Karlsruher Institut für Technologie

A novel search for di-Higgs events in the $\tau^-\tau^+ + \mathrm{b\bar{b}}$ final state in pp collisions at 13 TeV at the LHC

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN (Dr. rer. nat.)

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

DISSERTATION

von

M.Sc. Janek Bechtel

aus Karlsruhe

Mündliche Prüfung: 12.02.2021

Referent: Prof. Dr. Günter Quast
Korreferent: Priv.-Doz. Dr. Roger Wolf

Institut für Experimentelle Teilchenphysik

Contents

CHAPTER 1

Introduction

The theory in which all known fundamental particles and their interactions are described is the Standard Model (SM) of particle physics. Developed in the second half of the 20th century, it proved to be a precise mathematical framework, structuring all fundamental particles into groups and postulating sets of symmetries from which the fundamental interactions between the particles can be derived. The SM has been greatly successful in the prediction of particles such as the W and Z boson, the gluon, or the charm and top-quark, paving the way to their experimental discovery.

The latest highlight in the success story of the SM is the Higgs boson, which was predicted already in 1964 [1–3], and finally discovered in 2012 at the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) [4, 5]. This discovery enabled the experimental access to a new sector in the SM, the Higgs sector, which could prove to be the key towards the discovery of new phenomena beyond the SM.

Many theories going beyond the SM postulate supersymmetry, an additional fundamental symmetry creating a relationship between two fundamental groups of elementary particles, fermions and bosons. Supersymmetric theories are popular extensions to the SM, as they not only solve many of the issues which are still present in the SM, but also make tangible predictions about the occurrence of additional particles and as such can be experimentally tested within the currently reachable energies of particle colliders. Nevertheless, even after abundant data by the LHC has been collected since its inauguration over ten years ago, no sign of these supersymmetric particles is anywhere to be found. If supersymmetry is thus manifested in our universe, it is very likely not manifested in its most minimal form. Experimental searches, which have successfully constrained large phase space of the minimal supersymmetric extension to the SM, will therefore be required to shift towards probing also non-minimal supersymmetric extensions such as the next-to-minimal supersymmetric extension to the SM (NMSSM). These non-minimal extensions lead to a large number of degrees of freedom in the form of free parameters of the model, which are often experimentally unconstrained, resulting in promising future discovery prospects.

An analysis is presented to search for signatures of decays of a heavy scalar Higgs boson into the discovered Higgs boson with a mass of 125 GeV in addition to another scalar boson. Both heavy and additional scalar boson can arise from the extended Higgs sector of the NMSSM. This search has not yet been conducted so far. It is published for the first time in the context of this thesis. As any of the Higgs bosons would decay into lighter particles almost instantaneously after their creation, their presence can be deducted only indirectly from these decay products. Especially heavy fermions such as b-quarks or tau leptons posses a strong coupling to Higgs bosons and represent prime candidates for the search. The data set used for the search amounts to an integrated luminosity of 137.2 fb^{-1} and was collected by the Compact Muon Solenoid (CMS) detector at the LHC between the years 2016 and 2018.

A main challenge of the analysis is the accurate prediction of all background processes which result in an event signature similar to the predicted signature of the NMSSM process. The performance of the analysis is ameliorated using data-driven methods for the prediction of the major backgrounds due to known SM processes as well as the prediction of the rate and the kinematics of events in which the production of light quarks or gluons at the LHC contaminates the selected events. Furthermore, the use of a neural network-based multiclassification utilizes the close-to-complete information of each selected event to allow conducting the search for signal events based on a multivariate discriminator, achieving an optimal separation of signal events from the individual sources of background.

The theoretical framework motivating the search is discussed in Chapter 2, explaining the Higgs mechanism in the context of the SM and its supersymmetric extensions. In Chapter 3, the origin of the data used for this search is discussed with a detailed description of the CMS detector. All necessary preparations to conduct the search, such as the simulation of signal events or the data-driven estimation of the backgrounds are discussed in Chapter 4. The strategy that is pursued for the optimal categorization of events and thus the final statistical inference and results of the search are given in Chapter 5. Finally, a conclusion will be given in Chapter 6.

CHAPTER 2

Extending the Standard Model of particle physics

2.1 The Standard Model of particle physics

The Standard Model (SM) of particle physics is the underlying theory to describe the fundamental and kinematic properties of elementary particles. The SM makes accurate predictions of the dynamics, creation and annihilation of these fundamental constituents of matter using a mathematical framework based on the underlying symmetries of the universe.

Within the SM, all elementary particles are categorized into two fundamental groups: Fermions and bosons. Fermions are characterized by having half-integer spin, while bosons possess an integer spin. In nature, matter is made up of fermions, while bosons mediate the fundamental forces between these matter constituents. Whether this assignment is a coincidence or a law of nature is one of many unsolved questions, driving the curiosity to search for extensions of the SM, as will be discussed in section 2.3.

Twelve fermions exist in the SM, which all carry a spin of ½: Six leptons, which are the electron, muon and tau lepton as well as their corresponding neutrinos, and six quark flavors, which are called up, down, charm, strange, top and bottom. Leptons and quarks are distinguished by their coupling to the fundamental forces: Leptons do not interact via the strong force, while quarks do. Due to the nature of the strong force, which will be discussed in the following, this leads to a fundamentally different behavior of leptons and quarks.

The macroscopic matter in our everyday lives is made of only three fermions: The up and down quark, which are the fundamental components of the protons and neutrons which in turn make up the nuclei of all atoms, and the electron. All other fermions are either unstable, decaying within fractions of a second to lighter particles, or, in the case of neutrinos, cannot be brought into the bound states necessary for macroscopic matter.

For each fermion a corresponding antiparticle exists with inverted quantum numbers. Particles and antiparticles will not be semantically distinguished in this thesis: The term

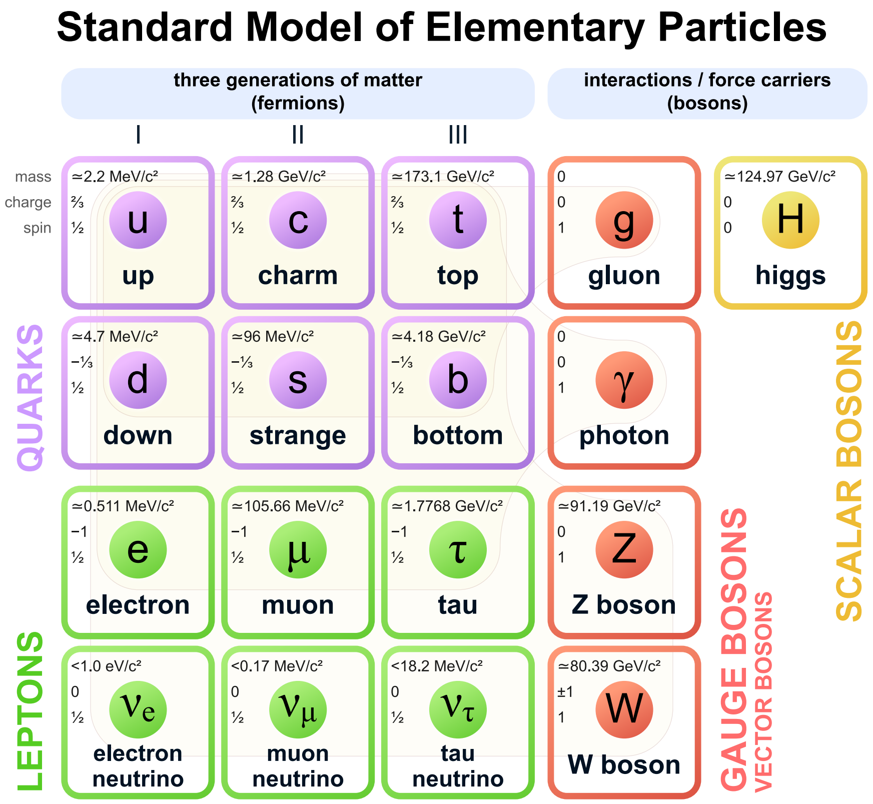


Figure 2.1: The particle content of the SM of particle physics [6]. Quarks (purple) and leptons (green) are the fermions corresponding to the constituents of matter. The gauge bosons (red) are mediators of the fundamental forces described by the SM. The Higgs boson is neither a matter constituent nor a mediator of a fundamental force and appears as a consequence of the Higgs mechanism, via which other particles obtain their masses.

electron will refer to either a negatively charged electron or a positively charged positron, the term up quark will refer to either an up- or an anti-up-quark, and so on.

The bosons of the SM are the gluons, the photon, and the W and Z bosons. They carry a spin of 1 and serve as mediators of the fundamental interactions described by the SM. Finally, the Higgs boson is the only spin-0 particle of the SM and is neither a matter constituent nor a mediator of a force. It appears as a direct consequence of the Brout-Englert-Higgs mechanism [1–3], which is a necessary ingredient of the SM to explain the massive W and Z bosons. The mechanism will in the following be abbreviated as Higgs mechanism. It is of special interest for this thesis and will be discussed in detail in section 2.2. An overview of all particles of the SM can be found in Figure 2.1.

Mathematically, the SM is a quantum field theory (QFT), combining quantum mechanics with special relativity in the framework of classical field theories. The mathematical construction of the SM starts by the postulation of underlying symmetries of the system and thus acquiring a Lagrangian that is invariant under the corresponding symmetry transformation groups.

The external symmetry of the SM is the Poincaré symmetry, which is the symmetry of special relativity and as such refers to the covariance of the system under the space-time transformations of translation, rotation as well as Lorentz boosts. The requirement of this symmetry ensures the Lorentz covariance of the SM.

The power of the SM comes from the additional postulation of the internal local gauge symmetries $SU(3)_C \times SU(2)_L \times U(1)_Y$. The symmetries are related to the three fundamental forces described in the context of the SM:

- The **electromagnetic force** is among the most familiar forces in our everyday lives. Due to its infinite range, it can be experienced macroscopically and is utilized in many technological applications. Its mediator, the photon, can even be observed by eye in the form of light. The electromagnetic force could already be understood in classical physics using Maxwell's equations, before the advent of quantum field theory. The understanding is expanded to the description of quantum effects, where the classical theory breaks down, in the QFT framework of the SM.

- The **weak force** is of similar strength to the electromagnetic force. However, as its mediators, the W and Z bosons, carry large masses, the weak force has exceptionally low range - two orders of magnitude below the diameter of a proton. This makes the interaction very weak and mostly unnoticeable in the macroscopic world. The weak force can be described in a common electroweak theory with the electromagnetic force [7–9], making them two aspects of the same fundamental force. In the context of the SM, the structure of this electroweak force is described by an $SU(2)_L$ symmetry in the space of the weak isospin, and a $U(1)_Y$ symmetry in the space of the weak hypercharge.

- The structure of the **strong force** is given by an $SU(3)_C$ symmetry in the color space. It is mediated by eight gluons, which couple to the color charge possessed only by quarks and gluons. Against the intuitive concept we have from our experience of macroscopic forces, the coupling constant of the strong force decreases at shorter distances between two color-charged objects. Particles participating in strong interactions can thus be described as free particles in the limit of the distance between two particles approaching zero. When going towards large distances however, the energy stored in the potential field between two particles increases linearly. This results in the creation of additional particles from the quantum vacuum when the energy of the strong potential exceeds their masses according to $E = mc^2$. This behavior ensures that only color-neutral objects can be directly observed, and thus makes the color charge a non-observable property of quarks and gluons.

Gravity as the fourth fundamental force, while also a very intuitive force in our everyday live, is the only force not yet included in a consistent way in a QFT approach with the other forces.

Due to the postulation of gauge symmetries, the SM is referred to as quantum gauge field theory. The term gauge corresponds to the mathematical formalism of the theory, containing degrees of freedom which do not correspond to a physical change of the system. If these degrees of freedom, e.g. the phase information of the fermion fields, can be chosen globally without a change to the Lagrangian of the SM, the theory is covariant under global gauge transformations.

By extending the requirement to a *local* gauge symmetry, e.g. allowing the phase to be different at any space-time coordinate, the covariance of the Lagrangian is broken but can be restored by the introduction of an additional degree of freedom in form of a gauge field. Each local gauge symmetry thus leads to the presence of a gauge field which can be identified with a gauge boson, a particle mediating the interaction. The gauge bosons are strictly required to be massless to preserve the covariance. In quantum electrodynamics (QED), requiring a local gauge symmetry with symmetry group $U(1)_{\mathrm{em}}$ leads to the prediction of a massless photon, while in quantum chromodynamics (QCD), local gauge symmetry with respect to the symmetry group $SU(3)_C$ leads to the prediction of eight massless gluons.

Utilizing these symmetries under local gauge transformations gave the SM its power in providing extremely accurate predictions in the description of the fundamental interactions. For electroweak interactions, which are accurately described by the $SU(2)_L \times U(1)_Y$ gauge symmetry, a problem appears due to the masses of the mediating W and Z bosons, as no mass term can be introduced to the Lagrangian without breaking local gauge symmetry. The description of the electroweak interaction via gauge theories, which had been extremely successful in the context of the QED and QCD seemed to fail - unless another mechanism would be developed which could give rise to the gauge boson masses while at the same time preserving the symmetry. In the early 1960's, a solution to this problem has been proposed with the introduction of spontaneous symmetry breaking via the Higgs mechanism [1–3], which will be discussed in the following.

2.2 Electroweak symmetry breaking and the Higgs mechanism

To understand its solution via the Higgs mechanism, first the electroweak gauge theory and the problem of the gauge bosons masses will be discussed in more detail.

The Lagrangian of the electroweak sector corresponding to the $SU(2)_L \times U(1)_Y$ gauge symmetry predicts the appearance of four gauge bosons [7–9]: Three bosons corresponding to the $SU(2)_L$ symmetry in the space of the weak isospin L, which are the three W bosons (W_1, W_2, W_3) and one boson corresponding to the $U(1)_Y$ symmetry in the space of the hyperspace Y, which is the B boson.

The four gauge boson fields as well as the weak isospin and the hypercharge do not yet correspond to the physical observables of the system. The physical fields of the W^+, W^-, and Z bosons as well as the photon (γ) can be obtained by rotation of (W_1, W_2, W_3, B)

in the space of $SU(2)_L \times U(1)_Y$ in which the charged gauge bosons W^+ and W^- arise from linear combinations of the (W_1, W_2) components, and the neutral gauge bosons Z and γ arise from a linear combination of the W_3 and B components. The angle of rotation mapping the physical fields to the W_3 and B fields is the weak mixing angle θ_W, connecting the fields as

$$\begin{pmatrix} \gamma \\ \mathrm{Z} \end{pmatrix} = \begin{pmatrix} \cos\theta_{\mathrm{W}} & \sin\theta_{\mathrm{W}} \\ -\sin\theta_{\mathrm{W}} & \cos\theta_{\mathrm{W}} \end{pmatrix} \cdot \begin{pmatrix} B \\ W_3 \end{pmatrix} \tag{2.1}$$

The electroweak theory thus predicts both the weak as well as the electromagnetic force. In this framework, also a peculiarity of the weak interaction is incorporated: The W^- bosons have been found couple only to fermions of left-handed helicity, while W^+ bosons only couple to fermions of right-handed helicity. The weak force is thus the only force not invariant under parity transformation, i.e. a transformation of the space coordinate $\vec{x} \to -\vec{x}$.

A weakness of the theory is the fact that no mass term can be added to the Lagrangian for either the gauge bosons or the fermions without breaking gauge invariance. The gauge symmetry is thus incomplete, and is completed with the addition of spontaneous symmetry breaking, to be discussed in the following.

An additional term with a new field ϕ and a kinetic term as well as a potential V is added to the electroweak Lagrangian as

$$\mathcal{L}_{\mathrm{Higgs}} = \partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi) \tag{2.2}$$

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2 \tag{2.3}$$

The new field ϕ is a scalar doublet in the space of the weak isospin with two complex components

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \tag{2.4}$$

in which the components ϕ^+ and ϕ^0 carry hypercharge $Y = 1$ and thus electric charge $Q = +1$ and $Q = 0$ respectively, according to the relation $Q = I_3 + \frac{Y}{2}$ with I_3 referring to the third component of the weak isospin. As the components are complex, ϕ carries four degrees of freedom. From the minimum of the potential V in Equation 2.3, an energy ground state can be identified at the vacuum expectation value v:

$$v = \sqrt{\frac{\mu^2}{2\lambda}} \tag{2.5}$$

If $\lambda > 0$ and $\mu^2 > 0$, the field thus has a real and non-zero energy ground state in which the symmetry is broken. A sketch of the Higgs potential and the broken symmetry in the energy ground state is shown in Figure 2.2.

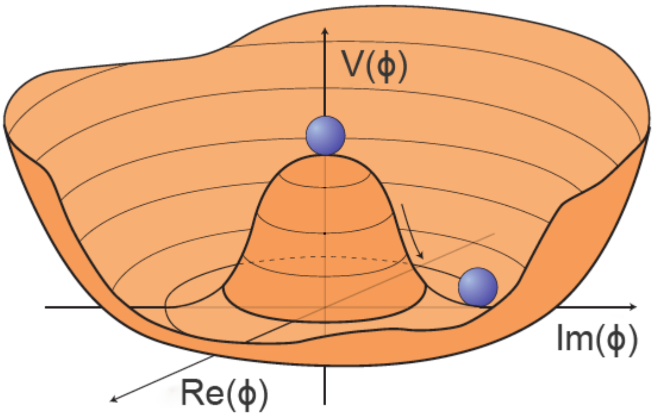


Figure 2.2: Illustration of the Higgs potential in case $\mu^2 > 0$. The rotational symmetry of the system before symmetry breaking is spontaneously broken by choosing any of the points in the minimum of the potential [10].

By respecting the radial symmetry of the system, ϕ can be expanded around the vacuum expectation value, in which the Higgs field H leading radially out of the minimum is introduced.

$$\phi = \begin{pmatrix} 0 \\ v + \frac{H}{\sqrt{2}} \end{pmatrix} \tag{2.6}$$

Here, the expansion is chosen to be limited to the lower component of the doublet, resulting in the $U(1)_{\mathrm{em}}$ group remaining unbroken and the presence of a massless photon. Introducing these additional fields, a mass term emerges from the coupling of the gauge bosons to the vacuum expectation value v. While usually such a mass term breaks the gauge invariance, in the Higgs mechanism the gauge invariance is restored due to the additional presence of the Higgs field and its couplings to the gauge bosons. The Goldstone field corresponds to the appearance of a massless Goldstone boson, however disappears in the unitary gauge choice of equation 2.6. The degrees of freedom lost by the gauge choice appear as additional degrees of freedom given by the longitudinal polarization of the massive gauge bosons.

Out of the four degrees of freedom of ϕ, three are committed to the W^+, W^- and Z bosons, which turn massive. A single degree of freedom remains corresponding to the radial excitations expressed by the Higgs field H. Also this field obtains a mass due to the coupling to the vacuum expectation value. The theory of electroweak symmetry breaking via the Higgs mechanism thus makes a tangible prediction of a massive neutral scalar boson - the Higgs boson. Over 50 years after its theoretical inception, a Higgs boson was discovered at the Large Hadron Collider in 2012 [4, 5].

The problem of the fermion masses is different to the problem of the gauge boson masses: While gauge bosons could not be massive under any circumstance without the inclusion of the Higgs mechanism, mass terms for fermions only break the $SU(2)_L$ symmetry due to the difference in coupling of the weak force to fermions of right-handed and left-handed helicity, imposing the need to describe the fermions as left handed $SU(2)_L$ doublets and right-handed $SU(2)_L$ singlets. The solution for the fermion masses are not yet included in the Lagrangian of Equation 2.2, however can also be dynamically created via the Higgs field. For this, a term corresponding to a Yukawa coupling is introduced [11] for a fermion in the following form, using the coupling to the isospin doublet ϕ of electrons as example,

$$\mathcal{L}_{\text{Yukawa}} = -y_e(\bar{\psi}_L \phi \psi_R + \bar{\psi}_R \phi \psi_L) \ . \tag{2.7}$$

The electron mass can be determined from the vacuum expectation value and the Yukawa coupling of the Higgs field to electrons y_e as

$$m_e = \frac{v y_e}{\sqrt{2}} \tag{2.8}$$

For the other fermions, similar terms are added to include all fermion masses into the Lagrangian.

2.3 The Higgs sector in supersymmetric extensions to the Standard Model

The SM of particle physics, while able to explain the observed phenomena at particle colliders with precision, is also known to be incomplete. Many appealing extensions to the SM impose supersymmetry, an additional symmetry creating a relationship between fermions and bosons. The main reasons motivating the study and search for theories involving supersymmetry are:

- In the SM, the theory of gravitation, general relativity, cannot be explained in terms of a quantum field theory. With supersymmetry imposed as a local symmetry, general relativity is naturally included [12].
- The hierarchy between the electroweak ($\mathcal{O}(10^2)$ GeV) and Planck ($\mathcal{O}(10^{19})$ GeV) energy scales in the SM is susceptible to quantum-loop corrections at the order of the Planck scale. To obtain the low observed value of the Higgs boson's mass, extensive fine-tuning of the quantum-loop corrections and its bare mass is required. In supersymmetry, the divergences leading to these large corrections are systematically canceled and thus a natural solution to the hierarchy problem is provided [13, 14].
- The running coupling constants of the fundamental forces of the SM do not unify at high energy. The running of the constants is altered in supersymmetry, allowing a unification of couplings and postulation of a grand unified theory (GUT) [15, 16].
- The SM fails to provide a candidate for dark matter, whose existence is implied by abundant astrophysical observations. In supersymmetry, additional elementary particles could provide such a candidate [17].

A minimal supersymmetric extension to the SM (MSSM) [18] adds the minimal amount of particles required for a supersymmetric model. The fields of the SM are adapted to superfields, introducing bosonic superpartners to all fermions, and fermionic superpartners to all bosons of the SM. In the MSSM, two Higgs doublets are required to give masses to both up- and down-type fermions. The additional Higgs doublet with respect to the SM gives rise to five Higgs bosons: two charged Higgs bosons $\mathrm{H}^{\pm}$, an additional heavy scalar boson H, a pseudoscalar boson A and the scalar Higgs boson which is associated with the discovered Higgs boson h_{SM}.

One of the main goals of the LHC, next to the observation of h_{SM}, is the search for such supersymmetric particles. While the former was successfully achieved and measuring the properties of h_{SM} is currently ongoing, after over ten years of LHC operation, no sign of supersymmetry has been observed up to now. A reason for this peculiar situation could be found in the fact that supersymmetry may not be realized in its most minimal form, as will be discussed below.

A majority of searches for supersymmetry parameterize their model in the context of the MSSM. The MSSM however does not parameterize all supersymmetric extensions for the SM, and has shortcomings which can be mitigated by further extensions to the model. Especially the Higgs sector, even though an additional Higgs doublet is added with respect to the SM, is highly restricted within the MSSM. Without quantum corrections, the mass of the lightest scalar Higgs boson within the MSSM is predicted to be below $m_{\mathrm{Z}} = 91\,\mathrm{GeV}$. The observed value of the Higgs boson mass of 125 GeV poses a problem within the MSSM, imposing large quantum corrections and thereby again introducing fine-tuning to a model which was originally motivated by natural avoidance of such measures.

Furthermore, in the MSSM the mass parameter μ appears in the superpotential of the Higgs doublets for up- and down-type fermions as $\mu(H_u^T \epsilon H_d)$. This value needs to be adjusted to a value close to the electroweak scale. The question why these two scales are so similar with respect to the Planck scale creates an additional unnaturalness in the model. If an additional complex Higgs singlet is introduced in an extension to the MSSM however, the relevant term can be generated in a natural way [19].

A model capable of solving these shortcomings is the next-to-minimal supersymmetric model (NMSSM) [20, 21]. Here, the additional singlet mentioned above is introduced, generating the $\mu-$term of the model dynamically. The additional Higgs singlet furthermore leads to large consequences to the Higgs sector.

The particle content of the NMSSM is shown in Table 2.1. The following discussion of the resulting NMSSM superpotential is restricted to its scalar components, giving rise to several physical spin-0 Higgs bosons which are of special interest for this analysis.

In the NMSSM, three scalar Higgs fields exist:

$$H_u = \begin{pmatrix} H_u^+ \\ H_u^0 \end{pmatrix} \quad , \quad H_d = \begin{pmatrix} H_d^0 \\ H_d^- \end{pmatrix} \quad , \quad S \tag{2.9}$$

Table 2.1: Supermultiplets of the NMSSM, decomposed into bosonic (spin-0 or spin-1) and fermionic (spin-½) components. The superpartners of the SM components are marked with a $\sim$. The ingredient entering with the extension towards the NMSSM is the complex Higgs supermultiplet $\hat{S}$.

Supermultiplets		Bosonic comp.	Fermionic comp.	$SU_C(3)$	$SU_L(2)$	$U_Y(1)$
quark / squark	$\hat{Q}$	$\tilde{Q} = (\tilde{u}_L, \tilde{d}_L)^T$	$Q = (u_L, d_L)^T$	3	2	1/3
	$\hat{u}$	$\tilde{u}_R^*$	$u_R^\dagger$	3	1	-4/3
	$\hat{d}$	$\tilde{d}_R^*$	$d_R^\dagger$	3	1	2/3
lepton / slepton	$\hat{L}$	$\tilde{L} = (\tilde{\nu}_e, \tilde{e}_L)^T$	$L = (\nu_e, e_L)^T$	1	2	-1
	$\hat{e}$	$\tilde{e}_R^*$	$e_R^\dagger$	1	1	2
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	1
	$\hat{H}_d$	$H_d = (H_d^0, H_d^-)^T$	$\tilde{H}_d = (\tilde{H}_d^0, \tilde{H}_d^-)^T$	1	2	-1
	$\hat{S}$	S	$\tilde{S}$	1	1	0
gluon / gluino		g	$\tilde{g}$	8	1	0
W boson / Wino		$W^\pm$, W^0	$\tilde{W}^\pm$, $\tilde{W}^0$	1	3	0
B boson / Bino		B^0	$\tilde{B}^0$	1	1	0

in which H_u and H_d are the complex scalar doublets also present in the MSSM, and S is an additional complex singlet. The scalar part of the NMSSM superpotential reads

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

in which the scalar components of the supermultiplets as defined in Table 2.1 are used. The matrix ϵ is defined as

$$\epsilon = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \ . \tag{2.11}$$

The superpotential is similar to the potential of the MSSM, however the μ term of the MSSM $\mu(H_u^T \epsilon H_d)$ is replaced by the latter two terms proportional to λ and κ. The parameters λ and κ refer to two dimensionless Yukawa couplings which are free parameters within the NMSSM. To allow for the much larger mass of the observed Higgs bosons with respect to the MSSM prediction, λ needs to be sufficiently large $\lambda \gtrsim 0.5$, while it is bound from above to allow for the NMSSM to be perturbative up the GUT scale, $\lambda \lesssim 0.8$ [20].

The scalar potential in supersymmetric theories has a specific form, represented by a sum of F- and D-terms V_F and V_D as

$$V(\Phi_i) = V_F + V_D = |F_i|^2 + \frac{1}{2}D^a D^a \tag{2.12}$$

with F and D given by

$$F_i^* = \frac{\delta W}{\delta \Phi_i} \tag{2.13}$$

$$D^a = -g_a \left(\Phi_j^* T_{ij}^a \Phi_j \right) \tag{2.14}$$

For the part describing the scalar Higgs potential, Φ refers to $\Phi = (H_u, H_d, S)$. Furthermore, g_a are the gauge couplings and T^a are the generators of the corresponding $U(1)_Y$ and $SU(2)_L$ gauge symmetries. For the Higgs potential of the NMSSM, the V_F and V_D terms are derived as

$$V_F = \sum_i \left| \frac{\delta W}{\delta \Phi_i} \right|^2 = |\lambda|^2 |S|^2 \left(H_u^\dagger H_u + H_d^\dagger H_d \right) + \left| \lambda \left(H_u^T \epsilon H_d \right) + \kappa S^2 \right|^2 \tag{2.15}$$

The V_D term is the same as in the MSSM and given by

$$V_D = \frac{1}{2} \sum_{i,j} g_a^2 (\Phi_i^\dagger T^a \Phi_i)(\Phi_j^\dagger T^a \Phi_j) = \frac{1}{2} g_2^2 \left| H_u^\dagger H_d \right|^2 + \frac{1}{8} \left(g_1^2 + g_2^2 \right) \left(H_u^\dagger H_u - H_d^\dagger H_d \right)^2 \tag{2.16}$$

in which g_1 and g_2 refer to the gauge couplings for the groups $U(1)_Y$ and $SU(2)_L$ respectively.

Superpartners of the SM particles have not been observed so far. Therefore, supersymmetry at low energy scales must be broken, which requires additional supersymmetry-breaking terms in the potential. As the exact mechanism of the breaking of supersymmetry is not known, all possible terms are considered which conserve matter parity and do not re-introduce quadratic divergences and thus the need for fine-tuning in the model. These terms are referred to as soft supersymmetry-breaking terms V_{soft}. In the scalar Higgs potential of the NMSSM, these soft terms are

$$V_{\text{soft}} = m_{H_u}^2 H_u^\dagger H_u + m_{H_d}^2 H_d^\dagger H_d + m_S^2 |S|^2 + \left(\lambda A_\lambda (H_u^T \epsilon H_d) S + \frac{1}{3} \kappa A_\kappa S^3 + \text{c.c.} \right) \tag{2.17}$$

in which the three mass terms corresponding to the three scalar fields appear, along with trilinear supersymmetry breaking parameters A_λ and A_κ. The complete scalar Higgs potential in the NMSSM is then derived from equations 2.15, 2.16 and 2.17 as

$$V = V_F + V_D + V_{\text{soft}} \tag{2.18}$$

The seven free NMSSM parameters appearing in the Higgs potential are therefore

$$\lambda,\ \kappa,\ A_\lambda,\ A_\kappa,\ m_{H_u}^2,\ m_{H_d}^2,\ m_S^2\ . \tag{2.19}$$

The mass parameters of the model $m_{H_u}^2$, $m_{H_d}^2$ and m_S^2 are not to be confused with the physical masses of the Higgs bosons, which arise from the diagonalized mixing matrices to obtain the mass eigenstates obtained during electroweak symmetry breaking.

The mechanism of electroweak symmetry breaking in the NMSSM is the same as in the SM. The complex scalar fields H_u, H_d and S can again be parameterized as expansions around the vacuum expectation values v_d, v_u and v_s, which are chosen to be positive and real:

$$H_d = \begin{pmatrix} \frac{1}{\sqrt{2}}(v_d + h_d + ia_d) \\ H_d^- \end{pmatrix} \tag{2.20}$$

$$H_u = e^{i\phi_u} \begin{pmatrix} H_u^+ \\ \frac{1}{\sqrt{2}}(v_u + h_u + ia_u) \end{pmatrix} \tag{2.21}$$

$$S = \frac{1}{\sqrt{2}} e^{i\phi_s} \left(v_s + h_s + ia_s\right) \tag{2.22}$$

Here, the h_d, h_u and h_s label neutral CP-even states, a_d, a_u and a_s neutral CP-odd states and H_d^- and H_u^+ charged states.
Similar to the Higgs mechanism in the SM, a change of basis can isolate massless Goldstone bosons by constructing linear combinations matching the mass eigenstates. The Higgs fields of the NMSSM comprise $4+4+2=10$ degrees of freedom, corresponding to the four degrees of freedom of the two complex isospin doublets, and the two degrees of freedom of the complex singlet. Of the ten degrees of freedom, three again are manifested as massless Goldstone bosons which get absorbed by the longitudinal degrees of freedom of the gauge bosons.

Thus, as opposed to the SM, seven degrees of freedom remain in the NMSSM, which can be expressed in mass eigenstates corresponding to physical Higgs bosons

$$\mathrm{h_{SM}},\ \mathrm{h_S},\ \mathrm{H},\ \mathrm{A_1},\ \mathrm{A_2},\ \mathrm{H^+},\ \mathrm{H^-}\ . \tag{2.23}$$

Of the seven Higgs bosons, two are charged ($\mathrm{H^+}$, $\mathrm{H^-}$), three are neutral and scalar (CP-even) ($\mathrm{h_{SM}}$, $\mathrm{h_S}$, H) and two are neutral and pseudoscalar (CP-odd) ($\mathrm{A_1}$, $\mathrm{A_2}$). The labels are defined such that the discovered Higgs boson with a mass of 125 GeV is labeled as $\mathrm{h_{SM}}$, the lighter of the additional scalar bosons is labeled as $\mathrm{h_S}$, and the heavier as H. Consequently, $\mathrm{A_1}$ labels the lighter, and $\mathrm{A_2}$ the heavier pseudoscalar boson. The index S of the lighter scalar boson $\mathrm{h_S}$ will in the following be used to indicate that this boson is expected to be very singlet-like in order to match the experimental constraints, as will be discussed in the following.

