass of the bb pair can help to separate the Z boson background from the signal events because the b-tagged jets in Z+jets events are not produced from a resonance.

4.5.3 Single top quark production

The single top quark production has a summed cross section of about 290 pb for pp collisions at $\sqrt{s} = 13$ TeV at the LHC [79, 80]. The LO Feynman diagrams for the two main production channels (t-channel and tW-channel) are shown in figure 4.15.

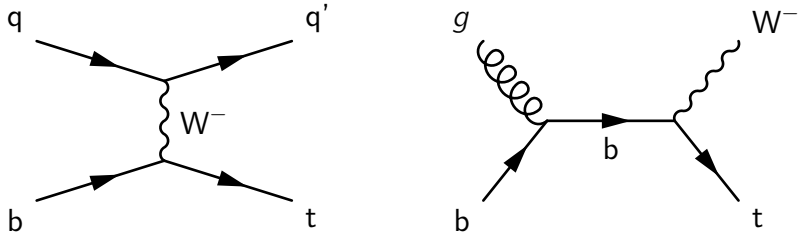


Figure 4.15: Leading order Feynman diagrams for single top quark production. On the left is the t-channel production where a top quark is produced via an W boson exchange between a b quark and another quark. On the right a top quark is produced in association with a W boson.

For this analysis, the single top quark production is a background similar to $t\bar{t}$ production but easier to suppress because the final state has fewer particles and the cross section is smaller. Depending on the decay of the involved W bosons, the possible final states at LO are $q + b + q\bar{q}$ or $q + b + l\nu$ for the t-channel and $b + l\nu + l\nu$, $b + l\nu + q\bar{q}$ or $b + q\bar{q} + q\bar{q}$ for the tW-channel. None of these final states have a pair of b quarks and a pair of tau leptons which means the single top background can be suppressed by the event selection applied in this analysis. However, single top quark events can still enter the analysis due to misidentification of jets as b-jets or as bb-jets. Similar to the $t\bar{t}$ background, an additional way to distinguish single top quark events from the signal events is provided by a kinematic fit.

There is a third possible production channel of single top quarks, the s-channel. This channel has a much smaller cross section compared to the other channels and is therefore not considered in this analysis.

4.5.4 QCD multijet production

The QCD multijet production has a cross section multiple orders larger than any of the background processes introduced in this chapter. During a pp collision many different kinds of scattering processes are happening where only the strong force is involved in the decay. A possible Feynman diagram of such processes is shown in figure 4.16. The produced quarks and gluons form hadrons and are later reconstructed as jets. In general, QCD multijet production is strongly suppressed by the event selection. QCD events only enter the analysis if jets are

misidentified as b-jets, τ_h or even electrons or muons. The probability of misidentification is low, especially for high p_T electrons or muons, but due to the very large cross section, QCD multijet production has to be considered in this analysis. QCD is mainly relevant for the resolved $\tau_h\tau_h$ channel where no electrons or muons are involved in the event selection. Additionally, it has a considerable contribution in the boosted $e\tau_h$ and $\mu\tau_h$ channels that can be explained by the relatively loose boosted tau identification WPs.

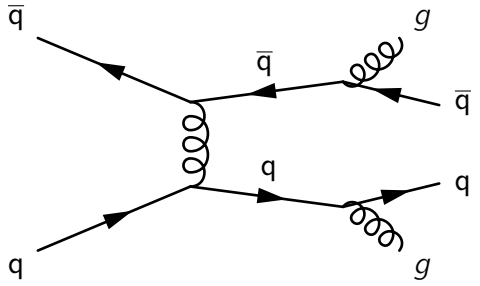


Figure 4.16: One possible Feynman diagram for QCD multijet production. Quarks and gluons can interact with each other only via the strong force producing final states with multiple jets.

Simulation of the QCD multijet production is computationally expensive and additionally the process is quite difficult to model. Therefore, in this analysis, QCD multijet production is fully estimated with the F_F method discussed in section 4.6.

4.5.5 W boson production in association with jets

The W boson production in association with jets has a cross section of about ten times higher than Z+jets production but is in general quite similar to the Z+jets process. Leading order Feynman diagrams for W+jets production are shown in figure 4.17.

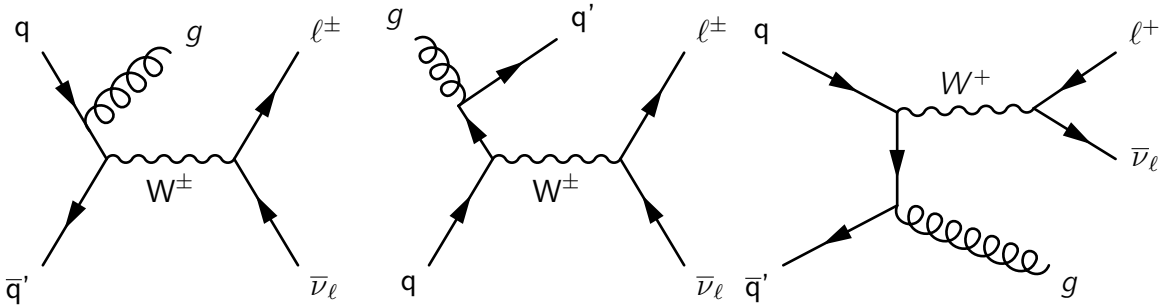


Figure 4.17: Feynman diagrams for W boson production in association with jets. The main production channel is through quark-antiquark annihilation and a following decay of the W boson into a lepton and the corresponding neutrino. One or more additional jets can be present in an event due to gluon or quark radiation during the hard process.

The only final state of the W boson relevant for this analysis is the decay into a lepton and its corresponding neutrino. Since only one lepton is present in such events, this process is

suppressed by the event selection. W+jets events only enter the analysis if more than one additional jet is misidentified as either a τ_h or a b-jet. Due to the large cross section, W boson production is still a considerable background process for the semi-leptonic $e\tau_h$ and $\mu\tau_h$ channels.

As explained, W+jets events mainly enter the analysis due to misidentification of jets as τ_h. Therefore this process is fully estimated with the F_F method (see section 4.6).

4.5.6 Diboson production

The production of two vector bosons has a relatively small cross section of a few pb and is regarded as a minor background for this analysis. Some LO Feynman diagrams for diboson production are shown in figure 4.18.

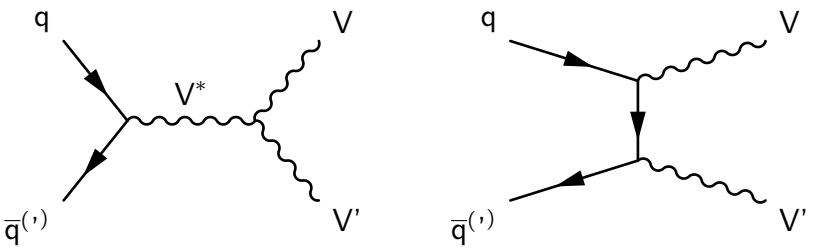


Figure 4.18: Leading order Feynman diagrams for diboson production. On the left the two vector bosons are produced via the s-channel and on the right via the t-channel.

Possible vector boson combinations are a pair of W bosons, a pair of Z bosons and a W and Z boson together. Diboson events can even have the same final state as the signal processes, for example, if a Z boson pair is produced and one Z boson decays into a pair of tau leptons and the other one into a pair of b quarks. However, as already mentioned the small cross section makes diboson production only a minor background contribution for any of the analysis channels.

4.5.7 Single Higgs boson production

Like diboson production, single Higgs production is one of the minor backgrounds in this analysis. The main production channels of Higgs bosons are gluon fusion, vector boson fusion and Higgs boson radiation. Feynman diagrams for all of them are shown in figure 4.19.

Gluon fusion is the production channel with the highest cross section of a few tens of pb. The other production channels have ten times smaller cross sections, not including the branching fraction for the $H_{SM} \to \tau\tau$ decay relevant for this analysis. There are more possible Higgs boson production channels, like being radiated from one of the top quarks in $t\bar{t}$ production ($t\bar{t}$ +H) but this channel has an even smaller cross section and is therefore not considered.

Only in the boosted $\tau_h\tau_h$ channel, single Higgs boson production is a relevant background contribution (see figure 4.12b). The reason is the same as already mentioned for the Z+jets production. The tau lepton pair is often boosted in one direction due to the massive initial particle like the Higgs boson. Further, especially Z+H production contributes the most because it can have the same final state as the signal processes.

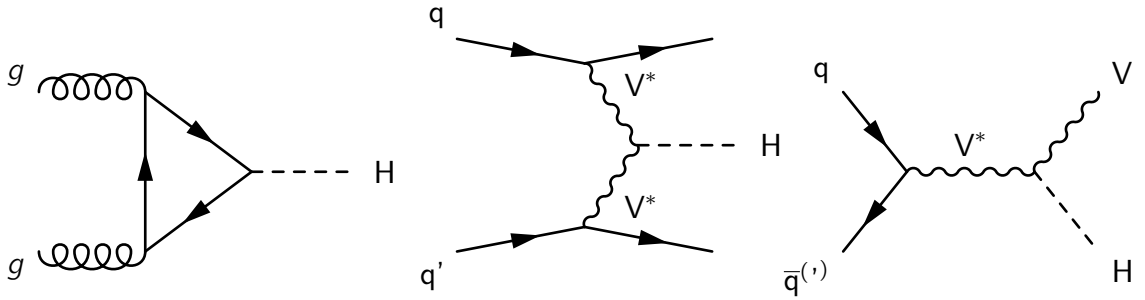


Figure 4.19: Feynman diagrams for single Higgs boson production. From left to right, the gluon fusion, vector boson fusion and Higgs radiation production channels are displayed.

4.6 Estimation of events with jets faking hadronic tau leptons

The algorithms used to identify hadronically decaying tau leptons from reconstructed jets were introduced in sections 3.4.3 and 3.4.4. However, analyses are still challenged by the misidentification of jets originating from light flavor quarks or gluons as hadronic tau leptons. The reason is not necessarily a bad misidentification rate of the algorithms but more the very high production rate of gluon and light quark induced jets at the LHC.

For this analysis the main background contributions from τ_h fakes (jets misidentified as τ_h) are from $t\bar{t}$, single t, QCD and W+jets events, which can be identified from figure 4.12. These contributions can be estimated using a data-driven method, called fake factor (F_F) method. This approach was already successfully used in the measurements of the $Z \to \tau\tau$ [81] and $H_{SM} \to \tau\tau$ [82] production cross sections as well as in the previous $X \to Y(bb)H_{SM}(\tau\tau)$ search [9].

The fundamental idea of the F_F method is illustrated in figure 4.20. Four regions are defined based on two different variables. For each variable a value is chosen that is used to orthogonally separate the regions. For the F_F method one of the variables is always the τ_h identification discriminant against jets and separates the signal regions (SR and SR-like) enriched with correctly identified τ_h candidates from the orthogonal application regions (AR and AR-like) enriched with τ_h candidates enriched with τ_h fakes from jets. The second variable is chosen to define a sideband determination region (DR) where the targeted background process is enriched. The details of which variables are used will be discussed in the following sections.

Events are classified into one of the four regions based on the reconstructed τ_h and the process dependent variable. Next, a very important assumption is made that the ratio of the event numbers between the two determination regions (SR-like and AR-like) is the same as the event ratio between the signal and application region and therefore

$$N_{SR} = N_{AR} \cdot \frac{N_{SR\text{-like}}}{N_{AR\text{-like}}} = N_{AR} \cdot F_F \tag{4.3}$$

holds. With this equation, the event number in the SR can be estimated from the calculated fake factor (F_F) applied to the event yield in the AR. For the full F_F method this approach is

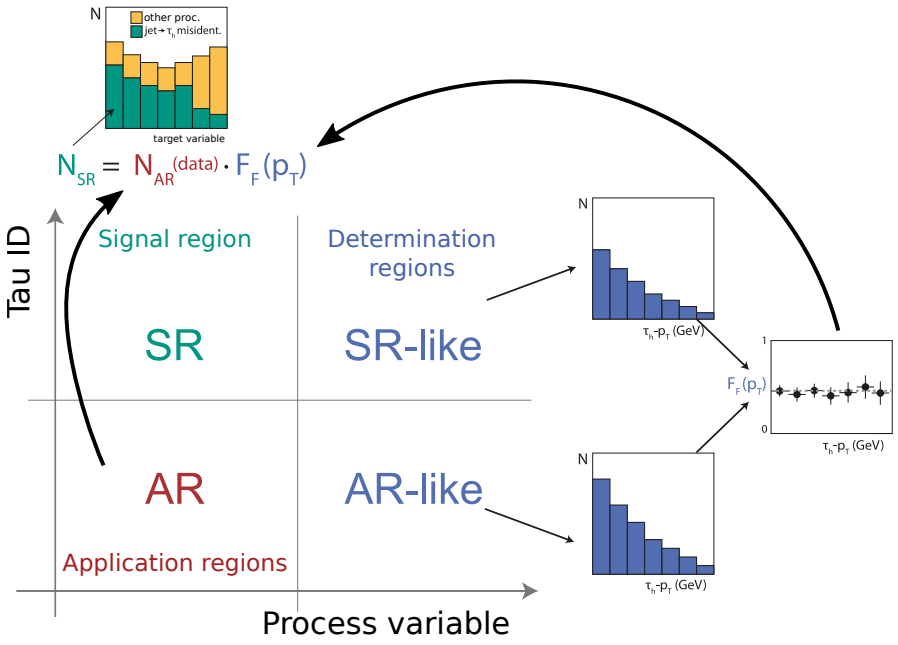


Figure 4.20: Sketch of the sideband approach used for the F_F method.

used multiple times for each relevant background process, which leads to the second assumption that needs to be made. The F_F's are applied in the AR to data and extrapolated to the SR. Additionally, the process composition in data is unknown, therefore, the fractions of the relevant processes are estimated from simulation. The process dependent F_F's are then only applied to a fraction of the data events in the AR,

$$N_{SR} = \sum_{proc} \text{frac}^{proc} \cdot N_{AR} \cdot F_F^{proc}. \tag{4.4}$$

The F_F method is specifically adapted for the needs of this analysis and will be described in a less abstract way in the following two sections for the resolved and boosted $\tau\tau$ analyses, respectively. An independent calculation is needed for both parts of the analysis because they depend on different τ_h identification algorithms.

4.6.1 Resolved $\tau\tau$ analysis

In the resolved $\tau\tau$ analysis the *DeepTau* algorithm is used to identify τ_h (see section 3.4.3). To be classified into SR or SR-like, events have to pass the medium WP and for AR and AR-like they have to fail the medium WP but still pass the vvvloose WP, which is the loosest *DeepTau* WP.

For the $e\tau_h$ and $\mu\tau_h$ channel only one τ_h is present and the F_F's are measured dependent on the p_T of this τ_h and on the number of jets in an event. On the other hand, two τ_h are present in the $\tau_h\tau_h$ channel. Therefore, for each τ_h an individual F_F is measured and only the corresponding τ_h has to fail the medium WP for the AR definition, the other τ_h has to pass the

WP. The F_F's in the $\tau_h\tau_h$ channel are measured dependent on the p_T of the respective τ_h and, like for the semi-leptonic channels, dependent on the number of jets.

The F_F's are measured for the relevant background processes split into three process categories, a QCD measurement, a W+jets measurement and a $t\bar{t}$ measurement. The single top contribution is considered in the $t\bar{t}$ measurement due to their large similarity. The measurements are explained in the following sections.

The fraction measurement is performed in the AR based on simulated events. The results for all three resolved channels are shown in figure 4.21. For all channels the measurement is split into two regions, for events with one or no b-tagged jet and for two or more b-tagged jets. The first category targets events with a bb-tagged AK8 jet because often such a jet can partially be reconstructed as a b-tagged AK4 jet. The second category targets the resolved bb pair reconstruction. Further, for the $e\tau_h$ and $\mu\tau_h$ channels the process fractions are measured in bins of the transverse mass (m_T) of the electron or muon and missing transverse momentum ($\vec{p}_T^{\,miss}$ or $\not{E}_T$) system, which is chosen due to its discriminating power between QCD and W+jets events. The transverse mass is an approximation of the W boson mass where the missing transverse momentum takes the place of the neutrino. In the $\tau_h\tau_h$ channel, on the other hand, the fractions are measured in bins of the invariant mass of the two τ_h due to its discriminating power between QCD and $t\bar{t}$ events. The process fractions are also measured twice for the two different F_F measurements of the $\tau_h\tau_h$ channel. They are very similar, therefore, only one of the measurements is shown in figure 4.21.

Due to the low fraction of the W+jets process in the $\tau_h\tau_h$ channel, it does not have any significant impact, so no individual F_F's are measured for this process in this channel and the QCD F_F's are applied instead.

The resulting event F_F is calculated by evaluating the measured process F_F's and the process fractions on an event-by-event basis as

$$\begin{aligned} F_F^{event} = {} & \text{frac}^{QCD}(m_T, N_{\text{b-jets}}) \cdot F_F^{QCD}(p_T, N_{jets}) \\ & + \text{frac}^{W+jets}(m_T, N_{\text{b-jets}}) \cdot F_F^{W+jets}(p_T, N_{jets}) \\ & + \text{frac}^{t\bar{t}}(m_T, N_{\text{b-jets}}) \cdot F_F^{t\bar{t}}(p_T, N_{jets}) \end{aligned} \tag{4.5}$$

for the semi-leptonic channels and for the $\tau_h\tau_h$ channel as

$$\begin{aligned} F_{F,AR}^{event} = {} & \left(\text{frac}^{QCD}(m_{vis}, N_{\text{b-jets}}) + \text{frac}^{W+jets}(m_{vis}, N_{\text{b-jets}})\right) \cdot F_F^{QCD}(p_T, N_{jets}) \\ & + \text{frac}^{t\bar{t}}(m_{vis}, N_{\text{b-jets}}) \cdot F_F^{t\bar{t}}(p_T, N_{jets}). \end{aligned} \tag{4.6}$$

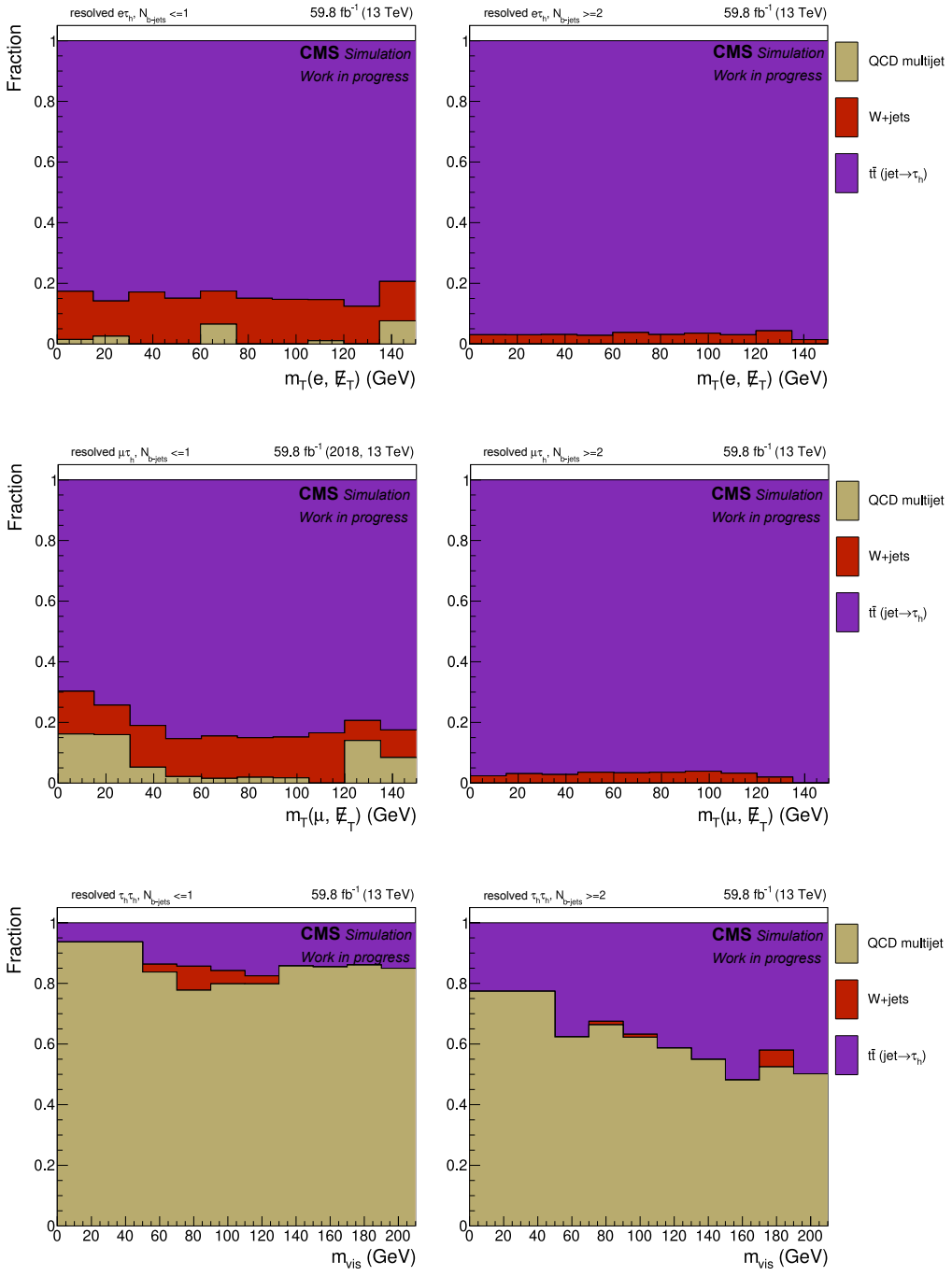


Figure 4.21: Process fraction measurements for the resolved $e\tau_h$ (top), $\mu\tau_h$ (middle) and $\tau_h\tau_h$ (bottom) channel. On the left side are the measurements of the fractions of QCD, W+jets and $t\bar{t}$ for events with one or no b-tagged jet, on the right side for events with two or more b-tagged jets. The sum of the process fractions is always one for each measured bin. The $t\bar{t}$ fraction also includes single top events. The fractions for the $e\tau_h$ and $\mu\tau_h$ channel are measured dependent on the transverse mass m_T of the lepton (e or μ) and the missing transverse momentum. In the $\tau_h\tau_h$ channel the measurement depends on the visible invariant mass of the two τ_h.

Since the $\tau_h\tau_h$ channel has two F_F measurements for two different AR definitions but only one SR, they have to be combined to avoid double contribution. The combined F_F is calculated as

$$F_F^{comb} = \frac{1}{2} \cdot (F_{F,AR_1}^{event} + F_{F,AR_2}^{event}). \tag{4.7}$$

After the F_F^{event}'s are applied as weights to data events in the AR using equation 4.3, there is still an overestimation expected in the SR because of other process contributions in the application region besides τ_h fakes. For example, events with real tau leptons, where the τ_h is not identified correctly, enter the AR. To avoid this effect, the simulated event yields of these processes are therefore subtracted from the data in the AR.

The weighted data events can be binned in any required variable because the F_F's are calculated on an event-by-event basis. The main application will be on the final discriminants (see chapter 5) of this analysis.

Estimation for QCD multijets

As can be seen from figure 4.21, the QCD F_F's are mainly relevant for the $\tau_h\tau_h$ channel. Their contribution in the $e\tau_h$ and $\mu\tau_h$ channels is only visible for the $N_{b\text{-jets}} \leq 1$ category. Nevertheless, F_F^{QCD} is calculated for all three channels.

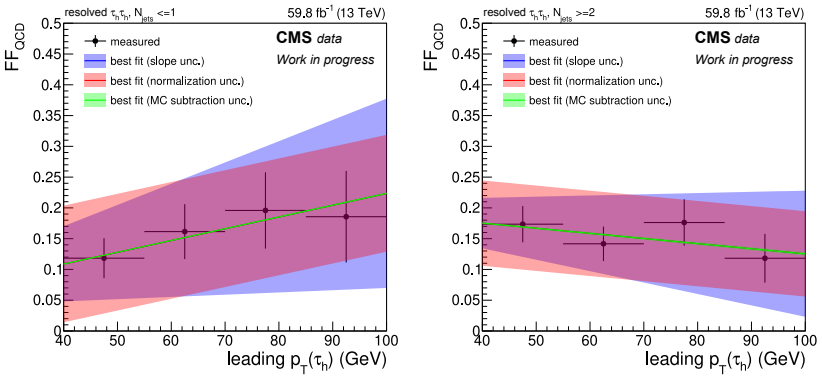
To define the DR for the F_F measurement, the charges of the two tau leptons are used. For the signal region the tau leptons need to be oppositely charged because they decay from a neutral particle. For QCD the reconstructed tau leptons are always jets misidentified as τ_h, electron or muons. This leads to the assumption that the reconstructed tau lepton pairs in QCD events do not have a preferred charge combination, at least at leading order. The DR is thereby defined to have tau leptons with the same charges. This cut already makes the DR enriched in QCD event because other relevant processes like $t\bar{t}$, Z+jets or diboson also have oppositely charged tau leptons.

In the $e\tau_h$ and $\mu\tau_h$ channels, additional cuts are applied to further enrich the DR with QCD events. To separate QCD from W+jets events, the transverse mass $m_T(e/\mu, \not{E}_T)$ is required to be smaller than $50\,\text{GeV}$ and to further remove contributions from $t\bar{t}$ events $N_{b\text{-jets}} = 0$ is required. The F_F^{QCD}'s are measured in categories of N_{jets} where the number of jets is set to zero, one and greater than one. After the applied cuts the QCD purity in the DR is between 60% to 75% depending on the region, category and channel.

For the $\tau_h\tau_h$ channel no additional cuts are applied. The DR region is already significantly enriched in QCD because W+jets and $t\bar{t}$ events are suppressed by the electron/muon veto in this channel. For statistical reasons, the F_F^{QCD} measurement is performed in only two N_{jets} categories. The zero and one jet categories are merged into one for this channel. After applying the cuts the QCD purity in the DR is between 70% to 95% depending on the region, category and considered τ_h.

The F_F^{QCD} measurements are performed for all categories independently and binned in the p_T of the τ_h. Since there still are some small remaining contributions from other processes in

the SR-like and AR-like data distributions of the DR, they are estimated via simulation and subtracted from the data. The resulting F_F^{QCD} bins from the ratio calculation are fitted with a linear function and this function is used for the estimation as $F_F^{QCD}(p_T)$. The results of the measurements in the $\tau_h\tau_h$ channel for the leading τ_h are shown in figure 4.22. The uncertainties of the measurement are discussed in section 5.2.3.



(a) $N_{jets} \leq 1$ category (b) $N_{jets} \geq 2$ category

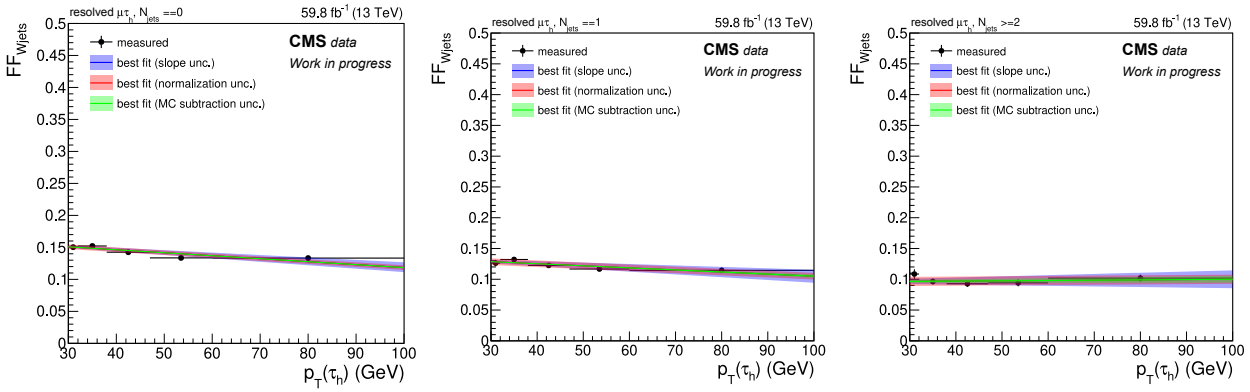
Figure 4.22: Results of the $F_F^{QCD}(p_T, N_{jets})$ measurement for the resolved $\tau_h\tau_h$ channel. The measurement is split into two N_{jets} categories in (a) and (b), and is derived in p_T bins of the τ_h. A linear fit is performed to extract the F_F functions. Two uncertainties are defined from the fit, namely a slope and a normalization uncertainty shown in blue and red, respectively. The third uncertainty (green) is related to the subtraction of (Monte Carlo) simulated background events from data.

Estimation for W+jets

The W+jets contribution is relatively small in all channels as can be seen from figure 4.21. The main contribution is in the $e\tau_h$ and $\mu\tau_h$ channels for the $N_{b\text{-jets}} \leq 1$ categories. Therefore, the F_F^{W+jets}'s are only derived for the semi-leptonic channels.

The DR for the W+jets F_F measurement is defined by inverting the bb pair selection of the signal region, which means no resolved b-jet pair and no bb-tagged AK8 jet should be present in an event. Neither of them are expected to be present in W+jets events. This cut already removes the majority of $t\bar{t}$ events. To further enrich the DR with W+jets events, a cut on the transverse mass $m_T(e/\mu, \not{E}_T)$ is applied to be higher than 70 GeV. With this the majority of QCD and Z+jets events can be separated. For Z+jets events this is the case because one of the leptons is not considered in the mass calculation and m_T is then usually smaller. The F_F^{W+jets}'s are measured in categories of N_{jets} where the number of jets is set to zero, one and greater than one. After applying the cuts, the W+jets purity in the DR is between 60% and 90%, depending on the region, category and channel.

The F_F^{W+jets} measurements are performed independently for all categories and binned in p_T of the τ_h. Like for the QCD F_F measurement, there still are some small remaining contributions from other processes in the SR-like and AR-like data distributions of the DR. These contributions are estimated via simulated events and subtracted from the data events. The resulting F_F^{W+jets} bins from the ratio calculation are fitted with a linear function, which is used for the estimation as $F_F^{W+jets}(p_T)$. The results of the measurements for the $\mu\tau_h$ channel are shown in figure 4.23. The uncertainties of the measurement are discussed in section 5.2.3.



(a) $N_{jets} = 0$ category (b) $N_{jets} = 1$ category (c) $N_{jets} \geq 2$ category

Figure 4.23: Results of the $F_F^{W+jets}(p_T, N_{jets})$ measurement for the resolved $\mu\tau_h$ channel. The measurement is split into three N_{jets} categories in (a-c), and is derived in p_T bins of the τ_h. A linear fit is performed to extract the F_F functions. Two uncertainties are defined from the fit, namely a slope and a normalization uncertainty shown in blue and red, respectively. The third uncertainty (green) is related to the subtraction of (Monte-Carlo) simulated background events from data.

Estimation for $t\bar{t}$

The main background contribution for this analysis comes from $t\bar{t}$ events. This is also the case for the F_F method, as can be seen in figure 4.21. Only in the $\tau_h\tau_h$ channel is the $t\bar{t}$ process not the dominant one.

The definition of the DR is more challenging for $t\bar{t}$ than for QCD or W+jets because the final state of the $t\bar{t}$ process is the same as for the signal processes and finding a $t\bar{t}$ enriched region orthogonal to the signal region is difficult. Therefore, the $F_F^{t\bar{t}}$ measurement is performed in two steps.

The first step is to calculate the $F_F^{t\bar{t}\,(sim)}$'s for the N_{jets} categories and dependent on the p_T of the τ_h from simulation. This calculation is performed directly in the AR and SR because no data is involved that could bias the measurement. The second step is to calculate an inclusive scale factor which is derived in a DR. This scale factor is used to scale the simulation based $F_F^{t\bar{t}\,(sim)}$'s to the observed data yield. The DR is defined by inverting the electron/muon veto in all of the channels, thereby allowing more leptons to be present in an event. This cut does not enrich the

selection of $t\bar{t}$ events but defines an orthogonal region where the scale factor can be calculated from data. The approach is the same as for the QCD and W+jets estimation. All process contributions besides $t\bar{t}$ with τ_h fakes are subtracted from the data yield and the $F_{F,incl}^{t\bar{t}\ (data)}$ between SR-like and AR-like is calculated inclusively without any dependence on a variable. The same is done for simulated $t\bar{t}$ events with τ_h fakes. The scale factor as the ratio between these two F_F's and is applied to the measured $F_F^{t\bar{t}\ (sim)}$'s to get the actual $F_F^{t\bar{t}}$ estimation,

$$F_F^{t\bar{t}}(p_T, N_{jets}) = F_F^{t\bar{t},sim}(p_T, N_{jets}) \cdot \frac{F_{F,incl}^{t\bar{t}\ (data)}}{F_{F,incl}^{t\bar{t}\ (sim)}}. \tag{4.8}$$

For the $e\tau_h$ and $\mu\tau_h$ channels the $F_F^{t\bar{t}\ (sim)}$'s are measured in four N_{jets} categories with one or zero, two, three and four jets in the events. This granularity is possible due to a large number of events. For the $\tau_h\tau_h$ channel only two categories are defined with two or fewer and three or more jets in the events.

For each category the resulting $F_F^{t\bar{t}}$ bins are fitted with a linear function, which is used for the estimation as $F_F^{t\bar{t}}(p_T, N_{jets})$. The results of the measurements for the $e\tau_h$ channel are shown in figure 4.24. The uncertainties on the measurement are discussed in section 5.2.3.

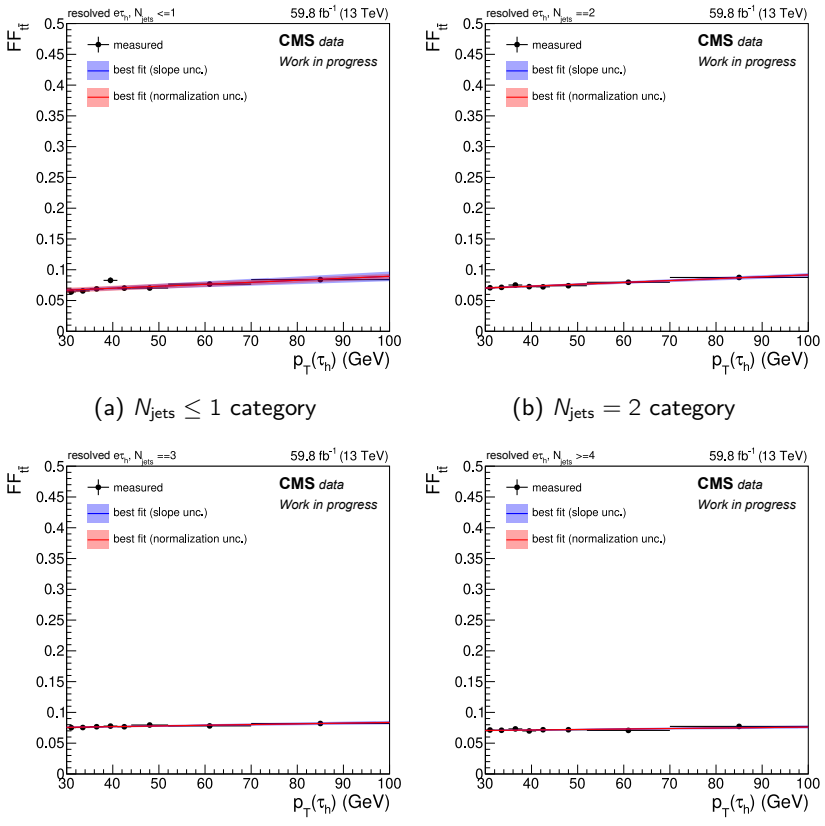
Corrections for the F_F measurements

For the F_F measurements two assumptions are made. The first assumption is that events can directly be extrapolated from the AR to the SR. Further, it is assumed that the F_F's measured in the DR can directly be extrapolated and applied in the SR. Both of these assumptions are not fully correct and additional correction factors to each F_F^{proc} need to be derived to account for it.

First, the F_F's are measured only dependent on the number of jets and the p_T of the τ_h. Other kinematic variables are not considered although they can have an effect on the extrapolation from the AR to the SR. One such variable is the p_T of the electron or muon in the $e\tau_h$ and $\mu\tau_h$ channels and the other τ_h in the $\tau_h\tau_h$ channel. The p_T of these leptons influences, for example, the triggers and have characteristic differences in the distribution of the QCD, W+jets and $t\bar{t}$ processes. Therefore, an inclusive non-closure correction is derived for this variable in the DR for all channels. For the non-closure correction the measured F_F^{proc}'s are applied in the application-like region and the extrapolation is compared to the data distribution in the signal-like region. The ratio of these two distributions is used as correction C for the F_F^{proc}'s when it is applied later in the analysis,

$$F_{F,corr}^{proc}(p_T^{\tau_h}, N_{jets}, p_T^{e/\mu/other\ \tau_h}) = F_F^{proc}(p_T^{\tau_h}, N_{jets}) \cdot C(p_T^{e/\mu/other\ \tau_h}). \tag{4.9}$$

A second non-closure correction is derived, with the first correction already applied, in the same way for all channels dependent on the mass of the τ_h. Studies showed that the F_F method has biases in the decay modes of the τ_h. To account for the decay modes directly is problematic



(a) $N_{\text{jets}} \leq 1$ category (b) $N_{\text{jets}} = 2$ category

(c) $N_{\text{jets}} = 3$ category (d) $N_{\text{jets}} \geq 4$ category

Figure 4.24: Results of the $F_F^{t\bar{t}}(p_T, N_{\text{jets}})$ measurement for the resolved $e\tau_h$ channel. The measurement is split into four N_{jets} categories in (a-d), and is derived in p_T bins of the τ_h. A linear fit is performed to extract the F_F functions. Two uncertainties are defined from the fit, namely a slope and a normalization uncertainty shown in blue and red, respectively.

because a second categorization next to N_{jet} would lead to very low event numbers in some of the categories. Therefore, the mass of the τ_h is used to derive a non-closure correction due to its high correlation with the decay modes.

Both non-closure corrections are shown in figure 4.25 for QCD in the $e\tau_h$ channel, for W+jets in the $\mu\tau_h$ channel and for $t\bar{t}$ in the $\tau_h\tau_h$ channel. The deviations of the correction values from one show the bias of the initially measured F_F's that only depend on the number of jets and the p_T of the τ_h. To better describe the extrapolation from the AR to the SR, these non-closure corrections are essential. Ideally, even more non-closure corrections are needed until the correction values get closer to one. However, this also increases the computational effort and is not included in the scope of this analysis.

For QCD an additional correction is derived to correct the extrapolation from the DR to the SR, namely the difference between QCD events with same-charge and oppositely charged tau

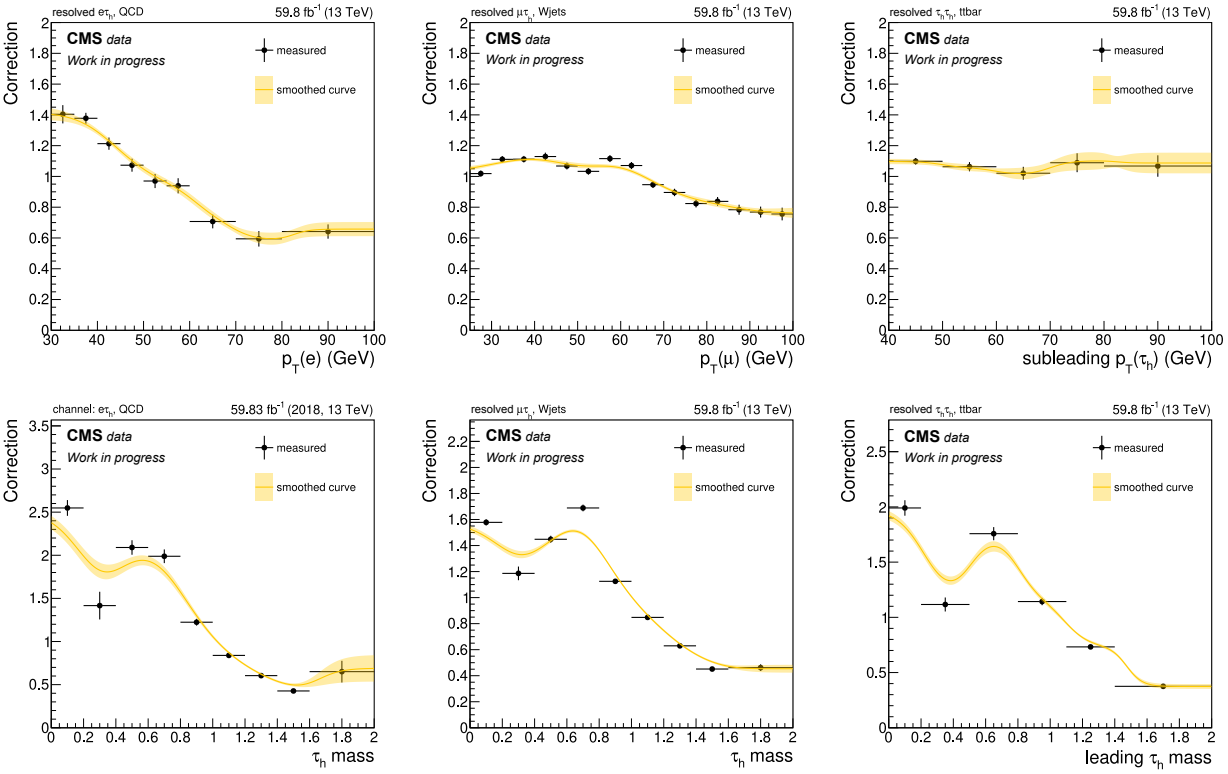


Figure 4.25: Results of the non-closure correction measurements dependent on the p_T of the electron, muon and second τ_h as well as the mass of the τ_h for the three resolved $\tau\tau$ channels. In the left column corrections applied to F_F^{QCD} are shown, in the middle column to F_F^{W+jets} and in the right column to $F_F^{t\bar{t}}$. Besides the histogrammed corrections also the smoothed functions with uncertainty bands are shown.

lepton pairs. In the semi-leptonic channels this is realized by defining four new regions as SR, AR, SR-like and AR-like. The new regions have the same cuts applied besides one, the isolation cut on the electron/muon. The isolation is inverted to select non-isolated electrons/muons with $Iso_{rel} \in [0.15, 0.3]$. In these four regions new F_F^{QCD}'s are measured. These F_F^{QCD}'s are then applied in the new DR and compared to estimation in the new SR. This correction is measured dependent on the visible invariant di-tau mass m_{vis}. In the $\tau_h\tau_h$ channel no electrons/muons are present, therefore a new DR is exploited where both τ_h fail the medium WP of *DeepTau* to compare same and opposite charged tau lepton pairs. Also in this channel the correction is measured dependent on m_{vis}.

For W+jets a similar correction is derived to correct the DR to SR extrapolation. The W+jets DR is influenced by the high transverse mass ($m_T^{e/\mu}$) cut and therefore the extrapolation into the SR can be biased. In case of W+jets, the extrapolation correction is derived from simulation. New F_F^{W+jets}'s are calculated where the cut on $m_T^{e/\mu}$ is dropped and after applying these new F_F^{W+jets}'s in the DR they are compared to the simulated W+jets events in the SR. Also here the correction is measured dependent on m_{vis}. Three DR to SR extrapolation corrections are

shown in figure 4.26, two for QCD in the $e\tau_h$ and $\tau_h\tau_h$ channels and one for W+jets in the $\mu\tau_h$ channel.

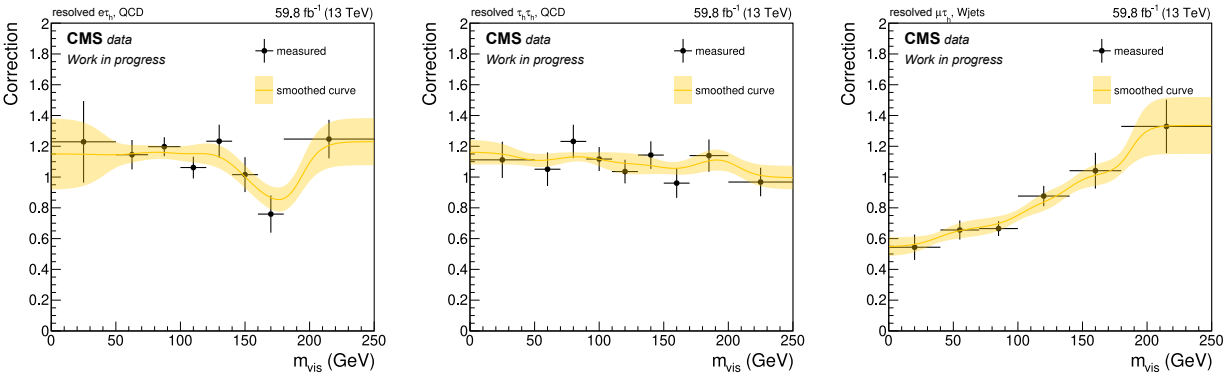


Figure 4.26: Results of the DR to SR extrapolation correction measurements dependent on the visible invariant di-tau mass m_{vis} for the three resolved $\tau\tau$ channels. The two left corrections are applied to F_F^{QCD} in the $e\tau_h$ and $\tau_h\tau_h$ channel, respectively. The right correction is applied to F_F^{W+jets} in the $\mu\tau_h$ channel. Besides the histogrammed corrections also the smoothed functions with uncertainty bands are shown.

All correction are derived as histograms which then are smoothed with a Gaussian kernel to get a smoothed correction function with a corresponding uncertainty. Such an approach is chosen to reduce effects of statistical fluctuations in the histograms. The uncertainties are discussed in section 5.2.3.

4.6.2 Boosted $\tau\tau$ analysis

In the boosted $\tau\tau$ analysis, a BDT based algorithm is used to identify τ_h (see section 3.4.4). The SR and SR-like are defined for events with τ_h candidates passing the loose WP and for the AR and AR-like, the events have τ_h candidates that fail the loose WP. The technical setup for the F_F^{proc} measurements and the construction of the final F_F^{event} is the same as described for the resolved $\tau\tau$ analysis. Therefore, this section will focus on the differences related to the definition of the SR, AR and DRs and the dependencies on variables like N_{jets}. In general, requiring $\Delta R(\tau_1, \tau_2) < 0.8$ significantly reduces the number of selected events, therefore, some statistical issues occur during the F_F measurements and are mitigated by reducing the number of N_{jets} categories for which the measurement is performed.

The process fraction measurement is performed in the AR based on simulated events after the boosted $\tau\tau$ event selection (see section 4.3.4). The results of the measurement can be found in figure 4.27 for the three boosted $\tau\tau$ channels. As already seen for the resolved $\tau\tau$ F_F fractions, the main contribution is from the $t\bar{t}$ (with single top) process.

Estimation for QCD multijets

For the measurement of the F_F^{QCD}'s the same approach is chosen as described for the resolved $\tau\tau$ analysis. The DR is defined by the charges of the two selected tau leptons which need to

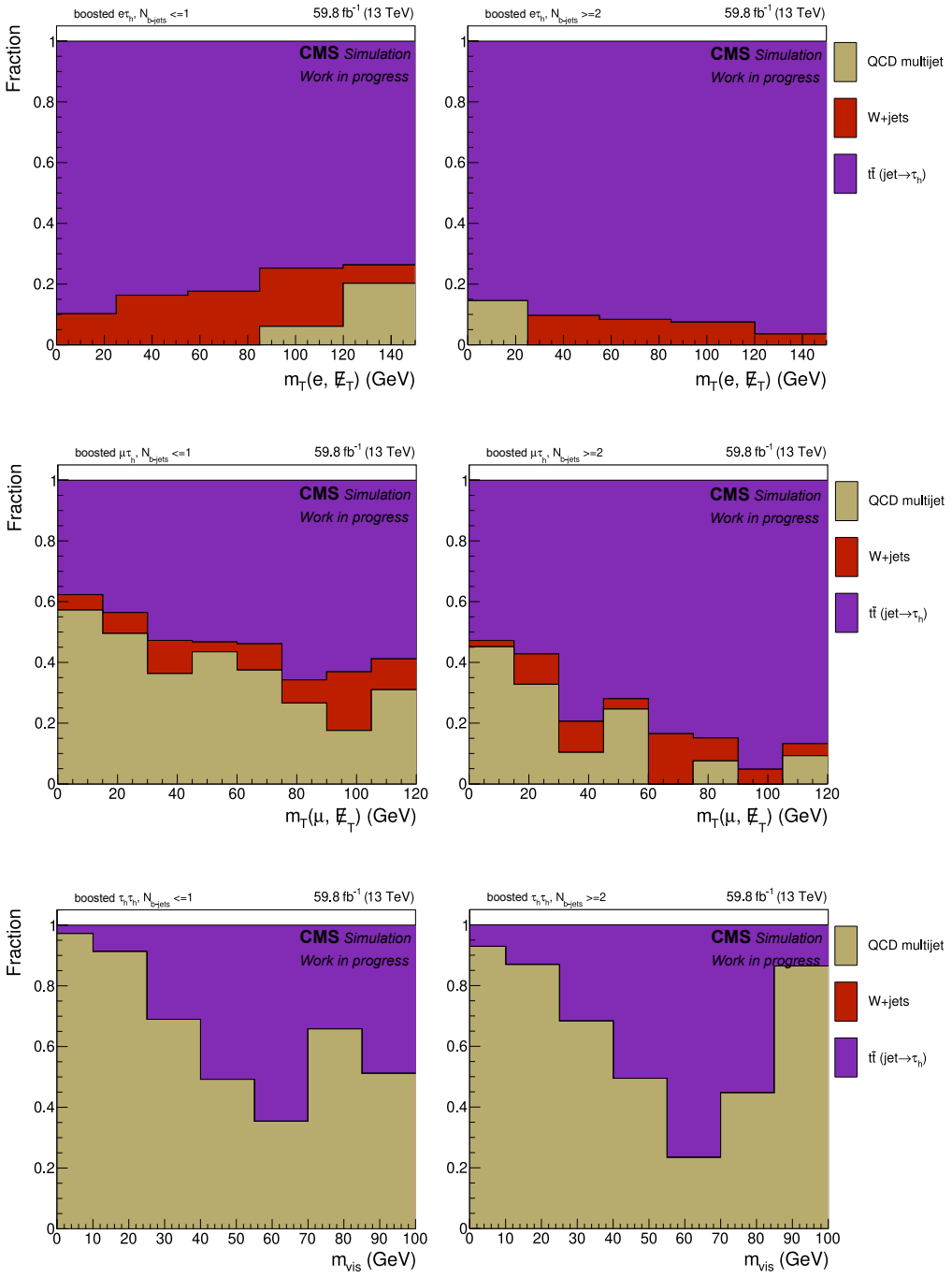
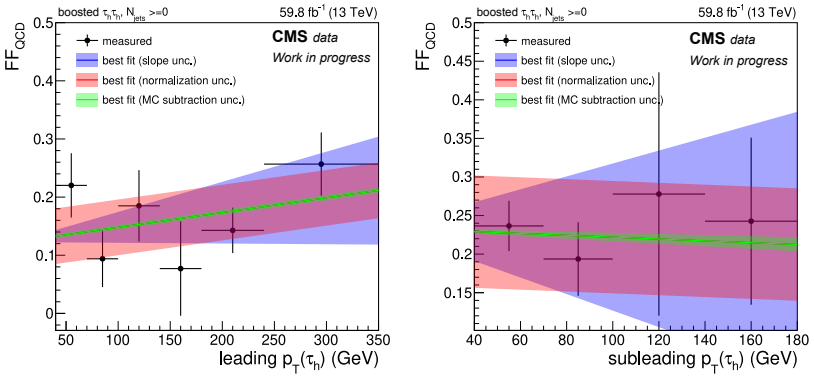


Figure 4.27: Process fraction measurements for the boosted $e\tau_h$ (top), $\mu\tau_h$ (middle) and $\tau_h\tau_h$ (bottom) channel. On the left side are the measurements for events with one or no b-tagged jet, on the right side for events with two or more b-tagged jets. The sum of the process fractions is always one for each measured bin. The $t\bar{t}$ fraction also includes single top events. The fractions for the $e\tau_h$ and $\mu\tau_h$ channels are measured dependent on the transverse mass m_T of the lepton (e or μ) and the missing transverse momentum. In the $\tau_h\tau_h$ channel the measurement depends on the visible invariant mass of the two τ_h.

be of the same charge. Due to small numbers of events, the categorization in N_{jets} changes to two categories for the $e\tau_h$ and $\mu\tau_h$ channels, namely one or zero jets and two or more jets. For the $\tau_h\tau_h$ channel no categorization in N_{jets} is defined. The results for the $\tau_h\tau_h$ channel are shown in figure 4.28.



(a) F_F^{QCD} for the leading τ_h (b) F_F^{QCD} for the subleading τ_h

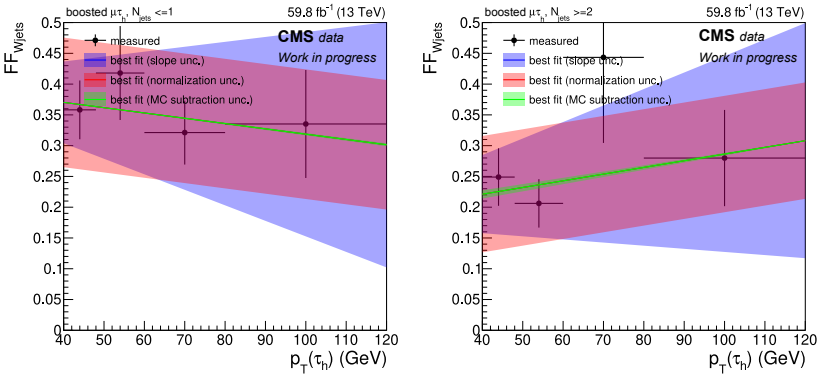
Figure 4.28: Results of the $F_F^{QCD}(p_T)$ measurements for the boosted $\tau_h\tau_h$ channel for each τ_h. The measurements are not split by N_{jets} categories due to small event numbers, and are derived in p_T bins of the τ_h's. A linear fit is performed to extract the F_F functions. Two uncertainties are defined from the fit, namely a slope and a normalization uncertainty shown in blue and red, respectively. The third uncertainty (green) is related to the subtraction of (Monte-Carlo) simulated background events from data.

Estimation for W+jets

The measurement of the F_F^{W+jets}'s is done in the same way as described for the resolved $\tau\tau$ analysis. The DR is defined by inverting the b-jet pair and bb-tagged AK8 jet selection. Similar to the F_F^{QCD} measurement before, due to small event yields the categorization in N_{jets} changes to two categories for the $e\tau_h$ and $\mu\tau_h$ channels, namely one or zero jets and two or more jets. For the $\tau_h\tau_h$ channel no F_F^{W+jets}'s are measured due to almost no contribution in this channel. Therefore, the same approach as for the resolved $\tau\tau$ analysis is chosen to use the F_F^{QCD}'s for W+jets events in the $\tau_h\tau_h$ channel. The results for the $\mu\tau_h$ channel are shown in figure 4.29.

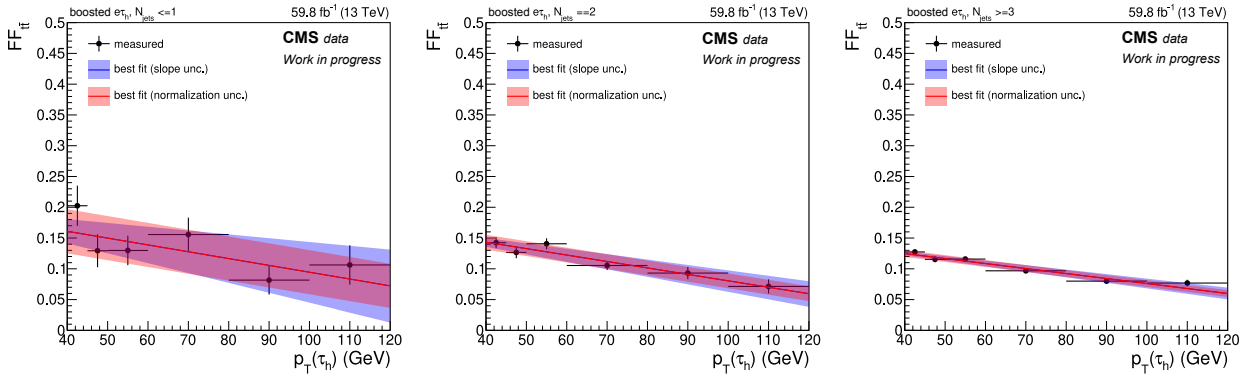
Estimation for $t\bar{t}$

Also for the $F_F^{t\bar{t}}$ measurement the same approach is used as in the resolved $\tau\tau$ case. The $F_F^{t\bar{t}}$'s are first measured with simulated events and then an inclusive simulation to data scale factor is applied like in equation 4.8. The DR is defined by inverting the electron/muon veto in the three boosted channels. In the $e\tau_h$ and $\mu\tau_h$ channels three N_{jets} categories are defined for one or zero jets, two jets and three or more jets. For the $\tau_h\tau_h$ channel no categorization in N_{jets} is used due to small event yields. The results for the $e\tau_h$ channel are shown in figure 4.30.



(a) $N_{\text{jets}} \leq 1$ category (b) $N_{\text{jets}} \geq 2$ category

Figure 4.29: Results of the $F_{\text{F}}^{\text{W+jets}}(p_{\text{T}}, N_{\text{jets}})$ measurement for the boosted $\mu\tau_{\text{h}}$ channel. The measurement is split into two N_{jets} categories and is derived in p_{T} bins of the τ_{h}. A linear fit is performed to extract the F_{F} functions. Two uncertainties are defined from the fit, namely a slope and a normalization uncertainty shown in blue and red, respectively. The third uncertainty (green) is related to the subtraction of (Monte-Carlo) simulated background events from data.



(a) $N_{\text{jets}} \leq 1$ category (b) $N_{\text{jets}} = 2$ category (c) $N_{\text{jets}} \geq 3$ category

Figure 4.30: Results of the $F_{\text{F}}^{\text{t}\bar{\text{t}}}(p_{\text{T}}, N_{\text{jets}})$ measurement for the boosted $e\tau_{\text{h}}$ channel. The measurement is split into three N_{jets} categories and is derived in p_{T} bins of the τ_{h}. A linear fit is performed to extract the F_{F} functions. Two uncertainties are defined from the fit, namely a slope and a normalization uncertainty shown in blue and red, respectively.

Corrections for the F_{F} measurements

As for the resolved $\tau\tau$ analysis, corrections need to be calculated to account for the mis-modeling related to the initial assumptions of the F_{F} method. The non-closure corrections are calculated to correct the extrapolation from the AR to the SR. This is done in the same way as described for the resolved $\tau\tau$ case by calculating the ratio between data in the SR-like and the applied

F_F^{proc} in the AR-like. Two non-closure corrections are derived, one for the p_T of the electron, muon or other τ_h in the $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$ channels, respectively, and one for the mass of the τ_h, also for all three channels. Both non-closure corrections are shown in figure 4.31, for QCD in the $\tau_h\tau_h$ channel, for W+jets in the $e\tau_h$ channel and for $t\bar{t}$ in the $\mu\tau_h$ channel.

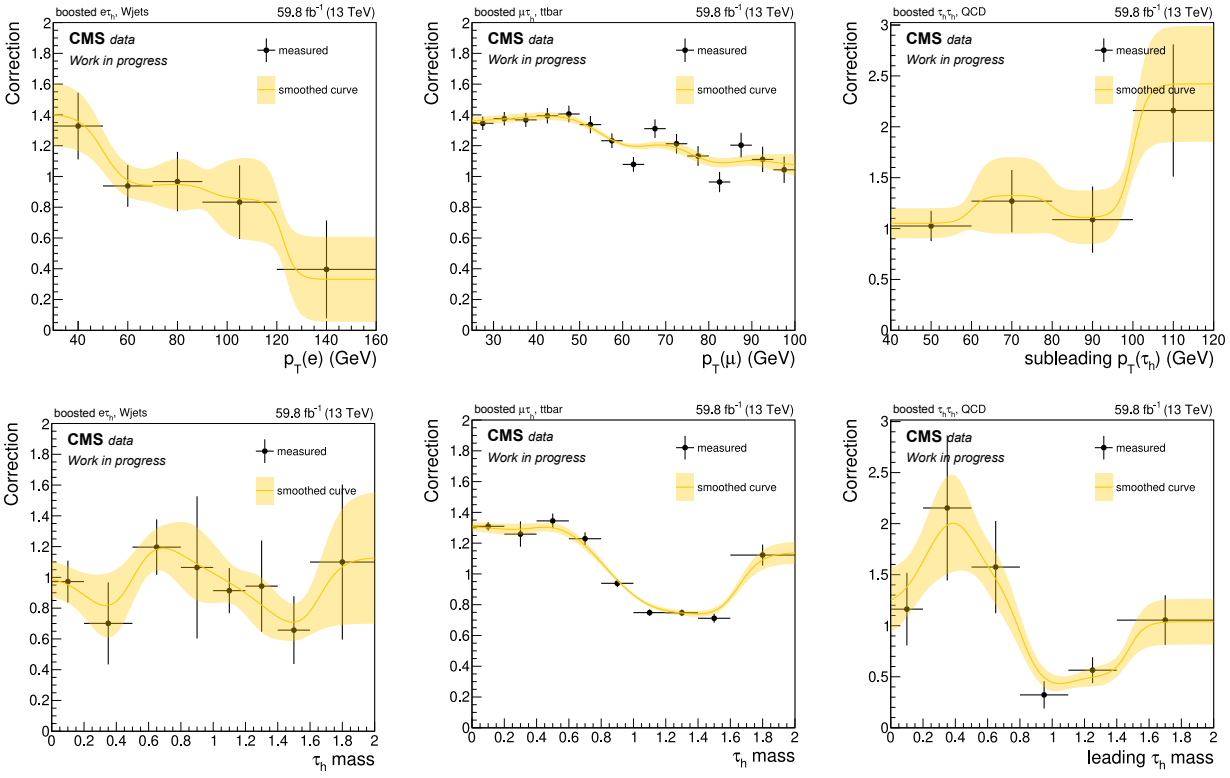


Figure 4.31: Results of the non-closure correction measurements dependent on the p_T of the electron, muon and second τ_h as well as the mass of the τ_h for the three boosted $\tau\tau$ channels. In the left column corrections applied to F_F^{W+jets} are shown, in the middle column to $F_F^{t\bar{t}}$ and in the right column to F_F^{QCD}. Besides the histogrammed corrections also the smoothed functions with uncertainty bands are shown.

Further, additional corrections dependent on the visible invariant mass m_{vis} are derived for QCD and W+jets targeting the extrapolation from the DR to the SR. The correction for W+jets is only relevant for the $e\tau_h$ and $\mu\tau_h$ channels and is derived in the same way as explained in the resolved $\tau\tau$ case. The cut on m_T is removed and new F_F^{W+jets}'s are measured. The estimation in the DR with these new F_F^{W+jets}'s is then compared to simulated W+jets events in the SR.

The DR to SR correction for QCD differs in the definition of the four orthogonal regions compared to the resolved $\tau\tau$ case. For the $e\tau_h$ and $\mu\tau_h$ channels, instead of the inversion of the electron and muon isolation, the $\Delta R(\tau_1, \tau_2)$ criterion is inverted to be greater than 0.8. This region can partially overlap with events selected for the resolved $\tau\tau$ analysis but since this correction is not applied in the resolved $\tau\tau$ analysis, this is not concerning. For the $\tau_h\tau_h$ channel,

again the unused DR is exploited where both τ_h fail the loose WP of the BDT discriminant. These DR to SR corrections are shown in figure 4.32.

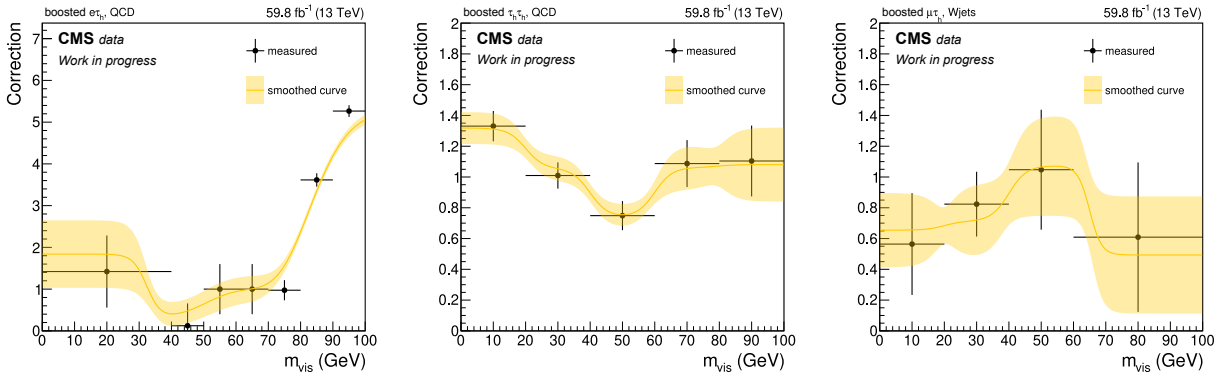


Figure 4.32: Results of the DR to SR extrapolation correction measurements dependent on the visible invariant di-tau mass m_{vis} for the three boosted $\tau\tau$ channels. The two left corrections are applied to F_F^{QCD} in the $e\tau_h$ and $\tau_h\tau_h$ channel, respectively. The right correction is applied to F_F^{W+jets} in the $\mu\tau_h$ channel. Besides the histogrammed corrections also the smoothed functions with uncertainty bands are shown.

4.7 Event simulation

The second method used to estimate background processes as well as signal processes is the Monte Carlo event simulation. Simulated events are used for all background contributions that are not covered by the F_F method. In sections 4.7.1 and 4.7.2, it will be discussed how the background and signal processes are simulated, respectively. Since simulated events usually differ from data, corrections are applied for multiple sources of discrepancies. The relevant corrections for this analysis will be discussed in section 4.7.3.

4.7.1 Simulation of background events

The simulation procedure is always similar, independent of the simulated process. The main part of the event simulation is performed by Monte Carlo (MC) event generators like *MadGraph5_aMC@NLO* [83], *Powheg* [84] and *Pythia 8.2* [85], which are state-of-the-art tools for generating simulated high energy physics events. A list of the MC event generators used for the different background processes is given in table 4.8.

The simulation starts with the pp collision. At such high energies as at the LHC, the protons in the pp collision do not interact with each other as a whole but rather their constituents called partons. The momenta of the protons are split among all partons, which means that the momentum fraction of the two colliding partons is not known. Therefore, parton distribution functions (pdfs) are used for the event simulation. The pdfs describe the probability of finding a parton with a given momentum fraction at a certain energy scale. One such pdf set is *NNPDF3.1* [86], which is frequently used for event simulations at the LHC.

Table 4.8: List of the MC event generators used for the simulation of the background processes and the order of perturbative QCD.

Process	Event generator	Order
$t\bar{t}$	*Powheg*	NLO
single top	*Powheg*	NLO
Z+jets	*MadGraph5_aMC@NLO*	NLO
Diboson	*Pythia 8.2*	LO
single H_{SM}	*Powheg*	NLO

Next, the hard scattering process is simulated, where the required partonic process with its final state is generated. This is, for example, the semi-leptonic $t\bar{t}$ process or Z boson production with a $Z \to \tau^-\tau^+$ decay. Following the simulation of the hard scattering process, additional steps are necessary to model a full event. This includes parton showering with *Pythia 8.2* [85], which models additional radiation and the showering of the final state particles to lower energies. At some point the energy gets low enough that the hadronization process starts and hadrons are formed from quarks and gluons.

In addition to the hard scattering process, soft interactions in a pp collision occur frequently. This process is called underlying event and is part of the simulation to get a better estimation of the data. Additionally, on average about 30 interactions happened during a single proton bunch crossing in data taken at CMS in 2018. This pileup is estimated by adding specifically simulated pileup events to the event.