2.4 Discovery prospects of NMSSM Higgs bosons

In the NMSSM, the singlet field is expected to mix with the two doublet fields as expressed by the λ term in equation 2.10. The mixing between the two fields can be parameterized by a mixing angle θ between the doublet and the singlet fields, and determines the coupling to SM gauge fields and fermions, and thus the production rate at a proton-proton collider as $\frac{\sigma_{\mathrm{singlet}}}{\sigma_{\mathrm{SM}}} = \sin^2\theta$. Reversely, $\mathrm{h_{SM}}$ would then have lower couplings to SM gauge fields and fermions proportional to $\cos^2\theta$. Large values of θ can thus be excluded, as all measurements of the observed couplings of $\mathrm{h_{SM}}$ to SM particles are so far consistent with a coupling modifier of one. A value of θ close to zero, and therefore the existence of a Higgs boson $\mathrm{h_S}$ which is dominated by the singlet field is however still possible. In this case, the couplings of $\mathrm{h_S}$ to all SM gauge fields and fermions are significantly reduced, suppressing the direct production via the collision of SM particles. The additional Higgs boson $\mathrm{h_S}$ can in this case even be very light, well within the kinematic reach of collider experiments such as the LHC or LEP, while still escaping the detection in direct searches.

The coupling of the $\mathrm{h_S}$ to other NMSSM Higgs bosons, e.g. the doublet-like scalar bosons H and $\mathrm{h_{SM}}$ is mediated by the self-coupling of the Higgs fields and not suppressed by a potentially small mixing angle between the singlet and doublet fields. A promising production mechanism of light $\mathrm{h_S}$ states is thus the production in decays of a heavy doublet-like Higgs boson H to $\mathrm{h_{SM}}$ and $\mathrm{h_S}$ [22]. This decay is largely unconstrained, and can have branching fractions of up to 50%, if the decay is kinematically possible given the masses of the bosons.

While overall suppressed by $\sin^2\theta$, the branching fractions of $\mathrm{h_S}$ into SM particles relative to each other are expected to be similar to the branching fractions of $\mathrm{h_{SM}}$. This means that, especially for low mass values of $\mathrm{h_S}$, the decay into a pair of b-quarks is expected to be the dominant fraction of $\mathrm{h_S}$ decays into SM particles. A promising final state is thus the search for a $\mathrm{H} \rightarrow \mathrm{h_{SM}}\mathrm{h_S}$ event, in which the $\mathrm{h_S}$ decays into a pair of b-quarks, and $\mathrm{h_{SM}}$ into a pair of tau leptons.

The decay of $\mathrm{h_{SM}}$ into tau leptons, even though its branching fraction is only around 1/10 of the branching fraction into b-quarks, creates a signature involving pairs of leptons, which helps the identification of such events over the large background of well-known physics processes occurring at the LHC, as will be discussed in the following chapters.

CHAPTER 3

The CMS experiment at the LHC

The Compact Muon Solenoid (CMS) detector [23] is a large particle detector located at the Large Hadron Collider (LHC) particle accelerator at the European Organization for Nuclear Research (CERN). The LHC, spanning 27 km in circumference, is the highest-energy particle accelerator in the world and located underground in Switzerland and France, close to the city of Geneva. The LHC accelerates bunches containing over 10^{11} protons to an energy of 6.5 TeV per proton. They are brought to collision at four points for a center-of-mass energy of 13 TeV. The bunches are spaced such that the collisions occur every 25 ns.

The CMS detector is build around one of the four collisions points. Here, of the 10^{11} protons per bunch, usually around 20-50 collide and a variety of particles emerge. The purpose of the CMS detector is to detect and record all such particles as accurately as possible. For this, a variety of detector submodules are used which will be explained in detail in the following.

3.1 CMS detector design

The goal of the CMS detector design is the hermetic detection of all products of the proton-proton collision, i.e. to have as much coverage as possible of the products in all spatial directions. It is therefore build symmetrically around the collision point. While the optimal design for a full coverage would be a globe around the collision point, the beam pipe of the LHC extending forward makes such a design impossible. Rather, a cylindrical design is used, with multiple layers of detector subsystems extending outwards from the beam pipe. A sketch of the CMS detector is shown in Figure 3.1.

The coordinate system used to label the extension of the subsystems in the CMS detector is chosen with the rotational symmetry of the detector in mind: Subsystems can be identified in their position using the (R, z, ϕ) coordinates for the radius, z-direction and azimuthal angle, with the collision point of the proton-proton interactions in the center of the coordinate system and the z-direction defined as the direction of the beam pipe. A projection of the detector layout onto the $R - z-$plane defined in such a way is shown in Figure 3.2. Due to the rotational symmetry, all detector subsystems cover the complete range of the azimuthal angle $\phi \in [-\pi, \pi]$.

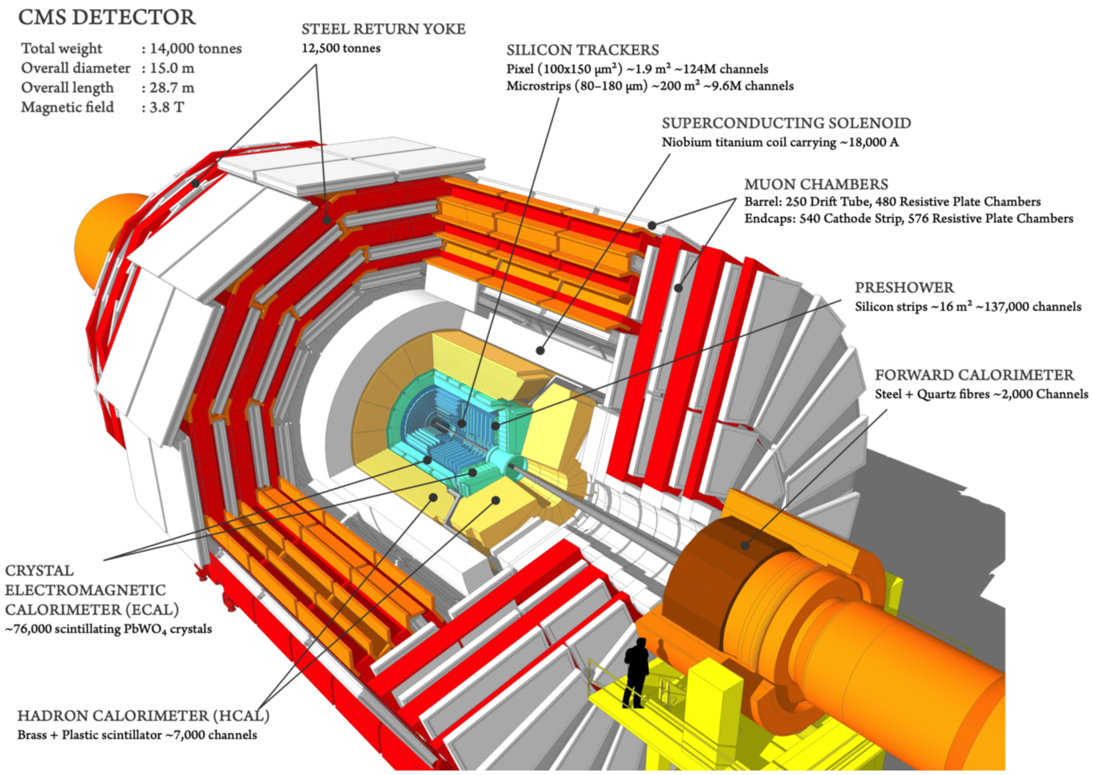

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm²) ~1.9 m² ~124M channels
Microstrips (80–180 μm) ~200 m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000 A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16 m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 3.1: Cutaway sketch of the CMS detector [24]. The grey beam pipe of the LHC containing the proton bunches lies in the center of the detector. Multiple layers of subsystems measure the particles emerging from the proton-proton collisions in the center of the detector as described in the text.

Furthermore, the polar angle θ is defined, also shown in Figure 3.2. To describe a particle's trajectory, instead of θ often the quantity

$$\eta = \ln(\tan(\frac{\theta}{2})) \tag{3.1}$$

is used, with η being called the pseudorapidity of the particle, as it approximates the rapidity for particles whose kinetic energy is much larger than their mass, which is often the case for the collision products. Using the η and ϕ of a particle together with the component of its momentum perpendicular to the beam line, the transverse momentum p_{T}, fully defines a particle's momentum vector $\vec{p}\,(p_{\mathrm{T}}, \eta, \phi)$. The perpendicular momentum component p_{T} can be most easily measured as the curvature radius due to the force acting on a charged particle in the magnetic field of the detector is proportional to this component, which will be discussed below.

The design of the CMS detector can be split into a barrel and an endcap region, with the circular-designed layers extending up to $|\eta| = 1.479$, and two endcaps closing off the CMS detector on both ends. The individual subsystems will now be discussed in more detail, beginning with the inner-most systems.

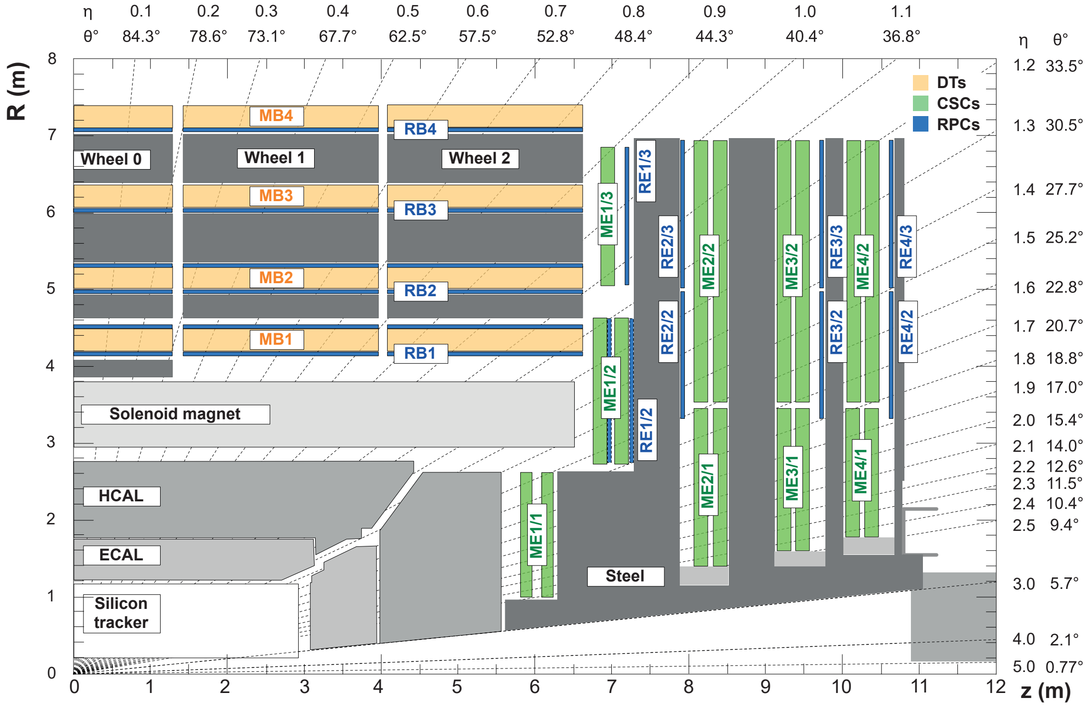


Figure 3.2: Longitudinal view of the upper right quadrant of CMS detector, projected onto the $R - z -$ plane [25].

3.1.1 Silicon Trackers

The task of the tracking system of the CMS detector is the accurate measurement of the particles' trajectories from the interaction vertex, while causing as little energy loss as possible to the particles themselves. It consist of an inner and an outer tracking system, which are called pixel and strip detector respectively. A sketch of the two tracking systems is shown in Figure 3.3.

The pixel detector is the smallest subdetector of CMS, and also the subsystem closest to the LHC beam pipe and the collision point. It consisted of three circular layers in the barrel region of the detector, starting at 4.4 cm from the beam pipe with the last layer 11 cm from the beam pipe, and two endcap disks extending the coverage to $|\eta| = 2.5$. After the 2016 run period, the pixel detector was upgraded to four barrel layers between 2.9 and 16 cm from the beam pipe, as well as three new endcap disks. The upgrade was mainly necessary as the instantaneous luminosity of the LHC Run-2 exceeded the design value of the original digital read-out chips [27], causing a loss of efficiency for events with a high particle rate. In addition, the fourth layer provides an additional point for the 3D reconstruction of the particle tracks. The smaller radius of the innermost layer furthermore provides a significantly improved reconstruction of the displacement of potential secondary vertices in the event. As will be discussed in section 3.2.7, this is

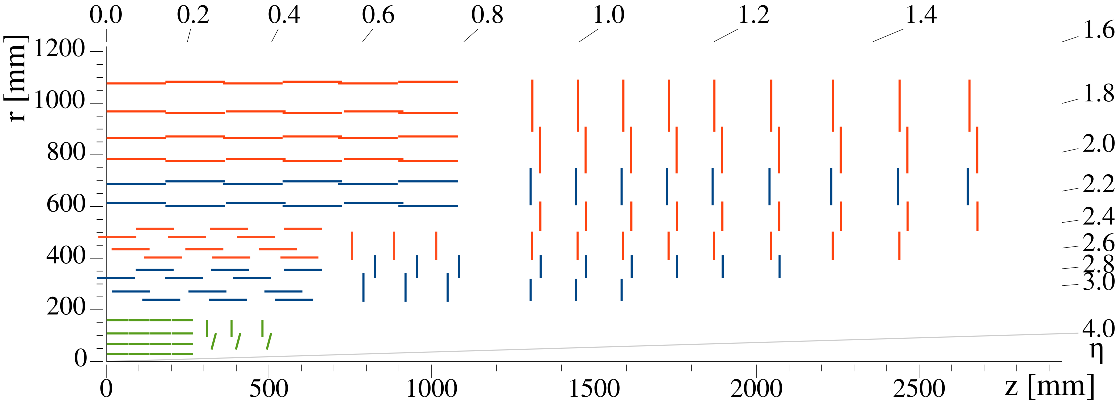


Figure 3.3: Sketch of the CMS silicon trackers after the upgrade between 2016 and 2017 has been performed, shown in the $R - z -$ plane of the CMS detector [26]. The z-axis represents the LHC beam pipe. The pixel detector is located closest to the beam pipe and shown in green. It consists of four layers in the barrel region and three endcap disks consisting of an inner and outer ring. The strip detector is shown in blue and red, and consists of four layers in the inner barrel and six layers in the outer barrel. Each endcap is closed of with nine wheels. Both the pixel and the strip detector rely on silicon chips to detect crossings by charged particles.

crucial for the identification of B hadrons, for which often a displacement of their decay in the order of the few millimeters can be reconstructed due to the relatively long lifetime of B hadrons.

The pixel detector consists of individual rectangular silicon chips ("pixels") of size $100{\times}150\,\mu m^2$, in which an electric signal is induced if the silicon chip is crossed by a charged particle (hit). The small size of the pixels allows for a high resolution of the hits, and thus a high resolution of the reconstruction of tracks, which can be reconstructed from the individual hits. The resolution of the hits is further improved by charge sharing: The induced electrons in the silicon experience a strong Lorentz drift due to the magnetic field of the CMS detector, allowing a precise hit reconstruction by using the charge distribution reconstructed in neighboring pixels. The achieved resolution is around 10 µm in the $R\phi$ direction and around 20 µm in the z direction [28]. The hit efficiency, referring to the probability to reconstruct a hit given that the pixel has been crossed by a charged particle is usually well above 99%, depending on the instantaneous luminosity delivered by the LHC.

The outer part of the tracking system is the silicon strip detector, structured in two barrels resulting in a total of ten layers of silicon strip modules and extending out to a radius of 1.3 m. In the endcap region of the detector, nine wheels on each side of the detector extend the coverage of the strip detector to $|\eta| = 2.5$.

The CMS tracking system is currently being upgraded for the start of Run-3 as well as the high-luminosity LHC. In the latter, especially high radiation tolerance of the tracking system will be required to allow an efficiency of the tracker up to the target integrated luminosity of $3000\,\mathrm{fb}^{-1}$ [29].

3.1.2 Crystal electromagnetic calorimeter

The task of the electromagnetic calorimeter (ECAL) of the CMS detector is the measurement of all predominantly electromagnetically interacting particles appearing in the collisions: electrons and photons. It is a compact and homogenous calorimeter made of over 75,000 lead tungstate scintillating crystals [30]. Lead tungstate has excellent properties for the use as both showering and scintillating material: It is radiation tolerant and has an exceptionally high density of $8.28\,\frac{\mathrm{g}}{\mathrm{cm}^3}$, resulting in a single small crystal with size $23 \times 2.2 \times 2.2\,\mathrm{cm}^3$ weighting almost 1 kg. The high density leads to a short radiation length of $X_0 = 0.89\,\mathrm{cm}$ and Molière radius of $r_M = 2.19\,\mathrm{cm}$ [31]. The length of a single crystal thus allows to contain around 26 radiation lengths, and therefore ensures the almost complete absorption of the electrons' or photons' energy, without the need for an additional absorber. Lead tungstate furthermore has a very fast response, with 99% of the light being collected within 100 ns.

As shown in Figure 3.4, the ECAL is separated between a barrel region, covering particles with $|\eta| < 1.479$, and an endcap region extending the coverage to $|\eta| = 3.0$. In front of the endcaps, two lead absorbers interlaced with scintillating layers make up the preshower detector and help to distinguish neutral pions, which decay into two photons, from prompt photons.

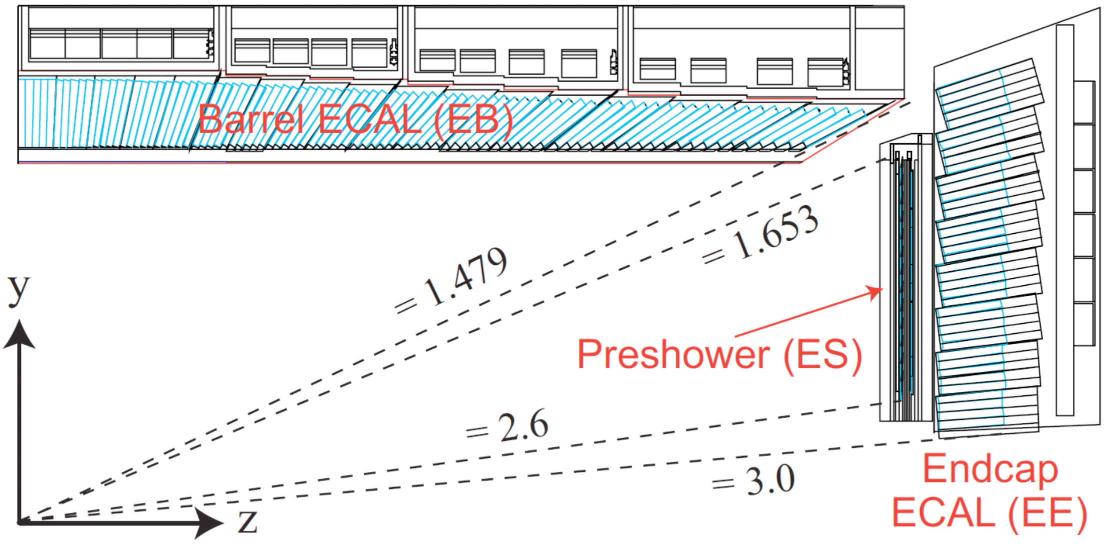


Figure 3.4: Sketch of the electromagnetic calorimeters of the CMS detector [31]. The split of the electromagnetic calorimeter in barrel ($|\eta| < 1.479$ and endcap ($|\eta| > 1.653$) region is indicated by the dashed lines.

The energy resolution of the ECAL as estimated from electrons of the $Z \rightarrow ee$ decay is around 1%. The resolution is roughly composed of two sources: As high energy photons and electrons pass through the crystals, they create an electromagnetic cascade in which the number of particles in the cascade is proportional to E. The higher number of particles improves the measurement resolution due to the statistical nature of the measurement to $\frac{\sigma_E}{E} \propto \frac{1}{\sqrt{E}}$. A second source to the resolution is a constant term of $\frac{\sigma_E}{E} \approx 0.3\%$ related to energy leakage or constant changes to the detector response due to radiation damage over time. This constant term dominates the energy resolution for high-energy photons and electrons.

3.1.3 Hadron calorimeter

Beyond the ECAL of the CMS detector lies the hadron calorimeter (HCAL). Its task is to absorb and measure the energies of remaining particles which are not fully stopped by the ECAL: predominantly hadronically interacting particles such as protons, neutrons, pions or kaons. The HCAL is the most hermetic part of the CMS detector, designed to capture the particles emerging from the proton-proton collision to the largest extend possible. This is achieved by covering a large sector in the forward region of the detector, extending out to $|\eta| = 5.0$. The HCAL stops and measures all particles except the muons, which only loose a minimal amount of energy to the material they traverse, and neutrinos, whose interaction probability with the detector material is so low that they cannot be directly measured at all by the CMS detector.

To achieve the stopping power, the calorimeter needs to be as large and dense as possible, which posed a challenge for the design of the CMS detector in the placement inside the superconducting solenoid as will be discussed below. In contrast to the ECAL, the HCAL is build as a sampling calorimeter, alternating layers of brass absorber material with active scintillating material [32]. It is split in a barrel ($|\eta| < 1.5$), endcap ($1.5 < |\eta| < 3.0$) and forward detector ($3.0 < |\eta| < 5.0$). The HCAL has a thickness, measured in nuclear interaction lengths in brass of $\lambda =$16.42 cm, between 5.8λ at $\eta = 0$ and 10λ for $|\eta| > 1.3$. Due to this relatively compact design, high energy hadron showers cannot be completely contained in the calorimeter, especially for showers developing deep in the calorimeter. To capture the tails of these showers, an outer HCAL component is placed beyond the superconducting solenoid.

The energy resolution is significantly worse than the resolution of the ECAL due to the dead absorber material in the calorimeter, the low number of interaction lengths, and the larger energy fluctuations of hadronic with respect to electromagnetic showers. It lies between 30% and 10% for particles with energies of $\mathcal{O}(10)$ GeV and $\mathcal{O}(100)$ GeV respectively.

3.1.4 Superconducting solenoid

The superconducting solenoid magnet is a central piece of the CMS detector around which the subdetectors are designed. Its task is to provide a strong magnetic field inside the CMS detector to bend the charged particles traversing the detector via the Lorentz force. The bending of the particles is necessary to measure their momentum: The radius of curvature of a particles' trajectory is proportional to the momentum component p_T of the particle perpendicular to the magnetic field. The magnetic field is oriented along the z-axis of the detector, resulting in a bending in ϕ-direction. The measurement of p_{T}, together with the measurement of ϕ and η, fully defines the particles' momentum vector.

The magnet is made of superconducting niobium-titanium coils which are cooled to 4.65 K at which the resistance of the material drops to zero, allowing the especially high homogenous magnetic field of 3.8 T inside the solenoid. The solenoid is enclosed in a 12,000 t steel yoke to capture the magnetic flux outside the solenoid. A measurement of the magnetic flux density caused by the magnet both inside the solenoid and in the return yokes is shown in Figure 3.5.

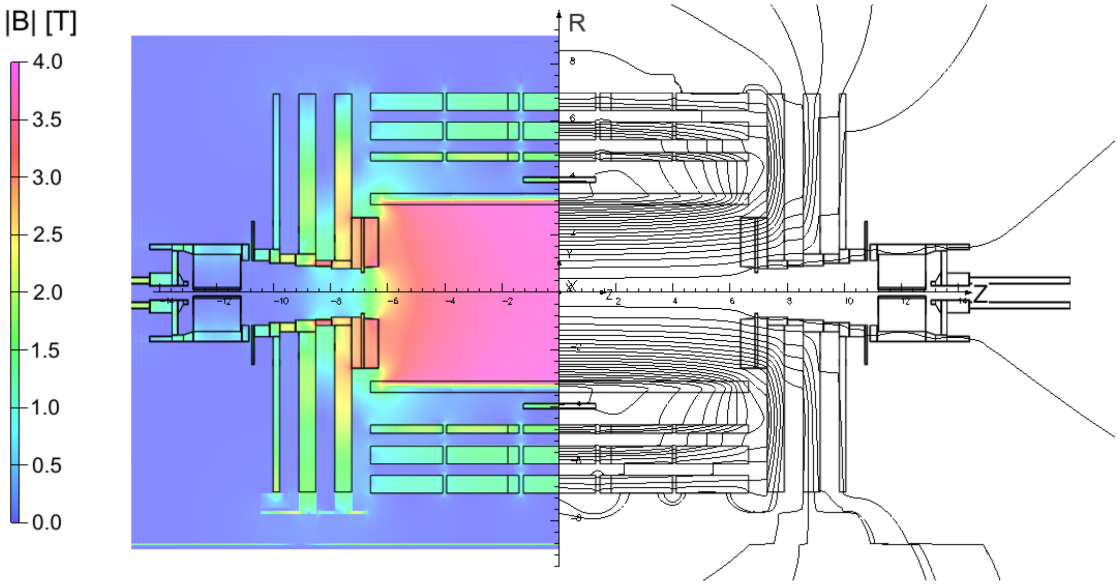


Figure 3.5: Measurement of the magnetic flux (left) as well as magnetic field lines (right) shown in the $R-z-$plane of the CMS detector. Around two thirds of the magnetic flux outside the solenoid is returned through the steel yoke [33].

3.1.5 Muon chambers

The detection of muons is among the most important tasks of the CMS detector. They are produced in a variety of interesting processes and are of interest also in this analysis as a decay product of tau leptons. Muons leave only minimal energy deposits even in dense materials of the calorimeters and are usually not stopped by these layers of the CMS detector. At the outer-most edges of the CMS detector, muon chambers are placed to allow the reconstruction of additional hits used to accurately track the muons' trajectories outside the solenoid.

The muon systems are shown also in Figure 3.2 and are separated into a barrel region, in which drift tubes cover the detection of muons up to $|\eta| < 1.2$, and an endcap region, in which cathode strip chambers extend the coverage to $|\eta| < 2.4$ [34].

The drift tube system consists of tubes with a width of 4 cm in which a stretched wire is located within a gas volume, containing a mixture of Argon and CO_2. If charged particles traverse the tubes, the gas is ionized and the free electrons move towards the positively-charged wire, allowing the reconstruction of a hit with a resolution of around 0.1 mm.

The cathode strip chambers consists of six layers of positively charged anode wires which are perpendicularly crossed by negatively charge cathode strips. Due to the perpendicular design, the chambers measure both the R coordinate via the wires as well as the ϕ coordinate via the strips. They are capable of providing precise space and time information even in the presence of the more heterogenetic magnetic fields and high particle rates present in the endcaps.

In both barrel and endcap detectors, resistive plate chambers provide a complementary triggering system. They consist of two parallel plastic plates, coated with conductive graphite, with opposite charge and high resistance, submerged in a gas volume. If the gas is ionized by a charged particle, a hit can be reconstructed with excellent time resolution. A pattern of such hits provides a fast estimate of the p_{T} of the muon and is used to make a triggering decision to store the event.

3.1.6 Trigger systems

With 40 million collisions per second and around 1 MB of information to be read out in a single event, the total data produced by the CMS detector amounts to around 40 TB/s. Storing this data is unfeasible, and even if possible would result in the storage of huge amounts of data containing relatively uninteresting events of low-energy scatterings of the two protons. The decision to store an event therefore relies on a triggering system, selecting the potentially interesting events for storage while discarding the rest.

The triggering system is made of two levels: The level-1 trigger is a fast hardware-based system, scanning the events for interesting signatures and reducing the event rate for further analysis down to around 100,000 events per second [35]. These events are sent to a computing cluster with several thousand CPUs, the high-level trigger, where the relevant information from different subdetectors is partially combined to form a more educated picture of the event. This higher-level trigger system is software-based and provides a flexible way to change the triggering requirements. The high-level trigger system reduces the event rate to $\mathcal{O}(100)$ events per second.

For these events, the full raw readout of all detector subsystems discussed above is stored and enters the next step, which is the reconstruction of physics objects from the raw detector data.

3.2 Event reconstruction at CMS

3.2.1 Track and vertex reconstruction

Using the individual hits obtained from the local reconstruction in the pixel and strip detectors, tracks are reconstructed [36]. Tracks refer to the estimation of the position and the momentum of charged particle candidates and their trajectories through the detector, taking into account the bending due to the magnetic field. The track reconstruction is a computationally challenging procedure due to the large amount of possibilities in combining hits to tracks, and is thus performed iteratively:

In the initial iterations, the tracks which are easiest to find due to their high p_{T} or proximity to the collision point are identified. The hits associated with these tracks can then be removed, which simplifies the combinatorial complexity of the following iterations. With each iteration, the number of hits is reduced, such that the final iterations are based on a limited number of hits, when hard-to-identify tracks, e.g. tracks displaced from the collision point with low p_{T}, are identified.

Using the reconstructed tracks, the reconstruction of track vertices is performed. The goal is to measure the location of all proton-proton interactions in the event. First, the tracks are selected which are consistent with being produced promptly in the proton-proton interaction by requiring a low impact parameter, i.e. a low perpendicular distance relative to the center of the beam crossing, at least two hits in the pixel and three in the strip detector, and a good quality of the track fit expressed by its χ^2.

Second, the tracks selected in such a way are clustered based on the z-coordinate of the track extrapolation to the beam spot. Due to the many degrees of freedom in the optimization of the clustering, finding the global optimum is a challenging task, which is performed using a deterministic annealing algorithm [37]. The candidate vertices identified by this clustering are finally subject to a fit, determining the position of the primary vertex as well as the likelihood of the clustered track to belong to this vertex.

The resolution of a vertex is strongly dependent on the number of tracks used for the vertex fitting, and ranges from 100 µm for vertices with only a few associated tracks, to around 10 µm for vertices with over 50 associated tracks. The efficiency of reconstructing a vertex is close to 100% if more than two tracks can be used for its reconstruction. After all track vertices have been located, they are sorted according to the $\sum p_{\mathrm{T}}^2$ of all particles stemming from the vertex. The vertex with the highest $\sum p_{\mathrm{T}}^2$ is identified as the primary vertex of the hard proton-proton collision of the event. The additional vertices are referred to as pile-up vertices.

3.2.2 The particle flow algorithm

At the CMS detector, the event description can be significantly improved by combining the information of the different subdetectors, i.e. the reconstructed tracks, the energy deposits in the calorimeters as well as hits in the muon chambers to identify the particle

responsible for this signature, and using this identification to reconstruct the particles' properties. This approach is called the particle flow (PF) algorithm [38].
The algorithm starts from the reconstructed tracks as discussed above, and matches the tracks to the energy deposits in the calorimeters. These deposits are clustered to

- detect neutral hadrons or photons which do not leave hits in the tracking system, and measure their direction and energy,
- separate these energy deposits from the energy deposits caused by charged particles,
- identify electrons in the electromagnetic calorimeter and collect all energy emitted by the electron via bremsstrahlung,
- improve the energy measurement for charged particles, especially for particles with tracks of low fit quality or high p_{T}.

The clustering is performed separately for the ECAL and HCAL, and for the endcap and barrel subdetectors. First, cluster seeds are identified as individual calorimeter cells which have recorded a significant energy deposit exceeding an energy threshold of several 100 MeV depending on the subdetector, and also exceeding the energy deposits of their neighbors. From the seeds, topological clusters are built by extending the cluster to all neighbors, i.e. calorimeter cells sharing at least a corner with the seeding cell, in which also an energy deposit exceeding a looser noise threshold is detected. For a single electron or photon, around 94% of the particles' energy is contained in a 3×3 cluster by extending the cluster to all direct neighbors of the seed. The topological clusters of individual seeds often overlap. It is assumed that the energy deposits in the individual cells arise from as many Gaussian energy deposits as there are seeds in the topological cluster. The individual clusters are then reconstructed using this Gaussian mixture model.

Due to the noise thresholds applied during the clustering, it is expected that the reconstructed energies are lower than the true particle energies especially if the true energy is low. For electrons and photons, this effect is calibrated using simulated photons, with corrections ranging up to 20% for low-energy photons. The calibration can finally be compared between the simulated photons and the photons in data using the abundantly produced neutral pions in their decay into photons $\pi^0 \to \gamma\gamma$ and fitting its known mass of 135 MeV using the invariant mass of the two photons.

For hadrons, which deposit energy in both the ECAL and the HCAL, the calibration is substantially different as it depends on the energy fractions deposited in the two calorimeters. The calibration is performed using simulated neutral K_L^0 hadrons and is dependent on the energy and η of the particle as well as its fractions deposited in the ECAL and HCAL.