Once the event is fully simulated, the interactions of the particles in the event with the CMS detector are simulated using *Geant4* [87]. This step involves detailed modeling of the detector geometry, material properties and generation of electrical signals to emulate the response of all the subdetectors to the passing particles.

At this stage, the simulated events include the full detector data equivalent to real data events and the same event reconstruction is applied. However, differences to real data still remain, for example, in identification efficiencies or energy reconstruction and need to be corrected, which will be discussed in section 4.7.3.

4.7.2 Simulation of signal events

Events for the signal processes, $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ and $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$, are centrally produced by the CMS collaboration. The simulation of the processes is based on the *Feynrules* implementation [88] of the NMSSM (see section 2.3). For the simulation of a single mass pair hypothesis the mass parameters for the X and Y bosons are set to the values of the hypothesis. Further, the width of these two bosons is set to a small value of 10^{-3} GeV so that the narrow width approximation [75] can be assumed for the limit estimate. The *Feynrules* implementation [88] is used in the *MadGraph5_aMC@NLO* simulation software [83] to simulate

the hard process at leading order. The branching fractions for the $X \to YH_{SM}$ decay and for the $Y(\to b\bar{b})H_{SM}(\to \tau^-\tau^+)$ and $Y(\to \tau^-\tau^+)H_{SM}(\to b\bar{b})$ decays, for the respective signal processes, are set to one in the simulation to produce only the corresponding final states of interest.

In total 574 different mass pair hypotheses are generated for each signal process to scan the unknown mass of the X and Y bosons. The mass grid is shown in figure 4.33 and ranges from $240 - 4000$ GeV for m_X and from $60 - 2800$ GeV for m_Y. For each mass pair hypothesis 400 thousand events are generated. For the analysis in this thesis only a subset of the hypotheses will be used to showcase the possibility of a combined, resolved and boosted, final state analysis strategy. This subset is indicated by the red colored mass pair hypotheses in figure 4.33. However, the final analysis of the CMS collaboration will include all of the mass pair hypotheses.

4.7.3 Corrections for simulated events

The simulated CMS detector response does not always align fully with the detector response of the real detector. Therefore, the following corrections are applied in this analysis.

Pileup reweighting

The number of pileup interactions in data depends on the instantaneous luminosity delivered by the LHC. However, this number is often unknown at the time simulated events are generated. This is mitigated by randomly adding additional pileup interactions to the hard scattering process following a Poisson distribution. For the Poisson distribution the expected number of pileup interactions for the given run period is used. Since the expected and measured pileup distribution usually does not match, a correction is needed.

The correction is calculated from the ratio between the measured pileup distribution during a run period and the pileup distribution used in the simulation [89]. By applying this correction, the simulated events are reweighted to better match the data.

Lepton reconstruction efficiencies

The analysis selection contains criteria for the identification, isolation and triggering of electrons and muons. For these, corrections are applied, derived from data and simulation efficiency measurements for these three criteria. For the measurements the tag-and-probe method is used, aiming for $Z \to ee/\mu\mu$ events. They can be selected with high purity, especially after applying a cut on the Z boson mass window around $m_Z = 91$ GeV. The efficiencies are calculated in an iterative manner by applying an already measured efficiency correction to the next measurement.

$$\varepsilon(\mathrm{ID}, \mathrm{iso}, \mathrm{trig}) = \varepsilon(\mathrm{trig}|\mathrm{iso}, \mathrm{ID}) \cdot \varepsilon(\mathrm{iso}|\mathrm{ID}) \cdot \varepsilon(\mathrm{ID}) \tag{4.10}$$

The already applied efficiency corrections are indicated by b in $\varepsilon(a|b)$ and the efficiency that should be measured by a. In the resolved $e\tau_h$ and $\mu\tau_h$ channels of the analysis all three corrections are applied, while for the boosted $e\tau_h$ and $\mu\tau_h$ channels no isolation cut is applied

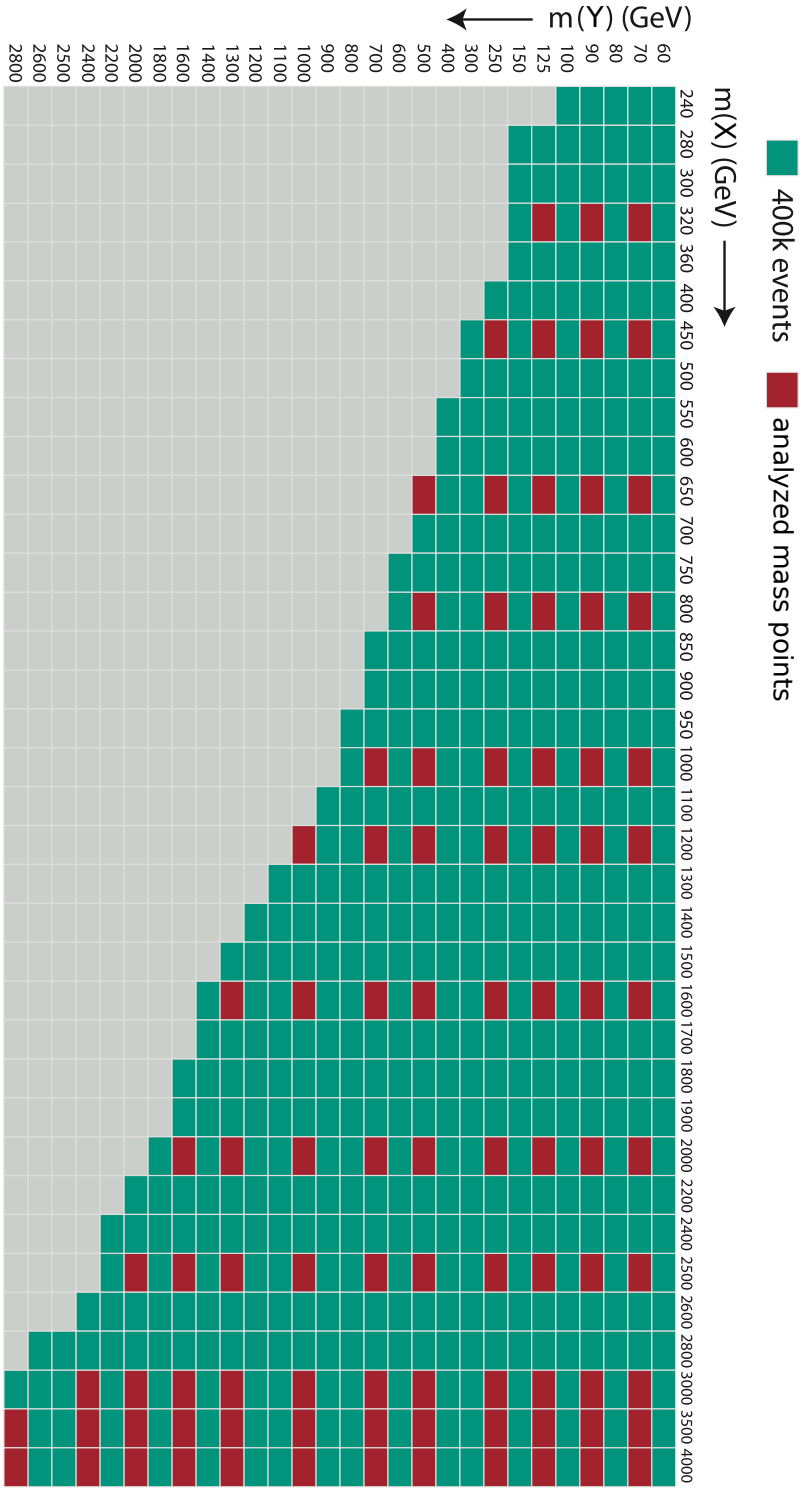


Figure 4.33: Two dimensional grid of produced signal samples valid for both signal processes. The grid points highlighted with green and red are officially produced by the CMS collaboration. The red grid points will be used in this analysis. The grid points in grey are not produced due to the resonant production constraint $m_X > m_Y + 125\,\text{GeV}$.

and therefore also no isolation correction is needed. The corrections are provided centrally by the CMS collaboration [40, 47].

For the identification of resolved and boosted τ_h and the di-tau triggers a similar approach is used. The efficiencies of data and simulation are measured with the tag-and-probe method in an $Z \to \tau\tau$ enriched control region in the $\mu\tau_h$ final state. The corrections are provided centrally by the CMS collaboration [55].

For the boosted $\tau_h\tau_h$ channel, a combination of an AK8 jet and a $\not{E}_T$ trigger is used. Specifically for these triggers, correction factors are measured in a dedicated study [90] with a reference trigger method. As reference trigger, a single muon trigger is used and the efficiency is calculated with

$$\varepsilon = \frac{N_{\text{sig. trig \& ref. trig.}}}{N_{\text{ref. trig.}}}. \tag{4.11}$$

Lepton and jet energy reconstruction

The energy correction for jets as well as an additional dedicated energy for b-jets are discussed in section 3.3.7 and section 3.4.5, respectively.

The electron energy correction is introduced in section 3.3.4. The energy correction for muons is small and can be neglected for this analysis.

The resolved and boosted τ_h energy corrections are measured separately due to the different identification algorithms. However, the measurement of the energy correction is performed in a similar way. The $\mu\tau_h$ final state in a $Z \to \tau\tau$ enriched control region is exploited to fit the τ_h energy dependent on the τ_h decay mode. For both resolved and boosted τ_h, the energy correction is provided centrally by the CMS collaboration [55, 58] and is below 2% for all decay modes.

Reweighting of the $t\bar{t}$ p_T spectrum

The main background of this analysis is the $t\bar{t}$ process, which is simulated at next-to-leading order with *Powheg* [84]. For the analysis, it is crucial that the kinematic distributions like the p_T of the top quarks are well modeled in simulation. However, in measurements of the $t\bar{t}$ production cross section [91, 92], it was identified that the p_T spectrum of the top quarks is harder in simulation compared to data. This discrepancy can partially be explained by missing higher order corrections but these are not applied in the *Powheg* event simulation.

In the measurements [91, 92] a correction function is derived which is used to reweight $t\bar{t}$ events dependent on the p_T of the two top quarks. For each top quark i the weight is calculated with

$$w(p_T^i) = \exp(0.0615 - 0.0005 \cdot p_T^i) \tag{4.12}$$

and then combined to a total event weight $w^{t\bar{t}} = \sqrt{w^t \cdot w^{\bar{t}}}$. This event reweighting is applied to all simulated $t\bar{t}$ events that enter the analysis, except the events where the selected τ_h is a misidentified jet. These events are estimated with the F_F method.

Efficiency of b-jet identification

The accurate modeling of the b-jet identification efficiency in simulated events is important for this analysis because the signal events, but also the main background ($t\bar{t}$), are mainly selected due to the b-jet pair selection described in section 4.3.3. However, the identification efficiency in simulation differs from data.

In the training of the final parametric neural network (PNN) classifier, the shape of the *DeepJet* discriminant is used as an input variable, therefore, the full discriminant shapes needs to be well modeled. The corresponding corrections are measured centrally by the CMS collaboration [59] with the tag-and-probe method in a phase space with two oppositely charged leptons and two additional jets targeting the $t\bar{t}$ events. The Z+jets contribution is reduced by removing events with a di-lepton mass close to the Z boson mass. One of the selected jets is used as the tag object requiring that it passes the medium WP of the *DeepJet* algorithm. Similar corrections are measured for mis-tagged jets induced by light quarks in a phase space specifically selecting Z+jets events.

The corrections are measured dependent on the value of the b-tagging discriminant D, the p_T and η of a single jet. An event weight is calculated based on the evaluation of the scale factor SF for each selected jet in an event,

$$w_{\text{event}}^{SF} = \prod_{i}^{N_{\text{jets}}} SF(D_i, p_{T,i}, \eta_i) \tag{4.13}$$

This way of correcting the full shape of the b-tagging discriminant has the effect that after applying the weight the yield of the selected events changes. To account for this, a reweighting ratio r is measured with the sum of the event weights before and after applying the b-tagging SF's and before applying event cuts related to the b-tagging discriminant,

$$r = \frac{\sum_i^{N_{\text{events}}} w_{\text{before},i}}{\sum_i^{N_{\text{events}}} w_{\text{after},i}} \tag{4.14}$$

For this analysis the reweighting ratios are measured for each simulated background process in each analysis channel. The results are listed in table 4.9. The ratios are applied together with the b-tagging scale factors for each event in the analysis as $w_{\text{event}} = r \cdot w_{\text{event}}^{SF}$.

Efficiency of $\mathcal{X} \to b\bar{b}$ AK8 jet identification

As for the b-jet identification efficiency also the efficiency of the *ParticleNetMD* algorithm for identifying boosted bb pairs needs to be estimated and corrected. The CMS collaboration provides measured corrections for the $\mathcal{X} \to b\bar{b}$ tagging of *ParticleNetMD* [66].

However, the provided working points are too tight for this analysis because together with the additional selection of a tau lepton pair almost no events would be left. For this reason a custom WP is used as mentioned in section 3.4.6. Dedicated corrections for this WP are

Table 4.9: Results of the reweighting measurement for the b-jet identification efficiency correction for the six analysis channels and each simulated background process.

Analysis	Process	Ratio in $e\tau_h$	Ratio in $\mu\tau_h$	Ratio in $\tau_h\tau_h$
resolved $\tau\tau$	$t\bar{t}$	1.002	1.004	1.005
	single top	1.002	0.998	1.009
	Z+jets	0.996	0.997	0.991
	Diboson	0.991	0.991	0.991
	single H_{SM}	0.987	0.989	0.989
boosted $\tau\tau$	$t\bar{t}$	0.973	0.968	0.921
	single top	1.055	0.999	0.892
	Z+jets	0.941	0.945	0.901
	Diboson	0.937	0.966	0.886
	single H_{SM}	0.939	0.938	0.923

planned to be derived for the full analysis of the CMS collaboration but are not available at the moment. Therefore, a different approach is chosen for this thesis. The correction factors are set to one for all events which have a bb-tagged AK8 jet and a 50% uncertainty on this correction is used in the final fits. This rather conservative approach should cover any inconsistencies between the data and simulation efficiencies for the $\mathcal{X} \to b\bar{b}$ identification.

4.8 Mass reconstruction algorithms

In this analysis two mass reconstructions are used. The first method is *FastMTT*. It is a simplified version of the *SVfit* algorithm [93], which is a likelihood based $\tau\tau$ system reconstruction method inspired by the matrix element method [94, 95]. Dedicated studies were conducted to investigate this method and the possibilities of replacing it with a neural network approach. For more details on both approaches readers are referred to [96]. In the following the second mass reconstruction method will be introduced.

4.8.1 Kinematic fit for mass of the $b\bar{b} + \tau^-\tau^+$ system

A mass estimation of the bb + $\tau\tau$ system, and hence of the X boson, can significantly help in the identification of signal events. The approach used for this analysis is based on the *HHKinFit* tool, which was originally developed for the reconstruction of a heavy X boson decaying into a pair of H_{SM} bosons [97, 98]. The method uses kinematic constraints that the decay products of the Y and H_{SM} bosons must satisfy due to the narrow width of these bosons and the precisely known mass of the H_{SM} boson of 125 GeV. These constraints are applied in a χ^2 fit of the bb + $\tau\tau$ + $\vec{p}_T^{\,miss}$ system to get a better estimation of the X boson mass. This method is applied in this analysis for events where a resolved b-jet pair is present.

A χ^2 function is introduced that allows one to vary the measured energies of each b-jet and each tau lepton within a range given by the resolution but is penalized if the varied energy values get too different from the initially measured energy. Additionally, the missing transverse energy $\vec{p}_T^{\,miss}$ is included as an estimate of the momenta of the neutrinos produced in the decay of the tau leptons. The sketch in figure 4.34 illustrates the typical constellation in signal events.

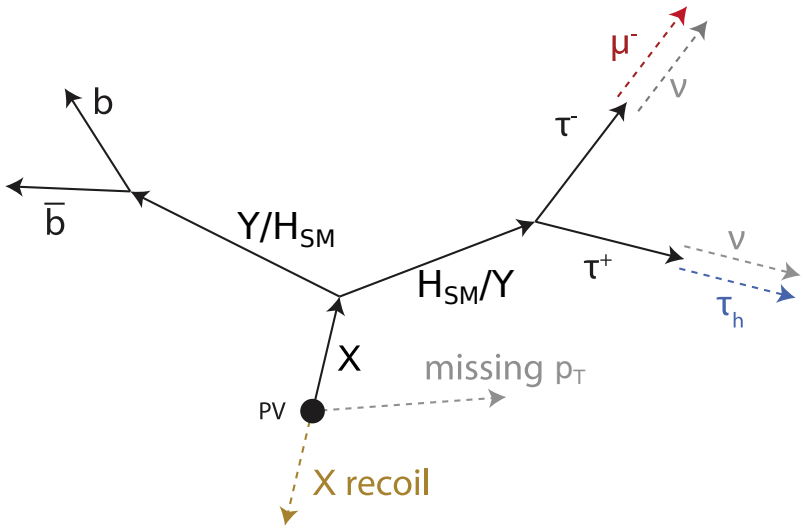


Figure 4.34: Sketch of a possible signal event in the $\mu\tau_h$ final state. The missing p_T vector is expected due to the missed neutrinos in the decay of the tau lepton pair. Further, the recoil vector is indicated because it is used in the kinematic fit.

Compared to the initial implementation of Ref. [97], this analysis has only one H_{SM} boson and the mass for the Y is unknown. Further, this analysis has two different signal processes in which the decay products of the Y and H_{SM} bosons change. This is challenging because, instead of one single kinematic fit, a range of multiple kinematic fits is performed scanning the m_Y mass hypotheses. The scanned m_Y values are the same as the produced signal mass hypotheses (see figure 4.33). Additionally, the number of kinematic fits doubles because the scan is repeated for each signal process. This results in 56 kinematic fits performed for each selected event in this analysis.

After all kinematic fits are performed the best fit is chosen as the one with the smallest χ^2 value and the best fit value of m_X^{KinFit} and the discrete value of m_Y^{KinFit} are used as mass estimations in the following analysis. Further, the χ^2_{min} itself is used for the training of the final classifier because real signal events are expected to have a relatively small χ^2_{min} value compared to background events in which the underlying event signature for the kinematic fit is not given.

In the following sections the individual steps of the kinematic fit will be explained.

Fit of the b-jet energy

For the fit it is assumed that the direction measurement (η and ϕ) of the b-jets is precise enough to be neglected in the fit. Therefore, only the energy of the b-jets is varied by the fit.

For a single fit the mass of the boson from which the b-jets are originating is fixed and a two body decay is assumed, which means that $m_{\mathrm{b_1,b_2}} \equiv m_{\mathrm{Y/H_{SM}}}$ and using the four vectors of the b-jets

$$\begin{aligned} m^2_{\mathrm{Y/H_{SM}}} &= m^2_{\mathrm{b_1,b_2}} = (p_{\mathrm{b_1}} + p_{\mathrm{b_2}})^2 \\ &= p^2_{\mathrm{b_1}} + p^2_{\mathrm{b_2}} + 2 \cdot p_{\mathrm{b_1}} \cdot p_{\mathrm{b_1}} \\ &= m^2_{\mathrm{b_1}} + m^2_{\mathrm{b_2}} + 2 \cdot E_{\mathrm{b_1}} \cdot E_{\mathrm{b_2}} - 2 \cdot \vec{p}_{\mathrm{b_1}} \cdot \vec{p}_{\mathrm{b_2}} \end{aligned} \tag{4.15}$$

At this step an approximation is applied that the mismeasurement of the jet energy and jet momentum is equivalent, at least in leading order, and the their ratio $\frac{\vec{p}}{E} = \beta$ is constant. The same constant behavior is assumed for the mass $\frac{m}{E} = \frac{1}{\gamma}$ with $\gamma = \frac{1}{\sqrt{1-\beta^2}}$ and equation 4.15 can be rewritten to only depend on the two energies of the b-jets and constant values as

$$m^2_{\mathrm{Y/H_{SM}}} = m^2_{\mathrm{b_1}} + \frac{E^2_{\mathrm{b_2}}}{\gamma^2_{\mathrm{b_2}}} + 2 \cdot E_{\mathrm{b_1}} \cdot E_{\mathrm{b_2}} \cdot C \tag{4.16}$$

with

$$C = 1 - \frac{\vec{p}_{\mathrm{b_1}} \cdot \vec{p}_{\mathrm{b_2}}}{E_{\mathrm{b_1}} \cdot E_{\mathrm{b_2}}} = \frac{m^2_{\mathrm{b_1,b_2}} - m^2_{\mathrm{b_1}} - m^2_{\mathrm{b_2}}}{2 \cdot E^{\mathrm{meas.}}_{\mathrm{b_1}} \cdot E^{\mathrm{meas.}}_{\mathrm{b_2}}} \tag{4.17}$$

which is a constant value that can be calculated from the measured b-jet energies. Equation 4.16 can now be solved as a quadratic equation of $E_{\mathrm{b_2}}$ which then only depends on $E_{\mathrm{b_1}}$,

$$E_{\mathrm{b_2}} = -E_{\mathrm{b_1}} \cdot \gamma^2_{\mathrm{b_2}} \cdot C + \sqrt{E^2_{\mathrm{b_1}} \cdot \gamma^4_{\mathrm{b_2}} \cdot C^2 - \gamma^2_{\mathrm{b_2}} \cdot (m^2_{\mathrm{b_1}} - m^2_{\mathrm{Y/H_{SM}}})}. \tag{4.18}$$

With this the χ^2 function for the minimization can be defined as

$$\chi^2_{\text{b-jets}} = \frac{(E_{\mathrm{b_1}} - E^{\mathrm{meas.}}_{\mathrm{b_1}})^2}{\sigma^2_{\mathrm{b_1}}} + \frac{(E_{\mathrm{b_2}} - E^{\mathrm{meas.}}_{\mathrm{b_2}})^2}{\sigma^2_{\mathrm{b_2}}}. \tag{4.19}$$

By replacing $E_{\mathrm{b_2}}$ with equation 4.18 the energy of the first b-jet can be determined by minimizing the $\chi^2_{\text{b-jets}}$ function. The resulting best fit energy $E^{\mathrm{fit}}_{\mathrm{b_1}}$ can then be inserted into equation 4.18 to get the best fit energy $E^{\mathrm{fit}}_{\mathrm{b_2}}$ of the second b-jet. The resolution σ_{b} of the b-jet energies is taken from the b-jet energy regression described in section 3.4.5.

Fit of the tau lepton energy

The tau lepton system is handled differently in the kinematic fit due to the neutrinos in the list of decay products of the Y or $\mathrm{H_{SM}}$. Therefore, a two body decay cannot directly be used to fit the energies. Nevertheless, the energies of the visible decay products like electrons, muons or τ_{h}, depending on the analysis channel, are defined with the same two body procedure as already for the b-jets in equation 4.18 to reduce one degree of freedom in the fit.

Instead of the individual tau lepton energies, the full event recoil in the transverse direction is fitted. The χ^2 function is defined as

$$\chi^2_{\text{recoil}} = \Delta\vec{p}^{\,T}_{\text{T,recoil}} \cdot \text{COV}^{-1}_{\text{recoil}} \cdot \Delta\vec{p}_{\text{T,recoil}} \tag{4.20}$$

which compares the difference $\Delta\vec{p}_{\text{T,recoil}} = \vec{p}^{\,\text{fit}}_{\text{T,recoil}} - \vec{p}^{\,\text{meas.}}_{\text{T,recoil}}$ between the measured recoil vector

$$\vec{p}^{\,\text{meas.}}_{\text{T,recoil}} = -(\vec{p}^{\,\text{meas.}}_{\text{T,b}_1} + \vec{p}^{\,\text{meas.}}_{\text{T,b}_2} + \vec{p}^{\,\text{meas.}}_{\text{T},\tau_1} + \vec{p}^{\,\text{meas.}}_{\text{T},\tau_2} + \vec{p}^{\,\text{meas.}}_{\text{T,miss}}) \tag{4.21}$$

and the best fit vector before the tau lepton decay

$$\vec{p}^{\,\text{fit}}_{\text{T,recoil}} = -(\vec{p}^{\,\text{fit}}_{\text{T,b}_1} + \vec{p}^{\,\text{fit}}_{\text{T,b}_2} + \vec{p}^{\,\text{fit}}_{\text{T},\tau_1} + \vec{p}^{\,\text{fit}}_{\text{T},\tau_2}). \tag{4.22}$$

The covariance matrix of the recoil is calculated from the covariance of the missing transverse momentum and the covariance matrices of the b-jets and visible tau lepton decay products as

$$\text{COV}_{\text{recoil}} = \text{COV}_{\vec{p}_{\text{T,miss}}} - (\text{COV}_{\text{b}_1} + \text{COV}_{\text{b}_2} + \text{COV}_{\tau_1^{\text{vis}}} + \text{COV}_{\tau_2^{\text{vis}}}). \tag{4.23}$$

The covariance matrices of the the four objects i can be calculated with

$$\text{COV}_i = \begin{pmatrix} \cos^2(\phi_i) & \sin(\phi_i)\cdot\cos(\phi_i) \\ \sin(\phi_i)\cdot\cos(\phi_i) & \sin^2(\phi_i) \end{pmatrix} \cdot \sigma^2_{p_{\text{T},i}} \tag{4.24}$$

and

$$\sigma_{p_{\text{T},i}} = \sin(\eta_i) \cdot \frac{E_i}{|\vec{p}_i|} \cdot \sigma_{E_i}. \tag{4.25}$$

For the b-jet energy resolution again the results from the b-jet energy regression are used and for the leptons it is assumed that their energy resolution is much better than for jets and therefore it is set to zero.

Full kinematic fit procedure

In summary the χ^2 function for the fit is defined from equations 4.18 and 4.20 as

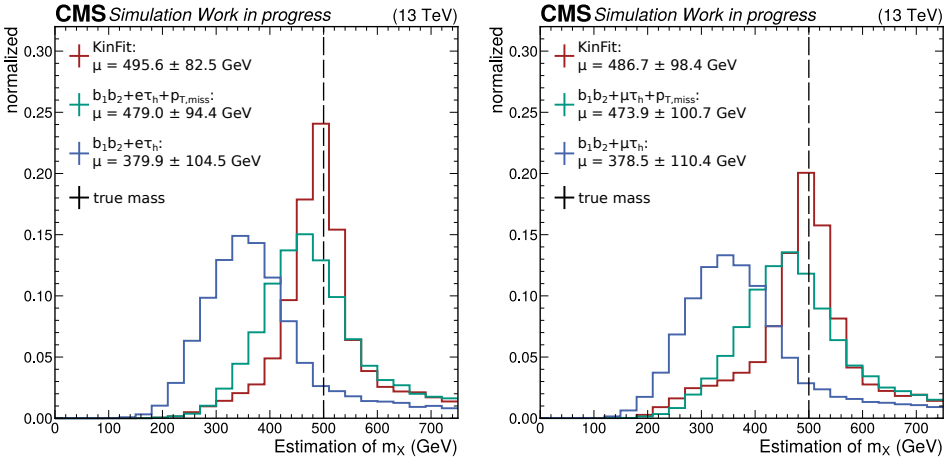
$$\chi^2 = \chi^2_{\text{b-jets}} + \chi^2_{\text{recoil}}. \tag{4.26}$$

Because the energies of the two b-jets and the energies of the two tau leptons are connected and during a single fit the masses m_{Y} and $m_{\text{H}_{\text{SM}}}$ are fixed, the χ^2 function has only two free parameters, which are $E_{\text{b},1}$ and $E_{\tau,1}$.

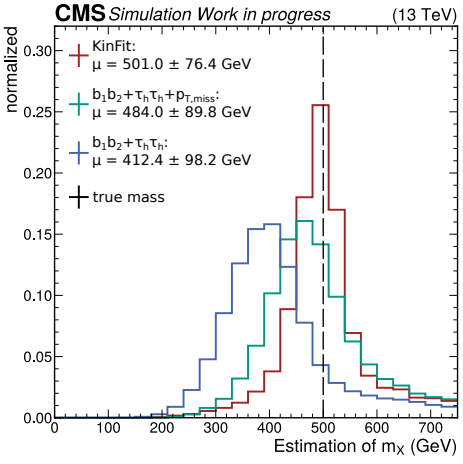
For the boundaries of possible energy variations of the four objects different approaches are used. For the b-jets the boundaries are set to $\pm 5\sigma_E$ of their energy resolution from the b-jet energy regression. For the tau leptons the lower boundary is set to the energy of the visible object $E^{\text{min}}_{\tau,1} = E^{\text{vis}}_{\tau,1}$ assuming that the neutrino energy is missing. For the upper boundary, equation 4.18 is used for the tau leptons with the visible energy of the second tau lepton $E^{\text{max}}_{\tau,1} = E_{\tau,1}(E^{\text{vis}}_{\tau,2})$.

The fit of this χ^2 function is performed in two steps. The first step is a minimization along one dimension of positive $(E_{\mathrm{b},1} - E_{\tau,1})$ values. The second step happens only if a minimum can be found in the first step. The second step is then a two dimensional minimization using the Newton method. The fit is regarded as converged if the value differences of the χ^2 or one of the two energies, ($E_{\mathrm{b},1}$ and $E_{\tau,1}$), between two steps of the fit is smaller than a threshold of $\epsilon = 0.01$.

The results of the kinematic fit applied in the resolved $\tau\tau$ analysis are compared in figure 4.35 to two simpler estimations of the X boson mass by calculating the invariant mass of the decay products, once just the visible bb+$\tau\tau$ system and once additionally with $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$. The true mass hypothesis is $m_{\mathrm{X}} = 500\,\mathrm{GeV}$. It is clearly shown that the kinematic fit provides a significantly better mass estimation and resolution.



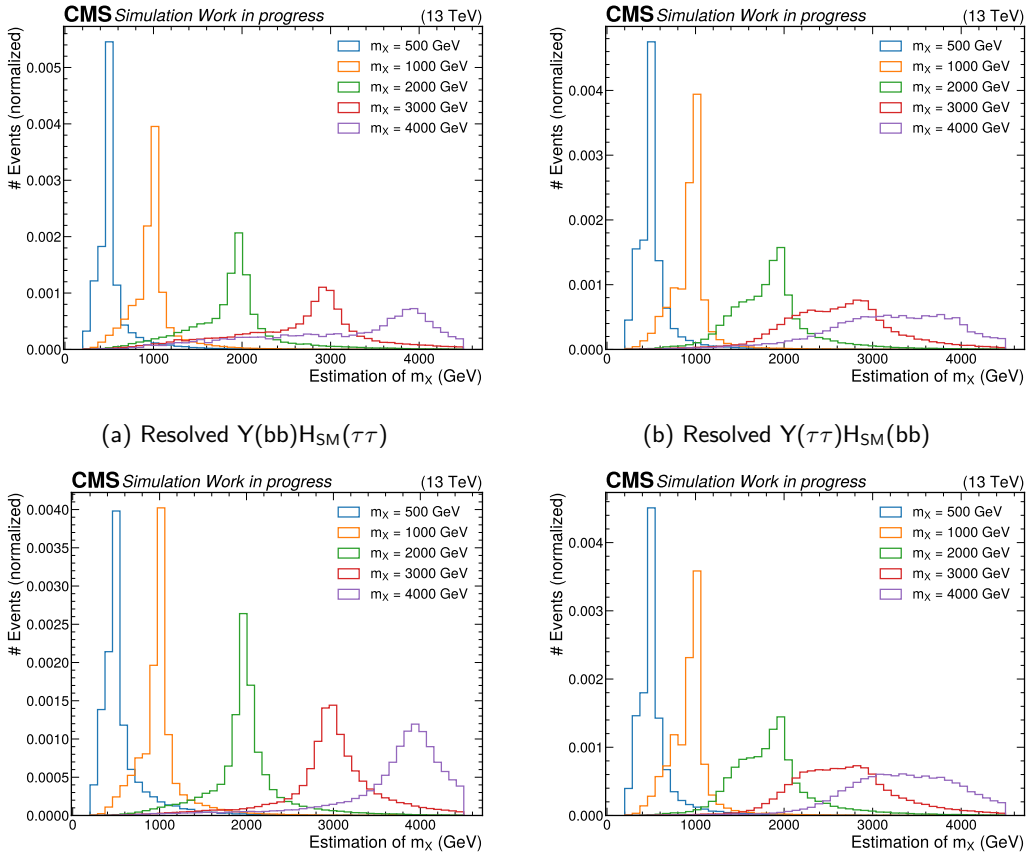
(a) Resolved $e\tau_{\mathrm{h}}$ channel (b) Resolved $\mu\tau_{\mathrm{h}}$ channel



(c) Resolved $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel

Figure 4.35: Comparison of the kinematic fit for the X boson mass estimation to a simple calculation of the invariant mass. The mean values of the distributions for the three resolved $\tau\tau$ analysis channels (a)-(c) are calculated and compared to the true mass of $m_{\mathrm{X}} = 500\,\mathrm{GeV}$.

Further, the mass estimations of the kinematic fits are compared for the two signal processes and the resolved and boosted $\tau\tau$ analysis. The distributions for five different mass pair hypotheses are shown in figure 4.36, where the Y boson mass is always $250\,\text{GeV}$ and the X boson mass is varied. For the distributions all three tau lepton pair decay channels ($e\tau_\text{h}$, $\mu\tau_\text{h}$ and $\tau_\text{h}\tau_\text{h}$) are combined. In all cases the estimation of the mass for lower m_X values yields similar results by finding the correct true mass. For higher m_X, the distribution starts to get broader, especially for the signal process where the Y boson decays into a pair of tau leptons. A reason for this could be that for such high energies the approach in which the energy boundaries are set for the fit is not suited anymore and gives the fit to much freedom, especially for the tau leptons, for which no direct energy constraint is present in the χ^2 function. The setting for the energy boundaries is not changed in any way compared to the initial implementation of the *HHKinFit* tool, where both boson are assumed to be H_SM bosons. For the signal process where the H_SM boson decays into a tau lepton pair the m_X is better estimated.



(a) Resolved Y(bb)$H_{SM}(\tau\tau)$

(b) Resolved Y($\tau\tau$)H_{SM}(bb)

(c) Boosted Y(bb)$H_{SM}(\tau\tau)$

(d) Boosted Y($\tau\tau$)H_{SM}(bb)

Figure 4.36: Comparison of kinematic fit results for different initial masses of m_X for the two signal processes on the left and right side. The resolved and boosted $\tau\tau$ pair reconstruction is shown in (a)-(b) and (c)-(d), respectively. For the Y boson, the mass is always set to 250 GeV. The shown distributions are a combination of the mass estimations of all three tau lepton pair decay channels, $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$.

5 Measurement strategy and results

The aim of a search is to find indications of BSM physics or, if no signal is found, provide exclusion limits. To achieve that, a phase space needs to be defined which is enriched in the hypothetical signal events. This implies that a clear separation between hypothetical signal events and expected background events is desired. The separation can be achieved by identifying kinematic variables, e.g. estimates of the invariant mass of particles, which show significantly different distributions for signal and background. For the resonant di-Higgs search conducted in this thesis, this could be, for example, the m_X estimation of the kinematic fit introduced in section 4.8.1. However, the analysis aims not to test a certain mass pair hypothesis but performed a full mass grid scan with very different phase spaces of the signal processes depending on the mass hypotheses. Therefore, a separation from background events cannot be achieved with a single variable for all tested mass pair hypotheses. Instead, parametric neural networks (PNNs) are trained to separate signal from background events for each individual mass pair hypothesis based on a set of input variables which describe the events. The details about the PNN classifier are explained in section 5.1. Next, the statistical framework is described in section 5.2, for which the foundations are taken from [99], and the final upper limit measurements of the search are evaluated in section 5.3.

5.1 Event classification with neural networks

The task of separating the signal and background processes for this analysis is performed with a PNN. This PNN classifies events into categories of physics processes by utilizing event information like, for example, the kinematic properties of the reconstructed b-jets or tau leptons. In the following the key features of these PNNs will be discussed.

5.1.1 Architecture and training process

The architecture of the networks is based on fully connected feed-forward layers with multiple neurons in each layer. Such an architecture is rather common and was already discussed in

the 1980's [100]. Nevertheless, such an architecture can still have, if not better, but at least a performance comparable to newer and more sophisticated architectures like, for example, graph neural networks [101].

The PNNs have, for all analysis channels, 39 input neurons corresponding to 37 event variables and two parameter variables. Before the variables are given to the input layer, they are transformed so that the variable distributions for all events have a mean value of zero and a standard deviation of one. Following the input layer, three hidden layers consisting of 500, 500 and 300 neurons are arranged, respectively. The output layer, located after the hidden layers, has eight neurons corresponding to eight process categories in which an event should be classified. A sketch of this network structure is shown in figure 5.1.

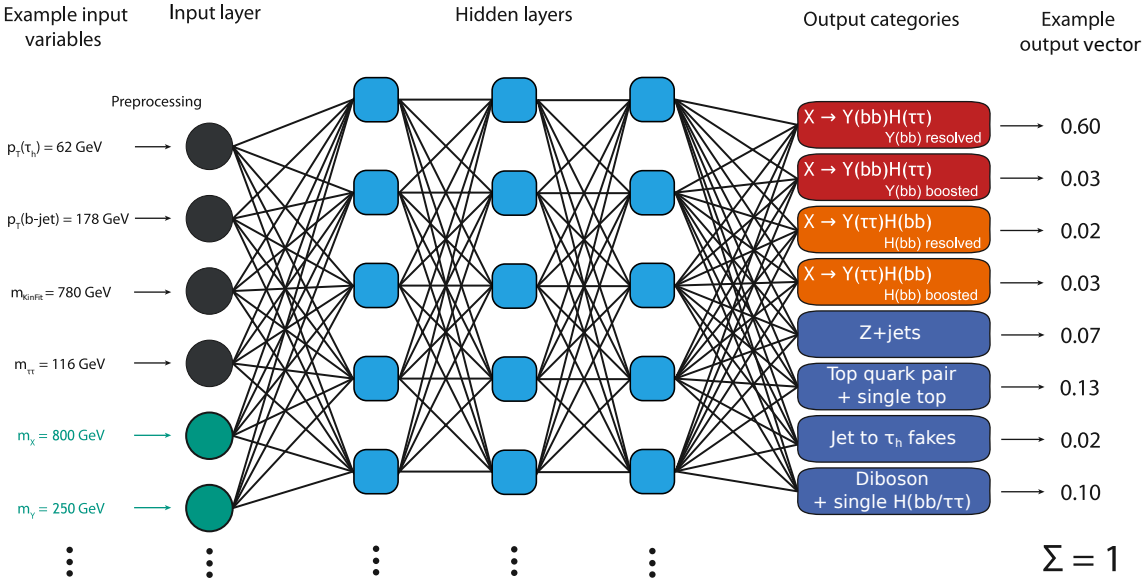


Figure 5.1: Sketch of the parametric neural network classification. The input layer receives an input variable vector for a single event. Part of the input vector are two parameter variables defining the probed mass pair hypothesis. After the processing through the hidden layers, the PNN has eight output nodes predicting the underlying physics process of the event. Four nodes are dedicated to the main background processes and two are for each signal process, respectively. The signal process nodes are split into nodes for resolved and boosted bb pair decay.

The neurons of a neural network are all set up in the same way. They take all values from the previous layer and calculate a weighted output value

$$f_{\text{out}}(\vec{x}) = \sum_{i}^{N_x} w_i \cdot x_i + b. \tag{5.1}$$

Each neuron has a trainable weight vector $\vec{w}$ with a length corresponding to the length of the values provided by the previous layer and a trainable bias parameter b. After each layer's output calculation, an activation function is applied. For the hidden layers the hyperbolic

tangent function is used, $f_{\text{act.}}(\vec{x}) = \tanh(f_{\text{out}}(\vec{x}))$, and for the output layer each node i gets a normalized prediction value calculated with the softmax activation function

$$f_{\text{act.},i}(\vec{x}) = \frac{\exp(x_i)}{\sum_i \exp(x_i)}. \tag{5.2}$$

The resulting output vector sums up to one and each of the values ranges between zero and one. Such a definition can be interpreted as a probability for an event to correspond to one of the processes. An event is classified into the process category where it gets the highest probability value.

The training process of a neural network is set up so that the trainable parameters are updated iteratively. The trainable parameters are the aforementioned weights and biases of the layers and an update is done after comparing the prediction of the neural network with the truth. For this comparison a loss function is calculated and in case of this analysis the categorical cross-entropy is used, which is defined as

$$L = -\sum_{n}^{N}\sum_{i}^{C} y^{(n)}_{\text{true},i} \cdot \log(y^{(n)}_{\text{pred},i}). \tag{5.3}$$

The first sum covers all events in a processed batch with N events. The second sum evaluates the truth $y_{\text{true},i}$ to prediction $y_{\text{pred},i}$ comparison for each output process category C. The truth labels y_{true} can be one for the correct category or zero for all wrong categories.

The goal of a training is to minimize the value of the loss function. This is achieved by using the *Adam* algorithm [102]. *Adam* is an extension of the stochastic gradient decent where a weight is updated via

$$w_{\text{new},i} = w_i - \text{lr} \cdot \frac{\partial L}{\partial w_i} \tag{5.4}$$

with an adjustable learning rate lr. The calculation of the gradients uses the backpropagation of error algorithm [100] to reduce the computational effort. In this analysis the initial learning rate is set to 10^{-4}.

During a training process of a neural network it can happen that the network starts to overtrain, which means that it only learns the data that is used in the training and cannot generalize it to other comparable datasets. Two ways to account for these problems are applied in this analysis. The first method is *Dropout* [103], which randomly excludes neurons during the training steps. This approach reduces the probability that individual neurons become too important, which is a sign of overtraining. In this analysis a dropout probability of 40% is used. This means that 40% of the neurons are randomly excluded in each training step. The second method is the L2 regularization, which is a penalizing term added to the loss function. The term calculates the squared sum of all training weights and thereby has a negative effect on the loss function if the weight values get too large. The scaling factor for this term, before it is added to the loss function, is chosen to be 10^{-3}.

As training data, all selected events are used. The data is split into half for each process and for each half an individual PNN is trained. After the training is finished, the PNNs are used to predict the process categories of the other dataset, respectively. During the training a second data split is performed to define a training (75%) and a validation (25%) dataset. The validation data is used to control the learning process. After each training step the loss function is calculated for this independent validation data without adjusting the training weights. If the loss of the validation and training datasets starts to deviate from each other, this is an indication of overtraining. Further, if the validation loss does not decrease for 50 training steps, the training is considered converged and the weight parameters with the lowest loss are used.

5.1.2 Classes of processes

The task of the PNN is to classify events into process categories which are mutually exclusive to each other and are enriched in events of a certain process. Having well separated signal and background processes helps to constrain systematic uncertainties related to a certain process during the statistical inference. Therefore, four background classes are defined targeting the main background contributions of the analysis and four signal classes targeting the two different signal processes and their different phase spaces for the Y or H_{SM} boson decay into the bb pair.

The first background category is defined for the F_F method. This process category is trained on measured data that fails the τ_h identification WPs for the respective resolved and boosted $\tau\tau$ analyses. The F_F method has its own systematic uncertainties that only affect the τ_h fake distributions, therefore, it is reasonable that it has an own category.

Next, the main background contribution has its own category, which is $t\bar{t}$ production. Single top production is added to this category due to the similarity of the final state with one or two top quarks. From both simulated processes the part of events that has jets faking the selected τ_h is removed since it is already covered by the F_F method.

The third background category is defined for Z+jets events. Since this process is of relevance especially for the boosted $\tau\tau$ analysis, it has its own category. Also for Z+jets, only events that have either real tau leptons or prompt electrons or muons that can also fake τ_h's are considered.

The fourth category is defined for all minor background contributions. This includes diboson production and single Higgs boson production. These processes are, in most of the analysis channels, almost negligible but have a small contribution of a few percent in the boosted $\tau_h\tau_h$ channel.

Each of the signal processes gets two separate categories. The idea is to separate the resolved and boosted reconstruction used to identify the bb pair, especially because both ways to reconstruct bb pairs can be present in an event. For the training of the PNN the simulated signal events are split based on the generator level information of the two b quarks. Each event with $\Delta R(b_1^{gen}, b_2^{gen}) > 0.5$ gets a "resolved" label and events with smaller ΔR distances between the two b quarks get a "boosted" label.

5.1.3 Parametrization of the neural network

One of the challenges of this analysis is that a large number of hypotheses needs to be tested. Ideally, for each hypothesis a dedicated neural network has to be trained, but the computational effort needed to achieve that is significant. Therefore, the previous analysis [9] grouped kinematically similar mass pair hypotheses together into a single training, thus reducing the number of trained neural networks by about a factor of six. Nevertheless, this was still a significantly large number.

A new approach is chosen for this thesis using parametric neural networks. This means that a single neural network is trained simultaneously on all mass pair hypotheses and two parameters are introduced that represent the two mass hypotheses for m_X and m_Y as additional input variables (indicated in figure 5.1). A study [104] tested this approach in a reduced setting of only one dimension, that is, having a fixed mass for m_X and a parameter for m_Y. This study showed that with this approach a comparable sensitivity in the final limit measurement can be achieved. In this analysis, the setup is extended to all produced mass pair hypotheses with a two dimensional parametrization.

The aim of this input parametrization is that the neural network learns to classify signal events in a certain specialized way for each tested mass pair hypothesis. On the other hand, background events do not have any defining mass pair hypothesis and should not depend on these two input parameters. This is mitigated in the training by randomly sampling mass pair values from the set of all mass pair hypotheses considered for each background event in each training update step. Using this approach, it is achieved that, regardless of the mass parameters, the separation between the background categories stays similar. Of course, the separation between signal and background events is still affected by the mass parameters because the kinematic information of the signals changes for different mass pair hypotheses.

5.1.4 Training data

For each analysis channel an individual PNN is trained. The number of events that can be used for each PNN training varies significantly between the process categories because of the rather tight event selection, especially for the boosted $\tau\tau$ channels. A list of unweighted event numbers that are used in the training is given in table 5.1.

The imbalance in training event numbers between process categories $N_{\text{evt,cat}}$ is mitigated by a balancing weight. Each process category gets a category weight w_{cat} that scales the importance of the events up or down depending on their deviation from the mean category event number $N_{\text{evt,tot}}/8$ of the full dataset.

$$w_{\text{cat}} = \frac{1}{N_{\text{evt,cat}}} \cdot \frac{N_{\text{evt,tot}}}{8} \tag{5.5}$$

The same procedure is applied for each signal process category to balance the importance of the different mass pair hypotheses. Although the total number of signal events is quite high, this is not the case for all individual mass pair hypotheses. Some of them only provide a few or

Table 5.1: Numbers of unweighted training events for each analysis channel, split for the eight process categories. The values represent the rounded mean of the two halves of the training datasets. For the signal categories the values are the sum of the selected events for all 574 mass pair hypotheses.

Category	res. $e\tau_h$	res. $\mu\tau_h$	res. $\tau_h\tau_h$	boost. $e\tau_h$	boost. $\mu\tau_h$	boost. $\tau_h\tau_h$
res. $Y(b\bar{b})$	156700	290900	217600	324700	517500	562700
boost. $Y(b\bar{b})$	24200	41000	28900	114400	210700	253800
res. $H_{SM}(b\bar{b})$	439200	509900	466100	143700	197000	176100
boost. $H_{SM}(b\bar{b})$	243100	257400	252800	113300	215600	262100
jet $\rightarrow \tau_h$ fakes	23000	48000	3000	440	930	130
$t\bar{t}$/ single t	113466	127500	11300	3470	7840	40
Z+jets	1560	3800	1680	580	930	1090
VV / single H_{SM}	2300	4800	2650	420	530	550

even no events, which is related to the split into resolved and boosted pairs of b quarks and tau leptons.

One of the negative effects of the tighter selection of the resolved b-jet pairs (see section 4.3.3), compared to the previous analysis [9], is the low number of training events for some of the background categories. Especially the boosted $\tau\tau$ categories suffer from this because they additionally have the tight cut on the ΔR between the two tau leptons. Nevertheless, a training with these few events could successfully converge, taking into account all the measures taken to prevent overtraining.

5.1.5 Input variables

In the PNN training, 37 variables are used which describe the physics of the events. These variables are related to the kinematic information about the tau lepton pair, the b-jet pair, the bb-tagged AK8 jet, the missing transverse momentum and more high-level variables calculated from this information like ΔR distances, invariant masses and the kinematic fit results. The full list of variables used in the training is given in table 5.2. If for an event a certain variable is not present, for example if no bb-tagged AK8 jet is reconstructed, a default value is set outside the variable range.

An analysis is performed to identify the most important input variables for the classification of events into a certain process category node using a Taylor expansion. This method derives Taylor coefficients taking the PNN as a function and expanding it around a specific input variable. These Taylor coefficients can be used as an indication for the sensitivity of a PNN output node to a certain input variable [105]. This Taylor coefficient analysis (TCA) is applied to each PNN training for all analysis channels. As an example, the TCA results are shown in

Table 5.2: All event variables used in the PNN training and classification with a short description.

Label	Description
njets	Number of jets
nbtag	Number of b-tagged jets
nfatjets	Number of AK8 jets
pt_1	p_T of the electron, muon or p_T-leading τ_h
eta_1	η of the electron, muon or p_T-leading τ_h
phi_1	ϕ angle of the electron, muon or p_T-leading τ_h
pt_2	p_T of the τ_h (p_T-subleading τ_h in $\tau_h\tau_h$)
eta_2	η of the τ_h (p_T-subleading τ_h in $\tau_h\tau_h$)
phi_2	ϕ angle of the τ_h (p_T-subleading τ_h in $\tau_h\tau_h$)
deltaR_ditaupair	ΔR distance of the tau lepton pair
m_vis	Visible invariant mass of the tau lepton pair
m_fastmtt	Mass estimate of the $\tau\tau$ system with *FastMTT*[1]
pt_fastmtt	p_T estimate of the $\tau\tau$ system with *FastMTT*[1]
eta_fastmtt	η estimate of the $\tau\tau$ system with *FastMTT*[1]
phi_fastmtt	ϕ angle estimate of the $\tau\tau$ system with *FastMTT*[1]
bpair_pt_1	p_T of the p_T-leading b-jet
bpair_eta_1	η of the p_T-leading b-jet
bpair_phi_1	ϕ angle of the p_T-leading b-jet
bpair_btag_value_1	*DeepJet* discriminant value of the p_T-leading b-jet
bpair_pt_2	p_T of the p_T-subleading b-jet
bpair_eta_2	η of the p_T-subleading b-jet
bpair_phi_2	ϕ angle of the p_T-subleading b-jet
bpair_btag_value_2	*DeepJet* discriminant value of the p_T-subleading b-jet
bpair_deltaR	ΔR distance of the b-jet pair
bpair_m_inv	Invariant mass of the b-jet pair
bpair_pt_dijet	p_T of the combined b-jet pair vector
fj_Xbb_pt	p_T of the bb-tagged AK8 jet
fj_Xbb_eta	η of the bb-tagged AK8 jet
fj_Xbb_phi	ϕ angle of the bb-tagged AK8 jet
fj_Xbb_msoftdrop	Softdrop mass of the bb-tagged AK8 jet
fj_Xbb_nsubjettiness_2over1	Estimate of the bb-tagged AK8 jet to be more one or two prong like
met	Missing transverse momentum
metphi	ϕ angle of the missing transverse momentum
mt_1	m_T of the electron, muon or p_T-leading τ_h and $\vec{p}_T^{\,miss}$
kinfit_mX	Estimated mass for the X with the kinematic fit
kinfit_mY	Estimated discrete mass for the Y with the kinematic fit
kinfit_chi2	Minimal χ^2 value of the kinematic fit

[1] *FastMTT* is a simplified version of the *SVfit* algorithm [93], which is a likelihood based $\tau\tau$ system reconstruction method inspired by the matrix element method [94, 95]

figure 5.2 based on a first order Taylor expansion for all PNN output nodes. The results are shown only for the resolved $e\tau_h$ channel, they are similar in the other channels. For each node the five input variables with the highest mean value of the Taylor coefficients $\langle t_i \rangle$ are listed.

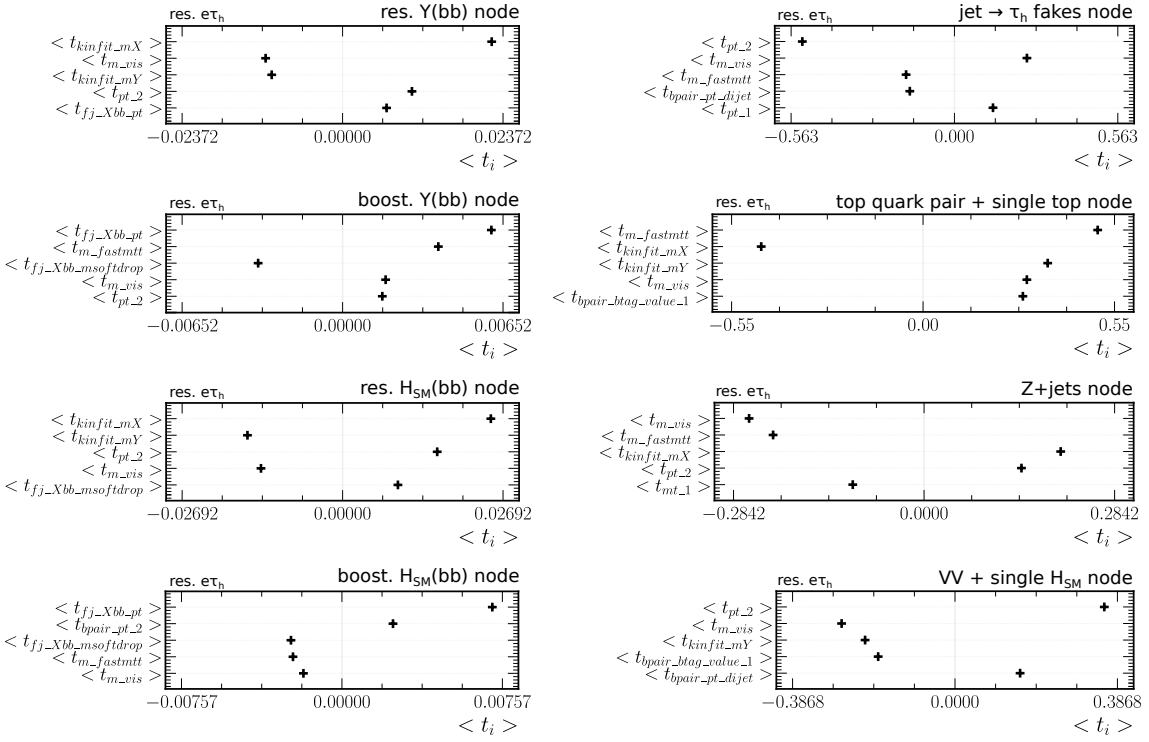


Figure 5.2: Results of the Taylor coefficient analysis for each output node of the PNN in the resolved $e\tau_h$ channel. On the left the four signal nodes are shown and on the right the four background nodes. Only the top five input variables are listed for each node, which have the highest mean value of the Taylor coefficients $\langle t_i \rangle$. Positive values indicate a correlation with the corresponding output node score, negative an anti-correlation.

The sign of $\langle t_i \rangle$ indicates the direction of the correlation between the input variable and the output node, where negative $\langle t_i \rangle$ values indicate anti-correlation. Further, the distributions of some of the most important variables are shown in figure 5.3 to illustrate the discriminative power of these variables between the different process categories.

First, regarding the four signal process nodes, the most important variables are mass estimates like the kinematic fit, *FastMTT* or the softdrop mass. This is expected because the difference of the two signal processes lies in the asymmetry of the Y and H_{SM} boson masses. The separation becomes worse when m_Y gets close to the H_{SM} mass or even has the same mass. In this case the PNN cannot separate the two signal processes and the classification is basically random in this respect. Another expected behavior is the importance of variables related to the bb-tagged AK8 jet for the two boosted bb categories. The existence of a well reconstructed bb-tagged AK8 jet in an event is a significant indicator of a boosted bb final state.

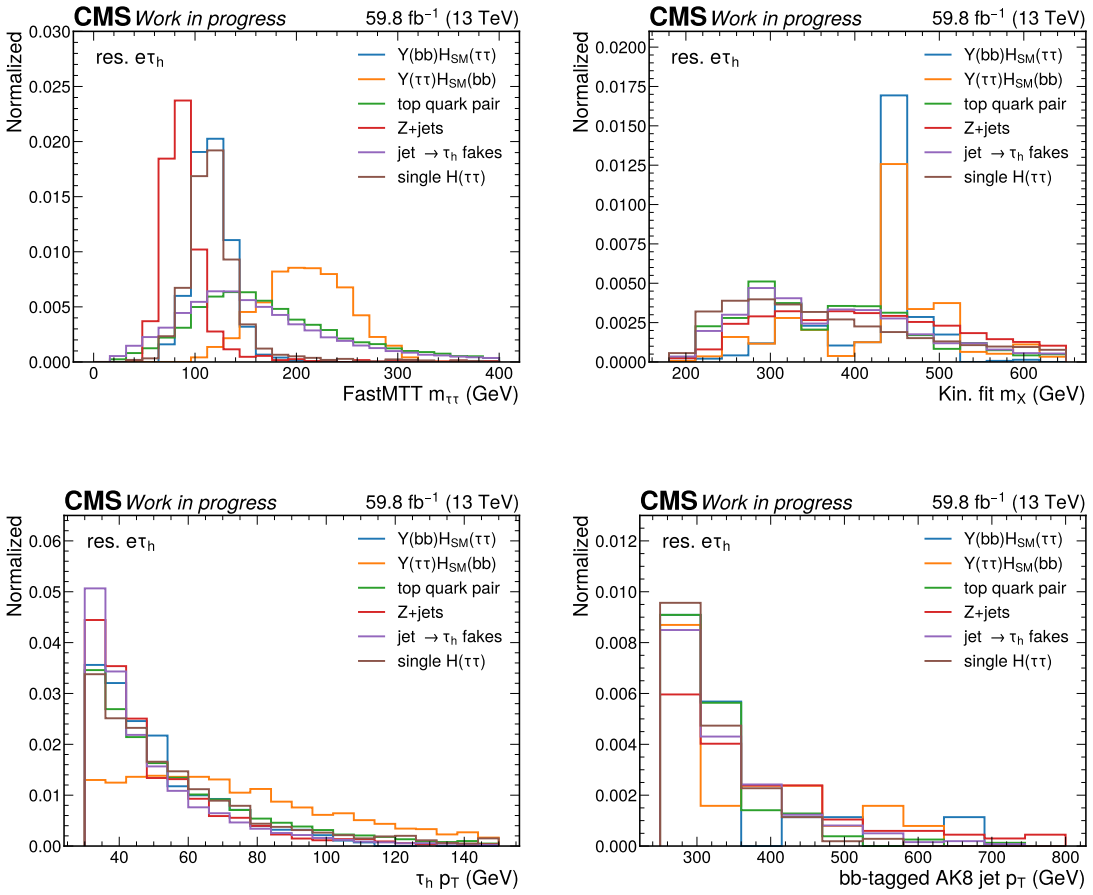


Figure 5.3: Distribution of important input variables split for the different process categories. The distributions are from the resolved $e\tau_h$ channel. For the signal distributions $m_X = 450\,\text{GeV}$ and $m_Y = 250\,\text{GeV}$ are chosen. The reconstructed mass of the $\tau\tau$ system (top left) shows dedicated mass peaks in the distributions for the Y, H_{SM} and Z bosons. The estimate of the kinematic fit for the m_X (top right) shows mass peaks for the two signal processes and a rather flat distribution for the background processes. The p_T of the τ_h (bottom left) shows different degrees of falling p_T spectra for the different processes. The p_T of the bb-tagged AK8 jet (bottom right) is statistically limited but shows a trend for the signal processes to have higher p_T.

For the background processes the mass reconstruction also plays a significant role for the separation from other processes. For example, the reconstruction of the $\tau\tau$ system results in peaking distributions for the Z and single H_{SM} bosons. On the other hand, the mass reconstruction is also relevant for the $t\bar{t}$ and single top category because it has rather flat mass distributions, which can be used to identify events from regions where no mass peaks are expected. For the category of the F_F method the same can be applied because it is enriched in $t\bar{t}$ events in which a jet fakes a τ_h. Additionally, the p_T of the τ_h is interesting for the F_F method

category because it has the distribution with the deepest falling course of all categories, as can be seen in figure 5.3.

The distribution of the kinematic fit estimation for m_X in figure 5.3 shows a deviation from the expected behavior. The expected mass peak at $m_X = 450\,\text{GeV}$ is clearly visible for both signal processes, however, additional smaller peaking structures can be identified. This could partially be related to a rather low number of events used for the histogram, but more likely this is an issue of the kinematic fit itself. Compared to the previous analysis [9] the number of scanned m_Y values is reduced in this analysis. This choice could have been too granular for a more exact mass estimate, for example, if the χ^2 values of the fits are close to each other. In that case neighboring mass hypotheses could be chosen which are further away from the truth.

5.1.6 Validation of the input variables

The variables used for the training of the PNNs are validated by performing goodness-of-fit tests on each of the variables. This is a necessary measure to ensure that the PNNs are not trained on information where the observed data and the prediction model are in disagreement. Such a disagreement would then be propagated to the output nodes of the PNNs. The goodness-of-fit test is a way to quantify how well the prediction model fits to the observed data by taking into account statistical and systematic uncertainties. For the goodness-of-fit a saturated model test [106] is performed on all input variables. This test is a more general, likelihood based approach compared to a χ^2 test.

The statistical model used for the final measurement will be introduced in the next section 5.2, but will already be used for the goodness-of-fit tests for the input variable validation. For simplicity the saturated model test will be introduced taking only the statistical uncertainties into account, which basically is equivalent to a χ^2 test. The likelihood is defined as

$$\mathcal{L} = \prod_i \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_i^2}} \exp\left(\frac{-(d_i - f_i)^2}{2 \cdot \sigma_i^2}\right) \tag{5.6}$$

where i is a single bin for which the observed data d_i is compared to the tested model prediction f_i, taking into account the statistical standard deviation σ_i of the data. The next step is to normalize this likelihood, following the Neyman-Pearson lemma [107], to the case where the model prediction is exactly the same as the data $f_i = d_i$. This is the saturated model and is defined as

$$\mathcal{L}_{\text{saturated}} = \prod_i \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_i^2}}. \tag{5.7}$$

Thereby, the normalization can be written as

$$\lambda = \frac{\mathcal{L}}{\mathcal{L}_{\text{saturated}}} = \prod_i \exp\left(\frac{-(d_i - f_i)^2}{2 \cdot \sigma_i^2}\right). \tag{5.8}$$

With all this, the test statistic of the goodness-of-fit test can be specified as

$$q = -2 \cdot \ln(\lambda). \tag{5.9}$$

The likelihood function becomes more complicated when taking systematic uncertainties and their correlations into account.

For the goodness-of-fit test of a single input variable the test statistic q_{obs} is calculated once with the observed data. To calculate the p-value of the test statistic, a Monte Carlo approach is used. Pseudo-datasets ("toys") are generated by randomly varying the prediction model within all its uncertainties. For a single goodness-of-fit test, 1000 toys are generated and their test statistic q_{toy} is calculated. By comparing the results of the test statistics of the observed data and the toys the p-value is derived as

$$p = \frac{N_{\text{toys}}(q_{\text{toy}} > q_{\text{obs}})}{N_{\text{toys}}}, \tag{5.10}$$

which is the fraction of toys that have a higher test statistic value than the observed data. The p-value can have values between zero and one and an input variable passed the goodness-of-fit test if the p-value is above 5%.

The distributions of the variables for the goodness-of-fit tests are binned in a way that the data distribution is equally populated in each bin. The number of bins for each variable distribution is initially set to ten, but some of the input variable distributions have very low event numbers. This is the case for several variables of the boosted $\tau\tau$ channels due to the tight boosted selection. Further, variables related to the bb-tagged AK8 jet are rather sparse in events because of the still rather tight WP of *ParticleNetMD*. For such variables the number of bins is reduced dependent on the total number of events for these variables while they have at least ten events in one bin.

In total 222 goodness-of-fit tests are performed for the six analysis channels. The calculated p-values for the full set of input variables are added in the figures A.1-A.6. Additionally, the results of three goodness-of-fit tests in the resolved $e\tau_{\text{h}}$, $\tau_{\text{h}}\tau_{\text{h}}$ and boosted $\mu\tau_{\text{h}}$ channels are showcased in figure 5.4 for one of the more important variables for the event classification, the kinematic fit estimate for m_X. The edge of the last bin of the equally populated m_X distributions is moved to 1000 GeV to make the lower bins better visible. In general, the last bin of the m_X distribution goes above 4 TeV. The p-values are calculated from the test statistic distributions with equation 5.10.

The results of the goodness-of-fit tests demonstrate good agreement between the observed data and the prediction model used in this analysis. A few variables show p-values below the defined 5% threshold, but they are not zero, which means that they are not totally outside the statistic distribution of the toy test. An example is shown in figure A.7. Further investigation of the distributions of these variables revealed that the reason for the small p-values is most likely

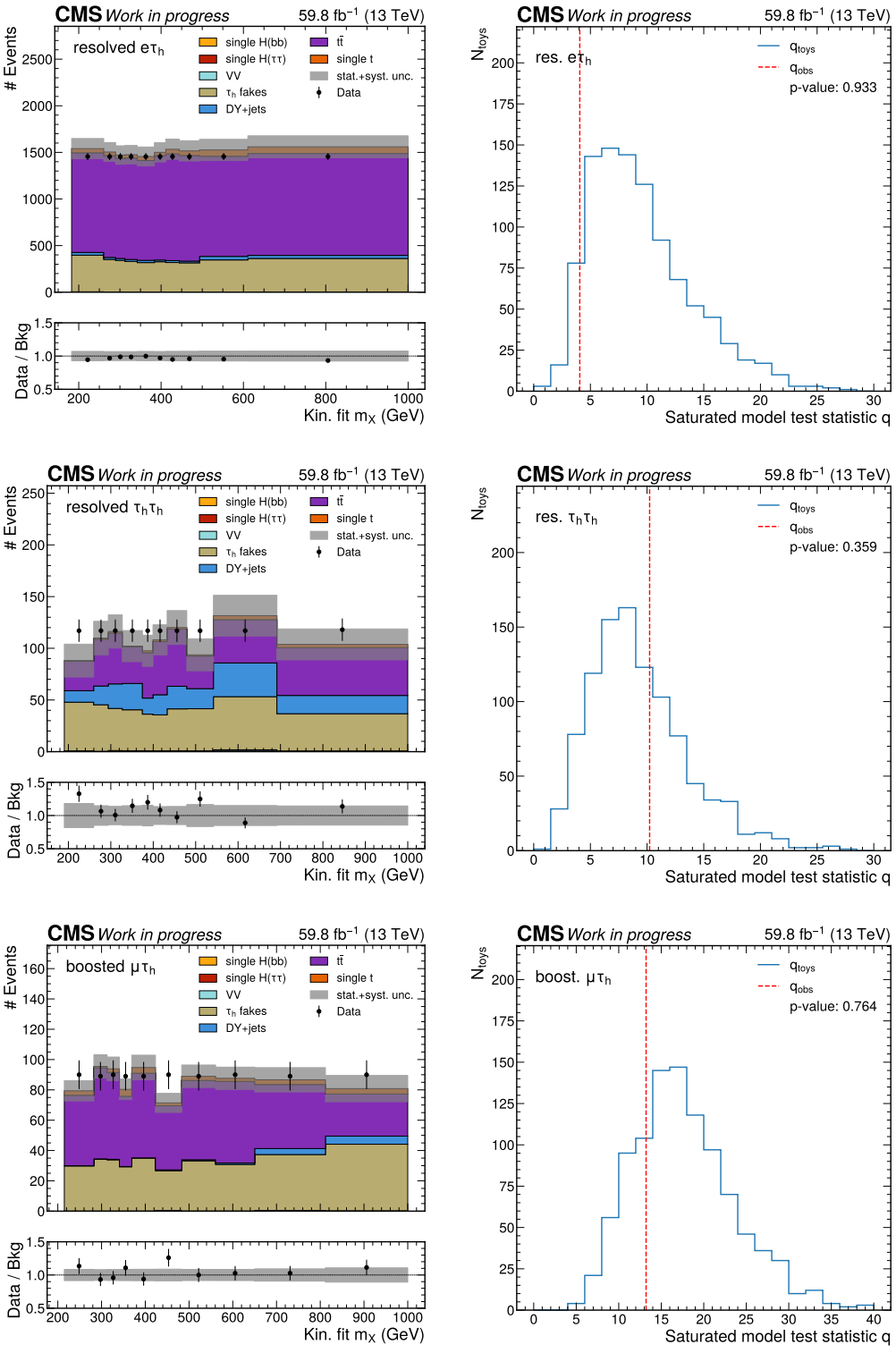


Figure 5.4: Distributions of the kinematic fit estimate for m_X with the corresponding goodness-of-fit tests. The top row shows the resolved $e\tau_h$ channel, the middle row the resolved $\tau_h\tau_h$ channel and the bottom row the boosted $\mu\tau_h$ channel. The distribution bins are chosen such that they are equally populated in data. For the saturated model test statistic (q_{toy}) 1000 pseudo-datasets ("toys") are generated and the observed value of the test statistic (q_{obs}) is indicated in red.

statistical fluctuations in single bins. Therefore, these variables will still be used in the PNN training.