The clusters are linked to the tracks first for the reconstruction of electrons and muons, which will be discussed in more detail below. Isolated clusters in the ECAL without an associated track are reconstructed as photon candidates. All clusters associated to electrons, muons and photons are then removed from the collection, and this reduced collection is used for the reconstruction of hadrons. Similarly as the reconstruction of photons, isolated clusters in the HCAL without an associated track are reconstructed as neutral hadrons. If a cluster in both the ECAL and the HCAL without associated track is found, precedence is given to the reconstruction of a photon corresponding to the ECAL cluster, as photons very frequently occur in hadronic jets and neutral hadrons are expected to emit only a small fraction of their energy in the ECAL. These ECAL clusters are however then assumed to belong to the same hadronic shower as the HCAL clusters. The HCAL clusters with associated tracks are reconstructed as charged hadrons. Remaining ECAL clusters crossed by tracks are also attributed to the hadron.

Hadrons produced in the CMS detector are often collimated in jets of particles. The clustering of the individual hadron candidates into jets will be discussed in section 3.2.5.

Particles produced in pile-up interactions create energy deposits with no affiliation with the primary interaction vertex. These reconstructed pile-up particles can distort the determination of the isolation of electrons or muons or the identification of tau leptons decaying hadronically to be discussed in the following, as well as the determination of global event quantities such as the energy sum from the primary vertex of the event. Furthermore, the measured energies of the hadrons from the primary vertex might be incorrect if hadrons from additional proton-proton collisions cross the same calorimeter cells. The hadrons from pile-up interactions thus need to be removed from the PF candidate collection. For charged hadrons, this removal is performed using the affiliation of the hadron track to a vertex given by its use in the vertex-finding fit for the given vertex by the charged-hadron subtraction algorithm [38]. For all neutral hadrons and photons this affiliation is not possible and they are kept in the PF particle collection. The mitigation of the effects originating from from pile-up will be discussed together with the jet reconstruction in section 3.2.5.

3.2.3 Reconstruction and identification of electrons

Electron reconstruction

The reconstruction of electrons is fully integrated in the PF framework [38] and starts from the clustering of energy deposits in ECAL clusters in the context of the PF algorithm. Electrons are charged, electromagnetically interacting particles, and can broadly be identified by the PF algorithm via a track in the tracking system and a cluster in the ECAL, but not the HCAL.

As most electrons emit a significant fraction of their energy as bremsstrahlung when interacting with the tracker material, the total energy of the original electron is expected to be distributed along several individual clusters. Due to the bending of the magnetic field along the ϕ coordinate, the emission of photons is expected to occur mainly along this coordinate. A superclustering is therefore performed by collecting individual ECAL clusters in a narrow η and broad ϕ region around the seeding cluster.

In the PF algorithm, the reconstructed tracks in the inner tracking system are linked with the clusters and superclusters in the ECAL. If an associated track is found for the clusters, the object corresponding to the cluster is labeled as electron, otherwise it is labeled as photon. To avoid the labeling of hadrons as electrons or photons, an additional loose requirement is imposed within the PF framework: The sum of energies in the HCAL in the direction of the ECAL supercluster is required to be less than 10% of the supercluster energy. If the object fails this requirement, the associated energies and tracks are unlinked from the electron reconstruction and used for the reconstruction of hadrons.

Even after the ECAL calibration discussed above, the reconstructed energy of the supercluster can miss the electrons energy due to several reasons: Energy can be lost in shower leakage due to soft photons, in gaps between the individual modules, due to defective ECAL readout, or lost via the interaction with the tracking system not captured by the superclustering.

The energy is thus corrected via a multivariate regression technique using boosted decision trees [39]. The target of the regression is the ratio between the true and reconstructed electron energy. The regression output can thus be directly applied as correction on the reconstructed energy value. In a second and third step, also an estimate of the event-by-event energy resolution of the measurement, and finally an additional smaller energy correction is derived, taking also the electron track information into account. The training of the regression is performed using simulated electrons weighted to a flat p_T spectrum to avoid biasing the regression towards specific energy values.

Finally, a fit of a Breit-Wigner function to the spectrum of the di-electron invariant mass spectrum in vicinity of the Z → ee peak is performed in dependence on the η of the electrons, as well as on the fraction of reconstructed energy in the highest-energy cluster divided by the total energy of the supercluster. The reconstructed electron energies are a free parameters in this fit. A final correction in the order of 1% is applied from this fit such that the observed di-electron events match the position of the Z peak in simulated events at the known value of the Z boson mass.

Electron identification

From the electron candidates reconstructed by the PF algorithm, a set of qualifying criteria are defined to select prompt electrons from the primary interaction vertex, and suppress electron candidates originating from conversions of a photon or electron candidates which are misidentified hadrons. Especially light hadrons collimated in a jet made of charged hadrons and π^0 mesons decaying into photons can cause a signature similar to an electron, with a track of a light charged particle in the direction of significant energy deposits in the ECAL.

A very efficient way to suppress such electron candidates is to require the candidate to be isolated by requiring very little activity by other particles in the area around the electron. For this purpose, a distance is defined in the $\eta - \phi -$ plane as

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}\ . \tag{3.2}$$

For the derivation of the isolation of the electron, the transverse momenta of all particles other than the electron itself are added in an isolation cone around the electron of size $\Delta R < 0.3$. The momenta of the other particles are available via particle candidates defined by the PF algorithm. For charged hadrons, only particles are considered that originate from the same primary vertex as the electron. For neutral hadrons and photons, where the absence of a track makes such a determination impossible, an estimation of the amount of p_{T} by the additional pile-up collisions $p_{\mathrm{T}}^{\mathrm{PU}}$ is subtracted, defining the isolation as

$$\mathrm{Iso} = \sum p_{\mathrm{T}}^{\mathrm{charged}} + \max\left(0, \sum p_{\mathrm{T}}^{\mathrm{neutral\ had}} + \sum p_{\mathrm{T}}^{\gamma} - p_{\mathrm{T}}^{\mathrm{PU}}\right) \tag{3.3}$$

To estimate $p_{\mathrm{T}}^{\mathrm{PU}}$, the two quantities ρ and A_{eff} are defined [40], in which ρ has the unit of energy over area and is measured for each event as the median transverse energy per unit area of the event. It has an almost linear dependence on the number of proton-proton collisions in the event. The effective area A_{eff} is the area of the isolation region around the electron, weighted by an η-dependent factor accounting for the fraction of the area susceptible to the pile-up energy density. The amount of energy due to pileup collisions in the isolation cone around the electron can then be calculated by multiplying the energy density with the effective area as

$$p_{\mathrm{T}}^{\mathrm{PU}} = \rho \cdot A_{\mathrm{eff}} \ . \tag{3.4}$$

To set the energy detected in the isolation code around the electron into perspective with the transverse momentum of the electron itself, the relative isolation is calculated by dividing the two quantities

$$\mathrm{Iso}_{\mathrm{rel}} = \frac{\mathrm{Iso}}{p_{\mathrm{T}}^{\mathrm{e}}} \tag{3.5}$$

For this analysis, only electrons fulfilling $\mathrm{Iso}_{\mathrm{rel}} < 0.15$ are considered for the event selection.

To further improve the identification of electrons, a multivariate electron classification is applied [40]. Boosted decision trees are trained on simulated Z → ee events, using a large number of event observables holding discriminating power over the probability of the electron candidate to originate from a genuine electron. The observables include the quality of the matching of the track to the supercluster, computed both at the ECAL surface and at the vertex, the energy deposited by the candidate in the HCAL compared to the energy in the ECAL, or the fraction of energy emitted by bremsstrahlung before reaching the calorimeter.

The boosted decision tree assigns a score for each electron candidate, with higher scores indicating a higher probability of the electron candidate to originate from a genuine electron. For this analysis, a selection based on this score is applied resulting in an efficiency of selecting genuine electrons of 90%, with a misidentification probability of around 1%.

3.2.4 Reconstruction and identification of muons

The track reconstruction for muons is performed independently from the PF algorithm. Due to the design of the CMS detector, finding muons can be performed with both high efficiency, due to the extended coverage of the muon detectors, and high purity given by the almost complete absorption of other particles in the calorimeters.
The muon tracks are reconstructed independently in the inner tracking and in the muon systems to define three different classes of muon tracks [41]:

- Tracker muons tracks are obtained by extrapolating the tracks reconstructed in the inner tracking system as discussed in section 3.2.1 to the muon detectors. If at least one hit in the drift tubes or cathode drift chambers loosely matches the extrapolated trajectory, the track is declared as a tracker muon track.
- Standalone muon tracks are derived purely by a track fit of the reconstructed hits in the muon chambers.
- Global muon tracks are derived by matching the standalone muon tracks with the tracker muon tracks by comparing the parameters of the tracks. If the tracks match, a common fit is performed to determine the global muon track. Muons reconstructed from a global muon track have an improved p_T resolution with respect to either standalone or tracker muon tracks, especially for high muon energies.

The requirement of a global muon track significantly decreases the probability of the muon to originate from a misidentified jet, which can occur for tracker muons if remnants from the hadron shower reach the innermost muon systems, and the probability of the muon to originate from a cosmic muon traversing the detector, which can occur if the muon is only reconstructed from a standalone muon track.

The collection of tracker, standalone and global muon tracks are fed into the PF algorithm described above, in which a set of quality requirements are imposed on the muon candidate.

At first, the relative isolation of the muon is computed as done for electrons using Equation 3.3. To estimate the p_T caused by neutral particles p_T^PU, a different method than the effective area method used for electrons is used. For muons, p_T^PU is estimated by computing the p_T of charged hadrons due to pileup, and scaling this contribution by a global factor of 0.5 to account for the fact that around half as much energy of the proton-proton collisions is emitted as neutral hadrons or photons rather than charged hadrons.

$$p_\mathrm{T}^\mathrm{PU} = 0.5 \cdot \sum p_\mathrm{T}^\text{charged, PU} \tag{3.6}$$

The relative isolation is then also defined via Equation 3.5. For this analysis, muons are required to have $\mathrm{Iso}_\mathrm{rel} < 0.15$ to be eligible for selection.

To further improve the identification of muons, the compatibility of the track with the muon segment is computed as probability-like score, taking into account the closeness of the extrapolated track to the hit in the muon chamber. In addition, a kink-finding algorithm is performed in which the track is split at several positions and a χ^2 value is computed reflecting the compatibility of the two tracks derived in such a way to belong to a single track. Three definitions of PF muons are created with decreasing efficiency and increasing purity:

- **Loose muons** are all PF muons reconstructed from a tracker or global muon track. About 99% of all muons produced at the CMS detector within its geometric acceptance qualify as loose muons.
- **Medium muons** are all loose muons, with the requirement that at least 80% of tracking layers traversed by the muon register a reconstructed hit. A score of the muon segment compatibility described above of at least 0.451 is required for a medium muon, if the muon is reconstructed only from a tracker muon track. If the muon is reconstructed from a global muon track, this requirement is relaxed to 0.303, however the global muon track must fulfill a track fit quality of $\chi^2/\mathrm{ndf} < 3$ and a position match between the global and tracker muon track of $\chi^2 < 12$, as well as a maximal χ^2 of the kink-finding algorithm of 20. For genuine prompt muons from the decay of a W or Z boson, the efficiency for identifying a muon as medium is around 99.5%.
- For **tight muons**, only muons reconstructed from a global muon track are eligible. The definition of tight muons aims to restrict the muon candidate to prompt muons produced in the proton-proton interaction by requiring a perpendicular distance along r of the extrapolation of the muon track to the primary vertex of the event of less than 2 mm, and a longitudinal distance along z of less than 5 mm. The fit quality of the global muon track must be $\chi^2/\mathrm{ndf} < 10$

For this analysis, the medium muon definition will be used to identify muon candidates arising from the decay of a tau lepton.

3.2.5 Reconstruction of jets

Quarks or gluons which are scattered from the protons during the proton-proton collision hadronize after around 10^{-24} s into many individual hadrons, such as pions, kaons, protons or neutrons. If the initial quark or gluon has a significant p_{T}, the hadrons will be collimated in a narrow cone in the same direction, which is called a jet. The hadrons to be clustered into jets are based on the collection of reconstructed hadrons given by the PF algorithm, in which charged hadrons associated to pile-up vertices have been removed as discussed in section 3.2.2.

An alternative approach to mitigate the influence of hadrons from pile-up is the pile-up per particle identification (PUPPI) algorithm [42]. The PUPPI algorithm assigns a weight between zero and one to each particle in the event, with a weight of one being a clear association of the particle to the primary vertex, and a weight of zero being a clear association to pile-up. For all charged hadrons with tracks matched to either a pile-up or the primary vertex, weights of either exactly zero or one are assigned. For charged hadrons with tracks not matched to any vertex, a weight of one is assigned if the extrapolation of the track to the beam pipe approaches the primary vertex within at most 0.3 cm. Otherwise, a weight of zero is assigned.

The advantage of the PUPPI algorithm over the straight-forward removal of all charged hadron candidates discussed above is that also neutral hadrons are assigned a weight, to be interpreted as the probability of the hadron to belong to the primary vertex. For this, the algorithm uses the fact that hadrons from the primary vertex are expected to be collimated into a few high-p_{T} jets, whereas hadrons from pile-up are expected to be homogeneously distributed. The weight is calculated based on the momentum sum of all particles from the primary vertex other than the neutral hadron itself in a cone of $\Delta R < 0.4$ around the neutral hadron. If the neutral hadron is surrounded by many charged hadrons matched to the primary vertex, the probability of it also belonging to the primary vertex is high. Vice versa, the probability is low if no other particle matched to the primary vertex can be found in the vicinity of the neutral hadron.

The jets used in this analysis are based on the hadron candidates in which a straight-forward removal of the charged hadrons from pile-up has been performed, while for the estimation of the total missing transverse energy of the event ($\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$), the jets as weighted by the PUPPI algorithm will be used.

Using all particles reconstructed by the PF algorithm, excluding only charged particles from pile-up, the particles are clustered into jets using the anti-k_t algorithm [43]. Anti-k_t is an iterative clustering, starting from the particle with the highest p_{T}. Two distances are defined: The distance between two particles i and j, d_{ij}, and the distance between particle i and the beam, d_{iB}, with the latter serving as a measure to define the stopping point of the algorithm.

$$d_{ij} = \min\left(\frac{1}{p_{\mathrm{T,i}}}, \frac{1}{p_{\mathrm{T,j}}}\right) \cdot \frac{\Delta_{ij}^2}{R^2} \tag{3.7}$$

$$d_{iB} = \frac{1}{p_{\mathrm{T,i}}} \tag{3.8}$$

$$\Delta_{ij} = \sqrt{(\mathrm{y}_i - \mathrm{y}_j)^2 + (\phi_i - \phi_j)^2} \tag{3.9}$$

The quantity R is a constant parameter influencing the cone sizes of the clustered jets. For this analysis $R = 0.4$ is used. The quantity y refers to the rapidity of the particle and ϕ to its azimuthal angle. In the first iteration, the distance d_{ij} of the p_{T}-leading particle is calculated with the closest PF candidate j. The two candidates are clustered, and the clustered jet will be used as i. PF candidates are clustered into the jet until the stopping criterion is reached, defined as $d_{ij} > d_{iB}$. At this point, the jet reconstruction is complete and the particles used in its clustering are removed for the reconstruction of additional jets.

Especially jets with low p_{T} can appear in the event due to detector noise. To suppress these unphysical jets, it is required that at least two particles are present in the jet, and that not all of the jets' energy is attributed to neutral hadrons or photons [44]. These requirements are fulfilled by over 99% of all physical jets while removing all but 2% of jets due to detector noise.

Finally, the energies of the reconstructed jets are calibrated both in data and simulation [45], e.g. by using events with two high-energetic jets traversing the detector in opposite directions, or using events containing a signature of a $\mathrm{Z} \rightarrow \mu\mu$ decay plus a single recoiling jet, where the p_T of the Z can be reconstructed with high precision and used to calibrate the recoiling jet.

3.2.6 Identification of hadronic tau lepton decays

Tau leptons decay into a jet of light mesons, i.e. pions or kaons, with a branching fraction of 65%. To identify tau lepton decays in the data, it is therefore a crucial, albeit challenging task to distinguish these jets from the abundant production of jets initiated by the hadronization of a quark or gluon discussed above. The reconstructed jets are used as input for the reconstruction of hadronically decaying tau lepton candidates, which will be noted as τ_h in the following.

The hadron-plus-strips [46] algorithm is used to take the specific signature of a τ_h into account, in which one or three charged hadrons are expected in a narrow cone, often accompanied by neutral pions π^0. The π^0 decays into two photons, which can produce additional electron-positron pairs and thus can be detected as an electromagnetic shower in the $\eta - \phi -$ plane, extended along the ϕ coordinate ("strip") in the ECAL due to the bending of charged particles in the magnetic field. For a full reconstruction of the τ_h candidate, all energy deposits in the strip need to be attributed to the π^0. The strip reconstruction is seeded by a high-energy photon or electron in the jet. The size of the strip is then chosen dynamically in the ranges $\Delta\phi \in [0.05, 0.3]$ and $\Delta\eta \in [0.05, 0.15]$ in an iterative procedure, with the maximal strip size inversely proportional to the p_T of the strip. The proportionality is chosen as a high p_T tau lepton decay will result in a narrower strip size, since its decay products will be boosted along a common axis.

The strips are then combined with the charged particles from the jet to create τ_h candidates in multiple hypotheses for the possible decay modes, depending on the multiplicity of charged and neutral constituents. Four decay modes are defined:

- One charged hadron
- One charged hadron + π^0's
- Three charged hadrons
- Three charged hadrons + π^0's

Per τ_h candidate, exactly one decay mode is assigned. To assign the decay mode, the charge of the possible constituents is required to be ±1 and all constituents must lie in a narrow signal cone with p_T-dependent radius between 0.05 and 0.1 in the $\eta - \phi -$ plane. For each decay mode, the reconstructed invariant mass of all constituents is required to lie in a certain mass window corresponding to the masses of a single charged pion or an intermediate ρ or a_1 resonance.

If still multiple decay modes are possible for the τ_h candidate, favor is given to decay modes with a larger number of charged constituents, higher p_T of the τ_h candidate, and larger number of neutral constituents in this order.

The τ_h candidates selected in such a way are heavily contaminated by mainly quark or gluon induced jets, but also electrons and muons. To suppress events in which these τ_h decays are misidentified, the `DeepTau` classifier [47] is applied on all τ_h candidates. For this classifier, event simulation is used to train a deep neural network in the discrimination of genuine τ_h candidates from misidentified jets, electrons or muons. The output of the neural network is a score, indicating the probability that the candidate is a genuine τ_h with respect to the three sources of misidentification.

For the classification, the neural network utilizes the low-level information delivered by the detector subsystems in vicinity of the τ_h candidate in a procedure similar to the use of machine learning for image recognition. Especially for the discrimination against jets, which often have a higher number of constituent hadrons and a broader cone size in which the particles are collimated, the hadronic activity around the τ_h candidate is among the most discriminating event features to be exploited by the multivariate classification. Furthermore, high level event information is also fed into the training, such as the decay mode of the τ_h, its 4-momentum, the distance of its individual constituents from the common axis, or the quality of their tracks.

Using the discriminators against jets, muons and electrons, several working points are defined with varying efficiencies and misidentification probabilities. For this analysis, the `medium` working point for the discrimination against quark or gluon induced jets is chosen with an efficiency to select genuine τ_h's of 70%, and a misidentification probability of quark or gluon jets of 1% for τ_h candidates with $p_\mathrm{T} < 100\,\mathrm{GeV}$, and over 80% and 0.8% respectively for τ_h candidates with $p_\mathrm{T} > 100\,\mathrm{GeV}$ [47]. For the discrimination against electrons and especially muons, both higher efficiencies and lower misidentification probabilities can be achieved, for which the chosen working points depends on the final state of the analysis as will be discussed in Chapter 4.

3.2.7 Identification of b-quark induced jets

As the most common decay of h_SM as well as potential additional Higgs bosons is the decay into a pair of b-quarks, substantial development has occurred in the challenging task of discriminating jets induced by a b-quark (b-jets) from the abundant production of jets induced by quarks or gluons. A main property to be exploited in the discrimination is the small but significant lifetime of hadrons containing a b-quark (B hadrons) in the order of $10^{-12}\,\mathrm{s}$, which leads to a displaced secondary vertex in the order of some mm to one cm, as illustrated in Figure 3.6.

The reconstruction of this displacement and potential secondary vertices is thus a powerful component for the identification of b-jets, and only possible due to the good resolution of the CMS tracking system as discussed in section 3.1.1. The reconstruction of secondary vertices is performed with the inclusive vertex finding [48] algorithm and utilizes the impact parameter, i.e. the perpendicular distance relative to the center of the beam crossing of the reconstructed tracks.

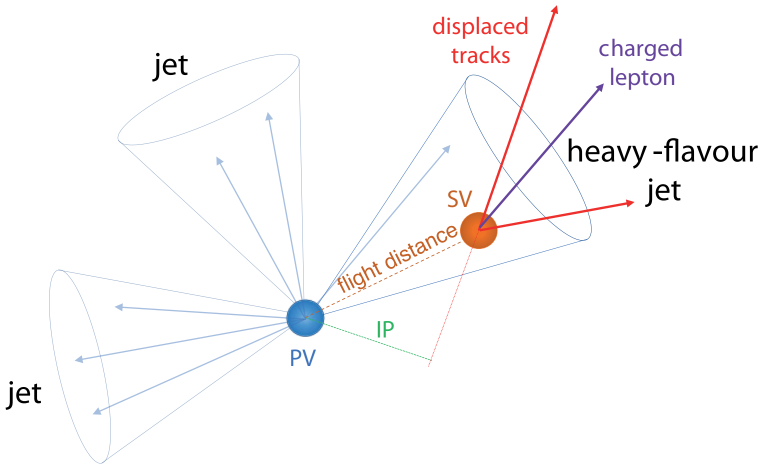


Figure 3.6: Illustration of a b-jet with a displaced secondary vertex (SV) [49]. The b-quark hadronizes within the primary vertex (PV) to form a B hadron. With a lifetime in the order of 10^{-12} s, the B hadron traverses the flight distance of several mm indicated by the orange dotted line before decaying at the SV into hadrons or leptons, with displaced tracks with respect to the direction of the B hadrons, resulting in a jet with a broad cone radius. The impact parameter (IP) is indicated by a green line.

Tracks with a three-dimensional impact parameter of more than 50 µm and a transverse impact parameter of at least 1.2 times its resolution are compared with all other tracks. If the extrapolations of the tracks approach each other closer than they approach the primary vertex, a candidate for a secondary vertex is found. All tracks associated to the vertex are then fitted similarly to the procedure performed for the primary vertex reconstruction. Only secondary vertices are considered in which the direct distance between the primary and secondary vertices is larger than the standard deviation of the measurement by a factor of 2.5 in the transverse plane, and by at least 0.5 when using the full three-dimensional distance vector. If two vertices reconstructed by this procedure share more than 70% of their tracks, they are merged into the same vertex. After the secondary vertex is reconstructed, tracks which are still more compatible with the primary than with the secondary vertex are re-attributed to the primary vertex and the fit is repeated.

An additional criterion discriminating b-jets from light quark or gluon induced jets is that, with a mass of 4.2 GeV, the b-quark is substantially heavier than the light quarks or the massless gluons, and usually much heavier than its decay products. The b-quark decay products thus are expected to have a significant momentum perpendicular to the original flight direction of the b-quark and a higher hadron multiplicity such that b-jets are expected to have a broader energy flux within its jet cone and contain more individual hadrons than light quark or gluon jets. Finally, around 20% of B hadron decays involve electrons or muons. The presence of an electron or muon in the jet cone can thus also serve as a weak indicator of a b-jet.

The `DeepJet` classifier [50] uses these features in the multivariate discrimination of b-jets against light quark or gluon induced jets. It is based on the properties of the charged or neutral constituents in the jet, as well as a set of quality metrics of the secondary vertex associated with the jet. Each of these three components enters a separate convolutional neural network. The output of the three networks is fed into a feed-forward neural network in combination with global high-level event quantities to achieve a discriminating score indicating the probability of the jet being initiated by the hadronization of a b-quark.

Several working points are defined for the discriminator, with the chosen working point in this analysis being the `medium` working point, defined at the point at which the misidentification probability of a light quark or gluon induced jet as b-jet is 1%. The efficiency to select a genuine b-jet at this point is 82% [50].

CHAPTER 4

Towards the search for di-Higgs events

The signal process of the search presented in this thesis is the gluon-fusion production of a heavy additional scalar Higgs boson H and its subsequent decay into a lighter Higgs boson h_S and the discovered Higgs boson h_{SM}. The bosons can result from the extended Higgs sector of the NMSSM as discussed in section 2.3. A signature is investigated in which $h_S \rightarrow$bb and $h_{SM} \rightarrow \tau\tau$. A sketch of this process is shown in Figure 4.1.

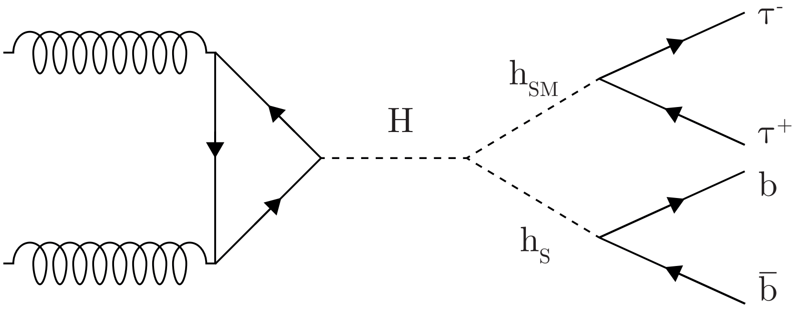


Figure 4.1: Feynman diagram of the signal process subject of the search presented in this thesis. An additional neutral scalar Higgs boson H decays into a pair of lighter, different-mass, scalar Higgs bosons h_{SM} and h_S. The h_{SM} refers to the discovered Higgs boson with a mass of around 125 GeV. The masses of H and h_S are unknown and free parameters of the search.

It is important to note that also the decay of the pseudoscalar boson $A_2 \rightarrow A_1 h_{SM}$ is possible within the NMSSM, and would result in a event signature which could not be distinguished in this analysis from the case in which all bosons are scalar. For abbreviation, only H and h_S will be used in the following, and can be understood as H/A_1 and h_S/A_2 respectively.

In the search, the decay of h_{SM} into tau leptons can be used as a tag of the signal process. Additional criteria are then applied to the jets in the events to also search for the signature of a resonant bb decay. As opposed to the mass of h_{SM}, and thus the invariant mass of the two tau leptons, the mass of the additional light Higgs boson h_S is not known. Therefore, the full spectrum of the invariant mass of the two b-jets is of interest for this analysis. The search is restricted to the resonant production of H and to cases in which the mass of H is larger than the sum of the masses of h_{SM} and h_S, as the signal process is otherwise suppressed.

4.1 Physics object selection

To extract the data set which will be used for the analysis, a set of selection requirements is applied to the collected data, building on the reconstructed electrons, muons, τ_h's and jets as discussed in section 3.2. The selection aims at extracting events containing two genuine tau leptons in addition to two genuine b-quarks, while optimizing the balance between efficiency, i.e. not to miss events with genuine $\tau\tau$+bb signatures and purity, i.e. making sure that as many selected events as possible contain a genuine $\tau\tau$+bb signature.

4.1.1 Selection of tau lepton decays

At first, a pair of physics objects corresponding to the decay products of two tau leptons is required. The tau lepton decays either to an electron (branching fraction of 17.8%), muon (17.4%), or light hadrons (64.8%). All of these decays are accompanied by neutrinos, which are undetectable within the CMS detector and therefore carry away a significant fraction of the initial tau lepton energy. The independent decays of the two tau leptons define six possible final states which are shown in Figure 4.2.

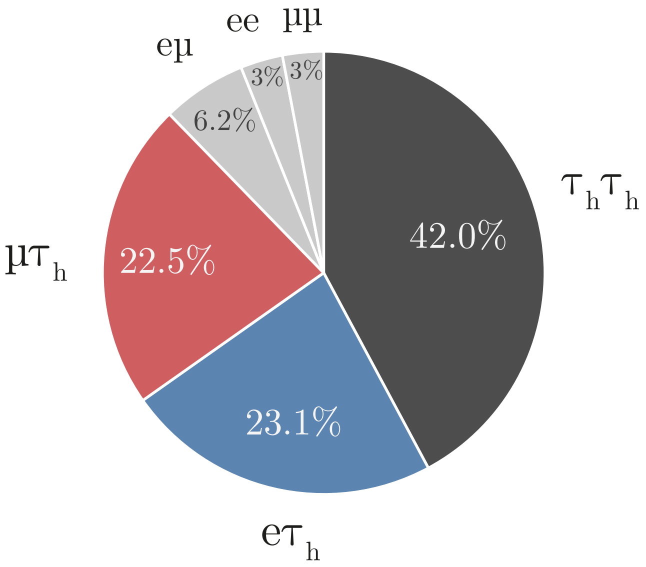


Figure 4.2: The possible final states of a $\tau\tau$ decay. The three final states with the highest branching fraction are considered for this analysis: $\tau_h\tau_h$, $e\tau_h$ and $\mu\tau_h$.

The three final states with the highest branching fraction are used for this analysis: $\tau_h\tau_h$, $e\tau_h$ and $\mu\tau_h$. The $e\mu$, ee and $\mu\mu$ final states do not only have a much lower branching fraction, and are therefore less abundant, they also suffer from a higher background contribution from top-quark pair decays in the $e\mu$ final state, and Z boson decays in the ee and $\mu\mu$ final states. The inclusion of these final states has been studied to impact the sensitivity of the analysis by less than 5% for $m(\mathrm{H}) < 300\,\mathrm{GeV}$, and less than 2% for $m(\mathrm{H}) > 300\,\mathrm{GeV}$.

For an electron, muon or τ_h to be selected to build the tau lepton pair, it needs to pass a set of selection criteria, which are

- to fulfill a certain identification criterion such as a multivariate identifier designed for the object,
- a low relative isolation,
- to fulfill a certain offline p_T threshold, and
- the match of the object to the physics object responsible for the decision of a single- or di-lepton trigger.

The specific requirements differ between electrons, muons and τ_h's and are discussed in more detail in the following.

Events used for the $e\tau_h$ ($\mu\tau_h$) final state are based on the set of events in which a single electron (muon) trigger or an electron(muon)+τ_h pair trigger has fired. In these events, a set of selection requirements shown in Table 4.1 is imposed.

Table 4.1: Minimal selection requirements imposed for eligible electron and muon candidates used for the di-tau lepton pair in the $e\tau_h$ and $\mu\tau_h$ final states. Depending on trigger configuration of the three run periods, more restrictive p_T requirements are imposed to remove events from the trigger turn-on region. The identification working points (WP) as well as the different isolation requirements for electrons and muons are defined in sections 3.2.3 and 3.2.4.

	Transverse Momentum	Pseudo-rapidity	Identification	Relative Isolation	Distance from PV
Electrons	$p_T > 25\,\mathrm{GeV}$	$\lvert\eta\rvert < 2.1$	90% eff. WP	$\mathrm{Iso}_{\mathrm{rel}} < 0.15$	$d_{xy} < 0.045\,\mathrm{cm}$ $d_z < 0.2\,\mathrm{cm}$
Muons	$p_T > 20\,\mathrm{GeV}$	$\lvert\eta\rvert < 2.1$	`medium` WP	$\mathrm{Iso}_{\mathrm{rel}} < 0.15$	$d_{xy} < 0.045\,\mathrm{cm}$ $d_z < 0.2\,\mathrm{cm}$

An additional selection is applied in which the p_T and $\mathrm{Iso}_{\mathrm{rel}}$ requirements are relaxed to $p_T > 10\,\mathrm{GeV}$ and $\mathrm{Iso}_{\mathrm{rel}} < 0.3$. Events in which an additional electron (muon) fulfilling these looser requirements is found in the $\mu\tau_h$ ($e\tau_h$) final state, i.e. events in which both electron and muon in addition to the τ_h candidate are found, are vetoed from the event selection to avoid overlap between the final states. If the additional lepton is contained within a jet, as it often occurs with muons in B hadron decays, the lepton is not an eligible candidate for the $\tau\tau$ system and its presence will not lead to an event veto. This is important as B hadron decays into muons are frequently expected to occur in the signal process, and an event removal based on the presence of leptons in jets would remove a large fraction of signal events.