In summary, the goodness-of-fit tests show that the prediction model in this analysis can describe the observed data within the uncertainties of the model. Nevertheless, it must be stated that these goodness-of-fit tests do not consider effects of higher dimensional correlations between the input variables, which is utilized by the PNN for the event classification. Such correlations can be tested by introducing a multidimensional binning with two or more variables at the same time, but this is not done in the scope of this analysis. Reasons for that are the high computational effort and the highly reduced event yields per bin in the multidimensional space, which is an issue for some of the input variables already in the one dimensional case.

5.2 Statistical inference and uncertainty model

The predictions of the PNNs are used to classify events into one of the process categories. An event is classified in the category for which it has the highest prediction score. This procedure results in eight orthogonal distributions enriched in a certain process. Studies of these eight discriminants showed that many of them have a very small event yield. This especially concerns the boosted $\tau_h\tau_h$ channel and in general the signal categories. Therefore, it is decided to not use any binning of these distributions and to use a single yield bin for each category. Considering all analysis channels, this approach introduces 24 background bins and 24 signal bins.

The measurement of a cross section σ or a limit on a cross section is accomplished by measuring the signal strength modifier defined as

$$\mu = \frac{\sigma_{\text{obs}}}{\sigma_{\text{exp}}}, \tag{5.11}$$

where σ_{exp} is the expected cross section times branching fraction from theoretical calculations based on a model like the SM or derived from simulation of the process. The measured cross section is σ_{obs} and if the expected cross section is correct, then the signal strength modifier should have a value close to one within uncertainties. To measure μ of the signal processes, all selected bins enter a binned likelihood function

$$\mathcal{L}(n|\mu \cdot s(\theta) + b(\theta)) = \prod_{i \in \text{bins}} \mathcal{P}(n_i|\mu \cdot s_i(\theta) + b_i(\theta)) \times \prod_{j \in \text{nuis.}} \mathcal{C}(\hat{\theta}_j|\theta_j). \tag{5.12}$$

This function derives the likelihood for an observed number of data events n by calculating the Poisson distributed probability

$$\mathcal{P}(n_i|\mu \cdot s_i(\theta) + b_i(\theta)) = \frac{(\mu \cdot s_i(\theta) + b_i(\theta))^{n_i}}{n_i!} \cdot \exp(-(\mu \cdot s_i(\theta) + b_i(\theta))) \tag{5.13}$$

for the observed number of data events n_i in each bin i for a given prediction model $\mu \cdot s_i(\theta) + b_i(\theta)$. The prediction model includes the signal event number s_i, scaled by a signal strength modifier μ, and the background event number b_i. Both, s and b, depend on the values of the nuisance

parameters θ. The nuisance parameters are used to incorporate the systematic uncertainties of the analysis. Each nuisance parameter θ_j is approximated by a probability density function $\mathcal{C}$, which reflects prior knowledge or constraints on the nuisance parameter. This prior is defined by the best estimate $\hat{\theta}_j$ of the value of θ_j.

The likelihood function $\mathcal{L}$ is maximized by varying all parameters, μ and θ, to obtain the best fit value of the signal strength modifier μ. Deviations from the prior of the nuisance parameters decrease the value of the likelihood function, which means that a good estimate of the priors is an important part of the likelihood maximization. Essential systematic uncertainties are discussed in the following. The uncertainty model is adapted from the previous analysis [9].

Usually, systematic uncertainties can be introduced as a rate or a shape changing variation of a certain uncertainty source. However, due to the decision to reduce the PNN category predictions to a single bin, effectively all shape uncertainties are reduced to rate uncertainties.

5.2.1 Uncertainties specific to signal events

Uncertainty sources of the signal event simulation are propagated to the signal estimation in the statistical model. These uncertainties arise primarily from the limited precision of the matrix element (ME) calculation for the hard interaction of the signal processes, which are $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ and $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$. The systematic uncertainties are defined for two sources.

The first source is related to the pdf. For signal events the *NNPDF3.1* [86] set is used in the simulation of the pp interaction. This set consists of about 100 individual pdf fits for which the pdf input parameters are varied within their uncertainties. The mean of these fits is used as a nominal pdf in the event simulation in *MadGraph5_aMC@NLO* [83]. An event weight is generated for each of the pdf fits and the standard deviation of these weight estimations, relative to the nominal pdf, quantify the pdf uncertainty. In the previous analysis [9] it was identified that the pdf uncertainty is flat and is around 18%. Therefore, in this analysis this uncertainty is introduced as an 18% rate uncertainty.

The second source of uncertainty is related to the re-normalization and factorization scale (μ_R and μ_F) at which the signal processes are produced. The scales are varied during the ME calculation in *MadGraph5_aMC@NLO* [83] by factors of 2 and 0.5 to define the up and down uncertainty variations. The nominal scales are dynamically chosen for each event to be

$$\mu_R = \mu_F = \sum_i m_{T,i}, \tag{5.14}$$

which is the sum of the transverse masses of all final state particles i. Usually in CMS analyses, two independent uncertainties are defined for μ_R and μ_F, but due to a technical issue in this analysis a combined uncertainty is used, where both scales are varied by a factor of 2 or 0.5 at the same time. The effect of this combined uncertainty goes up to 20%.

5.2.2 Uncertainties specific to simulated events

Uncertainty sources of simulated signal and background events are mainly related to the differences between simulation and the measured data. Corrections related to the reconstruction and identification of objects like electrons, muons, jets or τ_h are already discussed in section 4.7.3. In this section the corresponding uncertainties will be summarized. All systematic uncertainties are correlated between the analysis channels, except for uncertainties related to triggers, τ_h energy scale and identification, and the F_F method.

Identification and isolation of electrons/muons

For the electron and muon identification, a global rate uncertainty of 2% for each of them is introduced in the $e\tau_h$ and $\mu\tau_h$ channels, respectively. The same is done for the isolation in the respective channels. The reason for using a rate uncertainty instead of a shape uncertainty is the small dependence of the efficiencies on the electron/muon p_T.

Electron/muon triggers

The uncertainty for the single electron or muon triggers is constructed the same way as for the identification of electrons or muons by introducing a 2% rate uncertainty for each of them.

Electron/muon energy scale

The uncertainty of the electron energy scale measurement is included as systematic uncertainties in this analysis. This includes one uncertainty related to the energy scale and one to the energy resolution. Since for muons no energy correction is applied, no uncertainty on the muon energy scale is propagated.

Identification of τ_h

In the $e\tau_h$ and $\mu\tau_h$ channels, systematic uncertainties are defined by propagating the uncertainties of the correction measurements performed by CMS [55, 58] and are binned in p_T of the τ_h, thereby introducing a shape dependence. For the $\tau_h\tau_h$ channel the same is done but binned in the decay modes of the two τ_h. Since different τ_h identification algorithms are used in the resolved and boosted $\tau\tau$ analyses, independent uncertainties are introduced for each of them.

Energy scale of τ_h

The approach for the τ_h energy scale uncertainty is basically the same as for the identification of τ_h. The uncertainties of the τ_h energy scale measurements performed by CMS [55, 58] are propagated to the analysis as systematic uncertainties, independently for the resolved and boosted $\tau\tau$ channels.

τ_h triggers

The systematic uncertainties for the di-tau triggers in the resolved $\tau_h\tau_h$ channel and the AK8 jet $+ \not{E}_T$ triggers in the boosted $\tau_h\tau_h$ channel are propagated from the corresponding correction measurements mentioned in section 4.7.3.

Jet energy scale

The calibration of jet energies in CMS involves various sources of uncertainty that are accounted for in the analysis as systematic uncertainties. These uncertainties are part of the jet energy scale (JES) corrections which are applied during data and simulation processing. A total of 28 uncertainty sources are defined for the 2018 run period. To simplify the analysis, these uncertainties are grouped into 11 merged sources by combining strongly correlated uncertainties. Such an approach is motivated by the fact that this analysis is not particularly sensitive to the jet energy. This uncertainty scheme covers the absolute jet energy scale calibration. Further, it takes into account the relative jet energy scale calibration which takes care of momentum imbalance in different detector parts and statistical limitation of the jet energy scale measurements. Another source is related to the jet flavor (b, c, gluon or light flavors) dependence of the energy calibration. Remaining differences between simulation and data are taken care of by non-closure correction with corresponding uncertainties.

Jet energy resolution

The jet energy resolution is usually smaller in simulation compared to data, therefore, an energy smearing is applied to the jets in simulated events. The corresponding systematic uncertainty is part of the statistical model in this analysis.

Top quark p_T spectrum

For $t\bar{t}$ events, a top quark p_T reweighting is applied as introduced in section 4.7.3. The corresponding up and down variations are defined by not applying the reweighting to the $t\bar{t}$ events and by applying it twice.

Identification of b-jets

For the b-jet identification, a shape calibration is applied to the *DeepJet* discriminant in this analysis. This calibration introduces systematic uncertainties related to the purity of the b-jet selection and the contamination from light flavor jets. Additional uncertainties account for statistical fluctuations in the measurement of the b flavor and light flavor calibration scale factors. For c-jets no dedicated scale factors are measured, however, an uncertainty is defined based on the uncertainty measured for b-jets. Since the calibration depends on the p_T of the jets, scale factors are derived for each JES variation and are applied during the calculation of the JES uncertainties.

Identification of bb-jets

As already described in section 4.7.3, no dedicated scale factors are applied to correct the differences in efficiency of the *ParticleNetMD* algorithm between data and simulation. Instead, a large uncertainty of 50% is introduced and applied to each event containing a bb-tagged AK8 jet.

Pileup

A reweighting procedure is applied to align the distribution of pileup interactions in simulated events with the observed distribution in data. To estimate the uncertainty associated with this method, the assumed inelastic pp cross section is varied by $\pm 4.6\%$ as recommended by CMS. The impact of this variation is then propagated to the distributions used in this analysis.

Luminosity

All simulated events are scaled to the integrated luminosity measured in the 2018 run period. The measurement of the luminosity in this run [108] introduces a systematic uncertainty in this analysis of 2.5%.

Cross sections of background processes

Each background process is scaled to a cross section obtained from a state-of-the-art theoretical calculation. These cross sections are subject to the precision of the calculations. To account for this systematic uncertainties are considered. The uncertainties for each background process are listed in table 5.3 and are a combination of the scale, pdf and strong coupling α_S uncertainties.

Table 5.3: List of cross section uncertainties for each background process in the analysis. For single H_{SM} boson events the uncertainty depends on the production channel of the H_{SM}.

	$t\bar{t}$	Z+jets	single top	VV	single H_{SM}
Rate uncertainty	4.4%	2%	2.7%	5.6%	$1.3\% - 3.6\%$

5.2.3 Uncertainties on the F_F method

The different measurements performed for the F_F method are subject to systematic uncertainties related to the involved simulated events and to statistical fluctuations. In the following, it is explained how these uncertainties are propagated to the uncertainty model of the analysis.

The first uncertainties are related to the process fraction measurement. To account for the estimation of the fractions in the application region, each individual process fraction is scaled by $\pm 7\%$, while the remaining fractions are adjusted proportionally to maintain the total fraction sum of one. This procedure results in three variations of the estimation of misidentified τ_h which are treated as systematic uncertainties.

Next, uncertainties for the measured F_F^{QCD}, F_F^{W+jets} and $F_F^{t\bar{t}}$ are defined. Each F_F is measured with a linear fit of the form $F_F(p_T^{\tau_h}) = a \cdot p_T^{\tau_h} + b$. The uncertainties of the fit parameters a and b are used to construct two sources of uncertainty. The first source is a normalization defined by the variation of b. The second source can change the shape in p_T based on the variation of the slope parameter a. For each F_F measurement two such variations are added as systematic uncertainties to the analysis.

Another source of uncertainty arises in the measurements of F_F^{QCD} and F_F^{W+jets} from subtracting events other than QCD and W+jets events in the QCD and W+jets DRs, respectively. Any inaccuracies in the subtracted simulated events could bias the F_F measurement. To account for this, the combined shape of the subtracted events is scaled by $\pm 7\%$, and the F_F measurement is repeated. The resulting variation is treated as a systematic uncertainty for the corresponding F_F.

Statistical uncertainties associated with the non-closure corrections and the extrapolation from the DR to the SR are propagated into the analysis as independent systematic uncertainty for each correction. Since these corrections are not directly fitted but smoothed using Gaussian kernel smoothing, the uncertainties are constructed from an envelope around the smoothed curve. This envelope is derived through a Monte Carlo approach, where the measured data points of the correction histogram are shifted within their statistical uncertainties and the smoothing is repeated for each of these variations. The resulting $\pm 1\sigma$ intervals of this envelope are used as a systematic uncertainty in the analysis.

All mentioned uncertainty sources are defined independently in each analysis channel for both resolved and boosted $\tau\tau$. An example of some of these systematic uncertainties is shown in figure 5.5. In general, the variations can range from 1% to 10%. Especially, the uncertainties of the fraction measurements are relatively large, which is the result of the 7% up and down variation during the fraction measurement.

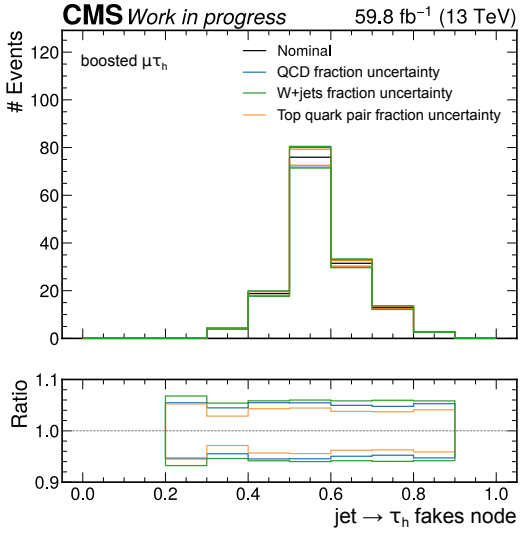
5.2.4 Statistical uncertainties

All background and signal estimation methods are inherently limited by their statistical precision. For simulated background and signal estimations, the number of simulated events constrains the accuracy of the prediction. Similarly, the F_F method is limited by the number of events in the application region. These limitations introduce systematic uncertainties that are statistical in nature and uncorrelated across all PNN categories and bins.

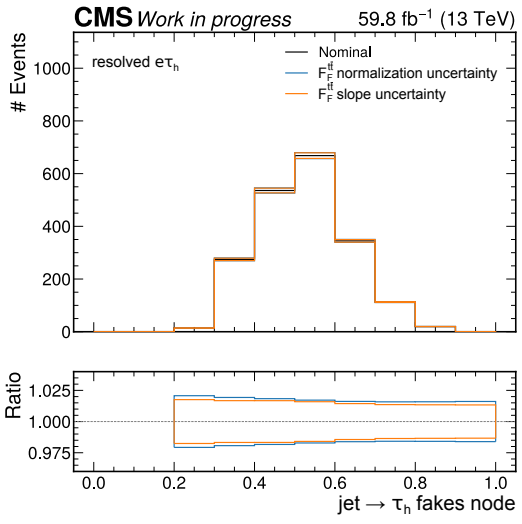
To account for these uncertainties, the Barlow-Beeston approach [109] is adapted. A single Gaussian nuisance parameter is introduced for each bin, allowing for variations in the predicted event yields. This method provides a simplified but effective framework for addressing statistical uncertainties. In this analysis, where the signal events are concentrated in a few bins with low background yield, these uncertainties can have a significant impact.

5.3 Results of the search

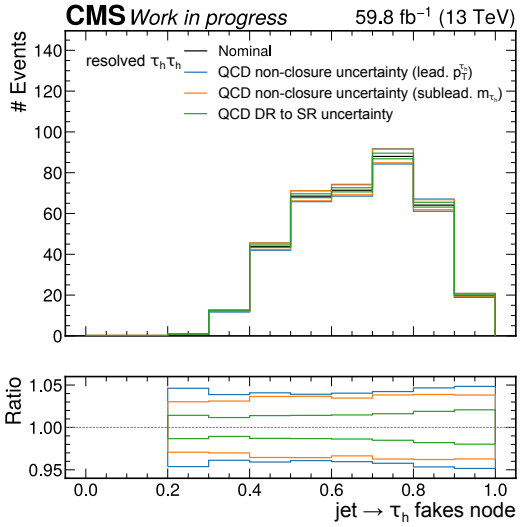
The conducted analysis has the goal to search for resonant di-Higgs production in bb and $\tau\tau$ final states, in particular in a new boosted phase space. A fit of the prediction model to data is performed to identify hypothetical excesses. In this analysis no excess is found. Therefore, upper limits of the production cross section times branching fractions are calculated for the two signal processes $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ and $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$, individually.



(a) Fraction uncertainties in boosted $\mu\tau_h$



(b) $F_F^{t\bar{t}}$ uncertainties in resolved $e\tau_h$



(c) F_F^{QCD} uncertainties in resolved $\tau_h\tau_h$

Figure 5.5: Variations of some F_F method uncertainties in the misidentified τ_h category. The binning is more granular than the distribution used in the final limit calculation. In (a) the fraction uncertainties of the boosted $\mu\tau_h$ channel are shown. Each variation is a result of a $\pm 7\%$ variation of the corresponding process in the fraction measurement. In (b) the $F_F^{t\bar{t}}$ measurement uncertainties of the resolved $e\tau_h$ channel are shown. The normalization and slope uncertainties are estimated from the linear fit of the $F_F^{t\bar{t}}$. In (c) the F_F^{QCD} correction uncertainties of the resolved $\tau_h\tau_h$ channel are shown. Variations of two non-closure corrections and one DR to SR correction are derived from the envelope of multiple smoothed curves. In general, the variations can range from $1\% - 10\%$.

For the upper limit measurement the test statistic is defined in a similar way as already introduced for the goodness-of-fit tests of the input variables (see section 5.1.5) by following

the Neyman-Pearson lemma [107] and using the likelihood function from equation 5.12.

$$q_\mu = -2 \cdot \ln \left(\frac{\mathcal{L}(\mu, \hat{\theta}_\mu)}{\mathcal{L}(\hat{\mu}, \hat{\theta}_{\hat{\mu}})} \right) \tag{5.15}$$

The parameters $\hat{\mu}$ and $\hat{\theta}_{\hat{\mu}}$ globally maximize the likelihood function $\mathcal{L}$ and thereby minimize the test statistic q_μ for a tested hypothesis μ. This test statistic is used in the CL_s method [110] to define a ratio of p-values, where once the signal hypothesis is tested with a signal strength modifier μ and once the background-only hypothesis is tested ($\mu = 0$). The CL_s value is defined as

$$CL_s = \frac{p_\mu}{1 - p_0} = \frac{\int_{q_{obs}}^{\infty} f(q_\mu | \mu, \hat{\theta}_\mu^{obs})}{\int_{q_{obs}}^{\infty} f(q_0 | 0, \hat{\theta}_0^{obs})} \tag{5.16}$$

where q_{obs} is the observed value of the test statistic in data and the p-values are calculated by integrating over the distributions f of the test statistics for the signal+background and background-only hypotheses. The CL_s method has advantages over relying only on the p-value p_μ of the signal hypothesis, as it avoids falsely excluding signal hypotheses when the analysis lacks sensitivity. This can happen when the signal strength modifier μ becomes very small. To set a limit on a signal hypothesis, the CL_s value must be $\leq \alpha = 0.05$, which defines the confidence level of the limit to 95%.

Before calculating the cross section limits for the two signal processes the validity of the prediction model is tested by performing goodness-of-fit tests including only background categories of the PNN output. These tests are performed for the 92 mass pair hypotheses for which a limit is calculated in this analysis (see figure 4.33). The results are shown in figure 5.6. Although some of the tests have a p-value smaller than 5%, the majority (around 78%) of the tested mass pair hypotheses show a good result. Therefore, it can be assumed that the fit model has no significant issues. However, the goodness-of-fit tests fail more often for low m_X. This observation could be investigated in a future study to further improve the fit model.

In figure 5.7 the four background categories are shown for the six analysis channels, evaluated for $m_X = 3000\,\text{GeV}$ and $m_Y = 1600\,\text{GeV}$. These distributions and uncertainties are the result of a maximum likelihood fit to the observed data. The background categories show very good agreement between the observed data and the prediction model. In almost all categories the most prominent process is $t\bar{t}$, however, the majority of $t\bar{t}$ events are still classified into the $t\bar{t}$ category. The other background processes are also primarily enriched in their dedicated categories. The largest confusion between background processes occurs for $t\bar{t}$ and jet $\rightarrow \tau_h$ fakes. The reason is that the misidentified τ_h events are enriched in $t\bar{t}$, as already identified from the fraction measurements of the F_F method (see section 4.6). For this mass pair hypothesis, an example of the efficiency and purity of the PNN categories in the resolved $\mu\tau_h$ channel is shown in figure B.8.

The signal categories for the same mass pair hypothesis are shown in figure 5.8. Across all tested mass pair hypotheses, including this example, no significant excess over the prediction model has been observed, therefore, no signal shape is drawn in the histograms. The categories

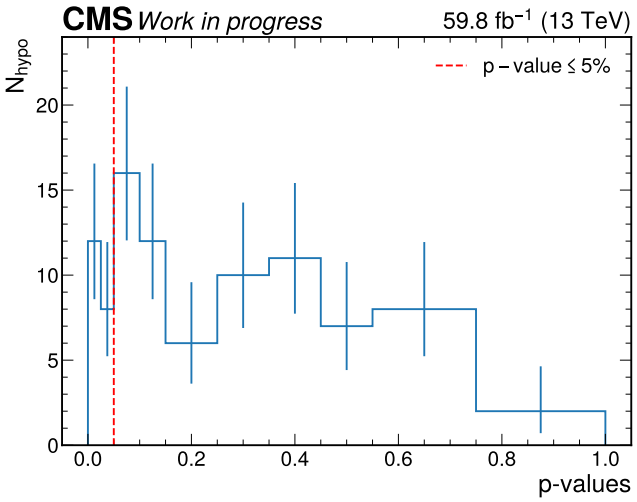


Figure 5.6: Goodness-of-fit results when only considering the background nodes of the PNNs. The p-values of the 92 tested mass pair hypotheses are histogrammed. For these 92 tests, 72 result in a p-value over 5%.

have very low event yields, even below a single event, which results in relatively large statistical uncertainties. Further, there are cases where no data events are present, however, this is not a problem for the limit calculation. For the shown mass pair hypothesis, it is expected that the Y boson decays in a resolved final state of tau lepton or b-jet pairs. For this reason, the boosted bb category of the $Y(bb)H_{SM}(\tau\tau)$ signal process is missing, even no background events are classified in this category. For the analysis channels in which data events are present, a reasonable agreement between the prediction model and data is visible.

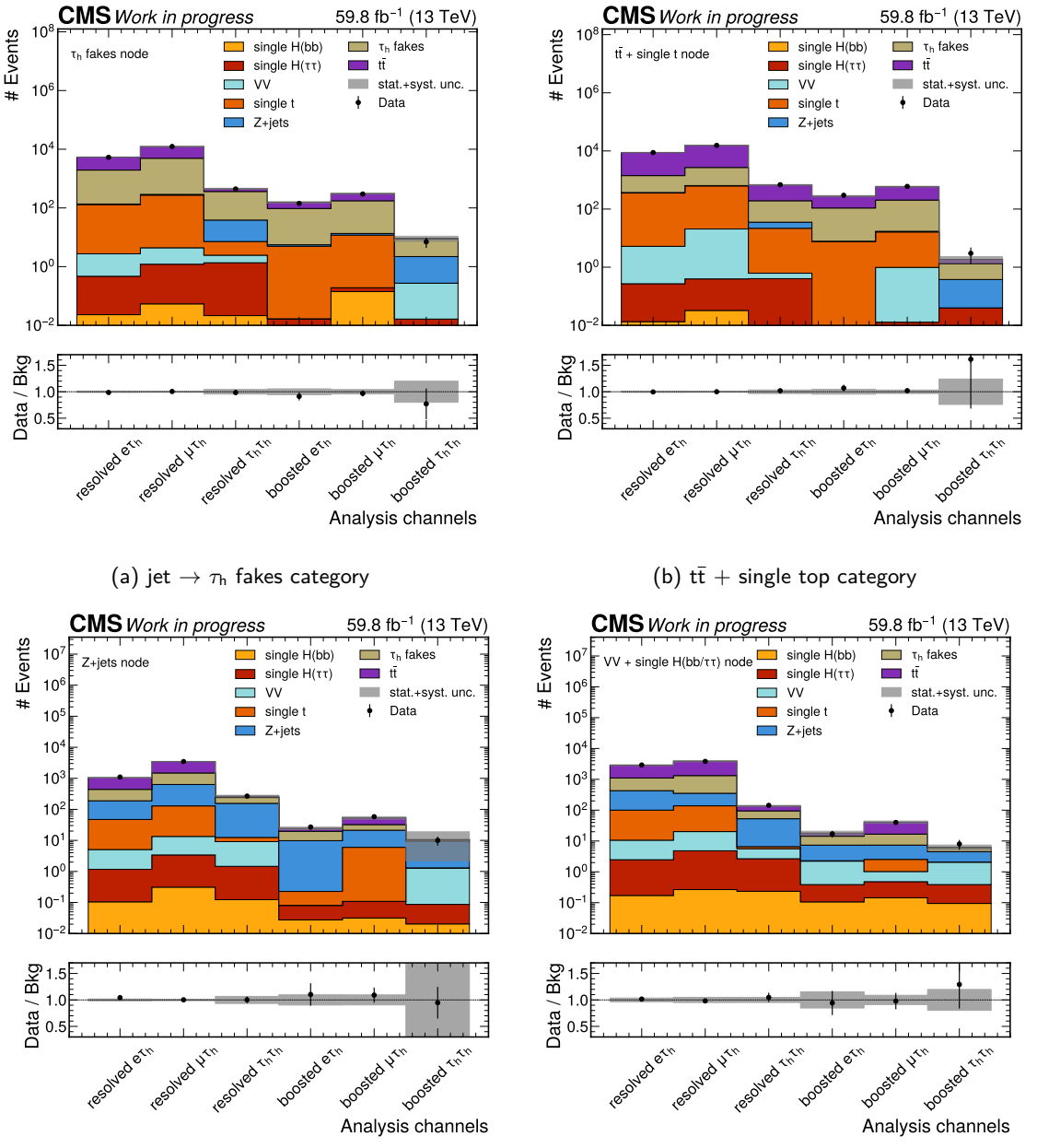
In the following the results of the limit calculations are discussed for the two signal processes independently. The $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ process is compared with the results of the previous analysis [9] to highlight the importance of the new boosted phase space, which is added in this analysis. Further, the new limit results for the $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$ process will be shown.

5.3.1 Upper limits for the $Y(bb)H_{SM}(\tau\tau)$ final state

To set exclusion limits on the product of the production cross section σ and branching fraction $\mathcal{B}$, a signal strength modifier μ is introduced, which is given by

$$\mu \propto \sigma(gg \to X) \times \mathcal{B}(X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)). \tag{5.17}$$

For this process, upper limits were measured in a previous analysis [9] that derived limits for the full Run-2 dataset of CMS, which corresponds to $137.2\,fb^{-1}$. This is a factor of about 2.3 more data than used in this analysis. Further, the NN used for classification, the final binning



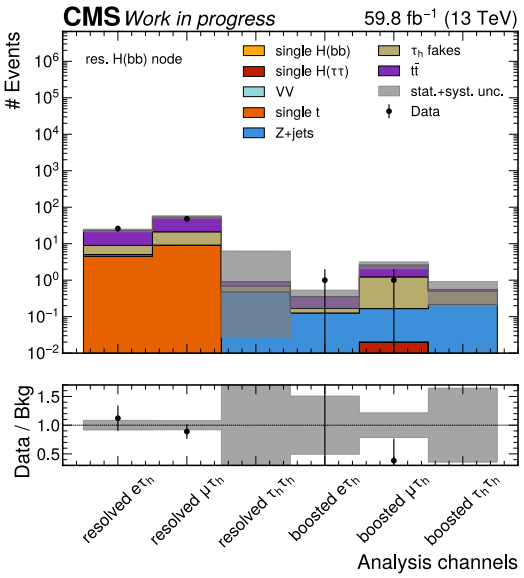
(a) jet $\rightarrow \tau_h$ fakes category

(b) $t\bar{t}$ + single top category

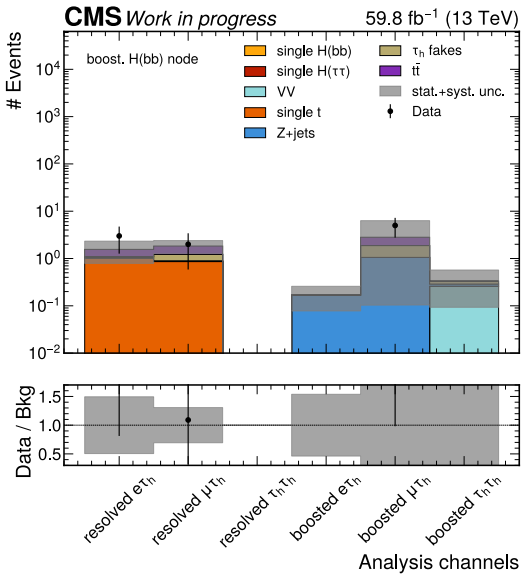
(c) Z+jets category

(d) Diboson + single H_{SM} category

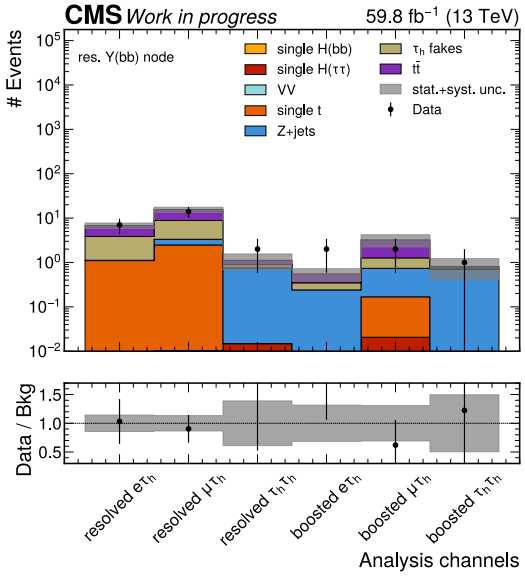
Figure 5.7: Background categories for all analysis channels based on the evaluation for $m_X = 3000\,\text{GeV}$ and $m_Y = 1600\,\text{GeV}$. In (a) the category of the F_F method is shown with the corresponding process named "τ_h fakes". In (b) the $t\bar{t}$ and single top category is shown, in (c) the Z+jets category and in (d) the diboson and single Higgs boson category. The uncertainties on the bins reflect the systematic and statistical uncertainties after a maximum likelihood fit to data.



(a) $Y(\tau\tau)H_{SM}(bb)$ with resolved bb category



(b) $Y(\tau\tau)H_{SM}(bb)$ with boosted bb category



(c) $Y(bb)H_{SM}(\tau\tau)$ with resolved bb category

Figure 5.8: Signal categories for all analysis channels based on the evaluation for $m_X = 3000\,\text{GeV}$ and $m_Y = 1600\,\text{GeV}$. In (a) and (b) the two categories of the $Y(\tau\tau)H_{SM}(bb)$ signal process are shown, in (a) for the resolved bb and in (b) for the boosted bb final states. In (c) the category of the $Y(bb)H_{SM}(\tau\tau)$ signal process is shown for the resolved bb final state. The boosted bb categories of this process is not present because no events are classified into this category. The uncertainties on the bins reflect the systematic and statistical uncertainties after a maximum likelihood fit to data.

of the NN predictions and the event selections were different compared to this analysis. The previous analysis contains more data and individually optimized NNs instead of a single PNN for all mass hypotheses. On the other hand, this analysis includes a boosted selection of the final state particle pairs, while the previous analysis only considered resolved tau lepton and b-jet pairs.

All these differences between this and the previous analysis are part of the comparison in figure 5.9. There, upper limits are shown for two different mass points of the X boson. Based on figure 4.6a, it can be concluded that for $m_X = 800\,\text{GeV}$, the final state particle pairs are most likely in a resolved state. This behavior is precisely reflected in figure 5.9a, where the expected upper limits for all analysis channels and for only resolved channels are close to each other, which means that the boosted categories do not contribute to the sensitivity of the analysis. Only for very small values of m_Y is the effect of the boosted categories visible, where both limit expectations diverge. Further, the limits for this m_X hypothesis in this analysis are still less stringent than the previously set upper limits. The difference of almost one order of magnitude can partially be explained by the smaller dataset analyze in this analysis. However, the more significant reason is the much simpler PNN-based setup of the final category discriminants. These findings are valid for all mass pair hypotheses where resolved final state particle pairs are expected.

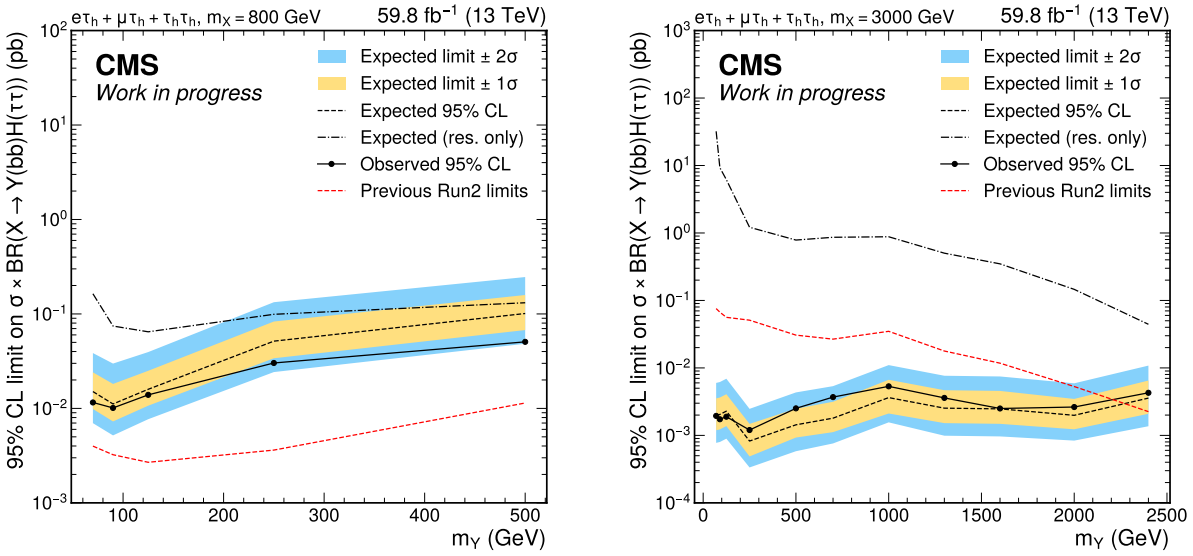
A different picture can be seen in figure 5.9b. For an X boson mass of $3000\,\text{GeV}$, almost all final state particle pairs are expected to be boosted. The upper limits calculated considering only resolved categories are much worse compared to both the previously set limits and the new limits with resolved and boosted signal events. Further, despite the fact that a smaller dataset is used and the setup of the classifier and definition of the final discriminants is less sophisticated, this analysis can already improve the previously measured upper limits by more than one order of magnitude. Of course, this is only valid for mass pair hypotheses for which boosted final state particle pairs are expected.

A summary plot of all 92 upper limit measurements is shown in figure 5.10. Further, in a more detailed view, the measurements are displayed in the appendix sections C.9 - C.11, together with a two dimensional summary histogram of the observed upper limits in figure C.12. If a signal were present, an excess of the observation from the expectation would be visible. The highest fluctuation is around 1.8 standard deviations (σ) for the mass pair hypothesis $m_X = 3500\,\text{GeV}$, $m_X = 1000\,\text{GeV}$. In summary, the observations are statistically consistent with the background-only expectation.

5.3.2 Upper limits for the $\mathbf{Y(\tau\tau)H_{SM}(bb)}$ final state

For the second signal process the measurement of the upper limits follows the same procedure. To set exclusion limits on the production cross section σ times branching fraction $\mathcal{B}$, a signal strength modifier μ is introduced, which is given by

$$\mu \propto \sigma(gg \to X) \times \mathcal{B}(X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})). \tag{5.18}$$



(a) Upper limits for $m_X = 800\,\text{GeV}$

(b) Upper limits for $m_X = 3000\,\text{GeV}$

Figure 5.9: Upper limits on $\sigma(gg \to X) \times \mathcal{B}(X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+))$ for two different mass hypotheses for the X boson. In (a) the limits for $m_X = 800\,\text{GeV}$ are shown and in (b) for $m_X = 3000\,\text{GeV}$. Besides the expected 95% confidence interval upper limits, the expected upper limits are shown when only resolved categories are considered. This is the closest setup that can be compared to the limit results from the previous analysis [9], which are shown as well.

A summary plot of all 92 upper limit measurements is shown in figure 5.11. Further, in a more detailed view the measurements are shown in the appendix sections D.13 - D.15, together with a two dimensional summary histogram of the observed upper limits in figure D.16. The highest fluctuation is around 2 standard deviations (σ) for the mass pair hypothesis $m_X = 320\,\text{GeV}$, $m_X = 70\,\text{GeV}$. In summary, the observations are statistically consistent with the background-only expectation.

Most of the mass pair hypotheses have asymmetric masses of the Y and H_{SM} bosons, however, the case where $m_Y = 125\,\text{GeV}$ is considered as well in this analysis. Studies showed that the PNNs cannot differentiate between the two signal processes for symmetric masses of the Y and H_{SM} bosons. Therefore, signal events are randomly classified into the signal categories. For the measurement of the upper limits this leads to very similar results, independent of the signal process. In figure 5.12 the results of this analysis are compared with other analyses, which targeted the symmetric mass pair combinations $X \to H_{SM}(bb)H_{SM}(\tau\tau)$. First, there are the results of the previous search [9], however, only the $m_Y = 120\,\text{GeV}$ mass point was part of the analysis. Furthermore, the CMS collaboration has conducted another search specifically targeting the $m_Y = m_{H_{SM}}$ case [71], but only the result with the 2016 dataset was published until now, corresponding to $35.9\,\text{fb}^{-1}$. The last measurement shown in the figure is the search

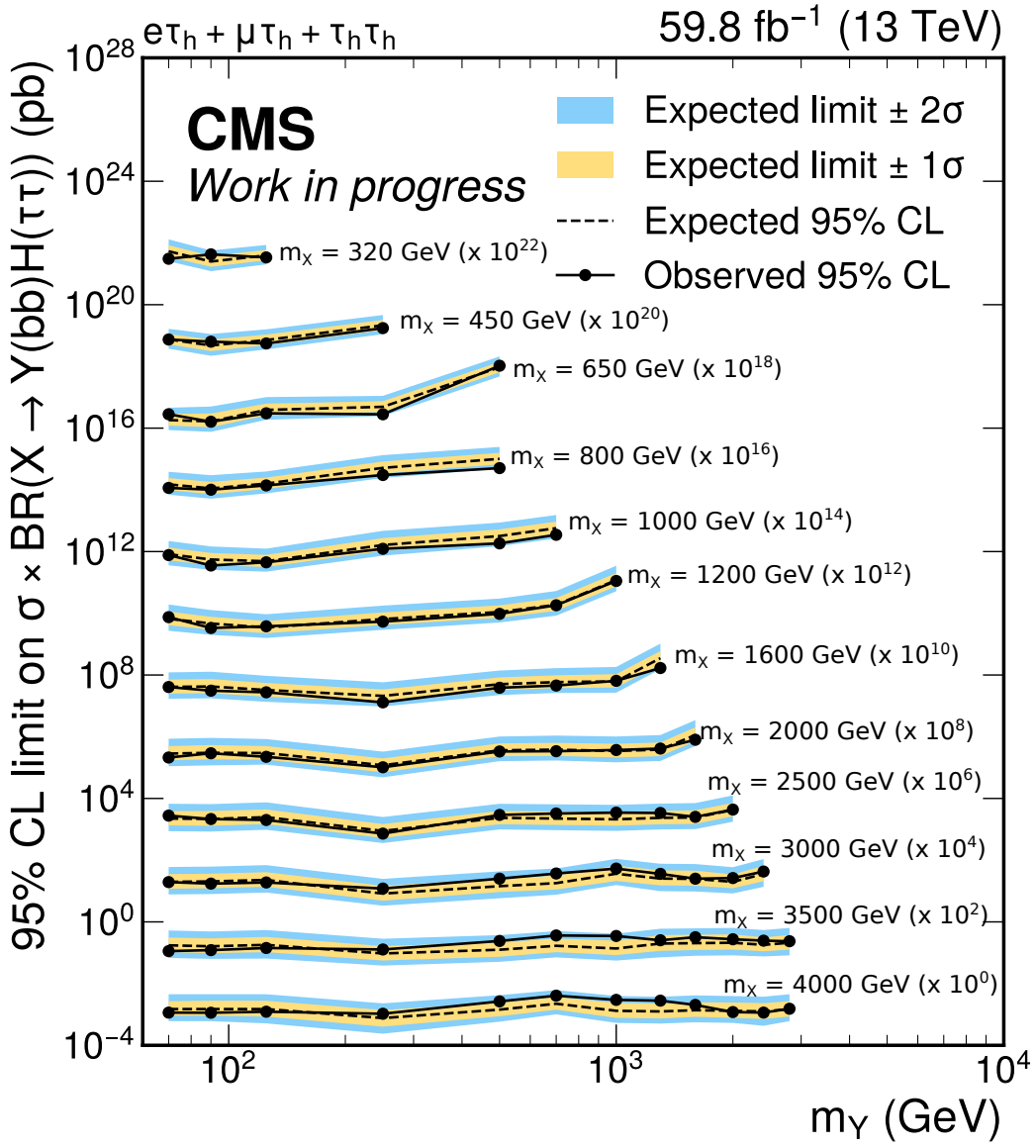


Figure 5.10: Observed and expected 95% confidence level upper limits on $\sigma(gg \to X) \times \mathcal{B}(X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+))$. A scaling (indicated in parentheses) is applied in order to display the results in a single figure. The expected limits are shown as a dashed line with the 68% and 95% confidence interval of the expectation given by the yellow and blue bands. The observation is shown by black points. No deviation beyond the 2 standard deviations level is found for all 92 tested mass pair hypotheses.

by the ATLAS collaboration using the full Run-2 dataset [72], which has an integrated luminosity of $139\,\text{fb}^{-1}$. The ATLAS search found an excess with 3.1σ (2σ) local (global) significance for a resonant mass of $m_X = 1000\,\text{GeV}$.

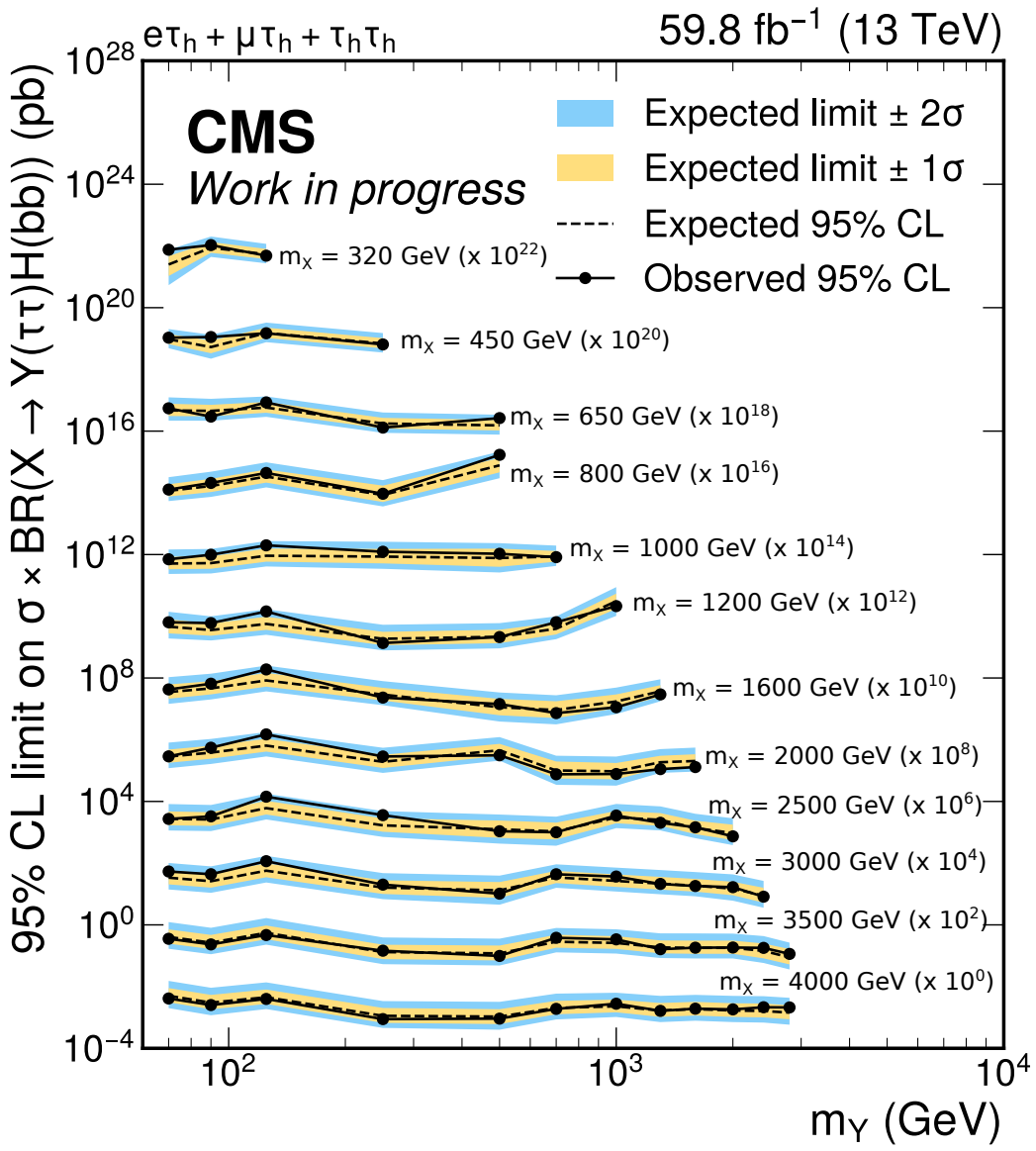


Figure 5.11: Observed and expected 95% confidence level upper limits on $\sigma(gg \to X) \times \mathcal{B}(X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b}))$. A scaling (indicated in parentheses) is applied in order to display the results in a single figure. The expected limits are shown as a dashed line with the 68% and 95% confidence interval of the expectation given by the yellow and blue bands. The observation is shown by black points. No deviation beyond the 2 standard deviations level is found for all 92 tested mass pair hypotheses.

The results of this analysis are less sensitive than the upper limits measured in the other searches for X boson masses lower than 1200 GeV. However, for higher masses this analysis can still maintain its sensitivity due to the newly added boosted categories, while the other searches

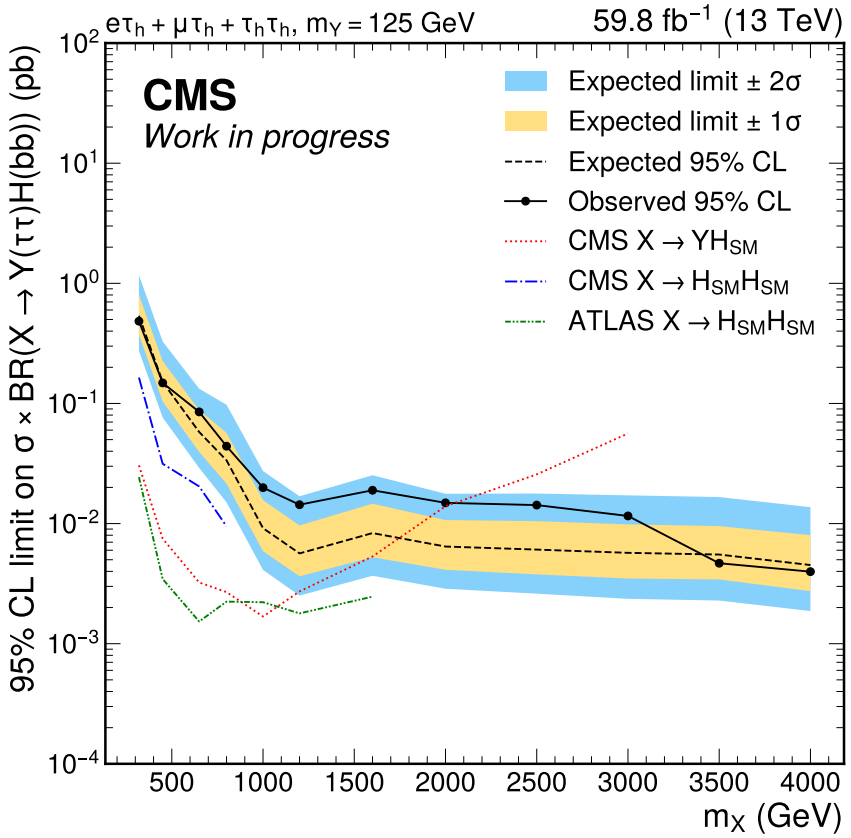


Figure 5.12: Observed and expected 95% confidence level upper limits on $\sigma(gg \to X) \times \mathcal{B}(X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b}))$ for the symmetric case when $m_Y = m_{H_{SM}}$. Additionally, other published results from the CMS and ATLAS collaborations are shown. The predecessor analysis [9] of this search is shown in red for $m_Y = 120\,\text{GeV}$. In darker blue, the limits from the CMS $X \to H_{SM}(bb)H_{SM}(\tau\tau)$ search [71] is shown, which considered the 2016 dataset with $35.9\,\text{fb}^{-1}$. In green, the limits obtained in the full Run-2 ATLAS $X \to H_{SM}(bb)H_{SM}(\tau\tau)$ search [72] is shown.

lose sensitivity because they are only focused on resolved final state topologies. Taking into account that this analysis still has room for improvement, for example by just adding more data from CMS Run-2, this comparison should change in the near future.

6 Summary and Outlook

Particle physics seeks to uncover the fundamental building blocks of the universe and their interactions. The discovery of the Higgs boson at the LHC in 2012 [4, 5] was a milestone, confirming the mechanism responsible for giving mass to SM particles. Current research topics, such as di-Higgs production at the LHC, aim to explore the Higgs sector further, like measuring the Higgs boson self-coupling and probe phenomena beyond the SM, potentially discovering new physics.

The goal of this thesis is to set an intermediate milestone in the search for resonant di-Higgs production. This thesis covers the cases where a heavy Higgs boson X decays into a pair of 125 GeV Higgs bosons (H_{SM}) as well as the decay into the H_{SM} boson and an additional light Higgs boson Y. A new event selection strategy is introduced that targets boosted topologies of the final state particles, which occur primarily for high m_X. By that, this analysis allows to test extensions of the SM Higgs sector as proposed, for example, by theories like the NMSSM. In this context, in this analysis, improved upper limits are set, compared to the previous CMS search [9], on the production cross section times branching fraction on the $gg \to X \to Y(b\bar{b})H_{SM}(\tau^-\tau^+)$ process for $m_X > 2\,\text{TeV}$ and $m_Y < 2\,\text{TeV}$. Further, the analysis is extended to additionally measure upper limits on the $gg \to X \to Y(\tau^-\tau^+)H_{SM}(b\bar{b})$ process, which has not been done before. This analysis provides a new baseline for searches for these processes, which will be extended in future searches with more data from the second and third runs of the LHC.

The challenging aspect of this analysis is the combination of the different possible bb and $\tau\tau$ final states. The kinematic properties of the signal processes vary significantly with the change of the mass hypotheses of the X and Y bosons. If a particle decays into two new particles, these two decay products can get a significant Lorentz boost, especially if the mass difference of the initial particle and the decay products is large. This can be referred to the $X \to YH_{SM}$ decay as well as the decay of the Y and H_{SM} boson into the final state particles (bb and $\tau\tau$). These boosted decay products are often too close to each other to be resolved individually. Therefore, a dedicated reconstruction and identification approach for each combination of resolved and

boosted topologies of the final state particles is required. Especially, boosted final states of the tau lepton and b-jet pairs were not explored in the scope of resonant di-Higgs searches before and started to emerge only recently due to modern identification algorithms such as *ParticleNet* [62].

Further challenges include the accurate estimation of background processes that have a similar event signature and the final event classification task, considering the large number of tested X and Y masses ($\mathcal{O}(100)$). One of the main background contributions are events in which jets are misidentified as hadronically decaying tau leptons (τ_h) and this background is estimated in a data-driven way using the fake factor (F_F) method. For this analysis, two sets of phase space specific F_F's have been measured targeting resolved and boosted $\tau\tau$ pairs, respectively. The challenge of the event classification is tackled by a parametric neural network, where the two signal mass hypotheses (m_X and m_Y) are used as input parameters for the neural network, thus reducing the training effort of individual neural networks for each mass combination to a single neural network.

Compared to the upper limit measurements published by ATLAS and CMS so far [9, 71, 72], this analysis is still lacking in terms of sensitivity in the resolved bb and $\tau\tau$ final state sectors of the two dimensional m_X and m_Y plane. However, this analysis aims to achieve a more comprehensive selection of signal events by including all possible final state topologies. This results in a comparable or even better sensitivity when looking at the boosted bb and $\tau\tau$ final state sectors. For example, the upper limit on the cross section times branching fraction for the mass combination ($m_\mathrm{X} = 3000\,\mathrm{GeV}$, $m_\mathrm{Y} = 250\,\mathrm{GeV}$) is measured to be $\sigma(\mathrm{gg} \to \mathrm{X}) \times \mathcal{B}(\mathrm{X} \to \mathrm{Y}(\mathrm{b\bar{b}})\mathrm{H_{SM}}(\tau^-\tau^+)) < 1.2\,\mathrm{fb}$, which is about 40 times lower than the previous result.

This analysis is a snapshot of the currently ongoing analysis in CMS. New developments of the *ParticleNet* algorithm targeting the boosted $\tau\tau$ pair identification will soon be available for analyses. Since in this analysis a relatively old approach is used to identify boosted $\tau\tau$ pairs, an improvement in sensitivity is expected. Further, more thorough studies of the kinematic fit and PNNs as well as a more refined binning of the final discriminants are also expected to contribute to a better sensitivity of the analysis. In the near future, more data will be added to the analysis, including the 2016 and 2017 data taking periods of the LHC Run-2 and from the currently ongoing Run-3 data taking. This alone will increase the analyzed dataset by five times. Finally, the results of this analysis only cover 92 of 574 available (m_X, m_Y) combinations, therefore, the coverage of X/Y masses will increase in the final CMS analysis.

Bibliography

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

[2] P. Higgs. "Broken symmetries, massless particles and gauge fields". In: *Physics Letters* 12.2 (1964), pp. 132–133. DOI: https://doi.org/10.1016/0031-9163(64)91136-9.

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

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

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

[6] D. Chung et al. "The soft supersymmetry-breaking Lagrangian: theory and applications". In: *Physics Reports* 407.1 (2005), pp. 1–203. DOI: 10.1016/j.physrep.2004.08.032.

[7] U. Ellwanger, C. Hugonie, and A. M. Teixeira. "The Next-to-Minimal Supersymmetric Standard Model". In: *Physics Reports* 496.1 (2010), pp. 1–77. DOI: 10.1016/j.physrep.2010.07.001.

[8] M. Maniatis. "THE NEXT-TO-MINIMAL SUPERSYMMETRIC EXTENSION OF THE STANDARD MODEL REVIEWED". In: *International Journal of Modern Physics A* 25.18n19 (2010), pp. 3505–3602. DOI: 10.1142/S0217751X10049827.

[9] The CMS collaboration. "Search for a heavy Higgs boson decaying into two lighter Higgs bosons in the $\tau\tau$bb final state at 13 TeV". In: *Journal of High Energy Physics* 2021 (11 Nov. 2021), p. 57. DOI: 10.1007/JHEP11(2021)057.

[10] Particle Data Group. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2022.8 (Aug. 2022), p. 083C01. DOI: 10.1093/ptep/ptac097.

[11] *Standard Model of Elementary Particles*. https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg. Accessed: 28.11.2024.

[12] S. L. Glashow. "Partial-symmetries of weak interactions". In: *Nuclear Physics* 22.4 (1961), pp. 579–588. DOI: https://doi.org/10.1016/0029-5582(61)90469-2.

[13] A. Salam and J. Ward. "Electromagnetic and weak interactions". In: *Physics Letters* 13.2 (1964), pp. 168–171. DOI: https://doi.org/10.1016/0031-9163(64)90711-5.

[14] S. Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (21 Nov. 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[15] S. D. Bass, A. De Roeck, and M. Kado. "The Higgs boson implications and prospects for future discoveries". In: *Progress of Theoretical and Experimental Physics* 3 (9 2021), p. 608. DOI: 10.1038/s42254-021-00341-2.

[16] H. Yukawa. "On the Interaction of Elementary Particles. I". In: *Progress of Theoretical Physics Supplement* 1 (1955), pp. 1–10. DOI: 10.1143/PTPS.1.1.

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

[18] M. Kobayashi and T. Maskawa. "CP-Violation in the Renormalizable Theory of Weak Interaction". In: *Progress of Theoretical Physics* 49.2 (1973), pp. 652–657. DOI: 10.1143/PTP.49.652.

[19] P. van Nieuwenhuizen. "Supergravity". In: *Physics Reports* 68.4 (1981), pp. 189–398. DOI: 10.1016/0370-1573(81)90157-5.

[20] H. Pagels and J. R. Primack. "Supersymmetry, Cosmology, and New Physics at Teraelectronvolt Energies". In: *Phys. Rev. Lett.* 48 (4 Jan. 1982), pp. 223–226. DOI: 10.1103/PhysRevLett.48.223.

[21] J. Ellis, S. Kelley, and D. Nanopoulos. "Probing the desert using gauge coupling unification". In: *Physics Letters B* 260.1 (1991), pp. 131–137. DOI: 10.1016/0370-2693(91)90980-5.

[22] U. Amaldi, W. de Boer, and H. Fürstenau. "Comparison of grand unified theories with electroweak and strong coupling constants measured at LEP". In: *Physics Letters B* 260.3 (1991), pp. 447–455. DOI: https://doi.org/10.1016/0370-2693(91)91641-8.

[23] E. Witten. "Dynamical breaking of supersymmetry". In: *Nuclear Physics B* 188.3 (1981), pp. 513–554. DOI: 10.1016/0550-3213(81)90006-7.

[24] S. Dimopoulos and H. Georgi. "Softly broken supersymmetry and SU(5)". In: *Nuclear Physics B* 193.1 (1981), pp. 150–162. DOI: 10.1016/0550-3213(81)90522-8.

[25] J. Casas and C. Muñoz. "A natural solution to the μ problem". In: *Physics Letters B* 306.3 (1993), pp. 288–294. DOI: 10.1016/0370-2693(93)90081-R.

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

[27] E. Lopienska. "The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022". In: (2022). URL: https://cds.cern.ch/record/2800984.

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

[29] The ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003.

[30] T. Sakuma and T. McCauley. "Detector and Event Visualization with SketchUp at the CMS Experiment". In: *Journal of Physics: Conference Series* 513.2 (June 2014), p. 022032. DOI: 10.1088/1742-6596/513/2/022032.

[31] M. Kim et al. "Web-based monitoring tools for Resistive Plate Chambers in the CMS experiment at CERN". In: *Journal of Instrumentation* 9.10 (Oct. 2014), p. C10031. DOI: 10.1088/1748-0221/9/10/C10031.

[32] The CMS Collaboration. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". In: *Journal of Instrumentation* 9.10 (Oct. 2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009.

[33] *CMS Tracker Detector Performance Public Results*. Twiki: https://twiki.cern.ch/twiki/bin/view/CMSPublic/DPGResultsTRK. Accessed: 14.11.2024.

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

[35] A. Benaglia. "The CMS ECAL performance with examples". In: *Journal of Instrumentation* 9.02 (Feb. 2014), p. C02008. DOI: 10.1088/1748-0221/9/02/C02008.

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

[37] J. Mans et al. *CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter*. Tech. rep. CERN, 2012. URL: https://cds.cern.ch/record/1481837.

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

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

[40] The CMS collaboration. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of Instrumentation* 13.06 (June 2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015.

[41] The CMS Collaboration. "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of Instrumentation* 15.10 (Oct. 2020), P10017. DOI: 10.1088/1748-0221/15/10/P10017.

[42] The CMS Collaboration. "The CMS trigger system". In: *Journal of Instrumentation* 12.01 (Jan. 2017), P01020. DOI: 10.1088/1748-0221/12/01/P01020.

[43] R. Frühwirth. "Application of Kalman filtering to track and vertex fitting". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 262.2 (1987), pp. 444–450. DOI: 10.1016/0168-9002(87)90887-4.

[44] K. Rose. "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2210–2239. DOI: 10.1109/5.726788.

[45] R. Frühwirth, W. Waltenberger, and P. Vanlaer. *Adaptive Vertex Fitting*. Tech. rep. CERN, 2007. URL: https://cds.cern.ch/record/1027031.

[46] The CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector". In: *Journal of Instrumentation* 12.10 (Oct. 2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003.

[47] The CMS collaboration. "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 16.05 (May 2021), P05014. DOI: 10.1088/1748-0221/16/05/P05014.

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

[49] M. Cacciari, G. P. Salam, and G. Soyez. "The anti-kt jet clustering algorithm". In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063.

[50] Y. L. Dokshitzer et al. "Better jet clustering algorithms". In: *Journal of High Energy Physics* 1997.08 (Sept. 1997), p. 001. DOI: 10.1088/1126-6708/1997/08/001.

[51] D. Bertolini et al. "Pileup per particle identification". In: *Journal of High Energy Physics* 2014.10 (Oct. 2014), p. 59. DOI: 10.1007/JHEP10(2014)059.

[52] The CMS collaboration. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *Journal of Instrumentation* 12.02 (Feb. 2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014.

[53] A. J. Larkoski et al. "Soft drop". In: *Journal of High Energy Physics* 2014.5 (May 2014), p. 146. DOI: 10.1007/JHEP05(2014)146.

[54] The CMS Collaboration. "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV". In: *Journal of Instrumentation* 10.06 (June 2015), P06005. DOI: 10.1088/1748-0221/10/06/P06005.

[55] The CMS Collaboration. "Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV". In: *Journal of Instrumentation* 13.10 (Oct. 2018), P10005. DOI: 10.1088/1748-0221/13/10/P10005.

[56] S. Navas et al. (Particle Data Group). "Review of particle physics". In: *Phys. Rev. D* 110.3 (2024), p. 030001. DOI: 10.1103/PhysRevD.110.030001.

[57] K. Androsov. *Identification of tau leptons using Deep Learning techniques at CMS*. Tech. rep. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2713735.

[58] The CMS Collaboration. "Identification of hadronic tau lepton decays using a deep neural network". In: *Journal of Instrumentation* 17.07 (July 2022), P07023. DOI: 10.1088/1748-0221/17/07/P07023.

[59] The CMS Collaboration. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *Journal of Instrumentation* 13.05 (May 2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011.

[60] E. Bols et al. "Jet flavour classification using DeepJet". In: *Journal of Instrumentation* 15.12 (Dec. 2020), P12012. DOI: 10.1088/1748-0221/15/12/P12012.

[61] The CMS Collaboration. *A deep neural network for simultaneous estimation of b quark energy and resolution*. Tech. rep. Geneva: CERN, 2019. URL: http://cds.cern.ch/record/2690804.

[62] H. Qu and L. Gouskos. "Jet tagging via particle clouds". In: *Phys. Rev. D* 101 (5 Mar. 2020), p. 056019. DOI: 10.1103/PhysRevD.101.056019.

[63] The CMS Collaboration. *Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques*. Tech. rep. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2707946.

[64] Y. Wang et al. "Dynamic Graph CNN for Learning on Point Clouds". In: *ACM Trans. Graph.* 38.5 (Oct. 2019). DOI: 10.1145/3326362.

[65] The CMS Collaboration. *Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment*. Tech. rep. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2683870.

[66] The CMS Collaboration. *Performance of heavy-flavour jet identification in boosted topologies in proton-proton collisions at* $\sqrt{s} = 13$ TeV. Tech. rep. Geneva: CERN, 2023. URL: https://cds.cern.ch/record/2866276.

[67] The ATLAS Collaboration. "Combination of Searches for Higgs Boson Pair Production in *pp* Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector". In: *Phys. Rev. Lett.* 133 (10 Sept. 2024), p. 101801. DOI: 10.1103/PhysRevLett.133.101801.

[68] The CMS Collaboration. "A portrait of the Higgs boson by the CMS experiment ten years after the discovery". In: *Nature* 607 (7917 July 2022), p. 60. DOI: 10.1038/s41586-022-04892-x.

[69] O. Aberle et al. *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*. Tech. rep. Geneva: CERN, 2020. DOI: 10.23731/CYRM-2020-0010. URL: https://cds.cern.ch/record/2749422.

[70] A. Dainese et al. *Report on the Physics at the HL-LHC, and Perspectives for the HE-LHC*. Tech. rep. Geneva, Switzerland: CERN, 2019. DOI: 10.23731/CYRM-2019-007. URL: https://cds.cern.ch/record/2703572.

[71] The CMS collaboration. "Search for Higgs boson pair production in events with two bottom quarks and two tau leptons in proton–proton collisions at $\sqrt{s} = 13\,\text{TeV}$". In: *Physics Letters B* 778 (2018), pp. 101–127. DOI: 10.1016/j.physletb.2018.01.001.

[72] The ATLAS collaboration. "Search for resonant and non-resonant Higgs boson pair production in the $b\bar{b}\tau^+\tau^-$ decay channel using 13 TeV pp collision data from the ATLAS detector". In: *Journal of High Energy Physics* 2023 (7 2023), p. 40. DOI: 10.1007/JHEP07(2023)040.

[73] K. Agashe et al. "Warped gravitons at the CERN LHC and beyond". In: *Phys. Rev. D* 76 (3 Aug. 2007), p. 036006. DOI: 10.1103/PhysRevD.76.036006.

[74] L. Fitzpatrick et al. "Searching for the Kaluza-Klein graviton in bulk RS models". In: *Journal of High Energy Physics* 2007.09 (Sept. 2007), p. 013. DOI: 10.1088/1126-6708/2007/09/013.

[75] D. Berdine, N. Kauer, and D. Rainwater. "Breakdown of the Narrow Width Approximation for New Physics". In: *Phys. Rev. Lett.* 99 (11 Sept. 2007), p. 111601. DOI: 10.1103/PhysRevLett.99.111601.

[76] M. D. H. Marz. "Studies of boosted topologies in the search for NMSSM inspired di-Higgs events in the tautau + bb final state with the CMS experiment". MA thesis. Karlsruhe Institute of Technology (KIT), 2022. URL: https://publish.etp.kit.edu/record/22129.