Events in the $\tau_h\tau_h$ final state are based on the triggering decision of a set of triggers designed for two τ_h candidates, of which both candidates need to be matched to the two legs of the trigger. The offline event selection that is additionally applied to the τ_h candidates is summarized in Table 4.2 for all three final states. The working points of the `DeepTau` classifier, discriminating the τ_h candidates against jets, electrons, and muons are chosen to optimize the ratio of the signal process with respect to the expected background. The correct selection of the working point for the discrimination of τ_h's against jets is the most crucial, as jets abundantly occur in the collisions. It has been studied that the use of the `Medium WP vs. jets` optimizes the expected significance of the analysis and yields up to 20% improved results in the $\tau_h\tau_h$ final state, and up to 10% in the $e\tau_h$ and $\mu\tau_h$ final states when compared to the `Tight WP vs. jets`.
The working points of the discrimination against electrons and muons are chosen to suppress the background of $Z \rightarrow ee/\mu\mu$ events and tighter discrimination against electrons (muons) is chosen in the $e\tau_h$ ($\mu\tau_h$) final state.

Table 4.2: Selection requirements imposed for eligible τ_h candidates used for the di-tau lepton pair. The requirements are dependent on the final state for which the τ_h is selected.

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

Tau lepton pair building

The electron, muon, and τ_h candidates fulfilling the selection requirements above are combined to build a pair of tau lepton decays. For the two constituents to be eligible for a $\tau\tau$ system, they need to be separated by at least $\Delta R(\eta, \phi) > 0.5$. If an additional electron, muon, or τ_h is present in the $e\tau_h$, $\mu\tau_h$, or $\tau_h\tau_h$ final state respectively, the event is not vetoed as no overlap between final states is expected. In this case, more than one di-tau lepton pair can be built. The pair with the most isolated electron or muon is chosen in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states, or with the τ_h identification classifier values closest to one in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final states. If the values are equal within 10^{-5}, the p_T of the electron or muon, or τ_h in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state is compared with preference given to higher p_T. If these are also equal within 10^{-5} GeV, the procedure is repeated for the (second) τ_h of the event.

The electron or muon need to match the physics object responsible for the triggering of the single lepton trigger used to select the event. In case the event was selected exclusively due to a lepton+τ_h pair trigger, or due to a di-τ_h trigger in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, both leptons used to build the $\tau\tau$ pair need to match the respective trigger legs. A match to the triggering physics object is achieved if the offline reconstructed particle has a distance of less than $\Delta R(\eta, \phi) < 0.3$ to the triggering object. Depending on the trigger that was used to select the event, a higher p_T threshold than defined in Tables 4.1 and 4.2 might be required, with the offline p_T threshold to be always at least 1 GeV above the design value of the trigger. This is done to remove events in the turn-on region of the triggers which lead to kinematic biases of the data, which are usually difficult to describe in simulated events. In the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, where the trigger turn-on region is larger than in single muon or electron triggers, the offline p_T needs to exceed the trigger design value by 5 GeV.

4.1.2 Selection of b-jets

The physics objects used for the selection of b-jets are the reconstructed jets as defined in section 3.2.5. The selection requirements for b-jets are summarized in Table 4.3. For comparison, also the requirements imposed on non b-jets are shown. Non b-jets have a stricter p_T requirement, however are not required to be as central as b-jets. The reason is that the valid identification of b-jets is highly dependent on the CMS tracking system, whereas non b-jets depend mainly on the hadronic calorimeter, which extends further in the forward direction of the CMS detector than the tracker.

Table 4.3: Selection requirements used for the selection of the b-jets. The $|\eta|$ requirement is < 2.4 in the 2016 run period, and < 2.5 in the 2017 and 2018 run periods. The `DeepJet` discriminator is described in section 3.2.7.

	Transverse Momentum	Pseudorapidity	Identification
b-jet	$p_\mathrm{T} > 20$ GeV	$\|\eta\| < 2.4(2.5)$	`DeepJet medium`
non b-jets	$p_\mathrm{T} > 30$ GeV	$\|\eta\| < 4.7$	

If at least two b-jets are found in the event, the two jets need to be separated by at least $\Delta R(\eta, \phi) > 0.4$. A bb system is then built from the two b-jets leading in p_T. If only one b-jet is found in the event, the event is still selected, as the efficiency of the b-jet identification is only at the level of 50-80%, depending on the kinematic properties of the jet. A missing identification of a true b-jet is therefore expected to commonly occur. In the case of only one b-jet, a bb system is built from the single b-jet with the

non b-jet with the highest `DeepJet` discriminator score, while still being separated by $\Delta R(\eta, \phi) > 0.4$ from the b-jet.
If there is no b-jet, or only one b-jet and no additional jet present in the event passing the requirements of Table 4.3, the event is removed from the event selection.

As events in which the bb system is build from only one b-jet are expected to be more often background events than events in which two b-jets could be selected, a flag is stored in which the information about the choice of jets used for the bb system is preserved. Furthermore, the exact values of the `DeepJet` discriminator scores of the jets is saved. Both the flag and the discriminator information are fed into the multivariate event classification described in the next chapter, allowing the classifier to infer about the validity of the bb system.

4.1.3 b-jet regression

In b-jets, the presence of leptonic decays and escaping neutrinos often lead to a bias towards a lower energy of the reconstructed jet. To improve the energy measurements of the b-jets for this analysis, a multivariate b-jet energy correction as developed for the measurement of the h_{SM} decay into two b-quarks [51] is applied to events both in data and simulation.
The regression exploits the available information about the jet, using 43 input variables such as the jet p_{T}, η, mass, or the energy composition split in electromagnetic and hadronic components in ring segments in the $\eta - \phi$ plane around the jet's cone axis. These variables are fed into a deep neural network trained on simulated $\mathrm{t\bar{t}}$ events containing genuine b-jets.
The improvement in reconstructed resolution, using the ratio of standard deviation and mean σ/μ of the di-b-jet mass as a figure of merit, is between 5-10% depending on the resonance masses the two additional Higgs bosons. In Figure 4.3, the location of the mean of the di-b-jet-mass distribution relative to the simulated value, as well as its standard deviation with and without applying b-jet regression are shown. After applying the regression, the reconstructed mass is shifted upwards towards the simulated value.
The regression is applied to all b-jets passing the medium working point of the b-jet identification as described above. If only one tagged b-jet exists in the event, the jet with the second-largest score of the b-jet identification classifier is used to build the di-b-jet system. In this case, the regression will also be applied to this jet.

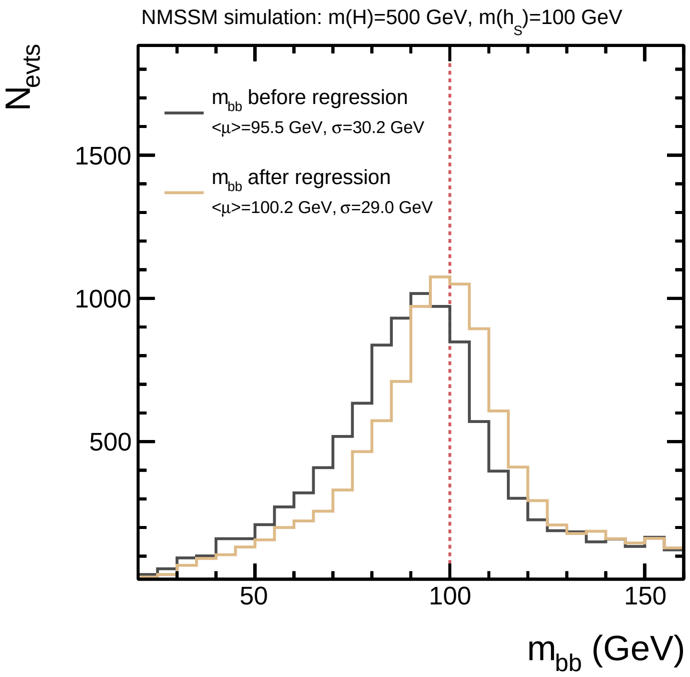


Figure 4.3: Effect of the b-jet regression applied to a simulated NMSSM signal sample of the process $\mathrm{H} \rightarrow \mathrm{h}_{\mathrm{SM}}(\tau\tau)\mathrm{h}_{\mathrm{S}}(\mathrm{bb})$ with $m(\mathrm{H}) = 500\,\mathrm{GeV}$ and $m(\mathrm{h}_{\mathrm{S}}) = 100\,\mathrm{GeV}$, shown as a function of the reconstructed di-b-jet mass m_{bb} before (black) and after (yellow) applying the regression. Applying b-jet regression results in a better agreement between the reconstructed and the simulated mass, while the standard deviation over mean $\sigma / < \mu >$ decreases by 5-10%.

4.2 Background processes

The events in data that contain a reconstructed pair of both $\tau\tau$ and bb and thus pass the selection criteria can enter the analysis due to a variety of known physics processes, which are considered as backgrounds for the search for the signal process.

In order to search for an excess of data events over the sum of these known processes, they need to be estimated as accurately as possible, with the limits of the accuracy reflected in systematic uncertainties. The estimation of background events is performed by either simulation as will be discussed in section 4.5.1, or by the data-driven τ-embedding and F_{F} methods discussed in sections 4.3 and 4.4. The choice of estimation method depends on the underlying physics process of the event.

The relevant physics processes are

- top-quark pair production ($\mathrm{t\bar{t}}$),
- Z boson production (Z),
- Quantum chromodynamics (QCD) multijet production,
- W boson production in association with at least one jet (W+jets),
- production of W or Z boson pairs and single-top quark production (VV) and
- single $\mathrm{h_{SM}}$ production (Higgs).

All physics processes as well as their inclusive occurrence in the analysis are given in Figure 4.4, indicating also the estimation method by which each corresponding process is derived.

Due to the presence of the isolated electron or muon in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states, the background compositions in these final states differ significantly from the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, as the isolated lepton requirement greatly suppresses the QCD multijet background. On the other hand, the background contribution from $\mathrm{t\bar{t}}$ is increased in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states, as both the decay of both top-quarks into tau leptons ($\mathrm{t\bar{t}}(\tau\tau)$) as well as the decay into a prompt electron or muon in addition to a tau lepton ($\mathrm{t\bar{t}}(\ell\tau)$) pose irreducible background sources in the semi-leptonic final states, whereas in $\tau_\mathrm{h}\tau_\mathrm{h}$ only the decay with two genuine leptons $\mathrm{t\bar{t}}(\tau\tau)$ does so.

About 42% of background events are estimated using data-driven estimation methods in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states, and over 94% in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state. All contributions will be discussed in the following.

4.2.1 Top-quark pair production

The Feynman diagrams of the leading order processes of $\mathrm{t\bar{t}}$ production at the LHC are shown in Figure 4.5. With a cross section of around 832 pb in proton-proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$ [52], top-quark pairs are abundantly produced at the LHC.

The presence of a top-quark pair in the event can result in a signature very similar to the $\mathrm{H} \rightarrow \mathrm{h_{SM}h_S} \rightarrow \tau\tau\mathrm{bb}$ signal. As the top-quark will decay almost exclusively (99.8%) to a b-quark and a W boson, $\mathrm{t\bar{t}}$ production is expected to result in two b-jets. Furthermore the W boson decays leptonically into a tau lepton, muon, or electron with a branching fraction of 33%. The overview of possible decays of the full $\mathrm{t\bar{t}}$ system is collected in Table 4.4.

Especially the cases in which both top-quarks decay into tau leptons, or cases in which one decays into a tau lepton and the other into an electron or muon pose a major source of background events in this analysis. The two sources cannot be distinguished from a $\mathrm{H} \rightarrow \mathrm{h_{SM}h_S} \rightarrow \tau\tau\mathrm{bb}$ event by selection requirements to the b-jets or tau lepton decay products alone as the leptons and b-jets are genuine. Therefore, $\mathrm{t\bar{t}}$ events make up by far the largest source of background in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states.

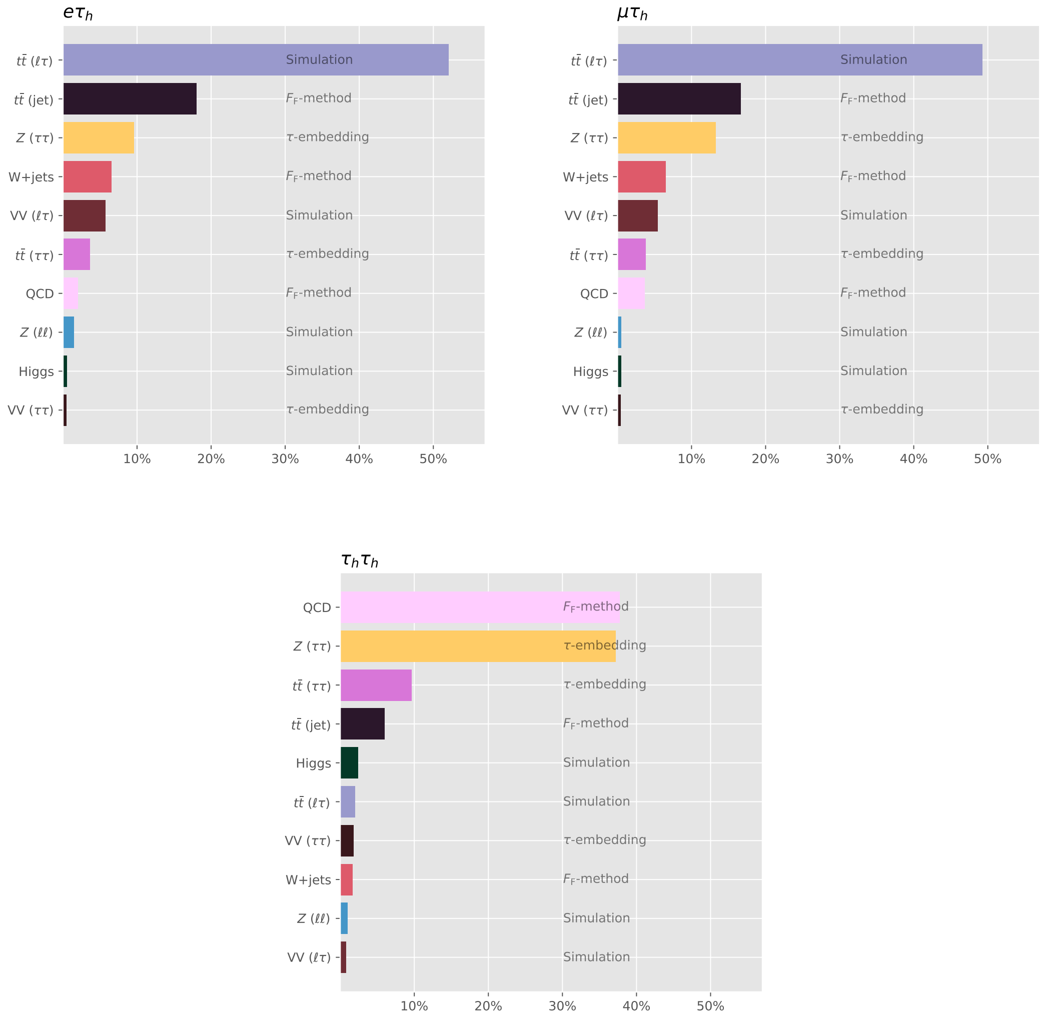
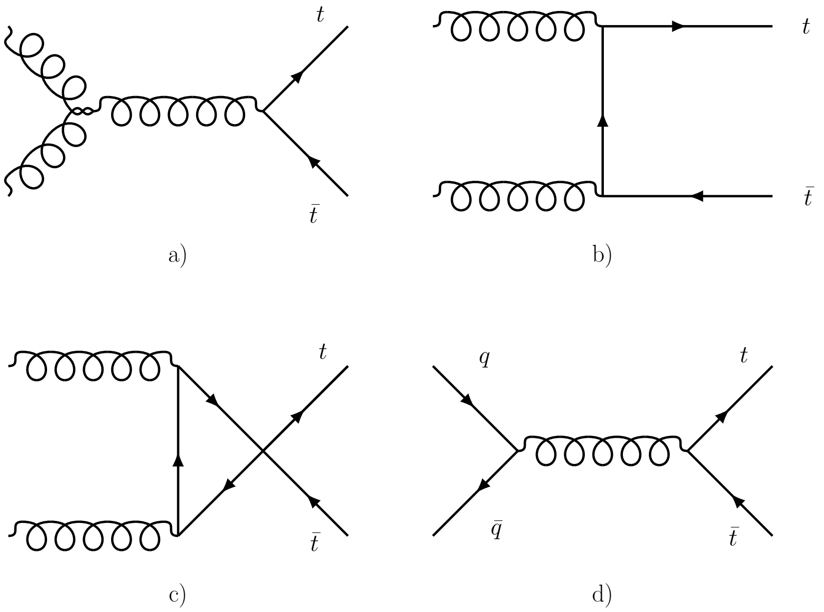


Figure 4.4: Inclusive composition of the total background after event selection in the $\mathrm{e}\tau_\mathrm{h}$ (top left), $\mu\tau_\mathrm{h}$ (top right), and $\tau_\mathrm{h}\tau_\mathrm{h}$ (bottom) final states. The background sources are sorted according to their contribution to the total background. The estimation method of the various background sources is indicated in the figures. The background compositions differ significantly between the semi-leptonic and full-hadronic final states, with $\mathrm{t}\bar{\mathrm{t}}$ production involving prompt leptons dominating the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states, and QCD multijet production dominating the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state. In the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states, in total around 58% of events are estimated from simulation, and around 42% from the data-driven methods of τ-embedding or the F_F method. In the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, less than 6% of events are estimated from simulation and over 94% from data-driven methods.



a) b)

c) d)

Figure 4.5: Leading-order Feynman diagrams of the top-quark pair production process. At the LHC, the production via gluon-gluon fusion (a-c) makes for around 90% of top-quark pair events, while the quark-antiquark annihilation (d) makes for around 10%.

Table 4.4: Branching fractions of decays of the two W bosons from top-quark pair decays in genuine tau leptons (τ), prompt muons or electrons (ℓ), or jets. Decays into tau leptons, muons, or electrons are accompanied by neutrinos which are omitted in the table. Every decay channel creates a significant source of background for this analysis. The leptonic decays (first three rows) take up only less than 10% of all possible $\mathrm{t\bar{t}}$ decays, however will create an event signature with genuine b-quarks and leptons and are hardly distinguishable from signal events.

W_1 decay	W_2 decay	Fraction of $\mathrm{t\bar{t}}$ decays (%)	Estimated by
τ	τ	1.2	τ-embedding
ℓ	τ	2.5	simulation
ℓ	ℓ	4.4	simulation
τ	jet	14.6	F_F method
ℓ	jet	28.5	F_F method
jet	jet	46.5	F_F method

During the multivariate event classification which will be discussed in section 5.1,$\mathrm{t\bar{t}}$ events can be separated from signal events due to their behavior in kinematic distributions: As $\mathrm{t\bar{t}}$ production is non-resonant, the kinematic distributions of e.g. the invariant bb or $\tau\tau$ masses show no resonant behavior, as opposed to the resonant decay of $\mathrm{h_S} \to \mathrm{bb}$ or $\mathrm{h_{SM}} \to \tau\tau$. Due to this, the kinematic fit discussed in section 4.6, in which the compatibility of the event with a resonant $\mathrm{h_{SM}} \to \tau\tau$ decay is quantified, can discriminate between $\mathrm{t\bar{t}}$ and signal events.

If one or both top-quarks decay hadronically, the resulting jets can be misidentified as the τ_h of the event. As opposed to the leptonic decays, this background can be reduced by the `DeepTau vs. jets` discriminator. However, due to the large fraction of $\mathrm{t\bar{t}}$ events decaying into jets as indicated in Table 4.4, also these decays pose a significant source of background. This background is modeled using the F_F method as discussed in section 4.4.

4.2.2 Z boson production

At the LHC, quarks and antiquarks can annihilate to create a fermion-antifermion pair mediated by a Z boson. In addition, but much less frequently, Z bosons can be produced via fusion of two vector bosons. Both mechanisms are shown in Figure 4.6.

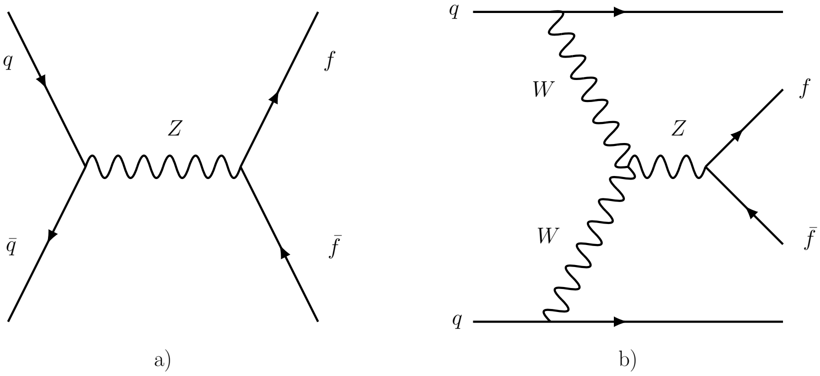


Figure 4.6: Leading-order Feynman diagrams of Z boson production and subsequent decay into a fermion-antifermion pair. The cross section of the Drell-Yan process (a) exceeds the cross section of the production via electroweak boson fusion (b) by three orders of magnitude.

If the Z bosons decays into two tau leptons or two b-quarks, such events can enter as background for the analysis. The process has a production cross section of 26 times larger than $\mathrm{t\bar{t}}$ production. In contrast to $\mathrm{t\bar{t}}$, this background is suppressed by the requirement of both a b-quark pair and a tau lepton pair, of which usually only one is present in the event. Events can still pass the selection criteria due the presence of a b-quark originating from another process or the misidentification of jets as b-jets or τ_h. As the process is resonant at the Z boson mass $m_\mathrm{Z} = 91\,\mathrm{GeV}$, the decay into tau leptons, with the invariant mass smeared out by the escaping neutrinos, is often compatible with the decay of $\mathrm{h_{SM}}$ into tau leptons. Thus it represents a major source of background in all final states. In a multivariate analysis, the slightly lower invariant mass of the resonance as well as the properties of the additionally reconstructed b-jets can help to discriminate these events from the signal process.

An additional minor background source is the decay of the Z boson into muons or electrons, and the subsequent misidentification of one of the leptons as the τ_h of the event. In these cases, no neutrinos are involved and the invariant mass is at 91 GeV, which is compatible with the $\mathrm{h_{SM}} \to \tau\tau$ decay. Such events are however strongly suppressed by the `DeepTau vs. electron/muon` discriminator.

4.2.3 W boson production in association with jets

A similar process to Z boson production is the production of a W boson shown in Figure 4.7. A W boson decays into quarks, or into a lepton and the corresponding (anti)neutrino. A W boson decay can thus be a source of an isolated tau lepton, muon, or electron. If an additional jet in the event, e.g. a jet recoiling from the W boson, is misidentified as a τ_h or a b-jet, the event may be selected and poses a significant source of background mainly in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states. These events are also estimated using the F_F method discussed in section 4.4.

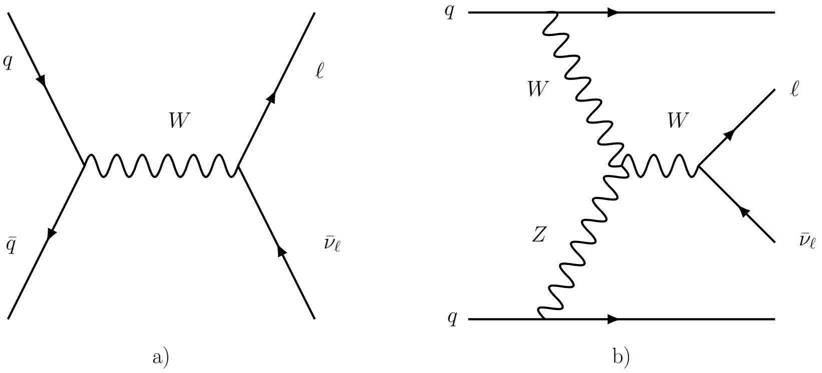


Figure 4.7: Leading-order Feynman diagrams of W boson production and subsequent decay into a pair of lepton and neutrino. The cross section of the quark-antiquark annihilation (a) exceeds the cross section of the production via electroweak boson fusion (b) by three orders of magnitude.

4.2.4 QCD multijet production

The dominant background in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state enters due to QCD multijet events in which both τ_h candidates and often even the b-jets are misidentified light quark or gluon induced jets. The label QCD refers to the fact that the scattering process and subsequent decay is exclusively mediated by the strong force. An exemplary Feynman diagram of this process is shown in Figure 4.8. Such processes result in multiple hadrons collimated in jets. The jets can be misidentified as hadronically decaying tau leptons or b-jets, or, less common, as electrons or muons.

This process is greatly suppressed during event selection by the identification requirements that are applied to the τ_h candidates and b-jets, however still enters the analysis due to its high cross section, which is twelve orders of magnitude above the maximally allowed cross section of the signal process. Due to the large value of the strong coupling constant α_s, many orders of QCD calculation would be necessary to correctly model QCD multijet events by purely using Monte Carlo simulation. Due to this and the large cross section at the LHC, QCD events are estimated from data via the F_F method as discussed in section 4.4.

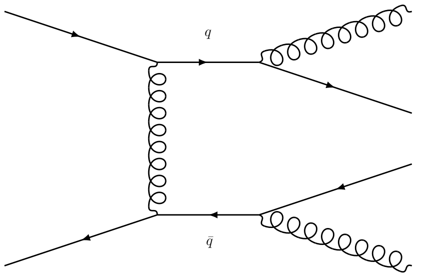


Figure 4.8: Feynman diagram of one of many processes resulting in a QCD multijet event.

4.2.5 Diboson and single-top production

The production of two vector bosons or the production of a single top-quark in association with a b-quark or a W boson enters this analysis as a minor background. The top-quark subsequently decays into a W boson and a b-quark, thus these processes can enter in any final state due to the many decay channels of the vector bosons including isolated leptons or b-quarks. The cross sections of these processes are however small in the order of $O(10)$ pb and thus contribute less than 10% of the total background. The small contribution by the production of a single top-quark is subsumed into diboson events. Both processes are sketched in Figure 4.9.

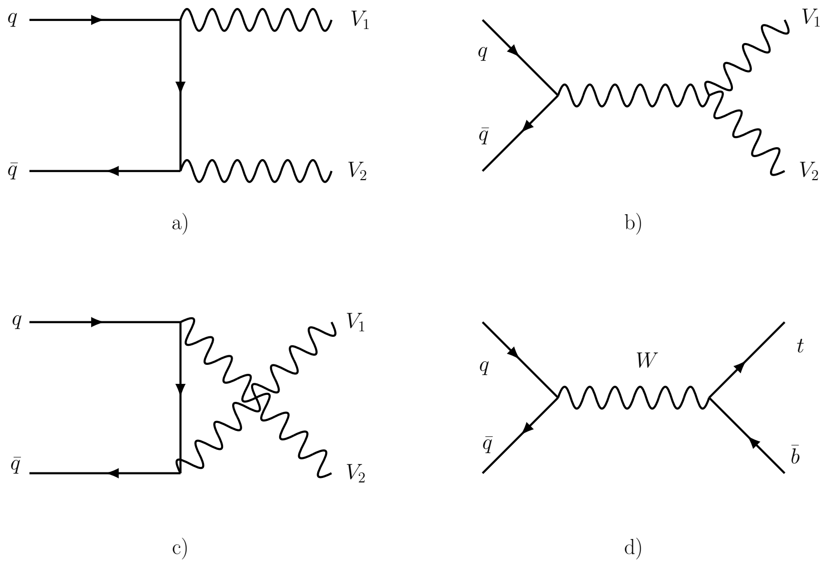


Figure 4.9: Leading-order Feynman diagrams of the diboson production process (a-c) as well as an example diagram of single-top production (d).

4.2.6 Single $\mathrm{h_{SM}}$ production

The production and decay of single $\mathrm{h_{SM}}$ bosons is a background process of this analysis. Even though the cross section of their production is small, if the $\mathrm{h_{SM}}$ decays into two tau leptons, the signature will overlap fully with the signal process. Here, the same decay is searched for, however $\mathrm{h_{SM}}$ is produced in a resonant decay of a heavier Higgs boson H in association with $\mathrm{h_S}$.

4.3 Data-driven estimation of genuine di-tau lepton events

The production and subsequent decay of Z bosons into two tau leptons represents a major source of background for any analysis targeting the decay of a Higgs boson into tau leptons, as the event topologies are very similar and can only be distinguished by the difference in mass. Furthermore, two genuine tau leptons can be produced via the decay of a top-quark pair, or via the production of two vector bosons to a smaller degree, as discussed in section 4.2.

These events are all mediated by the weak force, via either a Z or W boson. The weak force has the property of coupling equally strong to all leptons. The rate and event kinematics of $\mathrm{Z/t\bar{t}/VV} \rightarrow \tau\tau$ decays is thus exactly the same as of $\mathrm{Z/t\bar{t}/VV} \rightarrow \mu\mu$ decays, except for the different signatures of the tau leptons and muons in the detector.

The muon is reconstructed with high efficiency and resolution. This is utilized by the τ-embedding method [53], where reconstructed di-muon events in data are selected, and the muons are replaced with simulated tau lepton decays to describe a genuine decay of a tau lepton pair. The remainder of the event, such as the underlying event, pile-up or jets recoiling against the Z or W boson, is taken as observed in the data.
This improves the modeling of many event quantities which are of relevance for this analysis, such as the p_T and invariant mass of additional (b-)jets in the events, the p_T of the top-quark pair or Z boson decaying into tau leptons, or global event quantities such as $\vec{p}_\mathrm{T}^{\,\mathrm{miss}}$. All these quantities require tuning and are subject to systematic uncertainties when fully simulated. With the use of τ-embedded events, these quantities are described as they are observed in the data without any tuning.

The τ-embedding method can be separated into four distinct steps, which are sketched in Figure 4.10 and will be discussed in the following.

4.3.1 Selection of di-muon events from CMS data

The first step of the τ-embedding technique is the selection of reconstructed di-muon events from the CMS data. As the CMS detector was designed for an excellent reconstruction of muons, this selection can be obtained with high purity and efficiency. Selection criteria such as stringent requirements of the relative isolation of the muons are avoided to circumvent biasing the final τ-embedded sample towards specific event signatures. Events are selected based on the firing of a trigger designed for di-muon events. Kinematic requirements of $p_\mathrm{T} >17$ (8) GeV for the muon (sub-)leading in p_T as well $m_{\mu\mu} > 20\,\mathrm{GeV}$

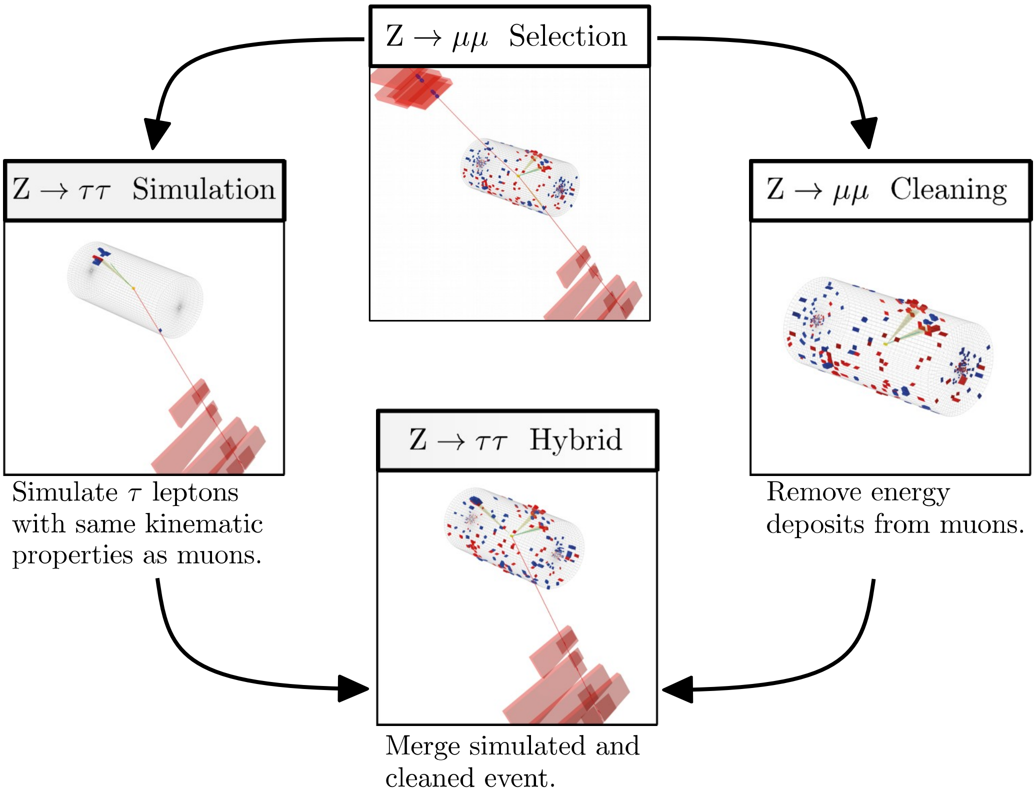

Z → μμ Selection
Z → ττ Simulation
Z → μμ Cleaning
Z → ττ Hybrid
Simulate τ leptons with same kinematic properties as muons.
Remove energy deposits from muons.
Merge simulated and cleaned event.

Figure 4.10: Overview of the τ-embedding technique [53]. Observed events with two muon candidates are selected and the energy deposits of the muons removed from the reconstructed event record. In their place, two tau lepton decays are simulated. The simulated event is then merged with the data event in which the muon records have been removed to achieve a hybrid event modeling the Z boson decay into two tau leptons.

are imposed on the two muon candidates. Furthermore the muons are required to be of opposite charge and to be reconstructed as global muons as defined in section 3.2.4. These requirements define the kinematic region in which the events can be used in a target analysis.

After di-muon selection, a very pure sample of event records is achieved, in which inclusively close to 100% of all selected events are genuine di-muon events, of which over 97% are Z $\rightarrow \mu\mu$ decays, with $\mathrm{t\bar{t}} \rightarrow \mu\mu + X$ and VV$\rightarrow \mu\mu + X$ decays making up the remaining 3%.

When using the embedded event samples in an analysis which imposes additional selection criteria, many of the selected di-muon events might be rejected. In the event selection relevant for this analysis, which is restricted to events that contain b-jets and di-tau masses broadly compatible with $\mathrm{h_{SM}} \rightarrow \tau\tau$, a much higher contribution of $\mathrm{t\bar{t}}$ events is expected with respect to the inclusive sample. The event composition for several selection criteria is shown in Table 4.5.

In the first column of Table 4.5, the invariant mass of the two muons $m_{\mu\mu}$ is required to be above 70 GeV. Events with lower mass usually play a negligible role as they often fail the triggering criteria of the analysis. In the second column, the presence of a b-jet is additionally required. This greatly enriches the fraction of $\mathrm{t\bar{t}}$ and VV events such that they make up over 15% of the remaining embedded event content. The majority of these events are prompt decays of the top-quarks or vector bosons into muons and as such, their replacement with a tau lepton decay is perfectly valid and will lead to a description of the same decays into tau leptons. However, if one of the top-quark or vector boson decays involve a tau lepton which subsequently decays into a muon, the event will cause an overestimation of the event content. This effect is covered by a systematic uncertainty dependent on the estimated fraction of selected $\mathrm{t\bar{t}}$ events in the specific category, which will be further discussed in section 5.2.3.

Table 4.5: Composition of selected di-muon events after two selection criteria. In the first column, the composition after selecting only events with an invariant di-muon mass of at least 70 GeV is shown. In the second column, the event composition with an additional requirement of at least one b-jet fulfilling the `medium` DeepJet classifier is shown.

	Fraction (%)	
	$m_{\mu\mu} > 70\,\mathrm{GeV}$	and $\mathrm{N_{b\text{-}tag}} > 0$
$\mathrm{Z} \to \mu\mu$	98.61	84.02
$\mathrm{t\bar{t}}$	0.65	13.69
VV	0.38	1.99
jet $\to \mu$	0.21	0.24
$\mathrm{Z} \to \tau\tau$	0.05	0.06

The same effect can occur for Z bosons decaying into tau leptons, which further decay into muons. These events are however largely removed by the mass requirements and therefore they play a negligible role in the analysis as shown in the last row of Table 4.5. The same is true for jets misidentified as muon, as shown in the second-to-last row of the same table.

The efficiency of the muon selection is measured as a function of the p_T and η of the reconstructed muon using the tag-and-probe method described in [54] and enters the analysis in form of p_T and η dependent scale factors as will be discussed in section 4.3.5.

4.3.2 Removal of the muon energy deposits

In the second step, the goal is to remove all energy deposits of the two muons from the reconstructed event record. This step is called muon cleaning. The muon leaves hits in the inner tracking system of the detector, small deposits in the electromagnetic and hadronic calorimeters and finally hits in the outer muon system of the CMS detector. The hits are removed based on their association to the reconstructed muon track. Deposits in the form of clusters in the calorimeters cannot be as clearly attributed to the crossed muon. They are defined by the intercept of the muon with a specific calorimeter cell, when interpolating the reconstructed muon track through the calorimeters. If deposits in crossed cells are found, they are set to zero in the event record.

The assignment of a specific calorimeter deposit to a specific particle can be ambiguous. Cases in which the deposit of the muon extends beyond the crossed calorimeter cell, or in which an additional particle caused parts of the energy deposit that was set to zero can lead to small reconstruction effects during the cleaning step. These can cause the appearance of low-energy photons or neutral hadrons with usually negligible importance for the analysis.

4.3.3 Simulation of tau lepton decays

The four-momenta of the muons that have been removed from the event record in the previous step are stored and propagated to the simulation of two tau lepton decays, after a correction for the mass difference between muon and tau lepton. The simulation of the tau lepton decays is performed using `PYTHIA 8.2` [55].
At this stage, the production of embedded events is being branched off to produce samples for four distinct $\tau\tau$ final states: $e\mu$, $\mathrm{e}\tau_\mathrm{h}$, $\mu\tau_\mathrm{h}$ and $\tau_\mathrm{h}\tau_\mathrm{h}$. For each final state, the complete sample of selected di-muon events is used to produce τ-embedded events exclusively in this final state. For this analysis, only the samples describing the $\mathrm{e}\tau_\mathrm{h}$, $\mu\tau_\mathrm{h}$, and $\tau_\mathrm{h}\tau_\mathrm{h}$ final states are used. The exclusive production of a single final state increases the number of available events, which is limited by the di-muon data, by a factor of the inverse branching fraction of the final state, so around 2.4 in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state and around 4 in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states.

In addition to the embedding of simulated tau lepton decays, events are created in which simulated electrons or muons are embedded into the event. These samples serve for validation and for the derivation of corrections for the τ-embedded events as will be discussed in section 4.3.5.
As the tau lepton decay always involves at least one neutrino, a significant fraction of the original tau lepton energy can be carried away undetected and the event could miss the selection requirements on the visible decay parts of the target analysis. To avoid this, requirements on the minimal p_T and maximal $|\eta|$ of the visible parts of the tau lepton decays are defined for the simulation of the tau lepton decay, depending on the final state. Only events passing these requirements are stored in the embedded event samples. To increase the probability of the event passing these kinematic requirements, the decays of the two embedded tau leptons is repeated $N = 1000$ times. If no trial passes the requirements, the event is discarded. Otherwise, the last passing trial is saved. To account for the bias on kinematic distributions introduced by this procedure, its efficiency is calculated by the ratio of all passing trials over the total number of trials $\epsilon = N_\mathrm{pass}/N$ and used as an additional weight representing the probability of this event to occur given the direction and energies of the muons.

This procedure shifts events from a lower-energy phase space to the phase space relevant for usual analyses of di-tau events and thus further enriches the number of events available for the estimation of SM di-tau lepton decays. Due to both the independent production of final states and the kinematic filtering, the number of τ-embedded events used for modeling is larger than the expected number of $\tau\tau$ events in data by a factor of 5 in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state and around 15 in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states.

4.3.4 Merging of simulated and reconstructed event records

In a final step, the simulated event record which only contains the two tau lepton decays is merged with the observed event record from which the muons have been removed during the cleaning step. This step is ideally done at the earliest possible stage of the reconstruction sequence to ensure that all reconstructed event properties are based on the complete di-tau lepton event, which would be the merging at the level of reconstructed hits in the tracking system and individual energy deposits in the calorimeters. Slight differences between the simulated detector geometry used during the simulation step and the geometry of the actual detector, used for reconstruction of the observed event complicate the merging on this level as they cause changes to the reconstruction of tracks. The tracks of the tau lepton decay products as well as the response of the trigger system to the event are thus calculated based on the simulated event, before merging with the remnant observed event. This can lead to differences in the efficiency of the triggering or identification of the tau lepton, which are corrected for as will be discussed in the following section.

4.3.5 Corrections applied to τ-embedded events

Limited efficiency of di-muon selection

Genuine di-muon events which do not cause a triggering decision or which do not fulfill the requirement of a global muon reconstruction are missed during the selection. This leads to an underestimation of the estimated di-tau lepton events.

To correct the event loss, the efficiency of the selection is measured as a function of the p_T and η of the muons. For this, the efficiency of the trigger and identification criteria is measured using the tag-and-probe method for each muon candidate individually. Figure 4.11 shows the resulting efficiency for CMS data recorded during the 2018 run period. Overall, a very high efficiency of the two criteria usually above 95% is achieved. The structure of the CMS detector appears via a measured efficiency loss in regions with incomplete coverage of the muon detection.

The correction is then performed by inverting the measured efficiency ϵ, and applying it as a weight w for the τ-embedded event sample during the analysis as

$$w(p_\mathrm{T}, \eta) = \frac{1}{\epsilon(p_\mathrm{T}, \eta)} \ . \tag{4.1}$$

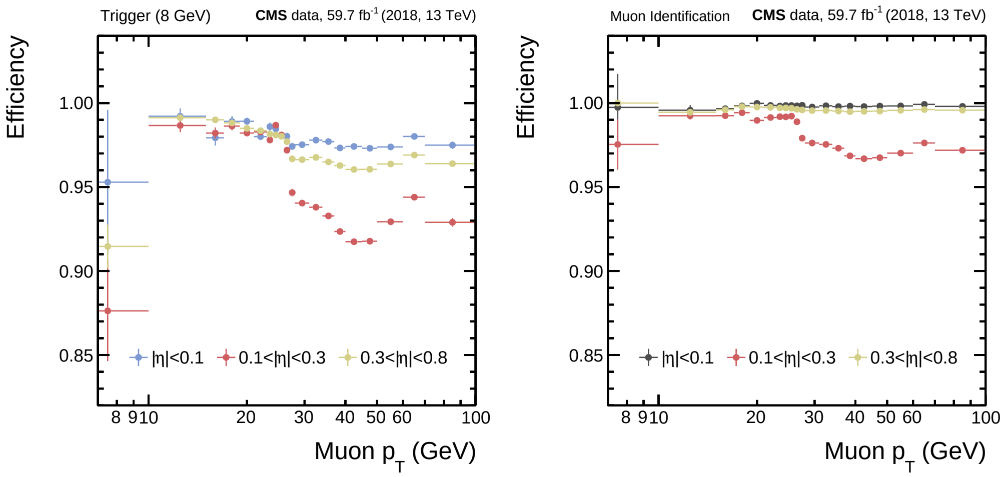


Figure 4.11: Efficiency of the $p_{\mathrm{T}} \geq 8\,\mathrm{GeV}$ leg of the di-muon trigger (left) and of the muon identification criterion (right), for events in the central barrel (blue), intermediate (red) and outer barrel region (yellow) of the CMS detector. The efficiency was measured using data of the 2018 run period. The intermediate region containing non-detecting material as shown in Figure 3.2 between two muon chambers of the barrel leads to an efficiency loss of the muon selection in this region. The efficiency loss is corrected for the application of τ-embedded events.

Correcting differences of electron, muon and τ_{h} reconstruction

As the tau lepton pair in a τ-embedded event is simulated, the efficiency of the reconstruction of the tau lepton decay products can differ from the observed efficiency in data. The same correction factors derived for the full simulation cannot be used, as the modeling of the underlying event can influence the efficiency, and, more importantly, as the triggering and tracking of the tau lepton decay products is performed before merging with the remaining event record from data as discussed in section 4.3.4. In general, this leads to a higher efficiency of the reconstruction of tau lepton decay products in embedded events.

For electrons (muons), a measurement of the efficiency of the identification and isolation requirements as well as the triggering of the lepton is performed using the tag-and-probe method with a selection aiming at $\mathrm{Z} \to \mathrm{ee}(\mu\mu)$ events. In the Z boson decay, two genuine electrons or muons are expected, and the selection can be obtained with high purity when restricted to events in which the di-lepton mass lies in a window of $[65, 115]\,\mathrm{GeV}$, which is around the Z boson mass of $m_{\mathrm{Z}} = 91\,\mathrm{GeV}$. To ensure that at least one lepton is genuine, strict requirements are imposed to one of the leptons, viz. a tight identification requirement, being well-isolated with $I_{\mathrm{rel}}^{\mathrm{e}/\mu} < 0.15$ as well a match of the lepton to the triggering object of a single-lepton trigger in the event. Furthermore a minimal $p_{\mathrm{T}}^{\mathrm{e}/\mu}$ between 23 and 28 GeV is required. This lepton is used to tag the event and is thus referred to as tag lepton.

The requirement on the second lepton can then be much looser, requiring only minimal

requirements on the p_{T} of the lepton. This lepton is used to probe the efficiency of its identification, isolation and trigger requirement and is referred to as probe lepton. The measurement is performed as a function on the p_{T} and η of the probe lepton.
In order to measure the three efficiencies separately, the probing requirement of the first measurement has to be applied in the second measurement.
This means that the identification (id) requirement is applied to the probe lepton for the measurement of the isolation (iso) requirement, which in turn are both applied for the measurement of the trigger (trig) requirement.

$$\epsilon(\text{id,iso,trig}) = \epsilon(\text{trig}|\text{iso,id}) \cdot \epsilon(\text{iso}|\text{id}) \cdot \epsilon(\text{id}) \tag{4.2}$$

To derive correction factors for τ-embedded events, e- or μ-embedded events are created by replacing the muons in the data event with simulated electrons or muons. Performing the measurement with these samples thus covers all effects introduced by the method, while at the same time accounting for the efficiency difference between the detector response of the simulation. The measurements and the resulting correction factors for the 2018 run period are shown in Figure 4.12.
For the measurement of the identification efficiency of hadronically decaying tau leptons, a selection enriching Z → $\tau\tau$ events is chosen in the $\mu\tau_{\mathrm{h}}$ final state by requiring the muon transverse mass to be below 60 GeV and the muons and τ_{h} to be within $|\Delta\eta| < 1.5$. The correction factor is derived from a maximum likelihood fit to the measured data, binned in the visible invariant mass of the tau lepton pair (m_{vis}). The cross section and energy scale of the background processes taken into account as nuisance parameters to the fit. All corrections for the muon reconstruction described above are applied before the fit.

A fit is performed independently for the four possible τ_{h} decay modes: One-prong, one-prong+π^0, three-prong and three-prong+π^0. A separate fit is also performed for all decay modes inclusively and instead as a function of the $p_{\mathrm{T}}^{\tau_{\mathrm{h}}}$. Both measurements are used in this analysis: In the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state, the decay mode dependent corrections are used as only τ_{h} candidates with $p_{\mathrm{T}}^{\tau_{\mathrm{h}}} > 40\,\mathrm{GeV}$ are used in the final analysis, after which no large $p_{\mathrm{T}}^{\tau_{\mathrm{h}}}$ dependence is expected anymore. In the eτ_{h} and $\mu\tau_{\mathrm{h}}$ final state, where the $p_{\mathrm{T}}^{\tau_{\mathrm{h}}}$ requirement is lower at 30 GeV, the $p_{\mathrm{T}}^{\tau_{\mathrm{h}}}$ dependent factors are used. Both correction factors are shown in Figure 4.13.

The measurement of the trigger efficiency of the τ_{h} is performed after the identification criterion is applied, similar to the measurement of the lepton triggers as

$$\epsilon(\text{id,trig}) = \epsilon(\text{trig}|\text{id}) \cdot \epsilon(\text{id}) \ . \tag{4.3}$$

The difference between the applied τ_{h} triggers with respect to the electron or muon triggers is that they always fire according to the presence of two physics objects, either as a lepton+τ_{h} pair trigger, or a di-τ_{h} trigger. The full efficiencies of these triggers is derived by measuring the efficiency of the τ_{h} leg of a $\mu + \tau_{\mathrm{h}}$ pair trigger in the $\mu\tau_{\mathrm{h}}$ final state, using again a selection enriching Z → $\tau\tau$ events.

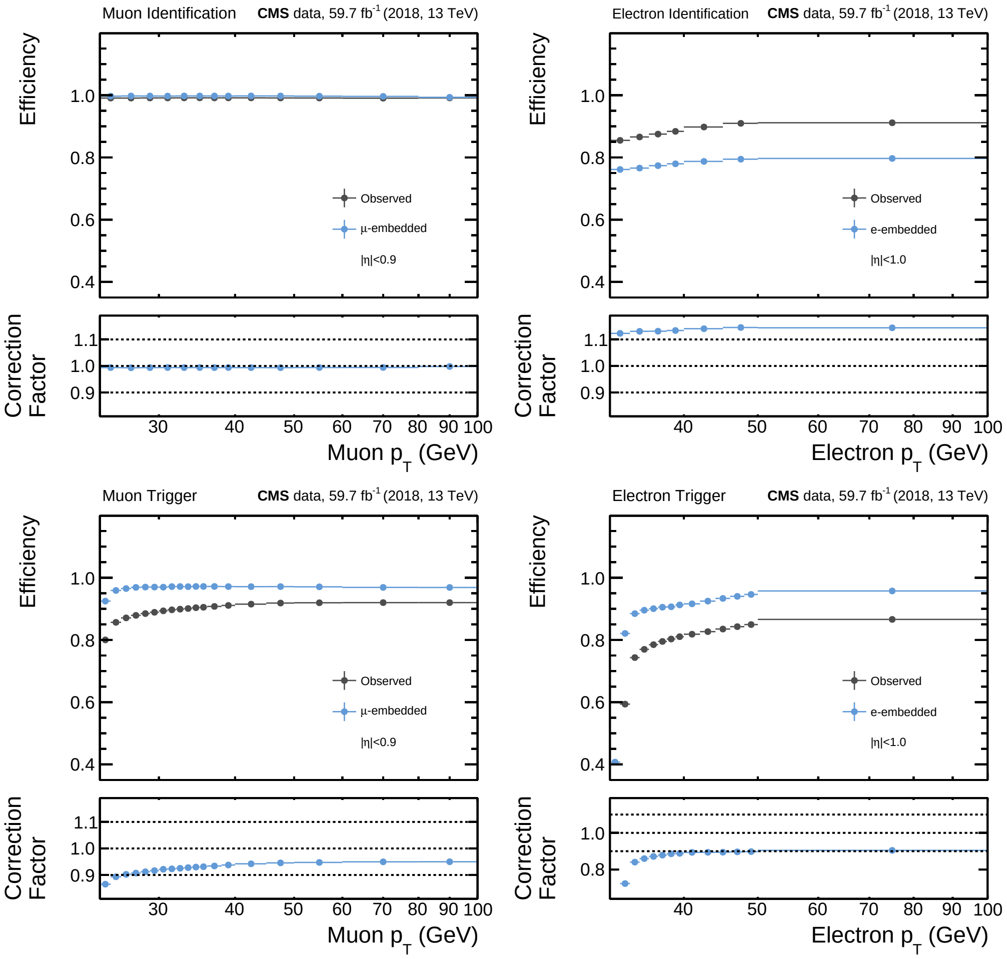


Figure 4.12: Efficiency of the identification (top row) and triggering (bottom row) of muons (left) and electrons (right) in μ/e-embedded events (blue), compared to data (black). The shown measurement is restricted to leptons in the central barrel region of the CMS detector, similar measurements are performed for the other detector regions. The correction factor to be applied on τ-embedded events, calculated as the data efficiency divided by the efficiency of the μ/e-embedded event samples, is shown in the lower panels of the corresponding figures.

The measurement is independently performed for the four main decay modes of the τ_h as a function of the $p_\mathrm{T}^{\tau_\mathrm{h}}$. A fit to the trigger turn-on curve is performed both for τ-embedded and for data events, with the ratio of the two fitted curves used as p_T-dependent correction factor. The measurements and resulting correction factors used in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state is shown in Figure 4.14. After the p_T-dependent correction, an additional correction in bins of η and ϕ is performed to account for a dependence of the efficiency on the CMS detector submodules.

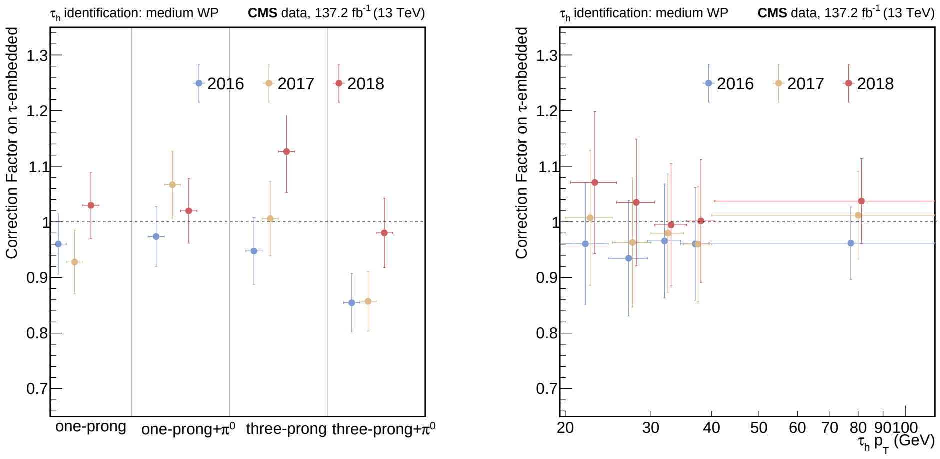


Figure 4.13: Correction factor for the identification of hadronically decaying tau leptons as measured for τ-embedded events using the `medium` working point of the `DeepTau` classifier. Two independent measurements are performed as a function of the τ_h decay mode (left) or p_T (right). The former is used in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, the latter in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states. The correction factor is measured independently for the 2016 (blue), 2017 (yellow), and 2018 (red) run periods. In the right figure, the measured points for 2017 and 2018 are shifted by 1-2 GeV to the right with respect to their bin centers to improve the visibility.

Calibration of electron and τ_h energy

Apart from the reconstruction efficiency, also the reconstructed energy of the simulated tau lepton decay products is different from the observed energy scale. As the energy reconstruction of the muons is very precise, differences in muon energy scale between simulated and observed muons are negligible compared to the differences for electrons and τ_h's and only corrections and corresponding uncertainties for the latter two are derived.

To correct the electron energy scale, a $\mathrm{Z} \rightarrow \mathrm{ee}$ control region is defined using well-isolated di-electron events. The observation in these events is estimated using e-embedded events. In these events, the description of the sharp Z boson peak allows to accurately determine the difference between the energy scale of electrons in embedded events and observed electrons. A maximum likelihood fit is performed to determine the best value of the energy scale shift, using the normalization of the embedded events as independent nuisance

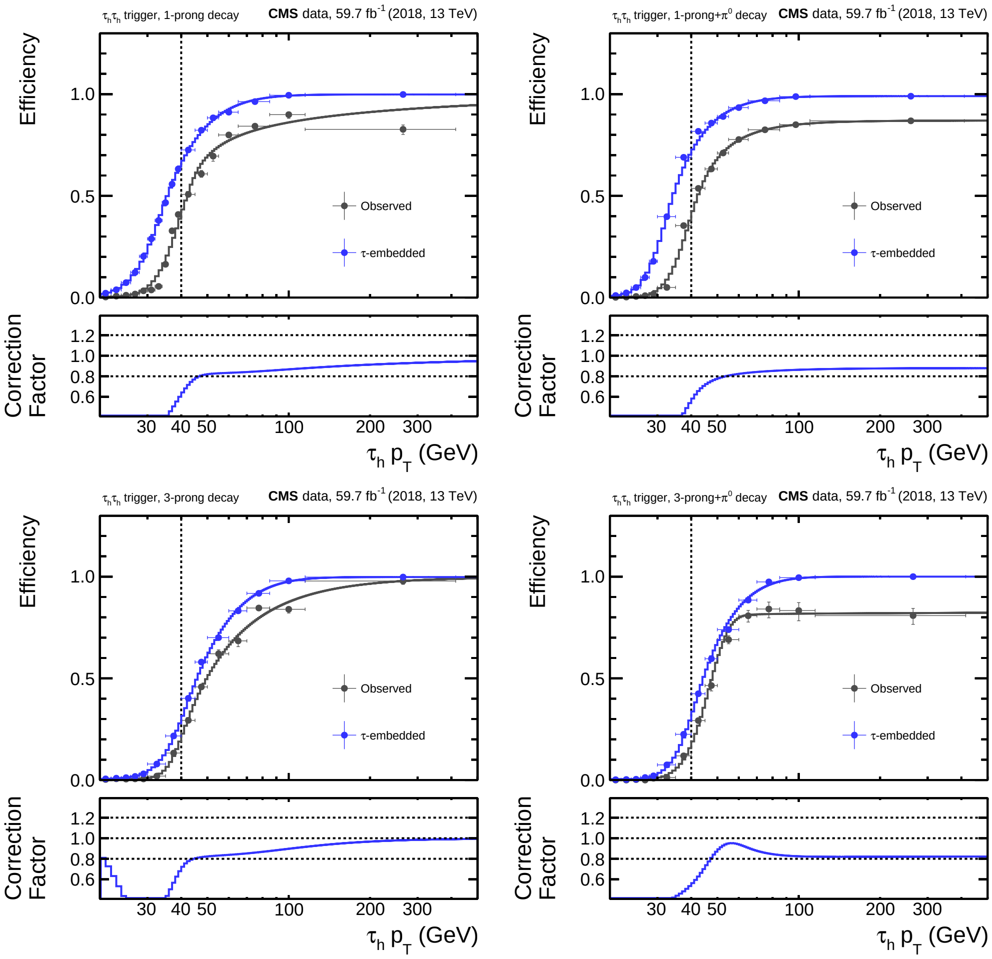


Figure 4.14: Efficiency of the triggering of hadronically decaying tau lepton pairs in τ-embedded events (blue), compared to data (black), as determined for events in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state of the 2018 run period. The measurement is performed separately for four decay modes of the τ_{h}: one-prong (top left), one-prong+π^0 (top right), three-prong (bottom left), three-prong+π^0 (bottom right). The measured values are shown as error points. The correction factor is derived from a fit to the measured values, which is shown as a line. The offline selection criterion of $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ is indicated by a dashed line.

parameter. The measurement is performed independently for electrons in the barrel region ($|\eta| < 1.479$) and electrons in the endcap region ($|\eta| > 1.479$) of the CMS detector, as well as for the three LHC run periods. The resulting best-fit values for the energy scale correction to be applied to embedded events are shown in Table 4.6.

Table 4.6: Corrections applied to the electron energy scale in τ-embedded events as determined from a fit in a Z → ee control region.

Run period	Barrel region (%)	Endcap region (%)
2016	-0.23 (±0.5)	-0.70 (±1.25)
2017	-0.07 (±0.5)	-1.13 (±1.25)
2018	-0.33 (±0.5)	-0.56 (±1.25)

For the measurement of the τ_h energy scale, again a $\mathrm{Z} \to \tau\tau$-enriched control region is used. This is achieved by requiring no b-jets in the event as well as a transverse muon mass below 50 GeV. Backgrounds other than genuine $\tau\tau$ events are estimated with the F_F method or simulated events which will be explained below. The identification corrections for muons and τ_h candidates, as well as their uncertainties are applied in the same way as for the main analysis. Furthermore, the normalization of the τ-embedded events and the remaining backgrounds enter the measurement as independent nuisance parameters. A binned maximum likelihood scan is performed as a function of m_vis. The measurement is performed independently for three τ_h decay modes. The measurement in which the τ_h decays into three charged pions is performed in combination for cases with or without an additional π^0.

An example of the maximum likelihood scan as well as the resulting distribution of m_vis is shown in Figure 4.15 for events of the 2018 run period, restricted to τ_h decays into three charged pions. The measurements for all run periods and decay modes are summarized in Table 4.7.

Table 4.7: Corrections applied to the τ_h energy scale in τ-embedded events as determined from a fit in a $\mathrm{Z} \to \tau\tau$ control region.

Run period	1-prong (%)	1-prong+π^0 (%)	3-prong (%)
2016	$-0.20^{+0.46}_{-0.46}$	$-0.22^{+0.22}_{-0.25}$	$-1.26^{+0.33}_{-0.51}$
2017	$-0.04^{+0.41}_{-0.42}$	$-1.20^{+0.52}_{-0.21}$	$-0.75^{+0.44}_{-0.46}$
2018	$-0.33^{+0.39}_{-0.39}$	$-0.57^{+0.37}_{-0.31}$	$-0.74^{+0.32}_{-0.32}$

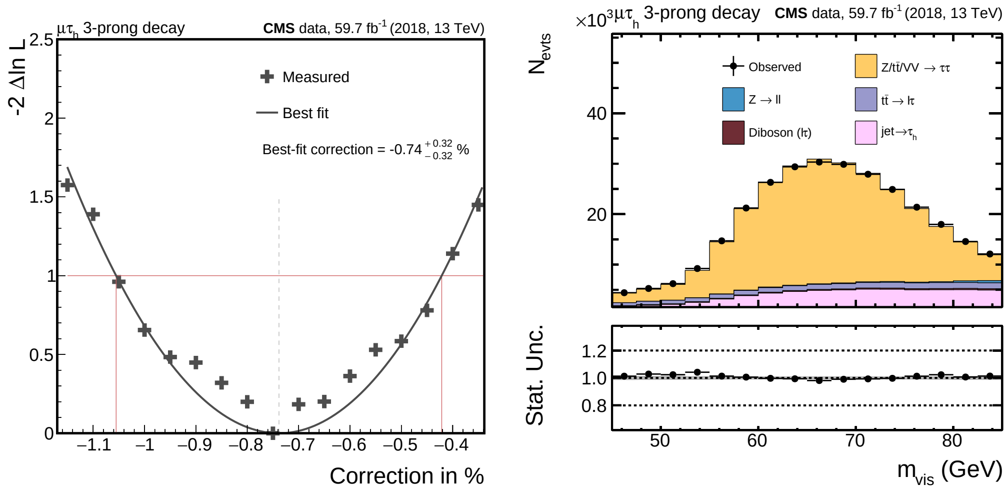


Figure 4.15: Left: Maximum likelihood scan to derive a best-fit value of the energy scale correction for genuine τ_h decays. A parabola (black line) is fitted to the measured values (black markers) for the $-2\Delta\ln(\mathrm{L})$ with the minimum defining the best-fit value used for the correction in the final analysis. The uncertainty is indicated by the red line. Right: The spectrum of m_vis after the correction of the τ_h energy scale has been applied to the τ-embedded events.

4.4 Data-driven estimation of jets misidentified as τ_h

Significant progress has been made in the identification of τ_h against jets, as was discussed in section 3.2.6. Still, a τ_h has a signature similar to the abundant production of light quark and gluon jets at the LHC and a significant background thus enters the analysis due to the misidentification of these jets as τ_h. As shown in section 4.2, these events make up over 20% of the backgrounds in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states, and over 40% in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, and are mainly caused by QCD multijet production, the production of single W bosons in association with jets, or hadronic decays of top-quark pairs.

The F_F method presented in the following is used to describe these backgrounds in a data-driven way. The general method was first introduced at CMS in the measurement of the Z boson cross section in [56] and was since used in multiple analyses related to the a measurement of the SM Higgs boson [57, 58], or searches of additional Higgs bosons in the di-tau final state [59]. The method is designed as a sideband-method, in which events from multiple regions orthogonal to the *signal region* (SR) are used to model the SR. The SR is the region given by the event selection of the analysis, defined in section 4.1. The method described in the following is adjusted specifically to allow the estimation of misidentified jets in the SR of this analysis. A sketch of the method is shown in Figure 4.16.