[77] M. Freudig. "Generator studies in the context of the NMSSM di-Higgs analysis with bb$\tau\tau$ final states at CMS". BA thesis. Karlsruhe Institute of Technology (KIT), 2022. URL: https://publish.etp.kit.edu/record/22140.

[78] M. Czakon and A. Mitov. "Top++: A program for the calculation of the top-pair cross-section at hadron colliders". In: *Computer Physics Communications* 185.11 (2014), pp. 2930–2938. DOI: 10.1016/j.cpc.2014.06.021.

[79] J. Campbell, T. Neumann, and Z. Sullivan. "Single-top-quark production in the t-channel at NNLO". In: *Journal of High Energy Physics* 2021 (2 2021), p. 40. DOI: 10.1007/JHEP02(2021)040.

[80] N. Kidonakis and N. Yamanaka. "Higher-order corrections for tW production at high-energy hadron colliders". In: *Journal of High Energy Physics* 2021 (5 2021), p. 278. DOI: 10.1007/JHEP05(2021)278.

[81] The CMS Collaboration. "Measurement of the $Z/\gamma^* \to \tau\tau$ cross section in pp collisions at $\sqrt{s} = 13\,\text{TeV}$ and validation of τ lepton analysis techniques". In: *The European Physical Journal C* 78 (9 2021), p. 708. DOI: 10.1140/epjc/s10052-018-6146-9.

[82] The CMS Collaboration. "Measurements of Higgs boson production in the decay channel with a pair of τ leptons in proton–proton collisions at $\sqrt{s} = 13\,\text{TeV}$". In: *The European Physical Journal C* 83 (7 2023), p. 562. DOI: 10.1140/epjc/s10052-023-11452-8.

[83] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *Journal of High Energy Physics* 2014 (7 2014), p. 79. DOI: 10.1007/JHEP05(2016)100.

[84] S. Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *Journal of High Energy Physics* 2010 (6 2010), p. 43. DOI: 10.1007/JHEP06(2010)043.

[85] T. Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024.

[86] R. D. Ball et al. "Parton distributions from high-precision collider data". In: *The European Physical Journal C* 77 (10 2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5.

[87] S. Agostinelli et al. "Geant4—a simulation toolkit". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[88] E. Conte et al. "Investigating light NMSSM pseudoscalar states with boosted ditau tagging". In: *Journal of High Energy Physics* 2016 (5 2016), p. 100. DOI: 10.1007/JHEP05(2016)100.

[89] The CMS Collaboration. "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2018 (7 2018), p. 161. DOI: 10.1007/JHEP07(2018)161.

[90] B. Schmiederer. "Trigger Efficiency Studies in the Scope of a Search for Next-to-Minimal Supersymmetric Standard Model Di-Higgs Production with the CMS Experiment". BA thesis. Karlsruhe Institute of Technology (KIT), 2024. URL: https://publish.etp.kit.edu/record/22233.

[91] The CMS Collaboration. "Measurement of differential cross sections for top quark pair production using the lepton + jets final state in proton-proton collisions at 13 TeV". In: *Phys. Rev. D* 95 (9 May 2017), p. 092001. DOI: 10.1103/PhysRevD.95.092001.

[92] The CMS Collaboration. "Measurements of $\mathrm{t\bar{t}}$ differential cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV using events containing two leptons". In: *Journal of High Energy Physics* 2019 (2 2019), p. 149. DOI: 10.1007/JHEP02(2019)149.

[93] L. Bianchini et al. "Reconstruction of the Higgs mass in events with Higgs bosons decaying into a pair of τ leptons using matrix element techniques". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 862 (2017), pp. 54–84. DOI: 10.1016/j.nima.2017.05.001.

[94] K. Kondo. "Dynamical Likelihood Method for Reconstruction of Events with Missing Momentum. I. Method and Toy Models". In: *Journal of the Physical Society of Japan* 57.12 (1988), pp. 4126–4140. DOI: 10.1143/JPSJ.57.4126.

[95] K. Kondo. "Dynamical Likelihood Method for Reconstruction of Events with Missing Momentum. II. Mass Spectra for 2→2 Processes". In: *Journal of the Physical Society of Japan* 60.3 (1991), pp. 836–844. DOI: 10.1143/JPSJ.60.836.

[96] M. Molch. "Reconstruction of the di-tau system in H $\rightarrow \tau\tau$ decays with machine learning methods". MA thesis. Karlsruhe Institute of Technology (KIT), 2023. URL: https://publish.etp.kit.edu/record/22201.

[97] M. Hoffmann et al. *HHKinFit – a kinematic fitting package to fit heavy Higgs decays*. Available on CMS information server: CMS-AN-2014/163. Accessed: 20.11.2024.

[98] The CMS Collaboration. "Searches for a heavy scalar boson H decaying to a pair of 125 GeV Higgs bosons hh or for a heavy pseudoscalar boson A decaying to Zh, in the final states with h $\rightarrow \tau\tau$". In: *Physics Letters B* 755 (2016), pp. 217–244. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2016.01.056.

[99] G. Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71 (2 2011), p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0.

[100] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (6088 1986), pp. 533–536. DOI: 10.1038/323533a0.

[101] E. Pfeffer et al. "A Case Study of Sending Graph Neural Networks Back to the Test Bench for Applications in High-Energy Particle Physics". In: *Computing and Software for Big Science* 8 (1 2024), p. 13. DOI: 10.1007/s41781-024-00122-3.

[102] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv* (2017). eprint: 1412.6980. URL: https://arxiv.org/abs/1412.6980.

[103] N. Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[104] R. Schmieder. "Studies of neural network architectures in the search for NMSSM inspired di-Higgs events in the $\tau\tau$+bb final state". MA thesis. Karlsruhe Institute of Technology (KIT), 2021. URL: https://publish.etp.kit.edu/record/22089.

[105] S. Wunsch et al. "Identifying the Relevant Dependencies of the Neural Network Response on Characteristics of the Input Space". In: *Computing and Software for Big Science* 2 (1 2018), p. 5. DOI: 10.1007/s41781-018-0012-1.

[106] S. Baker and R. D. Cousins. "Clarification of the use of CHI-square and likelihood functions in fits to histograms". In: *Nuclear Instruments and Methods in Physics Research* 221.2 (1984), pp. 437–442. DOI: 10.1016/0167-5087(84)90016-4.

[107] J. Neyman and E. S. Pearson. "On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London Series A* 231 (1933), pp. 289–237. DOI: 10.1098/rsta.1933.0009.

[108] The CMS Collaboration. *CMS luminosity measurement for the 2018 data-taking period at* $\sqrt{s} = 13$ TeV. Tech. rep. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2676164.

[109] R. Barlow and C. Beeston. "Fitting using finite Monte Carlo samples". In: *Computer Physics Communications* 77.2 (1993), pp. 219–228. DOI: 10.1016/0010-4655(93)90005-W.

[110] A. L. Read. "Presentation of search results: the CLs technique". In: *Journal of Physics G: Nuclear and Particle Physics* 28.10 (Sept. 2002), p. 2693. DOI: 10.1088/0954-3899/28/10/313.

Appendix

A Goodness-of-fit results for the input variables of the PNNs

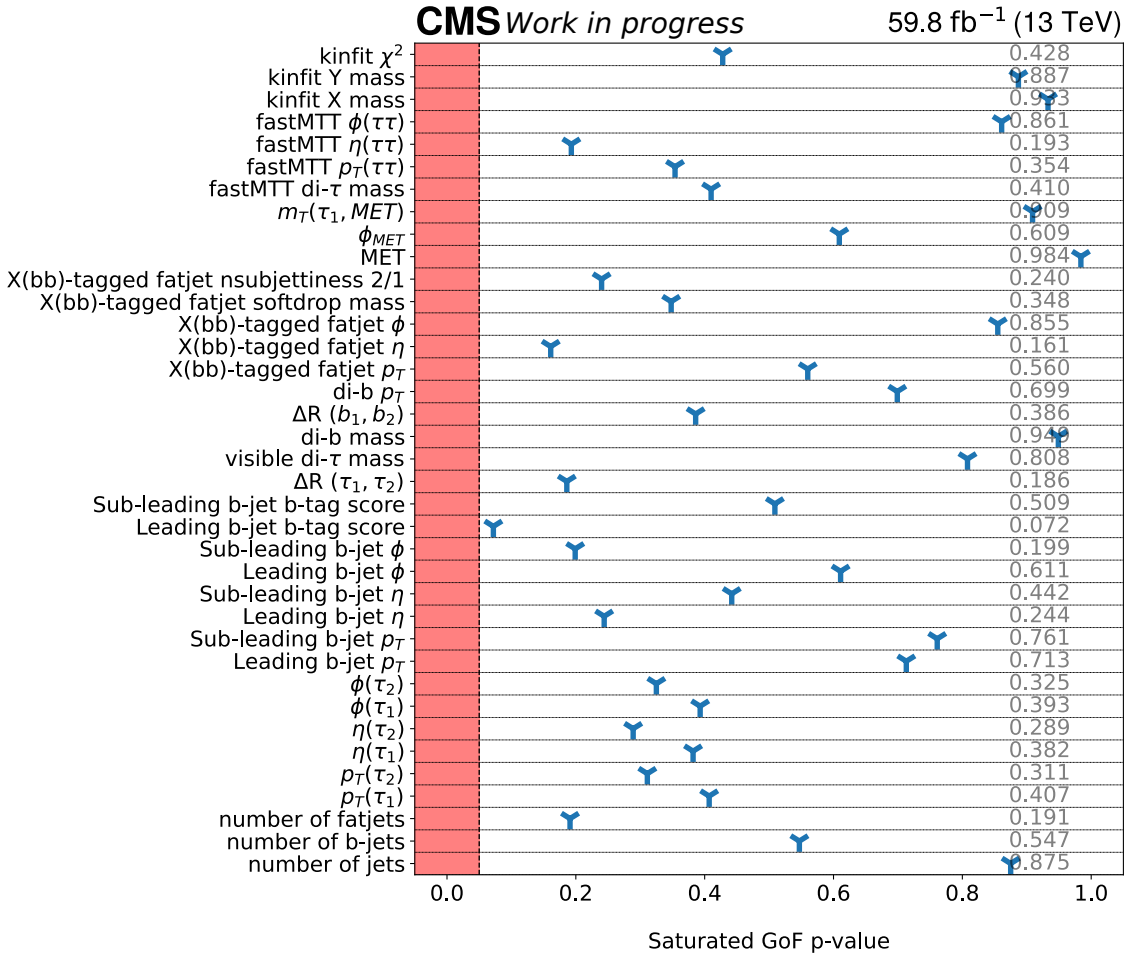


Figure A.1: One dimensional goodness-of-fit tests with p-values for the resolved $e\tau_h$ channel.

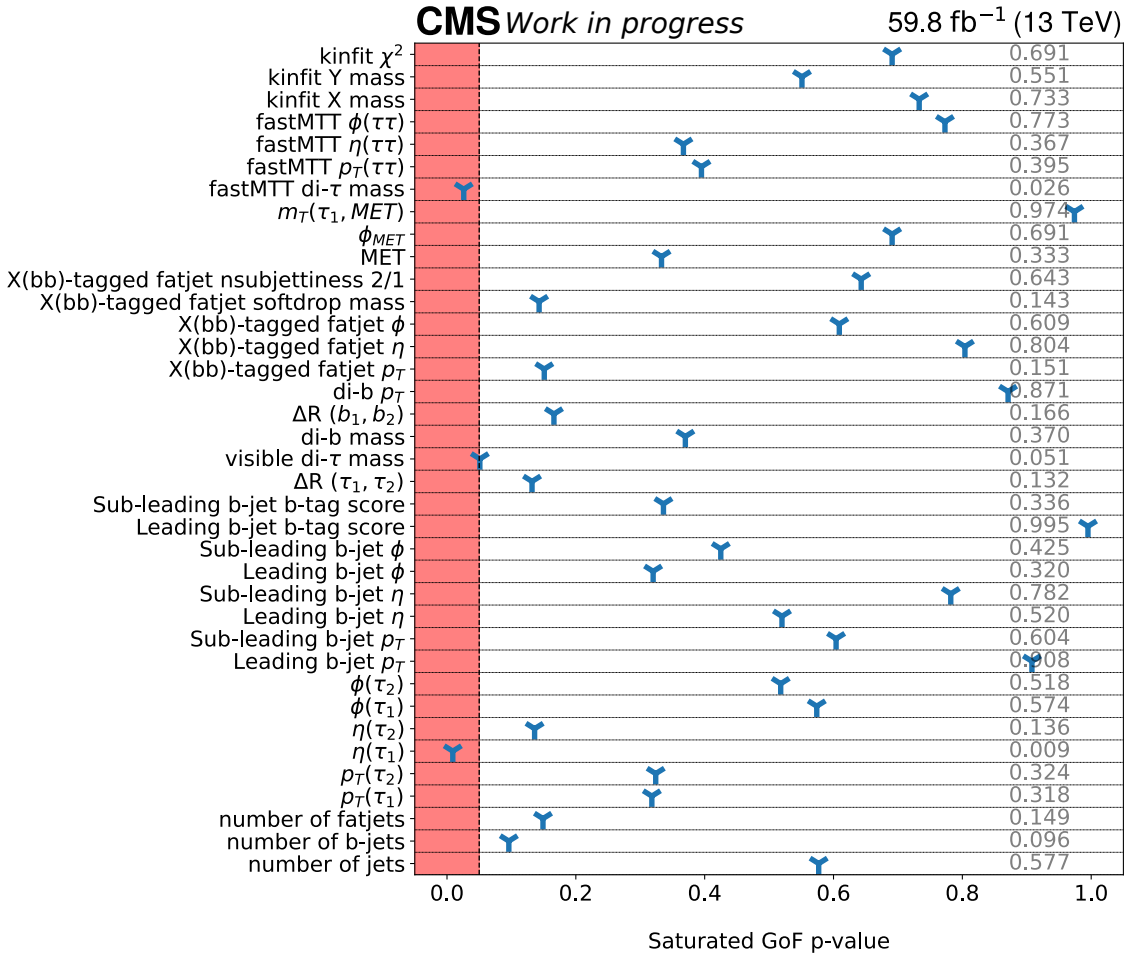


Figure A.2: One dimensional goodness-of-fit tests with p-values for the resolved $\mu\tau_h$ channel.

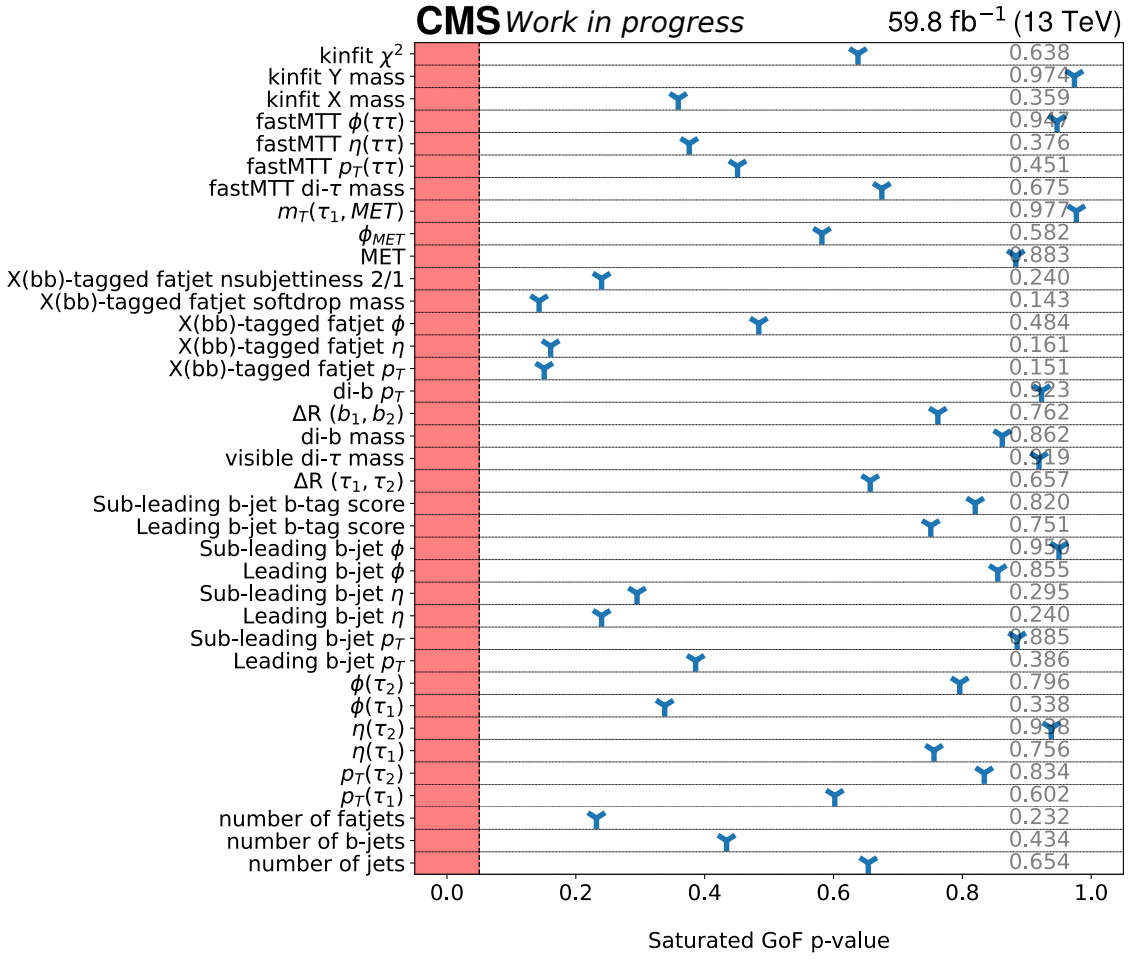


Figure A.3: One dimensional goodness-of-fit tests with p-values for the resolved $\tau_h\tau_h$ channel.

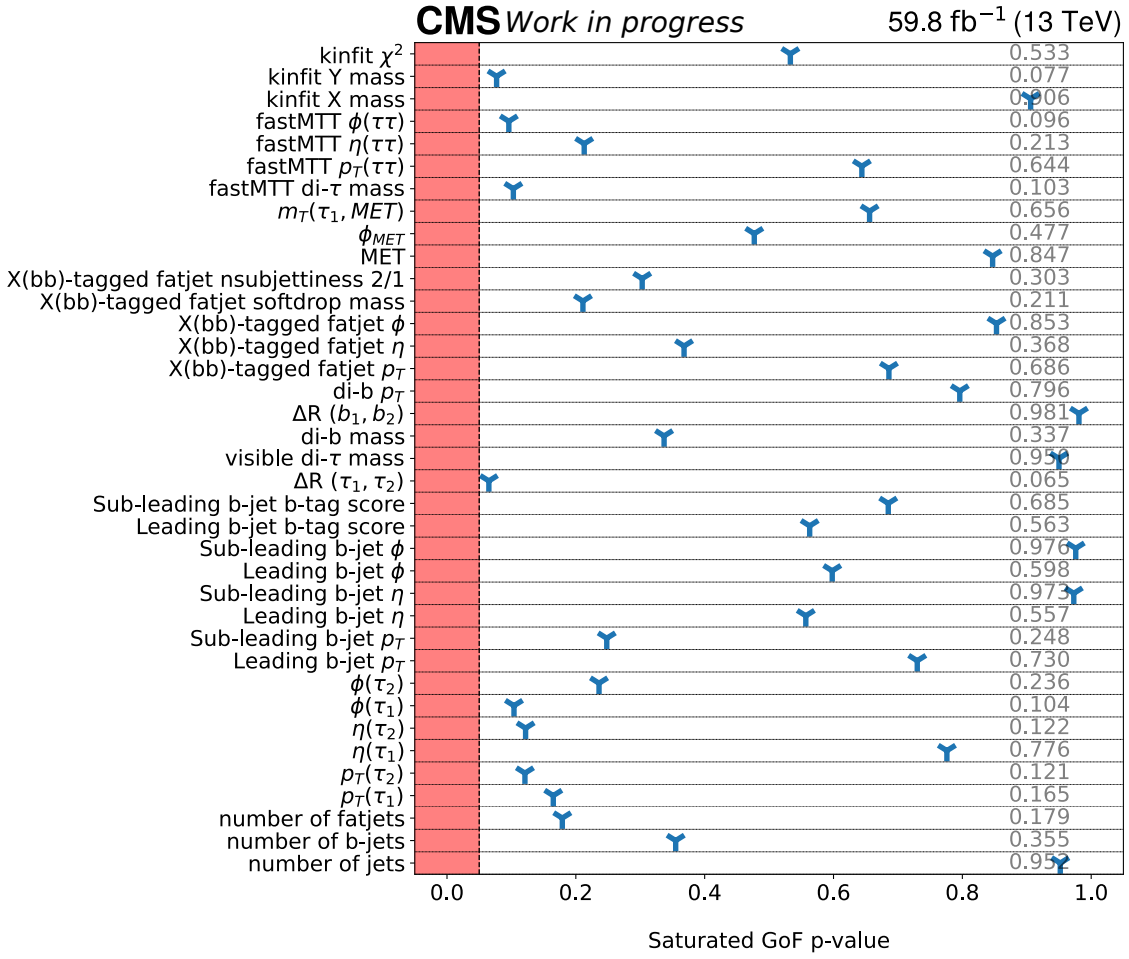


Figure A.4: One dimensional goodness-of-fit tests with p-values for the boosted $e\tau_h$ channel.

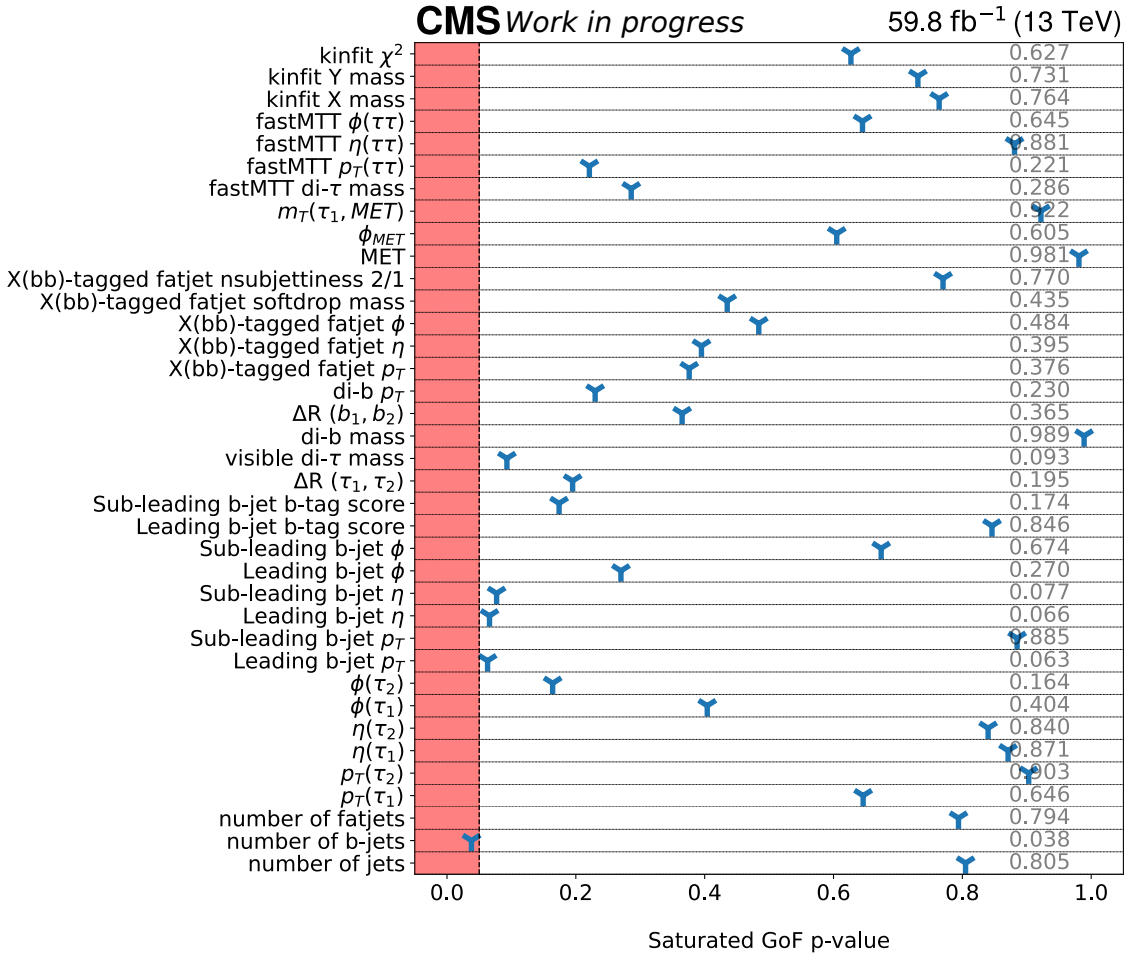


Figure A.5: One dimensional goodness-of-fit tests with p-values for the boosted $\mu\tau_h$ channel.

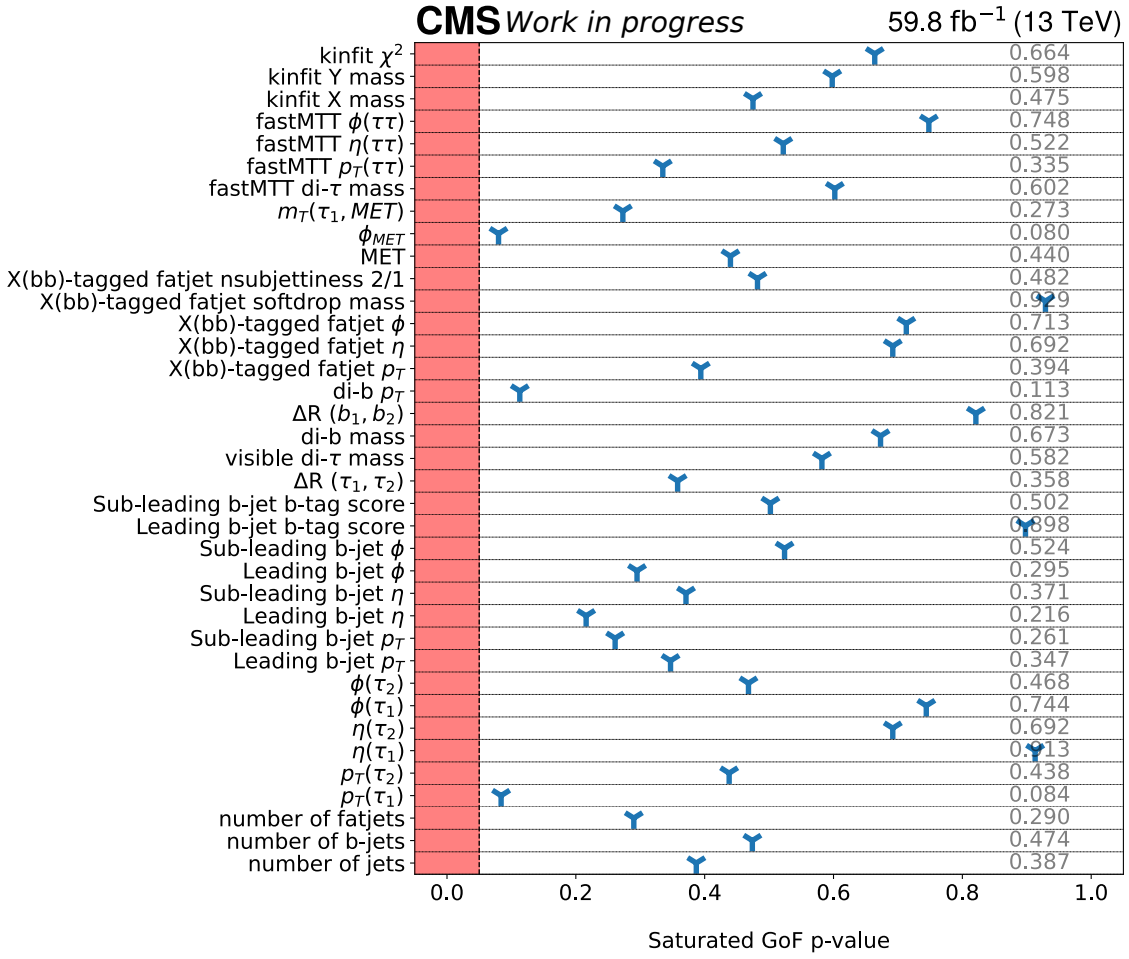


Figure A.6: One dimensional goodness-of-fit tests with p-values for the boosted $\tau_h\tau_h$ channel.

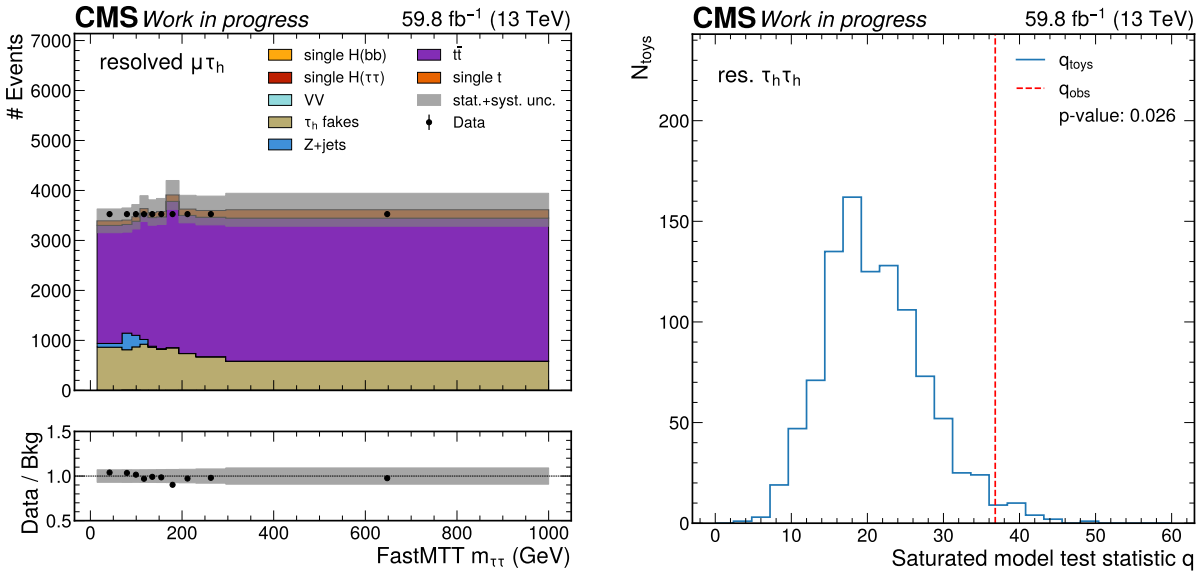
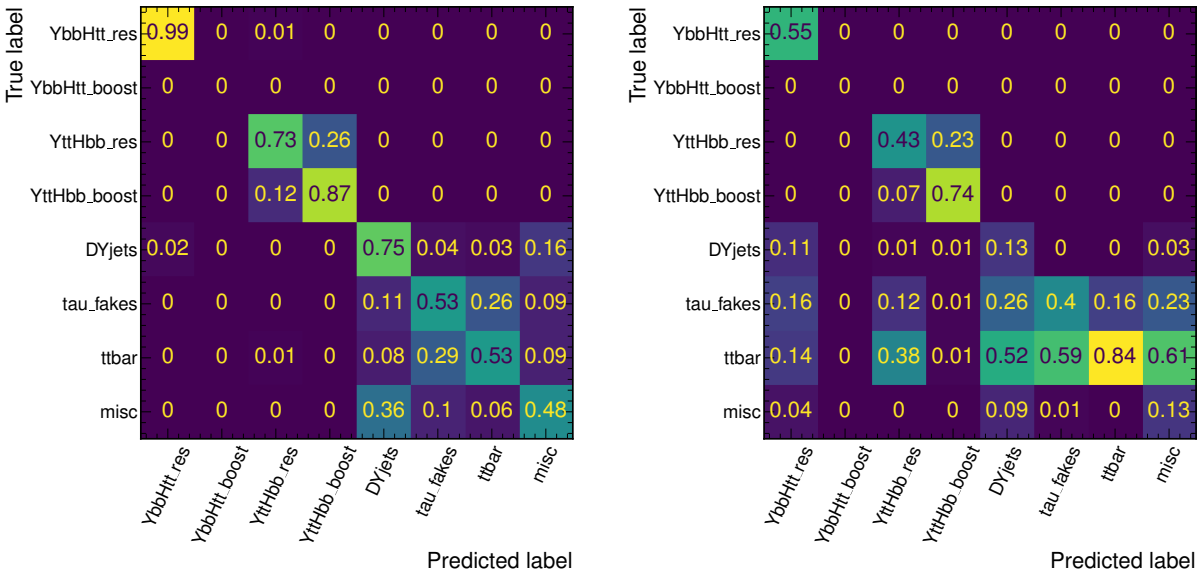


Figure A.7: Distributions of the *FastMTT* estimate for $m_{\tau\tau}$ in the resolved $\mu\tau_h$ channel with the corresponding goodness-of-fit tests as an example variable with a p-value below 5%. The distribution bins are chosen such that they are equally populated in data. For the saturated model test statistic (q_{toy}) 1000 pseudo-datasets ("toys") are generated and the observed value of the test statistic (q_{obs}) is indicated in red.

B PNN performance



(a) Efficiency of the PNN categories

(b) Purity of the PNN categories

Figure B.8: Efficiency (a) and purity (b) of the parametric neural network training shown an as example for the training in the resolved $\mu\tau_h$ channel, evaluated on signal masses $m_X = 3000\,\text{GeV}$, $m_Y = 1600\,\text{GeV}$. Both efficiency and purity are calculated based on raw event numbers. The contribution of the processes in the actual categories have to be calculated by considering the corresponding process cross sections and generator weights. The "misc" category correspond to the VV and single H_{SM} category.