The relevant sideband regions of the method are labeled *application region* (AR) and *determination region* (DR). The AR is defined with exactly the same event selection as the SR, with the only difference being the quality of τ_h candidates, expressed as working points of the τ_h discriminator. The `DeepTau vs. jets` discriminator used in

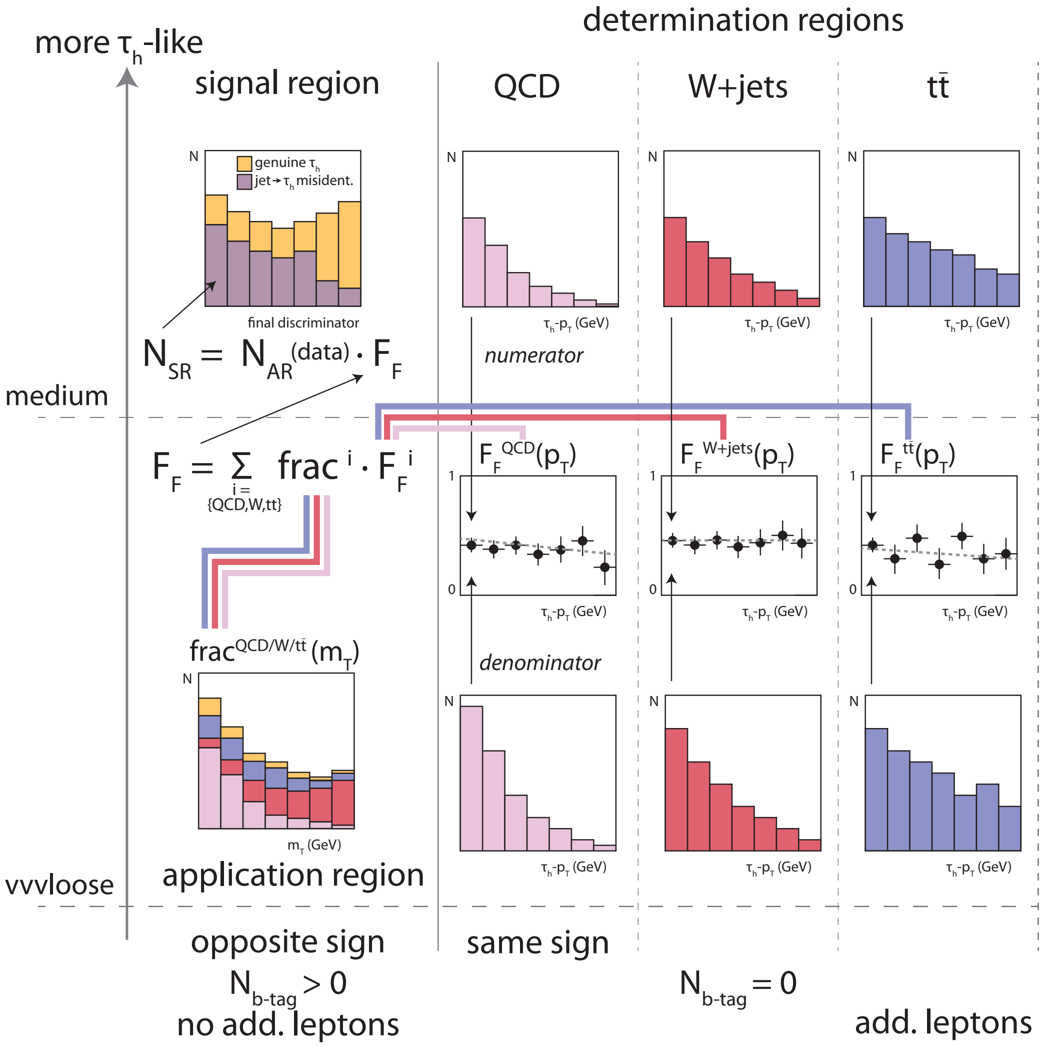


Figure 4.16: Schematic overview of the F_{F} method used to describe events entering the analysis due to jet→ τ_{h} misidentification. Three determination regions are defined (right side), each with a selection requirement ensuring both the orthogonality to both signal region (SR) and application region (AR) as well as enriching the desired process. For each process, a ratio F_{F}^{i} is calculated by dividing the number of SR-like events by the number of AR-like events. In the AR, the expected event composition of the three processes is derived. The final factor F_{F} can then be computed as the sum of the individual sources weighted by their expected fractions. This factor F_{F} is finally multiplied with the observed events in the AR to estimate the events entering due to jet → τ_{h} misidentification in the SR.

this analysis is described in section 3.2.6.

For the discriminator, several working points with varying trade-off between efficiency and purity of the discrimination are defined. Events in which the τ_h candidate fulfils the `medium` working point enter the SR and are used for the search. Inverting the requirement to `!medium` creates an orthogonal event region and defines the AR. A complete inversion would however select mainly events in which the discriminator clearly rejects the τ_h candidate as genuine and results in large differences between the τ_h candidates of the AR and the SR. Therefore, the AR is built from events which fail the `medium` working point, however pass the loosest working point of the discriminator, which is labeled `vvvloose`.

- SR: Events fulfill `medium` working point of `DeepTau vs. jets`
- AR: Events fulfill `vvvloose and not medium` working point of `DeepTau vs. jets`

The AR thus enriches events in which the τ_h candidate could not be clearly identified as either jet or τ_h. Events in which the τ_h is genuine make up less than 20% of events in the AR and are estimated using simulated events and subtracted from the observed events in the AR.

The extrapolation factor F_F, after which the method is named, is used to connect the AR to the SR and defined as

$$N_\mathrm{SR}(\mathrm{data}) = N_\mathrm{AR}(\mathrm{data}) \cdot F_\mathrm{F}(p_\mathrm{T}, N_\mathrm{jets}, ...)\ . \tag{4.4}$$

The factor F_F depends on the event kinematics. It is, in first order, derived as a function of the p_T of the τ_h candidate as well as the number of jets in the event, as the extrapolation is highly dependent on these quantities. In second order, also the dependence on other event characteristics enters for the F_F derivation in the form of non-closure corrections. Furthermore, the factors are derived independently depending on the physics process creating the jet responsible for the misidentification. The derivation of F_F is performed in the DRs which will be discussed in detail in the next sections.

The measured factors are applied on an event-by-event basis for each event in the AR. Here, the probability of the event to be caused by one of the three processes QCD multijet, W+jets or $\mathrm{t\bar{t}}$ production is included. The probability estimate is based on the relative fraction frac^i of the process i, binned in two event properties on which this fraction is largely dependent. The event properties are the transverse electron or muon mass $m_\mathrm{T}^{\mathrm{e}/\mu}$ in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states in combination with the number of b-jets in the event. The $m_\mathrm{T}^{\mathrm{e}/\mu}$ was chosen due to its discriminative power between QCD multijet and W+jets events. In the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, where W+jets events play a much smaller role, the event property used for the binning is m_vis. The relative fractions are estimated in the AR using τ-embedded and simulated events and are shown in Figure 4.17. They are used to weight the three independent contributions as

$$F_\mathrm{F} = \mathrm{frac}^\mathrm{QCD} \cdot F_\mathrm{F}^\mathrm{QCD} + \mathrm{frac}^\mathrm{W+jets} \cdot F_\mathrm{F}^\mathrm{W+jets} + \mathrm{frac}^\mathrm{t\bar{t}} \cdot F_\mathrm{F}^\mathrm{t\bar{t}} \tag{4.5}$$

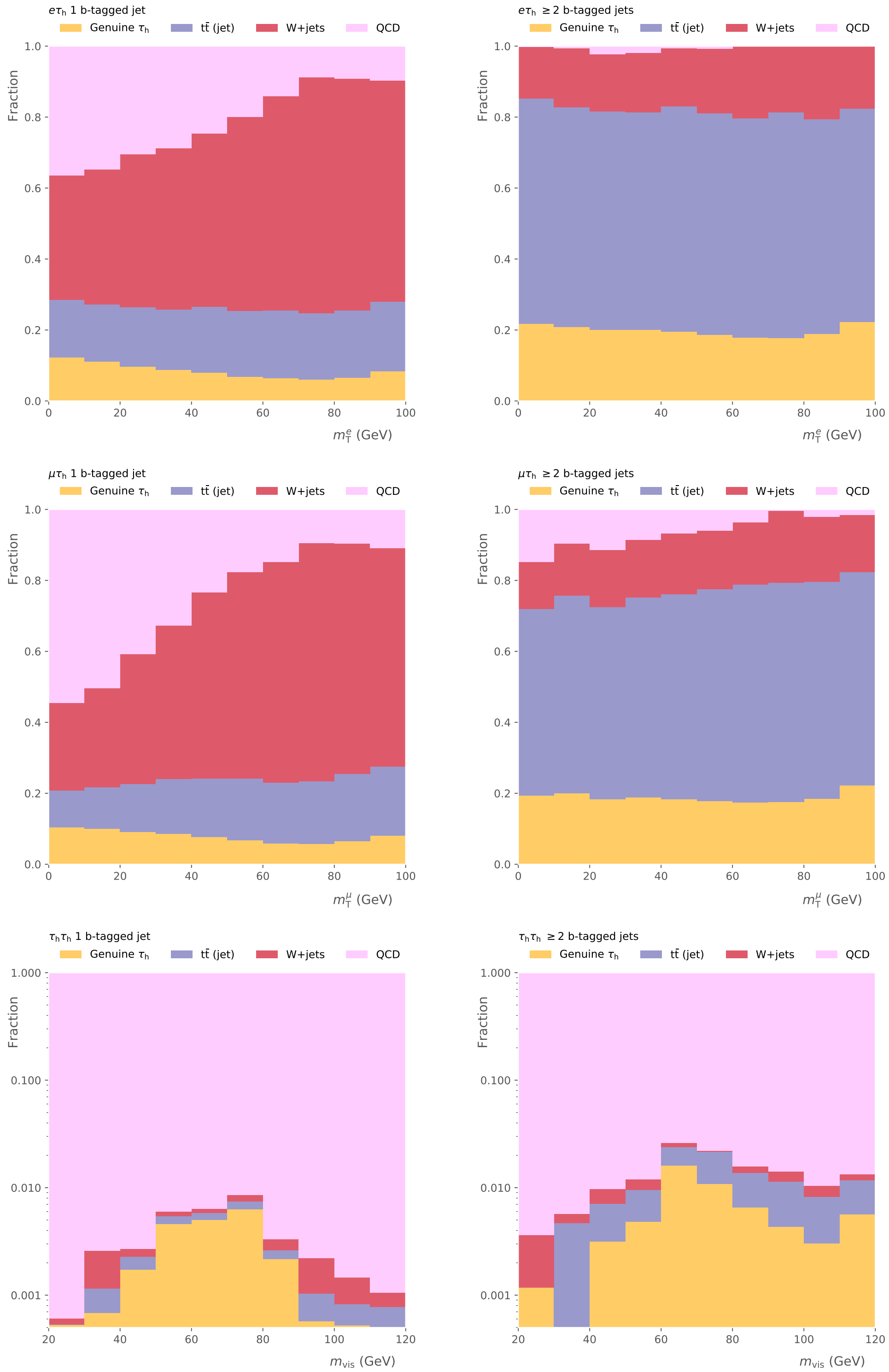


Figure 4.17: Fractions of the three processes QCD (pink), W+jets (red), and $t\bar{t}$(jet) (purple) in the AR of the $e\tau_h$ (top row), $\mu\tau_h$ (middle row) and $\tau_h\tau_h$ (bottom row) final states, as estimated for the 2018 run period. The fractions are derived independently for one b-jet (left) or at least two b-jets in the event (right), as well as a function of the transverse lepton mass in the $e\tau_h$ and $\mu\tau_h$ final states, or of m_{vis} in the $\tau_h\tau_h$ final state. Events in the AR which are genuine, i.e. did not enter due to a jet→τ_h misidentification (yellow) are subtracted from the event distributions in the AR. The AR of the $\tau_h\tau_h$ final state is heavily dominated by QCD multijet events such that other contributions make up less than 3% of the events. Note that a logarithmic scale is used in the bottom row, whereas a linear scale is used in the first two rows.

To apply the estimation method, the events in the AR are binned in the final discriminator of the analysis. In this case, the final discriminator is a neural net output score as will be discussed in the next chapter. Using equations 4.4 and 4.5, the estimation for the SR is extracted from the data events in the AR, which are weighted by their event-by-event value of F_{F}.

4.4.1 Estimation of QCD multijet events

In events entering the analysis due to QCD multijet production, only the strong force is involved in the hard interaction vertex. Both τ_{h} candidates, or both the lepton and the τ_{h} candidate are therefore assumed to enter the analysis due to a misclassification of a jet. It is the leading background process in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state, as was shown in Figure 4.4. For the definition of the DR of this process, the fact is utilized that the charges of the two reconstructed tau lepton decay products are, at first order, uncorrelated. In QCD multijet events, roughly as many events are expected in which the two decay products have equal charge than events in which both have opposite charge. For all other processes, such as the signal process or genuine tau lepton sources $\mathrm{Z/t\bar{t}/VV} \rightarrow \tau\tau$, only oppositely-charged tau lepton pairs are expected, and their opposite charge is a requirement of the event selection for the SR.

The DR is therefore defined similar to the SR, however the reconstructed tau lepton pairs are required be of same-sign charge. Furthermore, a requirement on the transverse mass of the electron or muon in the $\mathrm{e}\tau_{\mathrm{h}}$ and $\mu\tau_{\mathrm{h}}$ final states of $m_{\mathrm{T}}^{\mathrm{e}/\mu} < 50\,\mathrm{GeV}$ is imposed to suppress events in which the electron or muon originates from a W boson decay. Finally, the requirement of a b-jet to be present is dropped.

The DR has a QCD multijet purity of greater 99% (97%) in the `!medium&&vvvloose` (`medium`) τ_{h} identification region in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state. In the $\mathrm{e}\tau_{\mathrm{h}}$ and $\mu\tau_{\mathrm{h}}$ final states a larger fraction of W+jets events enters the DR. The overall QCD purity is greater than 75% (80%) for the two τ_{h} identification regions. All impurities are estimated from τ-embedded samples or simulation and subtracted from the data. The remaining data are assumed to be purely QCD multijet events, and $F_{\mathrm{F}}^{\mathrm{QCD}}$ is calculated as a function of the p_{T} of the τ_{h} of the event by dividing the event yield of the `medium` by the `!medium&&vvvloose` τ_{h} identification region. In the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state, where two τ_{h} are present, each event is used twice, with the measurement being performed for each τ_{h} candidate independently. The calculation is performed independently for events with zero, one, or at least two jets in the event. The distribution of events in the DR and the resulting measurement of $F_{\mathrm{F}}^{\mathrm{QCD}}$ are shown in Figure 4.18. The measured values are subject to a third-order polynomial fit in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state, and linear fit in the $\mathrm{e}\tau_{\mathrm{h}}$ and $\mu\tau_{\mathrm{h}}$ final states. The resulting fit function is used as estimate of $F_{\mathrm{F}}^{\mathrm{QCD}}(p_{\mathrm{T}}^{\tau_{\mathrm{h}}})$.

In the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state, only $F_{\mathrm{F}}^{\mathrm{QCD}}$ is calculated as the total F_{F} is heavily dominated by this factor. For the small fraction of W+jets and $\mathrm{t\bar{t}}$ events in the AR, $F_{\mathrm{F}}^{\mathrm{QCD}}$ is also applied.

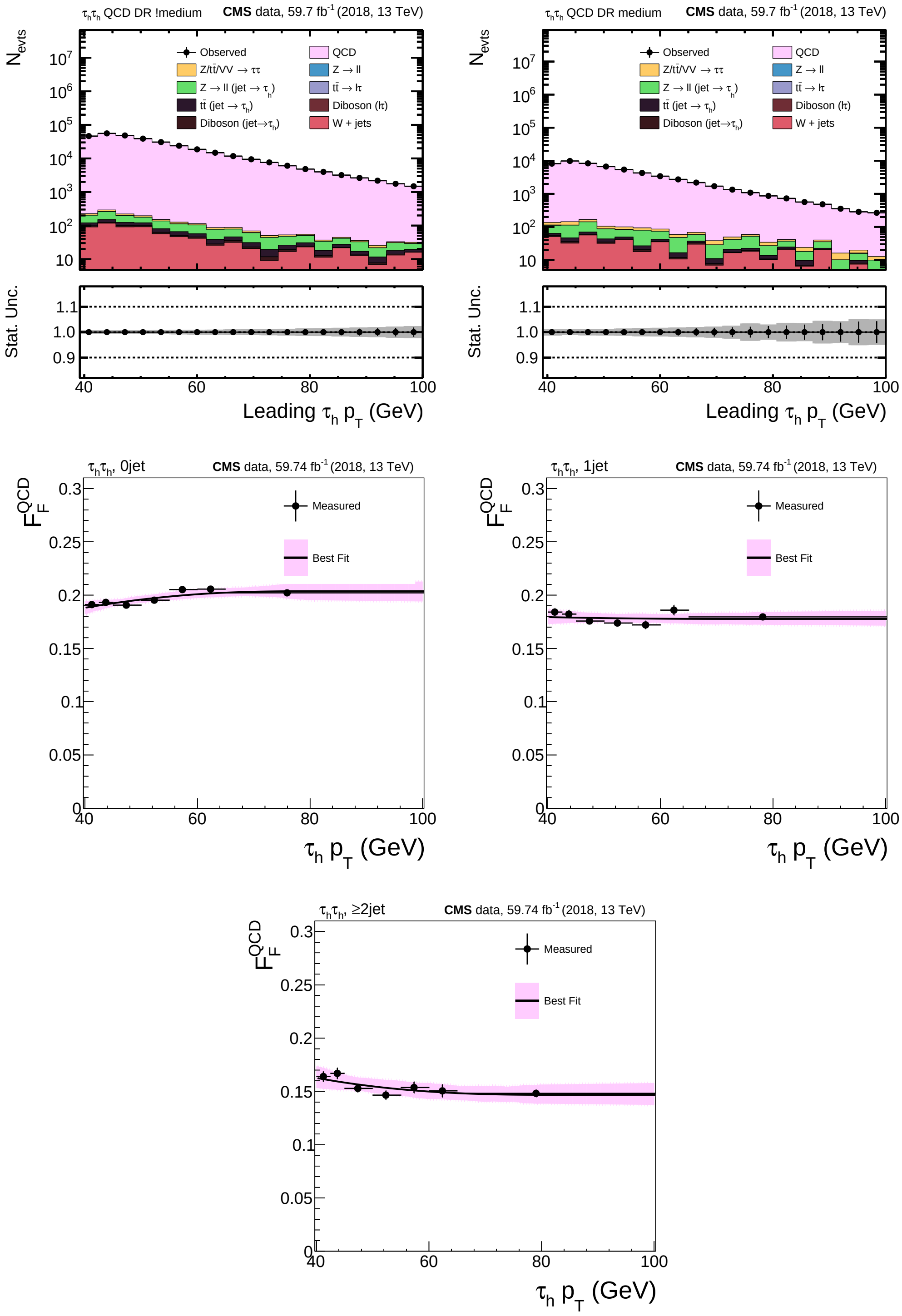


Figure 4.18: Events in the QCD determination region of the $\tau_h\tau_h$ final state using data from the 2018 run period. Events fulfilling the AR-like τ_h identification working point (top left), and the SR-like τ_h identification working point (top right) are shown. The right QCD histogram divided by the left one yields F_F^{QCD}. Events not produced by QCD multijet production make up less than 1-3% of the events in the DR and are subtracted from the data. The remaining data are assumed to be purely QCD multijet events. The estimation and observation thus agree perfectly by construction. The resulting values of F_F^{QCD} are calculated by dividing the SR-like events by the AR-like events in the QCD DR and given in the lower three figures. The calculation is performed independently for events with 0 jets (left), 1 jet (right) or ≥2 jets (bottom).

4.4.2 Estimation of W+jets events

Events in which a W boson is produced in association with at least one recoiling jet can lead to events with isolated electrons or muons from the W boson decay. If the accompanying jet is misidentified as a τ_h, the event enters the analysis. To define a DR for such events, the fact that usually no b-jet is expected in these events is utilized. While for the SR, the requirement of at least one b-jet is imposed, the DR uses only events in which no b-jet is present. To further enrich the content of W+jets events, a requirement of $m_\mathrm{T}^{\mathrm{e}/\mu} > 70\,\mathrm{GeV}$ is imposed. The distribution of events in the DR and the resulting measurement of $F_\mathrm{F}^{\mathrm{W+jets}}$ are shown in Figure 4.19. A linear fit is performed on the measured values.

The resulting purity of W+jets events in the determination region is around 86% (80%) in the `!medium&&vvvloose` (`medium`) τ_h identification region in both final states. The contribution of processes other than W+jets is estimated using simulated and τ-embedded events and subtracted from the distributions. The remaining data are assumed to be purely W+jets events.

4.4.3 Estimation of $\mathrm{t\bar{t}}$(jet) events

A prevalent source of background events is the production of a top-quark pair, in which one top-quark decay involves an isolated electron, muon, or τ_h, while the other top-quark decays hadronically. The misidentification of the resulting jet as an additional τ_h leads to the selection of the event.

For $\mathrm{t\bar{t}}$ events, the definition of an adequate DR poses a greater challenge than in the case of QCD multijet or W+jets events, in which a pure enrichment of the process was possible while still remaining orthogonal to the signal region. For $\mathrm{t\bar{t}}$ events however, also multiple b-jets and lepton signatures are expected just as for signal. Therefore, in first order, simulated $\mathrm{t\bar{t}}$ events are used to derive the $F_\mathrm{F}^{\mathrm{t\bar{t}(sim)}}$ as a function of $p_\mathrm{T}^{\tau_\mathrm{h}}$. A selection on the simulated sample is applied to only use events from the $\mathrm{t\bar{t}}$ simulation in which the τ_h of the event is a misidentified jet.

To validate this estimate on the data, a region is defined in which the presence of additional muon or electron candidates, next to the electron or muon used to build the $\tau\tau$ pair are required. These events are vetoed in the signal region to ensure the statistical independence of the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states and thus an orthogonal control region is built. This region is heavily dominated by $\mathrm{Z} \to \ell\ell$ events. A requirement on the electron or muon to be poorly isolated, and of at least one b-jet present in the event enriches the fraction of $\mathrm{t\bar{t}}$(jet) events to around 50% (30%) in the `!medium&&vvvloose` (`medium`) τ_h identification region. This region contains much less events and is less pure than the DR used for QCD and W+jets events. It is used to extract a single global value $F_{\mathrm{F,\,global}}^{\mathrm{t\bar{t}(data)}}$. The global data value is divided by the globally measured value from simulation and

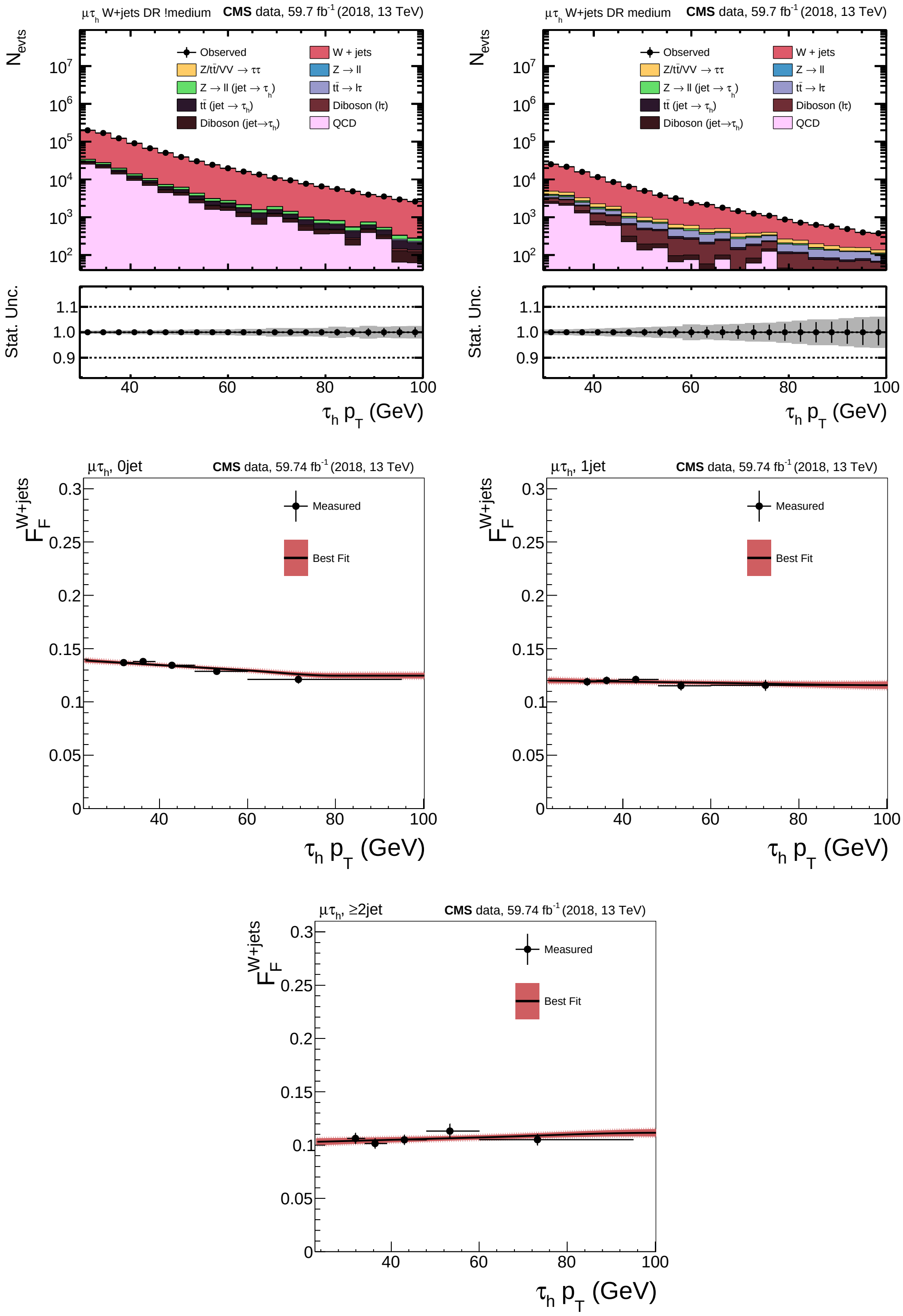


Figure 4.19: Events in the W+jets determination region of the $\mu\tau_\mathrm{h}$ final state using data from the 2018 run period. Events fulfilling the AR-like τ_h identification working point (top left), and the SR-like τ_h identification working point (top right) are shown. The right W+jets histogram divided by the left one yields $F_\mathrm{F}^\mathrm{W+jets}$. The non-W+jets events create impurities of up to 20% and are subtracted from the data. The remaining data are assumed to be purely W+jets multijet events. The estimation and observation thus agree perfectly by construction. The resulting values of $F_\mathrm{F}^\mathrm{W+jets}$ are calculated by dividing the SR-like events by the AR-like events in the W+jets DR and given in the lower three figures. The calculation is performed independently for events with 0 jets (left), 1 jet (right) or ≥2 jets (bottom).

applied as

$$F_{\mathrm{F}}^{\mathrm{t\bar{t}}}(p_{\mathrm{T}}) = F_{\mathrm{F}}^{\mathrm{t\bar{t}(sim)}}(p_{\mathrm{T}}) \cdot \underbrace{\frac{F_{\mathrm{F,\ global}}^{\mathrm{t\bar{t}(data)}}}{F_{\mathrm{F,\ global}}^{\mathrm{t\bar{t}(sim)}}}}_{\mathrm{SF}} \quad . \tag{4.6}$$

The global factor SF usually lowers the simulation-based estimation by 5-10%. The measurement is performed independently for events with at most one, or at least two additional jets. The measurement of $F_{\mathrm{F}}^{\mathrm{t\bar{t}}}$ in the $\mu\tau_{\mathrm{h}}$ final state of the 2018 run period is shown in Figure 4.20. A linear fit is performed on the measured values.

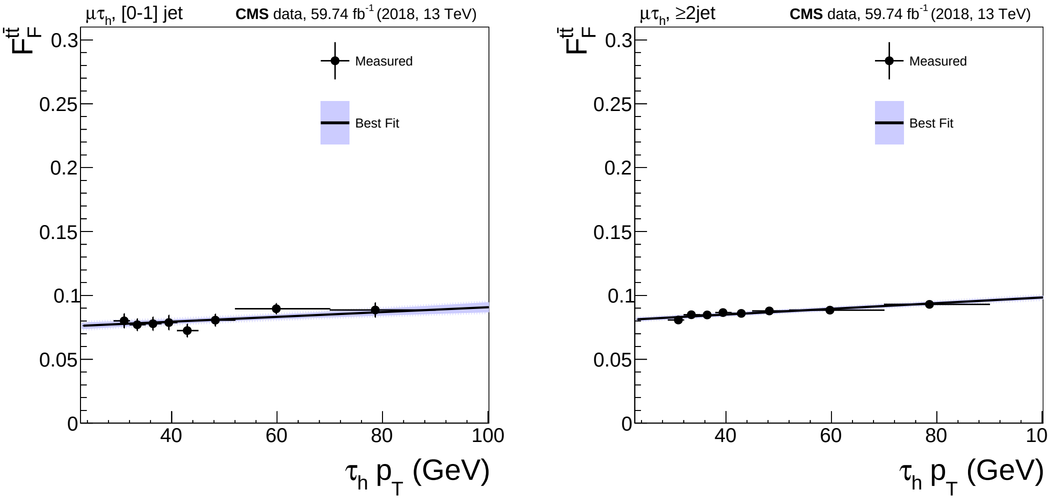


Figure 4.20: Values measured for $F_{\mathrm{F}}^{\mathrm{t\bar{t}}}$ for the $\mu\tau_{\mathrm{h}}$ final state using data collected in the 2018 run period. The measurement is performed independently for events with 0 or 1 additional jets (left) and ≥ 2 additional jets (right). The p_{T}-dependent factors have been determined purely from $\mathrm{t\bar{t}}$ simulation and a scale factor as determined from data in a control region with additional leptons is applied globally to achieve the shown values for $F_{\mathrm{F}}^{\mathrm{t\bar{t}}}$.

4.4.4 Corrections applied to the F_{F} measurement

For the measurement of the F_{F}^{i}, the assumption is made that the extrapolation of events from the AR to the SR depends only on the number of jets in the event and on the p_{T} of the reconstructed τ_{h} candidate. Furthermore, it is assumed that the behavior of the extrapolation is the same in the DR and in the SR. Both assumptions are not true, and therefore deviations in the modeling due to this need to be incorporated into the method as either additional corrections, systematic uncertainties, or both.

For the F_{F} method as applied in this analysis, multiple corrections are used in order to take additional kinematic dependencies of the F_{F}^{i} into account, and to correct for differences of the F_{F}^{i} between the DR and SR.

Corrections in the $e\tau_h$ and $\mu\tau_h$ final states

In the semi-leptonic final states, two additional dependencies on the event kinematics are considered as corrections: First, the trigger responsible for selecting the event has a major influence on F_F^i. Events selected by a lepton+τ_h pair trigger result in significantly higher values (up to 70%) of F_F^i than events selected by a trigger designed for a single isolated lepton. Secondly, a dependency on the p_T of the electron or muon is introduced in the form of a non-closure correction. Usually, a low-p_T lepton increases the probability of the lepton to be a misidentified jet itself in the case of QCD multijet events. For W+jets and $t\bar{t}$ events, also a dependence on the lepton p_T is expected due to the distinctive spectrum of the W boson decay products. The correction for the dependency on the trigger as well as on the lepton p_T is derived in a single measurement, using the fact that the triggers select different ranges of lepton p_T. For F_F^{QCD}, an additional non-closure correction is applied in dependence of the relative lepton isolation as defined in equation 3.5.

In addition to these additional dependencies, two corrections are introduced to take into account the differences between the QCD and W+jets DRs and the SR. For QCD, F_F^{QCD} of same-sign events is not exactly the same as of the opposite-sign events in the SR. A transfer factor is determined as a function of m_{vis} by comparing F_F^{QCD} in the same-sign region with an opposite-sign event selection. In order to ensure the orthogonality with the SR, the comparison is performed on events in which the lepton fulfills the relative isolation requirement $\text{Iso}_{rel}^{e/\mu} \in [0.15, 0.25]$.

For the W+jets DR, especially the high transverse mass requirement $m_T^{e/\mu} > 70\,\text{GeV}$ of the DR can influence the F_F^{W+jets} measurement. A correction is derived by comparing simulated W+jets events in which the requirement is removed to simulated events with the $m_T^{e/\mu} > 70\,\text{GeV}$ requirement applied. The comparison is also performed as a function of m_{vis}, with the observed differences applied as correction. For $F_F^{t\bar{t}}$, no extrapolation correction is necessary as the measurement is anyhow performed in a SR-like selection on simulated $t\bar{t}$ events. All corrections are shown in Figure 4.21 for the measurements in the $\mu\tau_h$ final state of the 2018 run period. The measurements are smoothed by a Gaussian kernel smoothing to avoid the propagation of statistical fluctuations into the F_F^i.

Corrections in the $\tau_h\tau_h$ final state

The treatment of the F_F^{QCD} in the $\tau_h\tau_h$ final state is very similar to the semi-leptonic final states. First, a correction is applied taking also the p_T of the second τ_h into account. An additional non-closure is then calculated as a function of m_{vis}. Finally, a correction addressing the extrapolation from the DR to the SR is used by comparing oppositely-charged τ_h pairs, in which both τ_h candidates fail the `medium` τ_h identification working point with the same-sign selection of the DR. This correction is also performed as a function of m_{vis}. All three corrections applied to F_F^{QCD} in the $\tau_h\tau_h$ final state are shown in Figure 4.22.