C Upper limits results for all tested mass pair hypotheses for Y(bb)$H_{SM}(\tau\tau)$

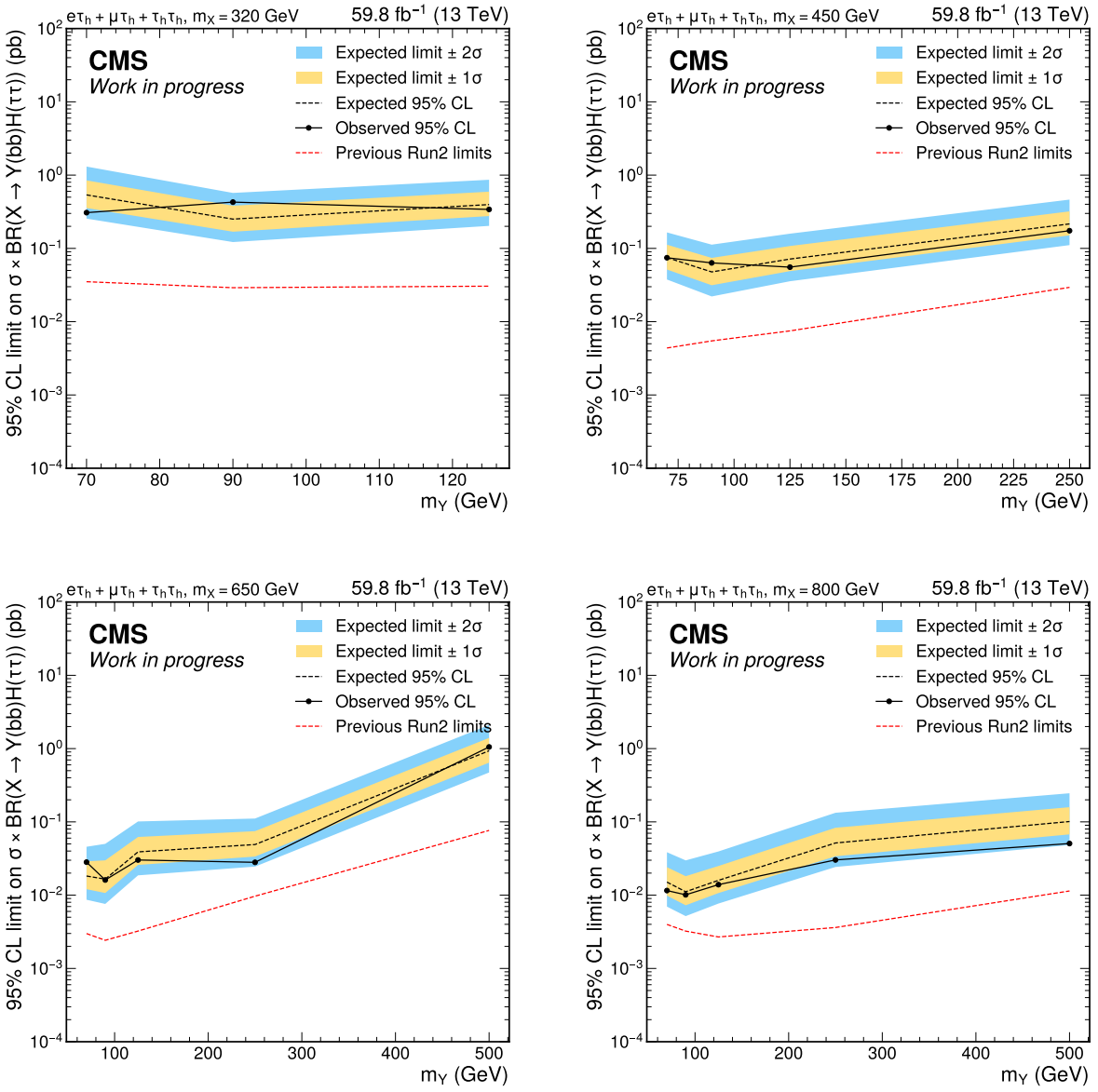


Figure C.9: Expected and observed 95% confidence level upper limits on the Y(bb)$H_{SM}(\tau\tau)$ signal process. The expected limits are shown as a dashed line with the 68% and 95% confidence interval of the expectation given by the yellow and blue bands. Besides the expected 95% confidence interval upper limits of this analysis the results from the previous analysis [9] are shown as well.

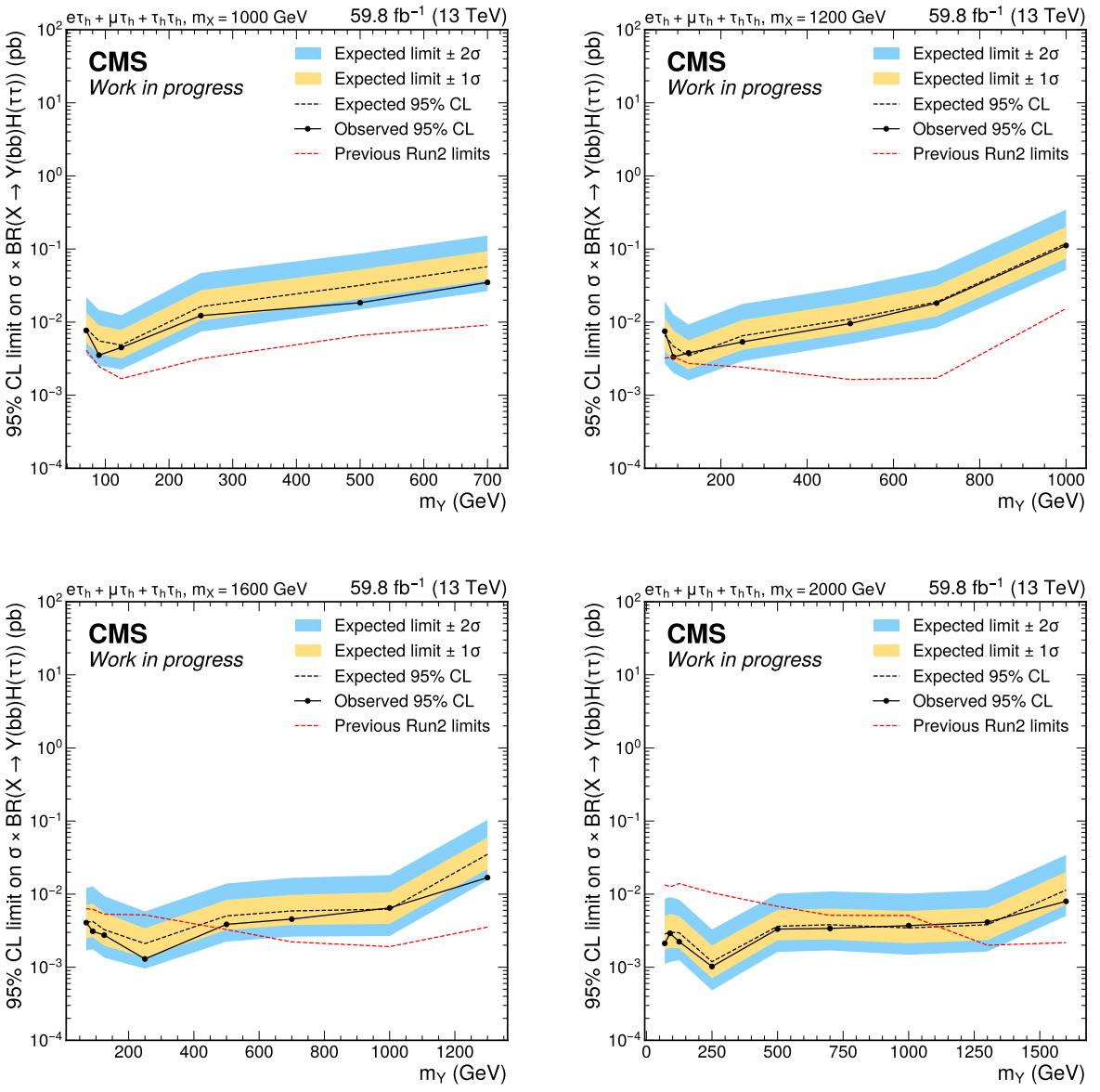


Figure C.10: Expected and observed 95% confidence level upper limits on the $Y(bb)H_{SM}(\tau\tau)$ signal process. The expected limits are shown as a dashed line with the 68% and 95% confidence interval of the expectation given by the yellow and blue bands. Besides the expected 95% confidence interval upper limits of this analysis the results from the previous analysis [9] are shown as well.

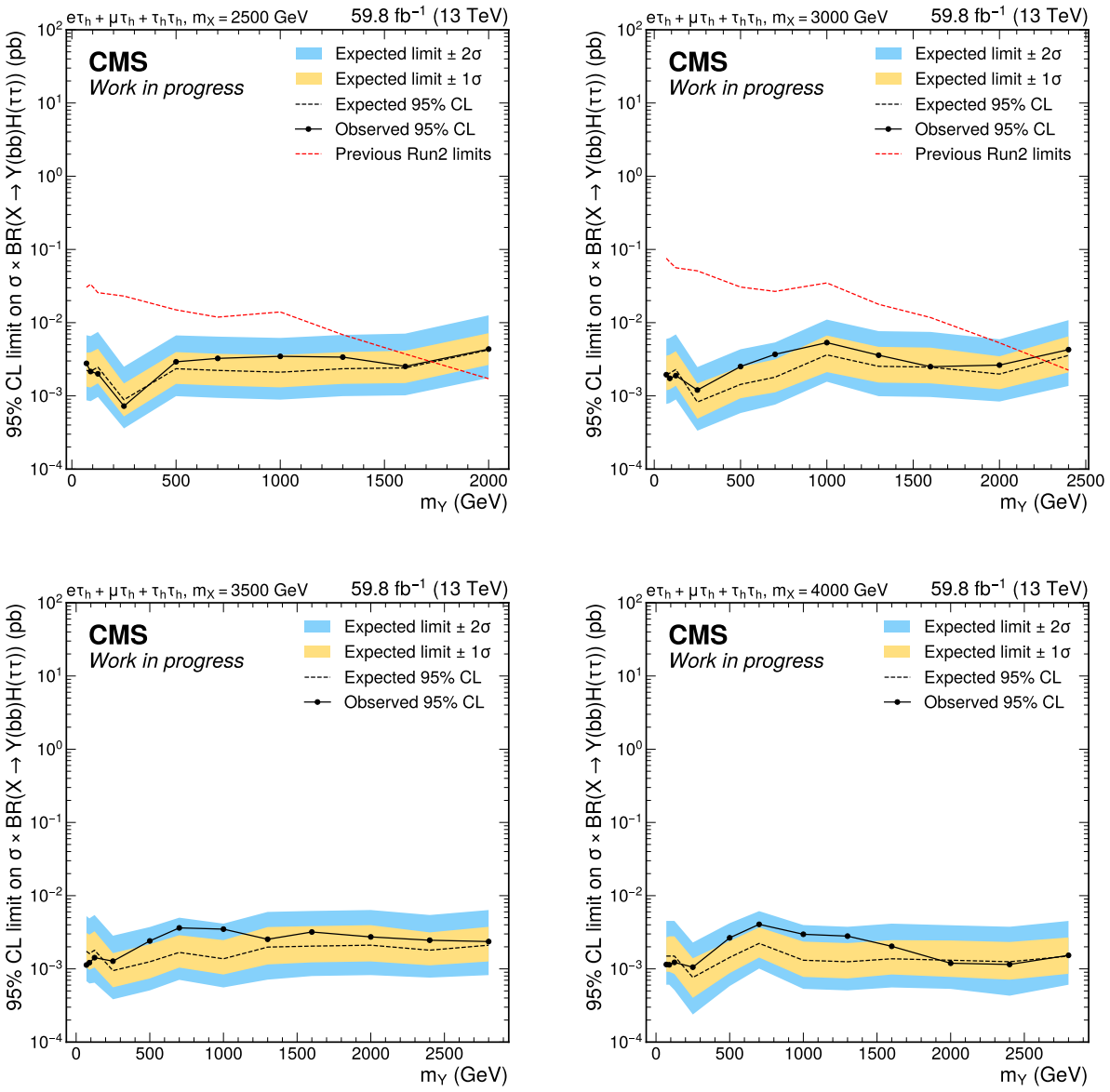


Figure C.11: Expected and observed 95% confidence level upper limits on the Y(bb)$H_{SM}(\tau\tau)$ signal process. The expected limits are shown as a dashed line with the 68% and 95% confidence interval of the expectation given by the yellow and blue bands. Besides the expected 95% confidence interval upper limits of this analysis the results from the previous analysis [9] are shown as well.

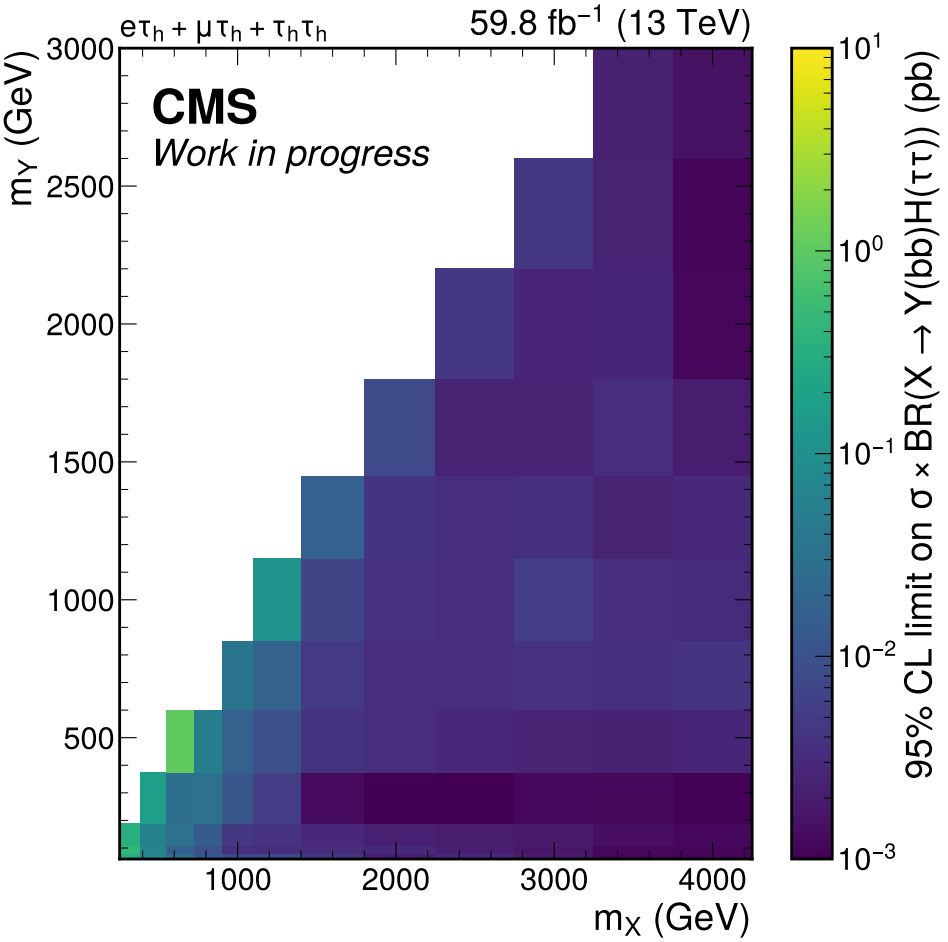


Figure C.12: Observed 95% confidence level upper limits on the $Y(bb)H_{SM}(\tau\tau)$ signal process for all mass pair hypotheses.

D Upper limits results for all tested mass pair hypotheses for $\mathbf{Y(\tau\tau)H_{SM}(bb)}$

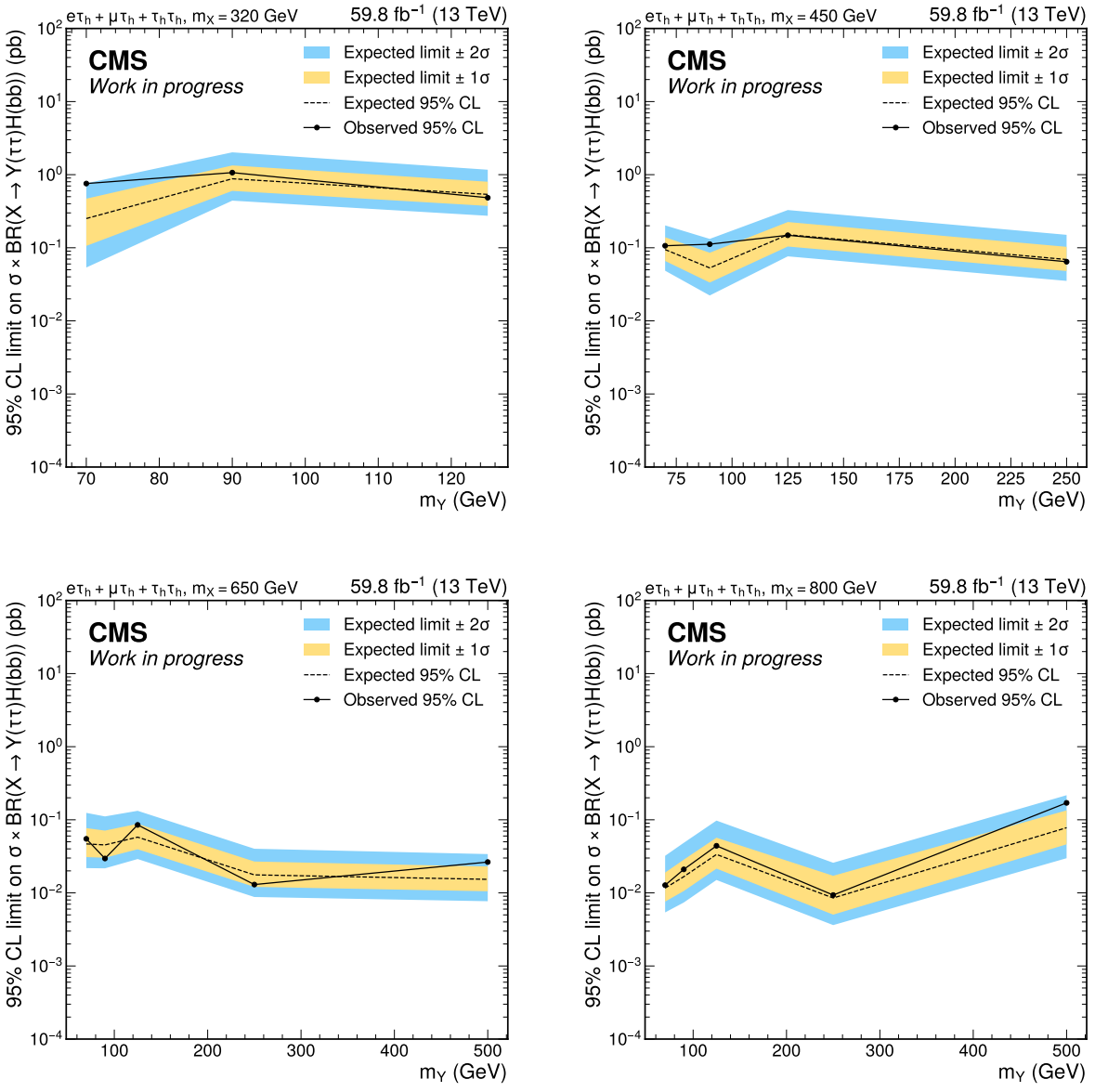


Figure D.13: Expected and observed 95% confidence level upper limits on the $Y(\tau\tau)H_{SM}(bb)$ signal process. The expected limits are shown as a dashed line with the 68% and 95% confidence interval of the expectation given by the yellow and blue bands.

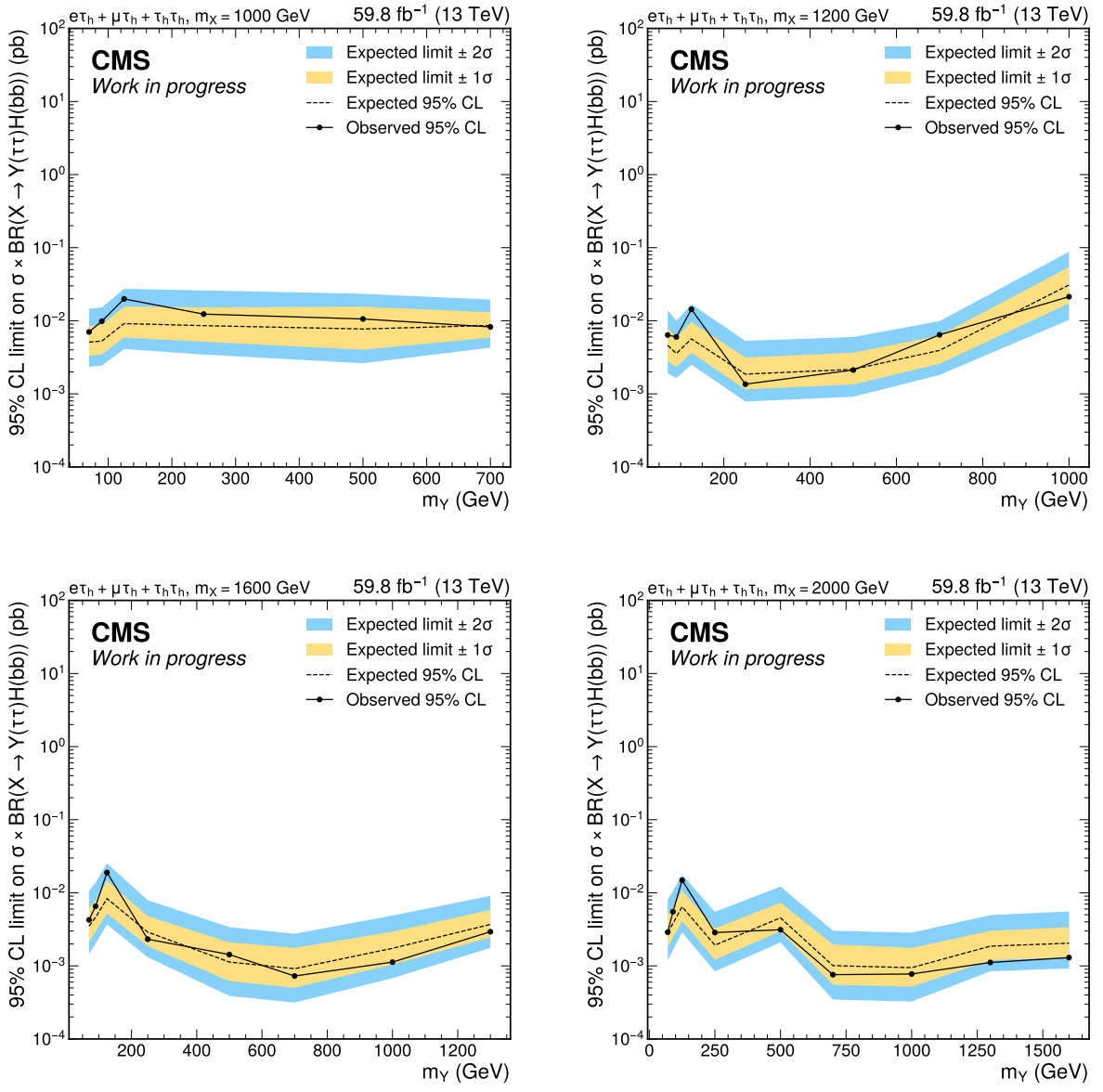


Figure D.14: Expected and observed 95% confidence level upper limits on the $\Upsilon(\tau\tau)H_{SM}(bb)$ signal process. The expected limits are shown as a dashed line with the 68% and 95% confidence interval of the expectation given by the yellow and blue bands.

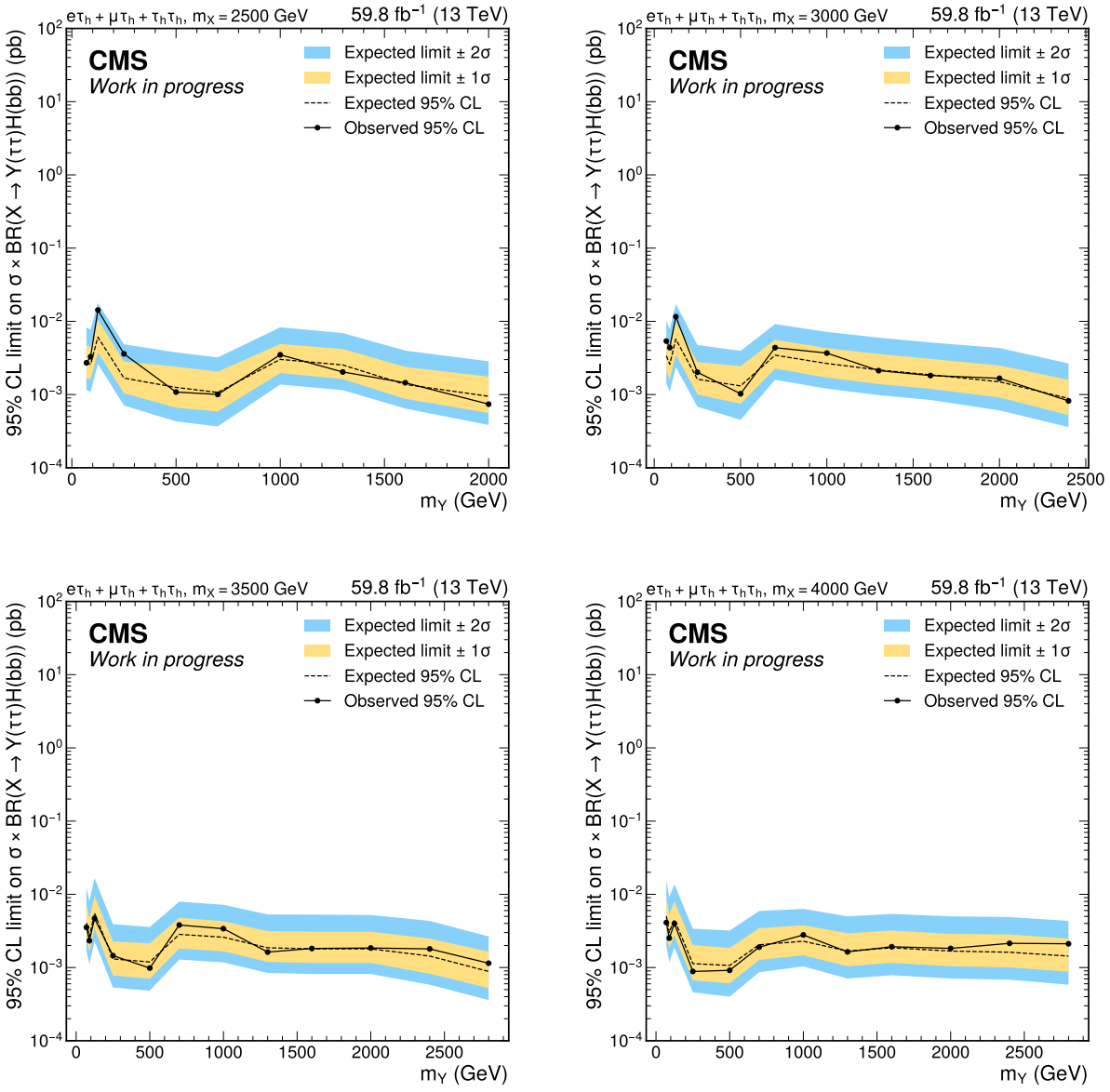


Figure D.15: Expected and observed 95% confidence level upper limits on the $Y(\tau\tau)H_{SM}(bb)$ signal process. The expected limits are shown as a dashed line with the 68% and 95% confidence interval of the expectation given by the yellow and blue bands.

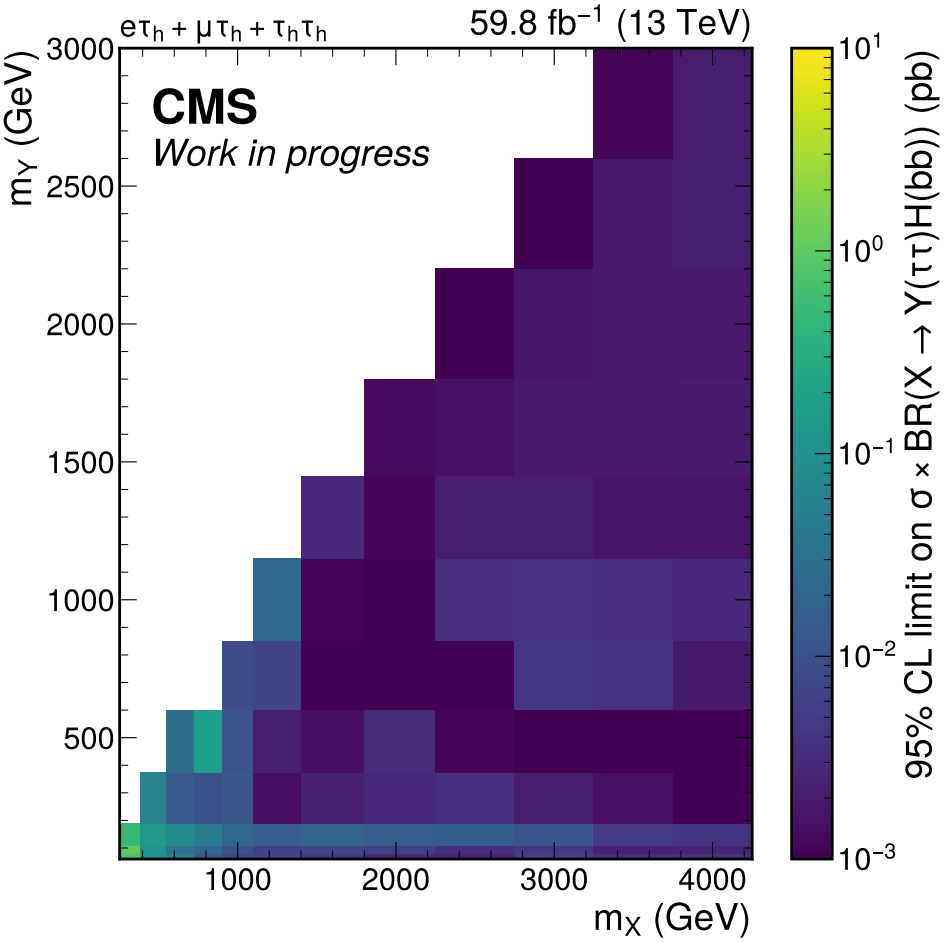


Figure D.16: Observed 95% confidence level upper limits on the $Y(\tau\tau)H_{SM}(bb)$ signal process for all mass pair hypotheses.

Danksagung

Meine Zeit als PhD Student wird wohl eine der wichtigsten Abschnitte im meinem Leben sein. Das hat vor allem damit zu tun, dass ich viele in dieser Zeit gelernt, viele neue interessante Bekanntschaften gemacht und mich dabei selbst weiterentwickelt habe. Ich würde mich gerne bei den Menschen bedanken, die das ganze möglich gemacht haben.

Der größte Dank geht dabei natürlich an Ulrich Husemann. Als ich nach meiner Masterarbeit von ihm und Matthias Schröder gefragt wurde, ob ich nicht Lust hätte noch zu promovieren, war ich erstmal unsicher, weil das damals kein direktes Ziel von mir war. Im Nachhinein war es aber die richtige Entscheidung. Ulrich gab mir immer genug Freiraum mich selber in meinem Thema zu entfalten, was ich sehr schätze, stand aber gleichzeitig immer beratend zur Seite, sobald ich Hilfe brauchte.

Als nächstes wäre Günter Quast zu erwähnen, der, wie schon bei meiner Masterarbeit, die Korreferenten Rolle für meine Dissertation übernommen hat. Auch Günter hat mich mehr oder weniger walten lassen bezüglich meiner Analyse und für diese vertrauen kann ich nur Danke sagen.

Ebenfalls bedanken würde ich mich gerne bei Roger Wolf. Die vielen und sehr hilfreichen Diskussionen über die verschiedene Aspekte meiner Analyse hatten einen signifikanten Anteil an dem Vorankommen und Finalisieren meiner Arbeit.

Was ich am meisten geschätzt habe während meins PhDs ist, die generelle Atmosphäre in Ulrichs's Arbeitsgruppe (aber auch am gesamten ETP). Die ist meiner Meinung nach herausragend, angefangen mit den Aufnahme neuer Leute (wie mich damals als neuer Masterand), als auch die Betreuung, wo man eigentlich von jedem bei Fragen direkt Hilfe bekommen hat. Dieser kollegiale Zusammenhalt ging auch über die Arbeit hinaus. Zu erwähnen ist da natürlich mein langjähriger Bürokollege Emanuel, mit den wir vieles gemeinsam erlebt haben, vom Erfinden von Golf Aufgaben für Erstsemester, über Diskussionen über unsere Analysen, bis hin zum (fast) täglichen 15 Uhr Kaffee. Genauso geht mein Dank am Michael, der immer weisend

zur Seite stand, wenn mal was nicht funktionierte bei meiner Analyse. Weiterhin geht mein Dank natürlich an die ehemahlige Gruppenmitglieder Karim, Philip, Sebastian und Jan, als auch and die gegenwärtigen Mitglieder Michele, Rufa and Moritz. Mit euch allen hat die Arbeit sehr viel Spaß gemacht! Ehemalige Master- und Bachelorstudenten will ich aufgrund der recht langen Liste nicht alle namentlich aufgezählt, aber jeder von ihnen waren ebenfalls wichtiger Bestandteil der Gruppengemeinschaft.

In diesem Zusammenhang würde ich gerne auch der gesamte τ-embedding Gruppe danken, zu der Ralf, Christian, Artur M., Olha, Artur G., Tim, Jan und mittlerweile auch einige neue Leute angehören. Allen voran geht mein Dank hier an Sebastian B., der sehr viel Arbeit in das Entwickeln der Analyse Frameworks gesteckt hat, die ich unteranderem verwendet habe. Aber auch die gemeinsame Entwicklung der analyse-spezifischen Teile der Frameworks war sehr angenehm und hätte nie so gut funktioniert ohne Sebastian's Expertise.

Zum Schluss würde ich gerne auch allem meiner Familie danken, die mich schon während meines gesamten Physikstudiums unterstützt und ermuntert hat weiter zu machen. Ein großes Dankeschön an meine Mutter Tatjana, meinen Stiefvater Viktor und meine Bruder Oleg!