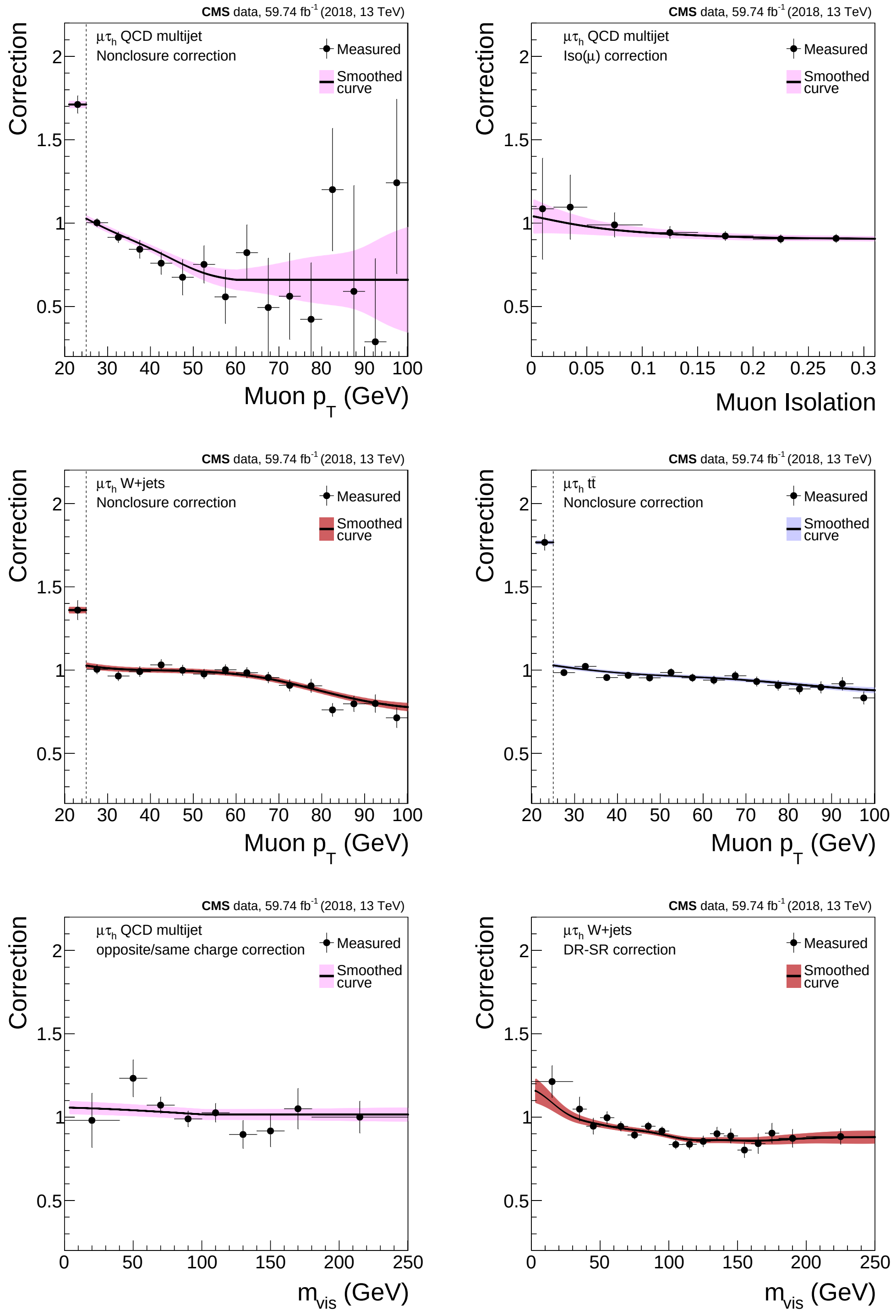


Figure 4.21: Corrections applied to the F_{F}^{i} as measured for the $\mu\tau_{\mathrm{h}}$ final state of the 2018 run period. Top row: Non-closure corrections applied to $F_{\mathrm{F}}^{\mathrm{QCD}}$, binned in lepton p_{T} (left) and isolation (right). Middle row: Non-closure corrections applied to $F_{\mathrm{F}}^{\mathrm{W+jets}}$ (left) and $F_{\mathrm{F}}^{\mathrm{t\bar{t}}}$ (right), both binned in lepton p_{T}. In the corrections binned in lepton p_{T}, the trigger selection of the di-tau lepton pair is indicated with a dashed line. Events with muon p_{T} below 25 GeV (left of dashed line) are selected by a lepton+τ_{h} pair trigger, events above 25 GeV by a single-lepton trigger. An up to 70% higher F_{F}^{i} value is observed for events selected by a lepton+τ_{h} pair trigger. Bottom row: DR-SR extrapolation corrections for $F_{\mathrm{F}}^{\mathrm{QCD}}$ (left) and $F_{\mathrm{F}}^{\mathrm{W+jets}}$ (right). The black lines represent the smoothed curve applied to the F_{F}^{i}, the colored bands represent the statistical uncertainties of the curves.

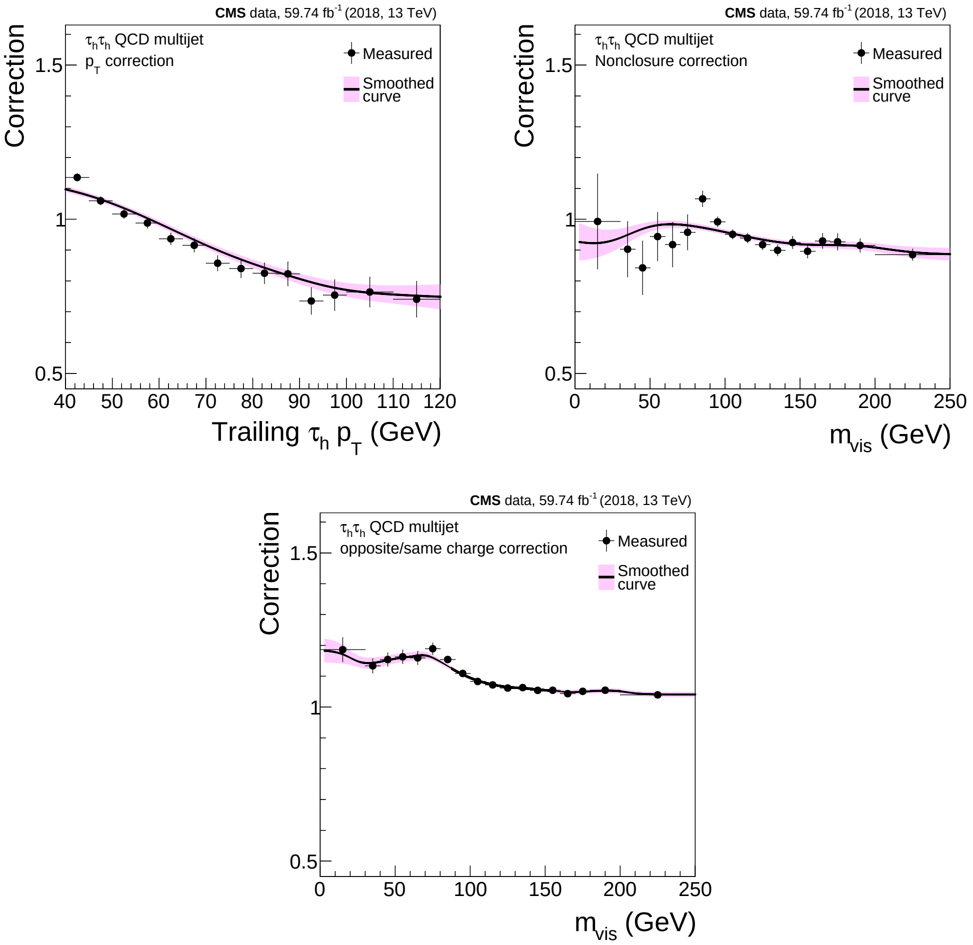


Figure 4.22: Corrections applied to $F_{\mathrm{F}}^{\mathrm{QCD}}$ measured in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state of the 2018 run period. Two non-closure corrections are derived, first as a function of the second τ_{h} of the event (top left), and as a function of m_{vis} of the di-tau lepton pair (top right). Finally, a correction for the extrapolation from the same-sign selection of the DR to the opposite-sign selection of the SR is derived (bottom).

4.5 Event simulation

4.5.1 Simulation of background events

As indicated in Figure 4.4, the majority of $\mathrm{t\bar{t}}$ and VV events in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final state entering the analysis are not included in either the τ-embedded event sample, nor are they covered by the estimation of the F_F method. For these events, as well as for small parts of the Z boson background and the background due to the production of single $\mathrm{h_{SM}}$ bosons, Monte Carlo event simulation is used.

The simulation is performed using the event generators `MadGraph5_aMC@NLO` [60], `POWHEG` [61] and `PYTHIA 8.2` [55]. For the parton density functions the sets `NNPDF 3.0` [62] and `NNPDF 3.1` [63] are used. The purpose of these event generators is the simulation of the hard interaction process of the event, i.e. the collision of the protons to a set of final state particles corresponding to a given process, e.g. $\mathrm{t\bar{t}} \rightarrow X$ or $\mathrm{h_{SM}} \rightarrow \tau\tau$ within the mathematical framework given by the SM. At the level of final state particles with calculated energy, momentum and direction, the response of the CMS detector to these particles is simulated using the `GEANT4` package [64]. At this stage, the full detector readout of the hard scattering process is available, in the same way as for an observed event.

As usually around 30 protons collide at each bunch crossing at the CMS collision point of the LHC, many additional collisions with usually low transverse momentum transfer called pile-up are present in the data. To model the pile-up as accurately as possible in the simulation, a random number of such events are simulated and added to the event, with the random number following a Poisson distribution. Problems usually enter as the exact number of pile-up is dependent on the operations of the LHC, and can also fluctuate during the run periods, whereas the underlying statistical distribution of the number of pile-up to be added to the events is a fixed parameter to be determined before event simulation. Due to this, the simulated number of proton-proton interactions only roughly matches the observed value in the data. To mitigate this, a reweighting procedure is applied which will be discussed in section 4.5.4.

With the pile-up events added to the simulated event record, the full reconstruction is performed on the simulated events and the response of the trigger system is simulated. Remaining differences in the efficiency of the reconstruction of simulated events to data are corrected, which will also be discussed in section 4.5.4.

Unlike data-driven methods, the total number of simulated events can be set freely, limited only by the significant computing effort of the simulation procedure. To estimate the total number of events in data of the specific process, the simulated events are divided by the total number of events produced, and scaled according to the cross section of the physics process which is modeled by the sample as well as the integrated data luminosity as

$$N_\mathrm{est} = L_\mathrm{int} \cdot \sigma \cdot \frac{1}{N_\mathrm{sim}} \; . \tag{4.7}$$

Both the luminosity and the cross section of the processes are known to a precision of around 2% depending on the run period and the process. The uncertainty on each is propagated to the final statistical inference as systematic uncertainties.

4.5.2 Simulation of NMSSM signal events

As opposed to the background processes discussed above, no centrally produced samples by the CMS collaboration of the signal process exist. The samples simulating the signal process gg → H → $h_{SM}(\tau\tau)h_S(bb)$ have been produced in the context of this thesis using the NMSSM implementation in `Feynrules` [65]. The model files from this implementation are used with `MadGraph5_aMC@NLO`, version 2.6.5 [60]. Here, the process is simulated at leading order using an effective coupling of the heavy Higgs boson H to gluons.

In total, 420 signal mass pairs are produced in a grid which scans the unknown masses of the additional Higgs bosons H and h_S. The grid is illustrated in Figure 4.23. For each grid point, between 100,000 and 500,000 events are generated, depending on the mass of the heavy scalar boson. The number of generated events is indicated in the figure. The grid spans from $m(h_S)_{min}$=60 GeV to $m(h_S)_{max}$=2800 GeV, and $m(H)_{min}$=240 GeV to $m(H)_{max}$=3000 GeV.

For each point of $m(H)$, samples are produced in a range from $m(h_S)_{min}$ to the value of $m(h_S)$ for which the masses of the SM Higgs boson (125 GeV) and the additional light scalar do not exceed the mass of the heavy scalar:

$$m(h_S) + m(h_{SM}) \leq m(H) \ . \tag{4.8}$$

For all samples, the branching fractions have been set to one for the decay H → $h_{SM}h_S$, and also for the decays $h_{SM} \rightarrow \tau\tau$ and h_S →bb. Events in which h_S decays into b-quarks, or in which any of the additional bosons decay into tau leptons, as well as non-resonant production of H or h_S in which the masses of H and h_S are similar are not considered in this analysis.

The produced signal samples have been compared to the official CMS production of the H → $h_{SM}h_{SM}$ process. For the validation, a signal sample of a heavy 650 GeV Higgs boson decaying into two different light scalar bosons with degenerate mass of 125 GeV has been produced with the same setup as used for the signal production. The comparison of distributions on the simulated stable hadron level of the two samples is shown in in the appendix in figure A.6. Additional comparisons of event information after detector reconstruction are shown in figure A.7. Both comparisons show excellent agreement between the privately and officially produced samples.

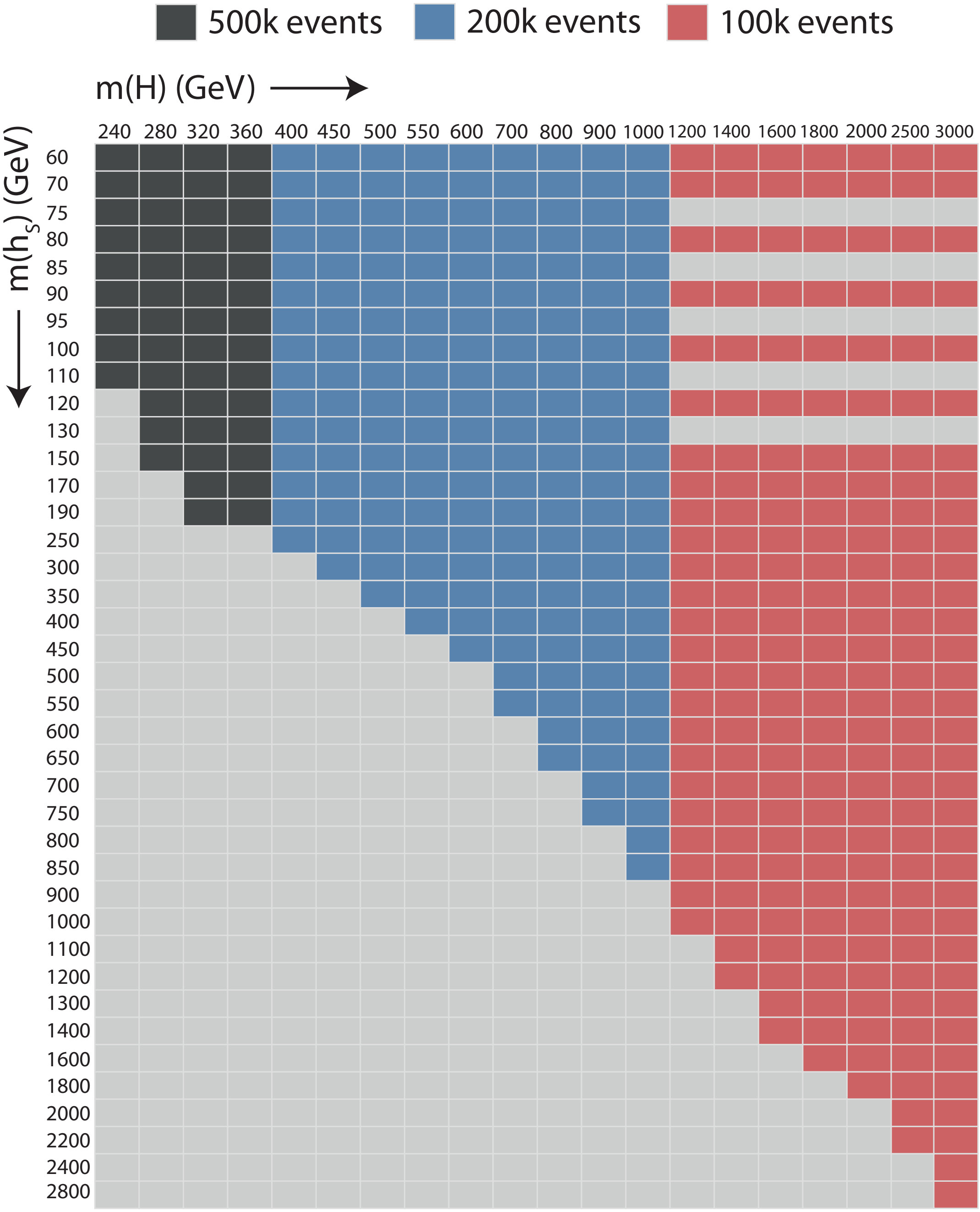


Figure 4.23: Two-dimensional grid of the simulation of signal processes. For each mass point of the heavy scalar boson $m(\mathrm{H})$ (x-axis), multiple event samples are simulated for different masses of the light boson $m(\mathrm{h_S})$ (y-axis), until the point where $m(\mathrm{h_S}) + 125\,\mathrm{GeV} > m(\mathrm{H})$. Omitted mass pairs are shown in grey. The acceptance of the process rises with a higher heavy boson mass. Therefore, 500,000 events are generated each for mass pairs with $m(\mathrm{H}) < 400\,\mathrm{GeV}$ (black) and 200,000 for events with $400 \leq m(\mathrm{H}) \leq 1000\,\mathrm{GeV}$ (blue). For $m(\mathrm{H}) > 1000\,\mathrm{GeV}$ (red), 100,000 events are generated and some low-mass points in the grid are omitted as they result in a very similar event topology than neighboring mass points. This grid is filled in total three times for each LHC run period, resulting in a total of 213 million generated events. The generation of 100 events takes around one hour depending on the computing system.

4.5.3 Technical aspects of the signal simulation

Due to the two unknown masses, a large number of mass pairs needs to be produced in a two-dimensional grid. With the chosen 420 mass points, and between 100,000 and 500,000 generated events per mass point, 71 million events were produced for each run period. As three run periods needed to be modeled, this number is multiplied by three to arrive at the total number of produced events of 213 million. With an average per-event duration of 36 s, over two million hours of CPU time are necessary to produce all signal samples. The simulation is split in two steps: In the first step, the hard proton-proton interaction is simulated, and subsequently the decay chain of the created particles and the detector response to all visible particles is simulated. An output file is created by this step which serves as input to the second step. Here, pile-up collisions are added as discussed in section 4.5.1, and the full event reconstruction is performed on the event.

The two steps differ greatly in their requirements to the computing system to be used for the production. While the first step has a 50% longer runtime per event, it requires as input only a small file containing the simulation code, and therefore can be run mostly independently from the network of the computing system. Running this step for 1000 events in a single computing job, requires only the transfer of around 10 MB at the beginning and around 100 MB in the end of the job, where the output is transferred to disk after six hours of runtime.

The second step deals with large inputs, not only the input file of the first step but also files containing simulated events of pile-up collisions which are added to the event in a step called pile-up mixing. This step therefore has significant network activity throughout its four hour runtime, with an average network traffic of 3.7 GB and over one third of jobs requiring over 10 GB.

The computing resources are chosen to optimize the demands of the two steps. The first step, requiring the majority of the total CPU time, can run ideally on resources such as the idle cores of desktop computers distributed across the office spaces of the Institute for Experimental Particle Physics. Running the input-heavy second stage would however cripple the network of this computing infrastructure, and computing sites specialized to input-heavy computing tasks were used for this step. These sites are the BwForCluster NEMO of the bwHPC initiative [66], specialized worker nodes at the institute, or the nodes at the GridKa cluster ForHLRII [67].

The total cumulative CPU time used for the simulation of the signal events is shown in Figure 4.24. The vast majority, 83%, of the full simulation was performed on the BwForCluster NEMO. The remaining simulation was performed on local institute resources, both worker nodes and desktop computers, and at the GridKa cluster.

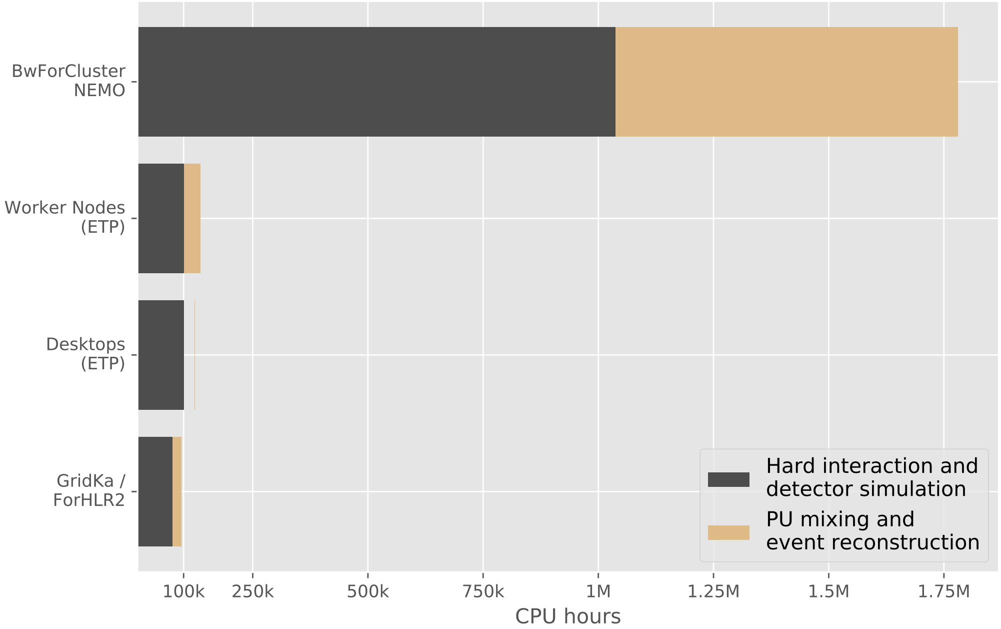


Figure 4.24: Cumulative CPU hours spent for the two steps of the event simulation: The simulation of the hard collision and detector response (black), and the pile-up mixing and event reconstruction (yellow). The computing effort was distributed on the high-throughput cluster BwForCluster NEMO (top row), local resources of the Institute of Elementary Particle Physics (ETP), split between worker nodes (second row) and desktop computers (third row), and the clusters at the GridKa (bottom row). In total, 2.14 million CPU hours were required for the production of 213 million signal events.

4.5.4 Corrections applied to the event simulation

Correcting differences in lepton reconstruction efficiency

Usually, the simulation of the CMS detector response does not perfectly match the detector response in the data. To account for this, the efficiencies of the applied selection criteria such as electron and muon identification, isolation, or triggering are measured using the tag-and-probe method. The efficiencies of the identification requirement of τ_h candidates and their triggering are measured in a fit in a $\mathrm{Z} \rightarrow \tau\tau$-enriched phase space in the $\mu\tau_\mathrm{h}$ final state.

In all cases, the procedure is the same as for τ-embedded events, explained in detail in section 4.3.5. The ratio of the observed efficiency in data with the efficiency of the simulated events is used as correction factor in the final analysis. The measurements for electrons and muons for the 2018 run period are shown in Figure 4.25. The correction factors applied for τ_h's are shown in Figure 4.26 for the identification correction and Figure 4.27 for triggering.

Correction of the electron and τ_h energy reconstruction

As discussed for τ-embedded events in section 4.3.5, the energies of simulated electrons and τ_h's are corrected. The muon energy scale correction is again negligible and neglected for this analysis. For electrons in simulation, the multivariate regression in a $\mathrm{Z} \rightarrow \mathrm{ee}$ selection as discussed in section 3.2.3 is applied as for the data. As discussed in this section, the scale of simulated electrons matches the scale in the data exactly by construction, as the electrons in data are calibrated to match the simulation. Small differences remain between the simulation and observed events, in particular the energy resolution is better in simulated events than in the data which is corrected in a resolution correction using a Gaussian smearing on an event-by-event basis.

For the τ_h's, the measurement is performed similar to the case of τ-embedded events using a maximum likelihood fit in a $\mathrm{Z} \rightarrow \tau\tau$ enriched control region in the $\mu\tau_\mathrm{h}$ final state. The resulting values of the measurement of the τ_h energy scale are given in Table 4.8.

Table 4.8: Corrections applied to the τ_h energy scale in simulated events as determined from a fit in a $\mathrm{Z} \rightarrow \tau\tau$ control region.

Run period	1-prong (%)	1-prong+π^0 (%)	3-prong (%)	3-prong+π^0 (%)
2016	$-1.0^{+0.7}_{-0.6}$	$-0.1^{+0.4}_{-0.3}$	$-0.8^{+0.7}_{-0.4}$	$0.1^{+1.0}_{-1.0}$
2017	$0.7^{+1.0}_{-0.8}$	$0.2^{+0.5}_{-0.4}$	$0.2^{+0.5}_{-0.5}$	$-0.5^{+1.6}_{-1.0}$
2018	$-1.6^{+0.7}_{-0.7}$	$-0.3^{+0.4}_{-0.4}$	$-1.1^{+0.5}_{-0.5}$	$0.1^{+1.1}_{-0.9}$

Reweighting of the top-quark pair p_T spectrum

For the estimation of the background due to $\mathrm{t\bar{t}}$ decays involving a prompt electron or muon and an additional genuine τ_h, $\mathrm{t\bar{t}}$ event simulation is used. The hard interaction process

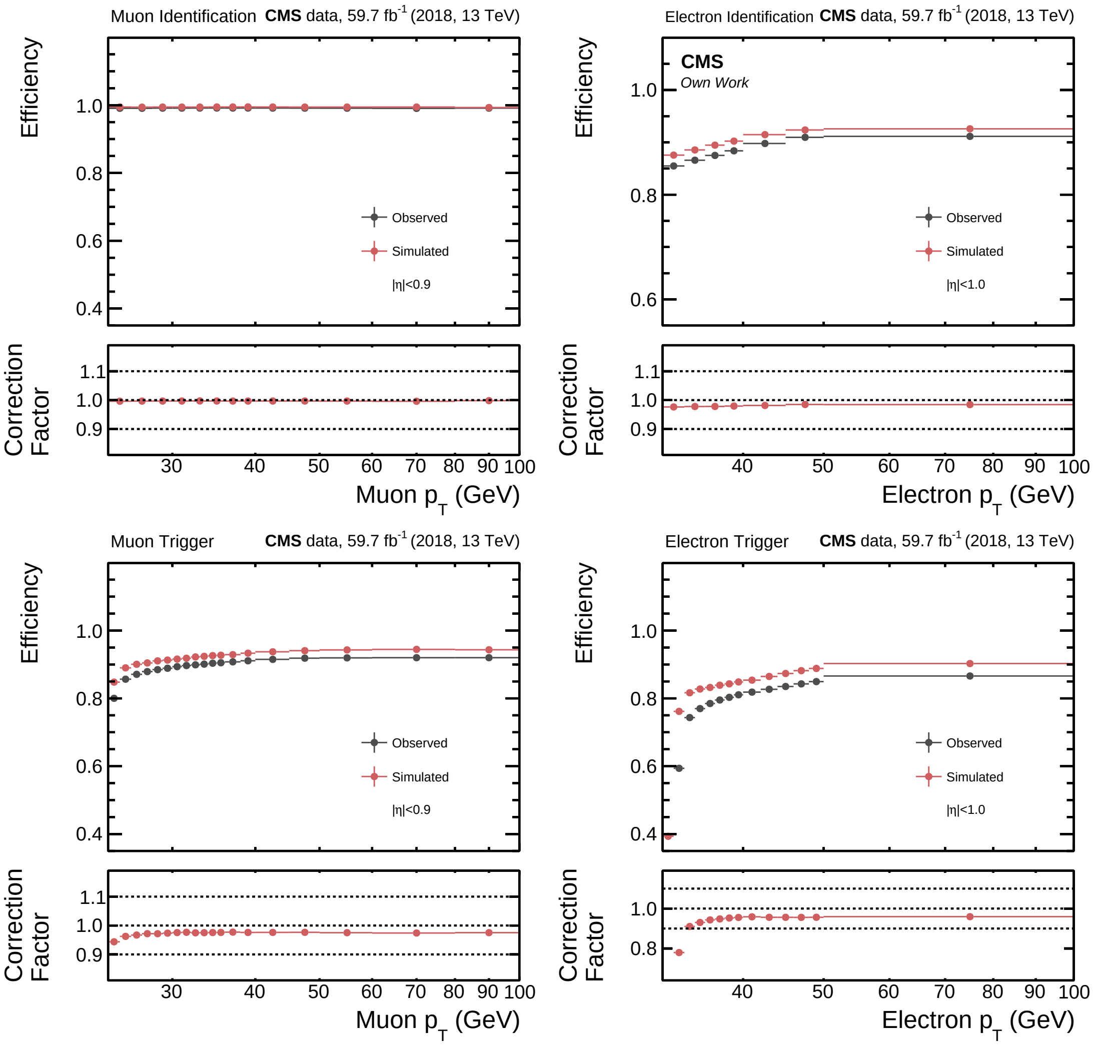


Figure 4.25: Efficiency of the identification (top row) and triggering (bottom row) of muons (left) and electrons (right) in simulated events (red), compared to data (black). Only the measurements in the central barrel region are shown. The correction factor to be applied on simulated events, calculated as the data efficiency divided by the efficiency of simulated events, is shown in the lower panels of the figures.

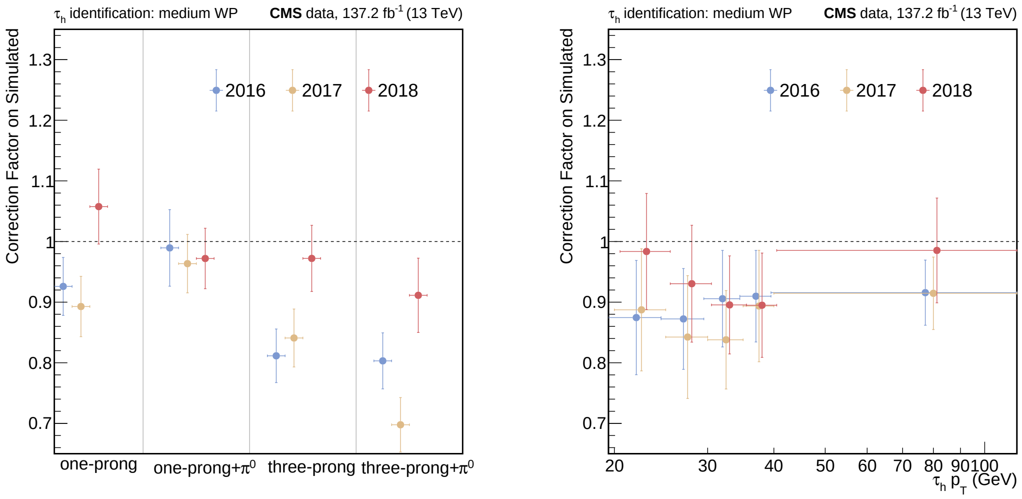


Figure 4.26: Correction factor for the identification of hadronically decaying tau leptons as measured for simulated events using the `medium` working point of the `DeepTau` classifier. Two independent measurements are performed in dependence of the τ_h decay mode (left) or p_T (right). The former is used in the $\tau_h\tau_h$ final state, the latter in the $e\tau_h$ and $\mu\tau_h$ final states. The correction factor is measured independently for the 2016 (blue), 2017 (yellow) and 2018 (red) run periods. In the right figure, the yellow and red points are shifted 1-2 GeV with respect to their bin centers to improve the visibility.

and thus the kinematics of the two top-quarks are simulated at next-to-leading order using `POWHEG` [61]. At NLO prediction, the p_T spectrum of the top-quarks is significantly harder, i.e. shifted towards higher p_T, compared to $t\bar{t}$ events observed in data.
As the analysis is sensitive to the p_T of the top-quark decay products, and thus the underlying top-quark spectrum, a p_T-dependent reweighting is applied. The reweighting has been determined with the measurements of the $t\bar{t}$ production in [68] and [69], and is parametrized based on the simulated information of an individual top-quark i as

$$w^i(p_T^i) = \exp(0.0615 - 0.0005 \cdot p_T^i) \tag{4.9}$$

with the total weight for the event being

$$w^{t\bar{t}} = \sqrt{w^{t} \cdot w^{\bar{t}}} \tag{4.10}$$

The reweighting is validated in the $t\bar{t}$ enriched control region determined by the neural network as will be discussed in the next chapter, with systematic nuisance parameters taking into account the magnitude of the reweighting.

For $t\bar{t}$ decays involving two tau leptons, or a jet which is misidentified as a τ_h, the description of the top-quark momentum is taken from the data via the τ-embedding or the F_F method and these corrections and corresponding uncertainties do not need to be applied.

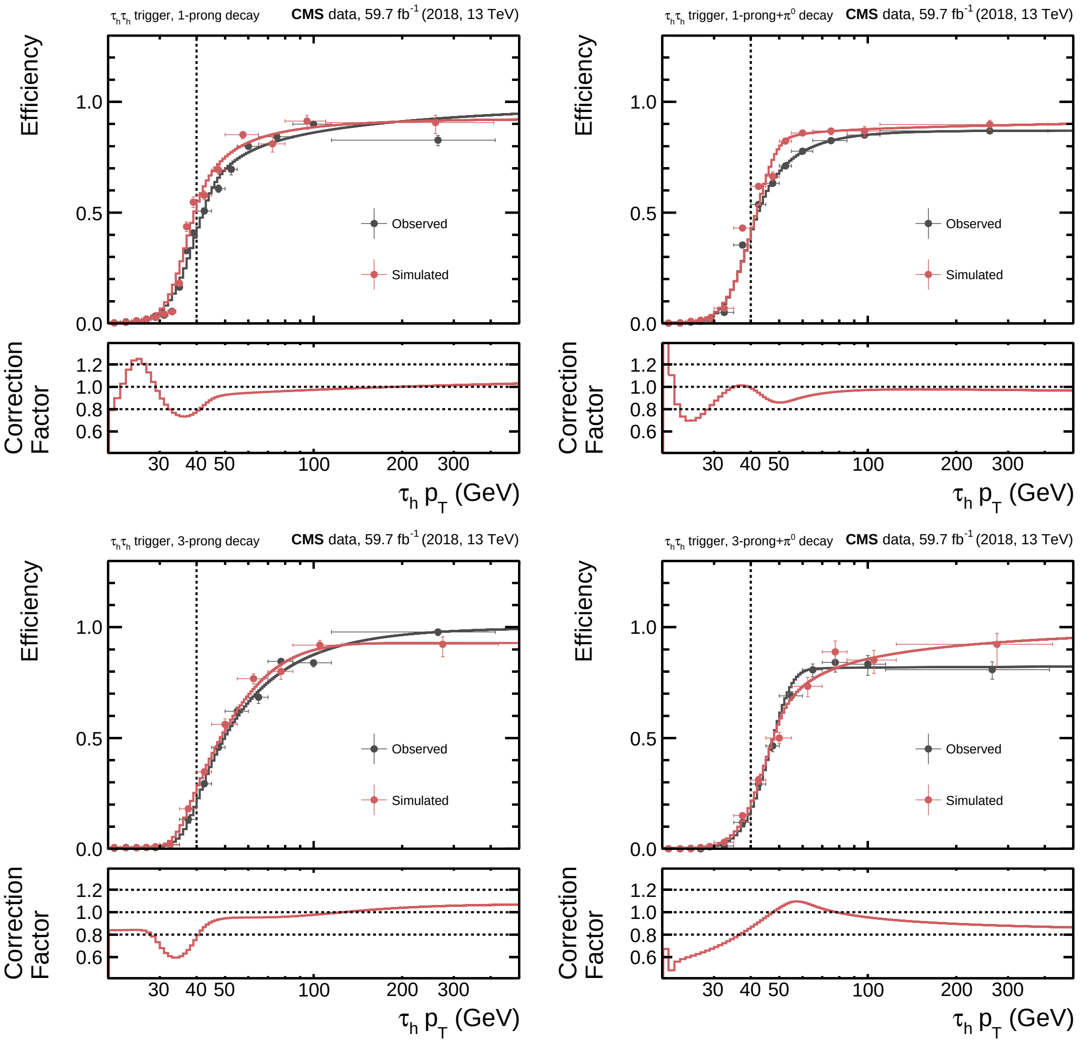


Figure 4.27: Efficiency of the triggering of hadronically decaying tau lepton pairs in simulated events (red), compared to data (black), as determined for events in the $\tau_h\tau_h$ final state of the 2018 run period. The measurement is performed separately for four decay modes of the τ_h: one-prong (top left), one-prong+π^0 (top right), three-prong (bottom left), three-prong+π^0 (bottom right). The measured values are shown as error points. The correction factor is derived from a fit to the measured values, which is shown as a line. The offline selection criterion of $p_T > 40\,\text{GeV}$ is indicated by a dashed line.

Efficiency of the b-jet identification

The correct description of the efficiency of the b-jet identification by simulated events plays a crucial role in this analysis. As was discussed in section 4.1, the `medium` working point of the `DeepJet` classifier is used to select b-jets. In the simulation, the efficiency of the b-jet identification can be different from the efficiency in the data. To account for this, the efficiency is measured using the selection criteria specific to this analysis, and comparing it with the efficiency and scale factors as measured centrally by the CMS collaboration [49]. From the comparison, scale factors are derived to account for the differences between data and simulation. The analysis-specific efficiency measurement of the simulated samples used for this analysis is shown in Figure 4.28.

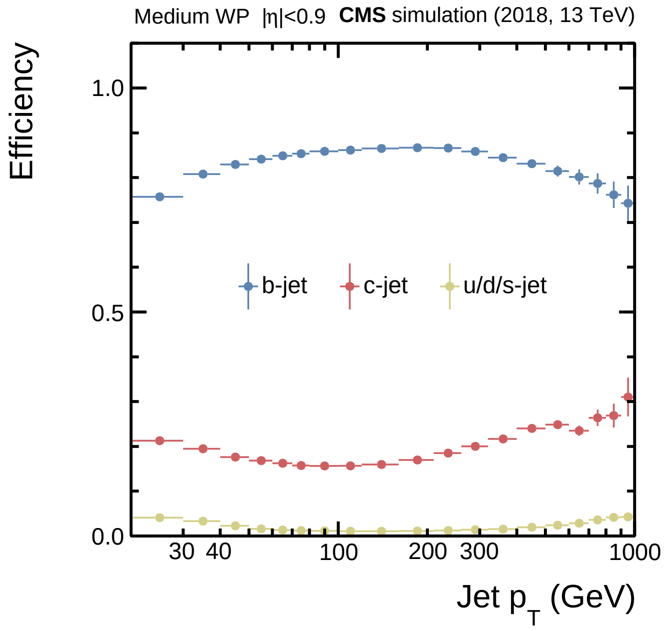


Figure 4.28: Analysis specific measurements of the efficiency and misidentification probability of the medium working point of the `DeepJet` b-jet identification for central jets ($|\eta| < 0.9$) as derived from simulated events used for the 2018 run period. The efficiency of true b-jets (blue points) to be tagged is between 70% and 90% depending on the p_T of the jet. The probability of misidentifying a c-jet as a b-jet (red points) is around 20%, the misidentification probability for jets initiated by lighter quarks (yellow points) is less than 5%. These analysis-specific efficiencies are compared with the centrally measured efficiencies to derive an efficiency correction for b-jets.

Applying the correction factor on the efficiency of b-jets is performed differently than e.g. the correction of the efficiency of the muon reconstruction in the $\mu\tau_h$ final state. The reason for this is that in this final state, the correct identification of the muon is a strict criterion for the selection of the event for the final analysis.

In the case of b-jets, also only one b-jet is sufficient to select the event, even thought two b-jets are expected in the signal process. A simple event scaling such as for the

muon identification is therefore not sufficient. To correct the efficiency, the "promotion-demotion" method is used. This means that simulated non b-jets can be promoted to b-jets in case the efficiency in simulation is lower than in data, and vice versa a b-jet can be demoted to a non b-jet if the efficiency is higher than in data.
The probability of the promotion or demotion depends on the ratio of the efficiencies measured in data and simulation $r = \epsilon_{\text{data}}/\epsilon_{\text{sim}}$ and is applied for each jet individually. The efficiencies are measured as a function of the p_{T} and η of the jet, $\epsilon = \epsilon(p_{\text{T}}, \eta)$.

- Demotion: If $r < 1$, the b-jet identification is removed with a probability $1 - r$ for any b-jet.
- Promotion: If $r > 1$, a b-jet identification is added for any non b-jet with a probability $\frac{r-1}{1/\epsilon_{\text{sim}}-1}$. Only jets fulfilling the p_{T} and η requirements used for b-jets given in Table 4.3 are eligible candidates for promotion.

Pile-up reweighting

As discussed in section 4.5.1, the number of pile-up in the data is highly dependent on the instantaneous luminosity provided by the LHC, and is usually not known when simulated events are produced. The simulation of the hard process is therefore combined with a random number of additional collisions, following a Poisson distribution. The parameter of the Poisson distribution is set to the expected number of pileup collisions based on the conditions of the run period, however the resulting distribution does not match the one observed in data. A reweighting is applied by comparing the two distributions and weighting the simulated events such that their distributions match.

In Figure 4.29, the observed distributions of the number of proton-proton interaction as measured by the CMS collaboration [70] is compared to the simulation. Simulated events follow a Poisson distribution, given by the random number generation during the pile-up mixing. The distribution in data is dependent on the LHC run conditions and shows distinct features. The ratio of the two distributions is used as a reweight factor applied on simulated events.

Correction of the $\vec{p}_{\text{T}}^{\text{miss}}$ and calibration of recoil of heavy resonance decays

After all corrections of the energies of the measured final state particles, such as electrons, τ_{h}'s, or b-jets, are applied, the total missing transverse energy ($\vec{p}_{\text{T}}^{\text{miss}}$) needs to be recalculated for the event. This is performed irregardless of the process.

In those simulated events in which a heavy resonance is created, such as the production of Z, W, H or h_{SM} bosons, an additional calibration of the recoil against these resonance decays is performed. In this analysis, the correction only concerns the rare cases of Z boson decays into two muons or electrons, in which one is misidentified as τ_{h}, the production of single h_{SM} bosons as well as the signal process $\text{H} \rightarrow \text{h}_{\text{SM}}\text{h}_{\text{S}}$. The production of W bosons as well as $\text{Z} \rightarrow \tau\tau$ events are estimated from data, and thus no recoil calibration needs to be performed.

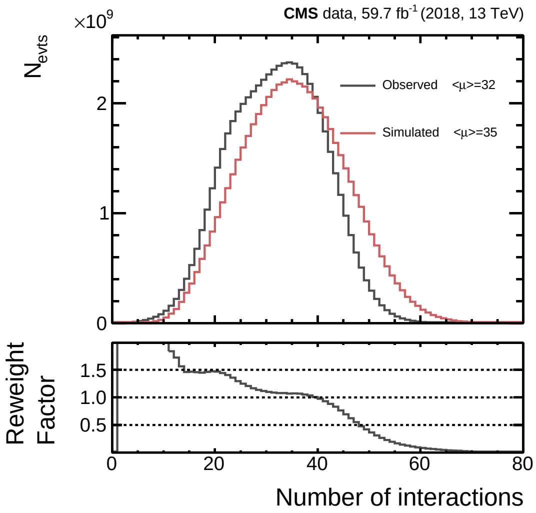


Figure 4.29: Distribution of the number of proton-proton interactions per bunch crossing. The observed profile for the 2018 run period is shown by a black line. The profile of the simulation used for the same run period is shown by a red line. The number of simulated events has been scaled to the data luminosity as done in the final analysis. As the exact profile of the observed data is not yet known during simulation, the profiles differ and need to be corrected using a reweighting procedure. The reweighting factors are shown in the lower panel of the figure.

For the calibration of the recoil, the resonance decay $\mathrm{Z} \rightarrow \mu\mu$ is used. Here, the p_T of the two muons can be reconstructed with high precision, resulting in the precise measurement of the Z boson momentum as $\vec{p}^{\,\mathrm{Z}} = (\vec{p}^{\,\mu_1} + \vec{p}^{\,\mu_2})$. As no neutrinos are present in the event, the true $\vec{p}_\mathrm{T}^{\,\mathrm{miss}}$ is zero and the reconstructed $\vec{p}_\mathrm{T}^{\,\mathrm{miss}}$ can be used to calibrate the recoil. The reconstructed $\vec{p}_\mathrm{T}^{\,\mathrm{miss}}$ in such events is split into two components, one parallel and one orthogonal to the reconstructed $\vec{p}^{\,\mathrm{Z}}$:

$$\vec{p}_\mathrm{T}^{\,\mathrm{miss}} = p_{||}^{\mathrm{miss}} \cdot \hat{e}_{||}^{\mathrm{Z}} + p_{\perp}^{\mathrm{miss}} \cdot \hat{e}_{\perp}^{\mathrm{Z}} \tag{4.11}$$

These two components are measured in dependence of the $\vec{p}_\mathrm{T}^{\,\mathrm{Z}}$ as well as the number of reconstructed jets in the event. As no dependence of the correction on the sign of the $p_{\perp}^{\mathrm{miss}}$ component is expected, the measurement is performed as a function of the absolute value $|p_{\perp}^{\mathrm{miss}}|$. The distributions for observed and simulated events with low $\vec{p}_\mathrm{T}^{\,\mathrm{Z}}$ and no reconstructed jets are shown in Figure 4.30. Especially in these cases with low $\vec{p}_\mathrm{T}^{\,\mathrm{Z}}$ and small hadronic activity, the recoil of the Z boson is not well described in simulated events and requires a correction.
From these values, a correction is then performed using a quantile mapping method: The probability density functions of the observed and simulated distributions of $p_{||}^{\mathrm{miss}}$ and $p_{\perp}^{\mathrm{miss}}$ are compared. A mapping function is derived connecting the points of the distributions with equal cumulative probability density.

When applying the mapping function to the event simulation for resonance decays used for the analysis, the true $\vec{p}_\mathrm{T}^{\,\mathrm{miss}}$ can be non-zero if neutrinos appear in the event. In

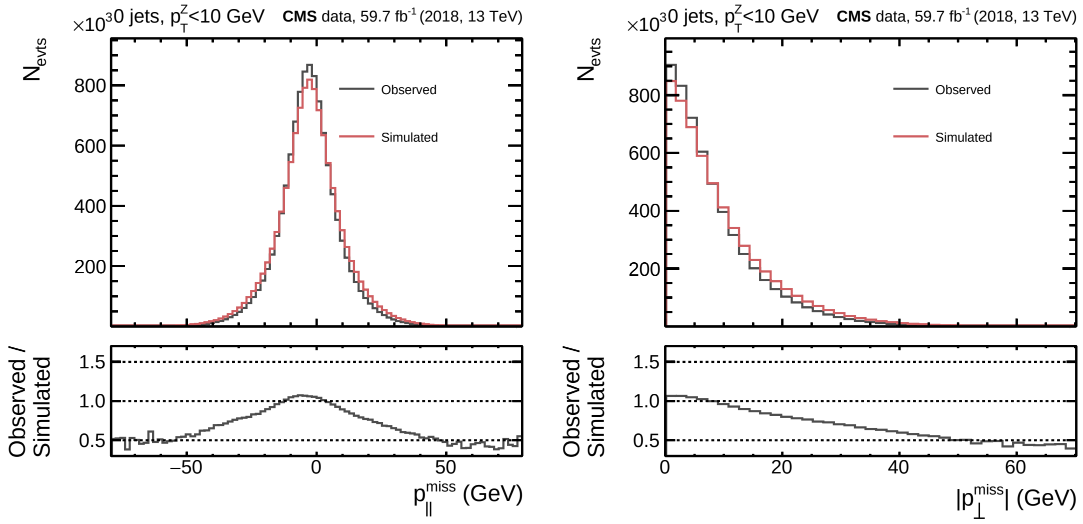


Figure 4.30: Measurements of the parallel and orthogonal components of $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ in observed $\mathrm{Z} \to \mu\mu$ events of the 2018 run period (black line) as well as simulated events (red line). The shown measurement is restricted to events with $|\vec{p}_{\mathrm{T}}^{\mathrm{Z}}| = |(\vec{p}^{\mu_1} + \vec{p}^{\mu_2})_{\mathrm{T}}| < 10\,\mathrm{GeV}$ which do not contain a reconstructed jet. The difference between the distributions of observed and simulated events is corrected in the simulation.

this case, the true $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ is subtracted using the simulated information of the sum of neutrino momenta $\vec{p}_{\mathrm{T}}^{\nu}$. As estimation of $\vec{p}_{\mathrm{T}}^{\mathrm{Z/H/h_{SM}}}$ as well as for the split in a parallel and orthogonal component, also the simulated information of the boson momentum is used. After calibration of the recoil, also the reconstructed $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ is corrected with the updated recoil vector.

L1 prefiring

After collection of the data of the 2016 and 2017 run periods, an issue with the timing of the electromagnetic calorimeter readout was discovered in which the L1 triggering decision of an event was falsely attributed to the previous proton-proton bunch crossing, and not the bunch crossing in which the triggering physics object was actually produced. This lead to a loss of events, as wrong events are stored in the data. This effect is not covered by simulated events. The effect is corrected in the simulation by the observed probability of the prefiring to occur, depending on the p_{T} and η of the event.

4.6 Deriving the bb$\tau\tau$ mass via a kinematic fit

The mass estimate of the bb$\tau\tau$ system and thus of the heavy additional Higgs boson H is improved using a kinematic fit. It builds on the `HHKinFit` tool designed for the search for a heavy scalar boson decaying to a pair of $\mathrm{h_{SM}}$, $\mathrm{H} \rightarrow \mathrm{h_{SM}h_{SM}}$ [71, 72].
The method utilizes the fact that the decay products of the $\mathrm{h_{SM}}$ and $\mathrm{h_S}$ bosons need to fulfill tight kinematic constraints due to the small width of the bosons, and the fact that the mass $m(\mathrm{h_{SM}}) = 125\,\mathrm{GeV}$ is known. These constraints are propagated to the bb$\tau\tau$ system.
The measured energies of the b-jets and the tau leptons can be varied by the kinematic fit to match these constraints, with a χ^2 cost function penalizing large differences between the fitted and the measured kinematics. This procedure is used to fit both the energies of the b-jets as well as find the most likely values of the true tau lepton energy before its decay into visible decay products and neutrinos, given the measured $\vec{p}_\mathrm{T}^\mathrm{miss}$ of the system and the momentum of the visible decay products.

An example sketch of the relevant decay products in a typical $\mathrm{H} \rightarrow \mathrm{h_{SM}}(\tau\tau)\mathrm{h_S}(\mathrm{bb})$ event in the $\mu\tau_\mathrm{h}$ final state is shown in Figure 4.31. For the kinematic fit of the b-jet energies, only the two reconstructed b-jets are used together with the constraint on the mass of $\mathrm{h_S}$. For the tau leptons, also the neutrinos are considered via the $\vec{p}_\mathrm{T}^\mathrm{miss}$ of the system by building the recoil vector of the heavy boson H. Both cases will be discussed in more detail in the following sections.

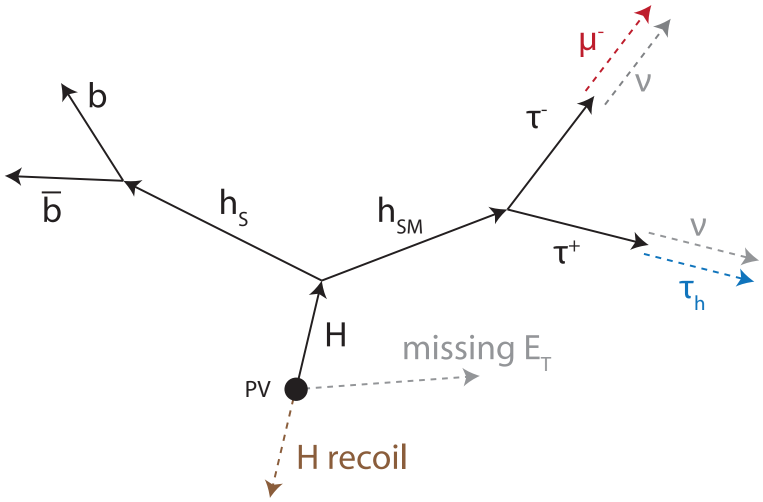


Figure 4.31: Sketch of an example event containing the decay $\mathrm{H} \rightarrow \mathrm{h_{SM}}(\tau\tau)\mathrm{h_S}(\mathrm{bb})$ in the $\mu\tau_\mathrm{h}$ final state. The transverse momenta of the particles involved in the decay are indicated by the arrows. Due to the neutrinos involved in the decay of the tau leptons, there is usually genuine $\vec{p}_\mathrm{T}^\mathrm{miss}$ in the event. The neutrinos from the tau decay and the total vector of $\vec{p}_\mathrm{T}^\mathrm{miss}$ are indicated by the gray arrows. Depending on the decay of the tau lepton, one or two neutrinos are involved which are summed up to a single dashed line in the sketch.

In Ref. [72], for which a similar fit was originally used, both masses of the $\tau\tau$ and bb systems are constraint to $m(\mathrm{h_{SM}}) = 125\,\mathrm{GeV}$. In this analysis however, the mass of the light Higgs boson $m(\mathrm{h_S})$ is a free parameter. This is considered in the kinematic fit for the mass of the heavy scalar boson $m(\mathrm{H})$. To use the kinematic fitting procedure of constraining the bb mass with varying nominal mass values, the fit is performed independently for 64 values of $m(\mathrm{h_S})$ ranging from 5 to 3000 GeV. For each of the 64 mass values, both a best-fit value $m_{\mathrm{H}}^{\mathrm{KinFit}}$ given the fixed values of $m(\mathrm{h_S})$ and $m(\mathrm{h_{SM}})$ as well as a χ^2 value of the fit are extracted. Only one result of the 64 fits is saved, chosen by the fit with the minimal value of χ^2_{min}. The discrete value of $m(\mathrm{h_S})$ and the best-fit value of $m_{\mathrm{H}}^{\mathrm{KinFit}}$ at which the minimum is reached are saved. Furthermore, the value of χ^2_{min} is saved as well. All three stored quantities are among the strongest discriminators in the multivariate analysis that will be discussed in the next chapter.

4.6.1 Fit of the b-jet energy

The measurement of the b-jet directions, i.e. η and ϕ of the b-jets, is assumed to be precise compared to the measurement of the energy and thus only the energy needs to be varied by the fit. The measured mass value of the two b-jets is then set to the fixed value of the hypothezised light Higgs boson mass $m(b_1b_2) \overset{!}{=} m(\mathrm{h_S})$:

$$\begin{aligned} m(b_1b_2)^2 &= (p_{b_1} + p_{b_2})^2 \overset{!}{=} m(\mathrm{h_S})^2 \\ \Rightarrow m(\mathrm{h_S})^2 &= (p_{b_1} + p_{b_2})^2 = p_{b_1}^2 + p_{b_2}^2 + 2p_{b_1}p_{b_2} \\ &= m_{b_1}^2 + m_{b_2}^2 + 2E_{b_1}E_{b_2} - 2\vec{p}_{b_1}\vec{p}_{b_2} \end{aligned} \tag{4.12}$$

Now, the approximation is made that a mismeasurement of the jet momentum to first order equals the mismeasurement of the jet energy. The ratio $\vec{p}/E$ can thus be assumed to be constant. The same is assumed for the ratio $m/E = 1/\gamma$. Equation 4.12 is then simplified to contain only the energy of one of the b-jets, with all other parameters, including the mass and energy of the other b-jet, being constant or arithmetically following by the variation of the first parameter in the fit.

In case of deriving the best-fit energy value of the first b-jet, $E_{b_1}^{\mathrm{fit}}$, the constraints are used to directly receive the energy of the second b-jet $E_{b_2}^{\mathrm{fit}}$, and the equation reads

$$m(\mathrm{h_S})^2 = m_{b_1}^2 + \frac{E_{b_2}^{2,\mathrm{fit}}}{\gamma_{b_2}^2} + 2E_{b_1}^{\mathrm{fit}}E_{b_2}^{\mathrm{fit}}\underbrace{(1 - \frac{\vec{p}_{b_1}\vec{p}_{b_2}}{E_{b_1}E_{b_2}})}_{c} \tag{4.13}$$

The constant c is set using the measured values of both two b-jets before the kinematic fit:

$$c = \frac{m(b_1b_2)^2 - m_{b_1}^2 - m_{b_2}^2}{2E_{b_1}^{\mathrm{meas}}E_{b_2}^{\mathrm{meas}}} \tag{4.14}$$

Solving the quadratic Equation 4.13 for $E_{b_2}^{\text{fit}}$ connects $E_{b_2}^{\text{fit}}$ directly to the value of $E_{b_1}^{\text{fit}}$,

$$E_{b_2}^{\text{fit}} = -E_{b_1}^{\text{fit}}\gamma_{b_2}^2 c + \sqrt{(E_{b_1}^{\text{fit}}\gamma_{b_2}^2 c)^2 - m(\text{h}_\text{S})^2 - m_{b_1}^2}\,, \tag{4.15}$$

and thus both b-jet energies are determined in parallel during the fit.
The correction factor $E_{b_{1,2}}^{\text{fit}}/E_{b_{1,2}}^{\text{meas}}$ is then applied to all components of the b-jet four-vector p_b. The χ^2 cost function for the change of the measured values of the b-jet four-vector is calculated as

$$\chi_{b_{1,2}}^2 = \frac{(E_{b_{1,2}}^{\text{fit}} - E_{b_{1,2}}^{\text{meas}})^2}{\sigma_{b_{1,2}}^2} \tag{4.16}$$

in which $\sigma_{b_{1,2}}$ is the b-jet resolution. This resolution is applied as a function of the $\eta_{b_{1,2}}$ and $E_{b_{1,2}}$ as measured in [72]. The fit finally yields the two best-fit values $E_{b_1}^{\text{fit}}$ and $E_{b_2}^{\text{fit}}$ as well as the two cost terms $\chi_{b_1}^2$ and $\chi_{b_2}^2$.

4.6.2 Fit of the tau lepton energy

The treatment of the tau leptons in the kinematic fit differs from the treatment of the b-jets due to the fact that neutrinos are involved in the tau decay, carrying away a significant fraction of the tau lepton energy.

The tau lepton decay products are approximated to be collinear with the direction of the tau leptons, as they are usually highly boosted if produced from a decay of a Higgs boson with much heavier mass than the tau lepton mass. Furthermore it is assumed that the measurement of the direction (η and ϕ) of the tau lepton decay products is accurate and its uncertainty can be neglected compared to the uncertainty on the energy reconstruction.
The two unknown parameters, i.e. the tau lepton energies E_{τ_1} and E_{τ_2} can again be connected using the known constraint of $m(\text{h}_{\text{SM}}) = 125\,\text{GeV}$ analogue to Equation 4.15, resulting in only one unknown parameter to be considered by the fit.

As the invariant mass of the visible decay products of the two tau leptons from a decay of h_{SM} are expected to lie below 125 GeV due to the energy loss by neutrinos, the p_T of the neutrinos is considered in the kinematic fit by using the measured sum and direction of missing p_T, $\vec{p}_{\text{T,miss}}^{\text{meas}}$. For the cost function of the fit, the total measured H recoil $\vec{p}_{\text{T,recoil}}^{\text{meas}}$ is compared to the best-fit value of this recoil $\vec{p}_{\text{T,recoil}}^{\text{fit}}$. The measured recoil can be calculated from the visible parts of the event by

$$\vec{p}_{\text{T,recoil}}^{\text{meas}} = -(\vec{p}_{\text{T,miss}}^{\text{meas}} + \vec{p}_{\text{T},b_1}^{\text{meas}} + \vec{p}_{\text{T},b_2}^{\text{meas}} + \vec{p}_{\text{T},\tau_1^{\text{vis}}}^{\text{meas}} + \vec{p}_{\text{T},\tau_2^{\text{vis}}}^{\text{meas}})\ . \tag{4.17}$$

The function used to derive the best-fit recoil is calculated from the best-fit values of the b-jet and tau lepton p_T as

$$\vec{p}^{\,\mathrm{fit}}_{\mathrm{T,recoil}} = -(\vec{p}^{\,\mathrm{fit}}_{\mathrm{T},b_1} + \vec{p}^{\,\mathrm{fit}}_{\mathrm{T},b_2} + \vec{p}^{\,\mathrm{fit}}_{\mathrm{T},\tau_1} + \vec{p}^{\,\mathrm{fit}}_{\mathrm{T},\tau_2}) \ . \tag{4.18}$$

The χ^2 cost function is derived from the covariance matrix of the recoil vector $\mathrm{COV}_{\mathrm{recoil}}$ and the residual vector $\Delta\vec{p}_{\mathrm{T,recoil}} = \vec{p}^{\,\mathrm{fit}}_{\mathrm{T,recoil}} - \vec{p}^{\,\mathrm{meas}}_{\mathrm{T,recoil}}$:

$$\chi^2_{\mathrm{recoil}} = \Delta\vec{p}^{\,T}_{\mathrm{T,recoil}} \ \mathrm{COV}^{-1}_{\mathrm{recoil}} \ \Delta\vec{p}_{\mathrm{T,recoil}} \ . \tag{4.19}$$

The covariance matrix of the recoil vector is estimated from the covariance of the $\vec{p}^{\,\mathrm{meas}}_{\mathrm{T,miss}}$ and the visible decay products by

$$\mathrm{COV}_{\mathrm{recoil}} = \mathrm{COV}_{\vec{p}_{\mathrm{T,miss}}} - (\mathrm{COV}_{b_1} + \mathrm{COV}_{b_2} + \mathrm{COV}_{\tau_1^{\mathrm{vis}}} + \mathrm{COV}_{\tau_2^{\mathrm{vis}}}) \tag{4.20}$$

in which the covariance matrix of a b-jet or tau lepton decay product i is given by

$$\mathrm{COV}_i = \begin{pmatrix} \cos^2(\phi_i) & \sin(\phi_i)\cos(\phi_i) \\ \sin(\phi_i)\cos(\phi_i) & \sin^2(\phi_i) \end{pmatrix} \sigma^2_{p_{\mathrm{T},i}} \tag{4.21}$$

and

$$\sigma_{p_{\mathrm{T},i}} = \sin(\eta_i)\frac{E_i}{|\vec{p}_i|}\sigma_{E_i} \ . \tag{4.22}$$

4.6.3 Fitting procedure and resulting mass distributions

With the χ^2 cost functions defined in equations 4.16 and 4.19, the total cost function is given by

$$\chi^2 = \chi^2_{b_1} + \chi^2_{b_2} + \chi^2_{\mathrm{recoil}} \tag{4.23}$$

As the energies of the two b-jets and two tau leptons are connected by the constraints of the fixed masses $m(\mathrm{h_{SM}})$ and $m(\mathrm{h_S})$ as shown in Equation 4.15, the final cost function is a function of only two variables E_{b_1} and E_{τ_1}.

This function is minimized first by searching for a minimum along the one-dimensional line of the positive $E_{b_1}-E_{\tau_1}$ axis. The minimal and maximal values of E_{τ_1} are determined using the visible decay products of the tau leptons, i.e. $E_{\tau_1,\mathrm{min}} = E_{\tau_1^{\mathrm{vis}}}$ and $E_{\tau_1,\mathrm{max}} = E_{\tau_1}(E_{\tau_2,\mathrm{min}} = E_{\tau_2^{\mathrm{vis}}})$ and Equation 4.15. For the b-jets, an interval of $\pm5\sigma$ of their resolution is used as boundary values for the fit.

Due to the tight constraints, especially the requirement to estimate the tau lepton energies in a way that their invariant mass matches 125 GeV, this search is not guaranteed to converge. For events in which the two visible tau decay products already have an invariant mass of above 125 GeV, the constrains can not be met and the fit fails. These events are removed from the analysis as they are expected to contain mostly background events.

If a minimum is found, the Newton method is used to determine a new search direction, in which another minimization is performed. This procedure is repeated until the convergence criterion is met. The fit is considered as converged if the differences in χ^2 or E_{b_1/τ_1}

after one step are below predefined thresholds of $\epsilon = 0.01$. The procedure is repeated for 64 mass values of $m(h_\mathrm{S})$ as discussed above, saving only the result at the value in which the total χ^2 is minimal.

In Figure 4.32, the correlation of the reconstructed $\tau\tau$ mass as estimated by the SVFit algorithm [73] and the minimal χ^2 of the kinematic fit is shown using data collected during 2018 in the $\mathrm{e}\tau_\mathrm{h}$, $\mu\tau_\mathrm{h}$ and $\tau_\mathrm{h}\tau_\mathrm{h}$ final states. For events with $m_{\tau\tau}$ close to 125 GeV, the kinematic fit shows good quality resulting in low values of the total χ^2. Selecting events in data according to a fit quality requirement of e.g. $\chi^2 \leq 10$ selects events close to the 125 GeV, while removing events with $m_{\tau\tau}$-values incompatible with this mass.

The resulting distribution of the mass derived by the kinematic fit, $m_\mathrm{H}^\mathrm{KinFit}$, is shown in Figure 4.33 in comparison to two mass estimators derived by a summation of the visible decay products, and the visible decay products plus $\vec{p}_\mathrm{T}^\mathrm{miss}$. The mass estimator derived by the kinematic fit shows a peak close to the simulated value of $m(\mathrm{H})$ with a greatly improved resolution compared to the non-fitted estimators.

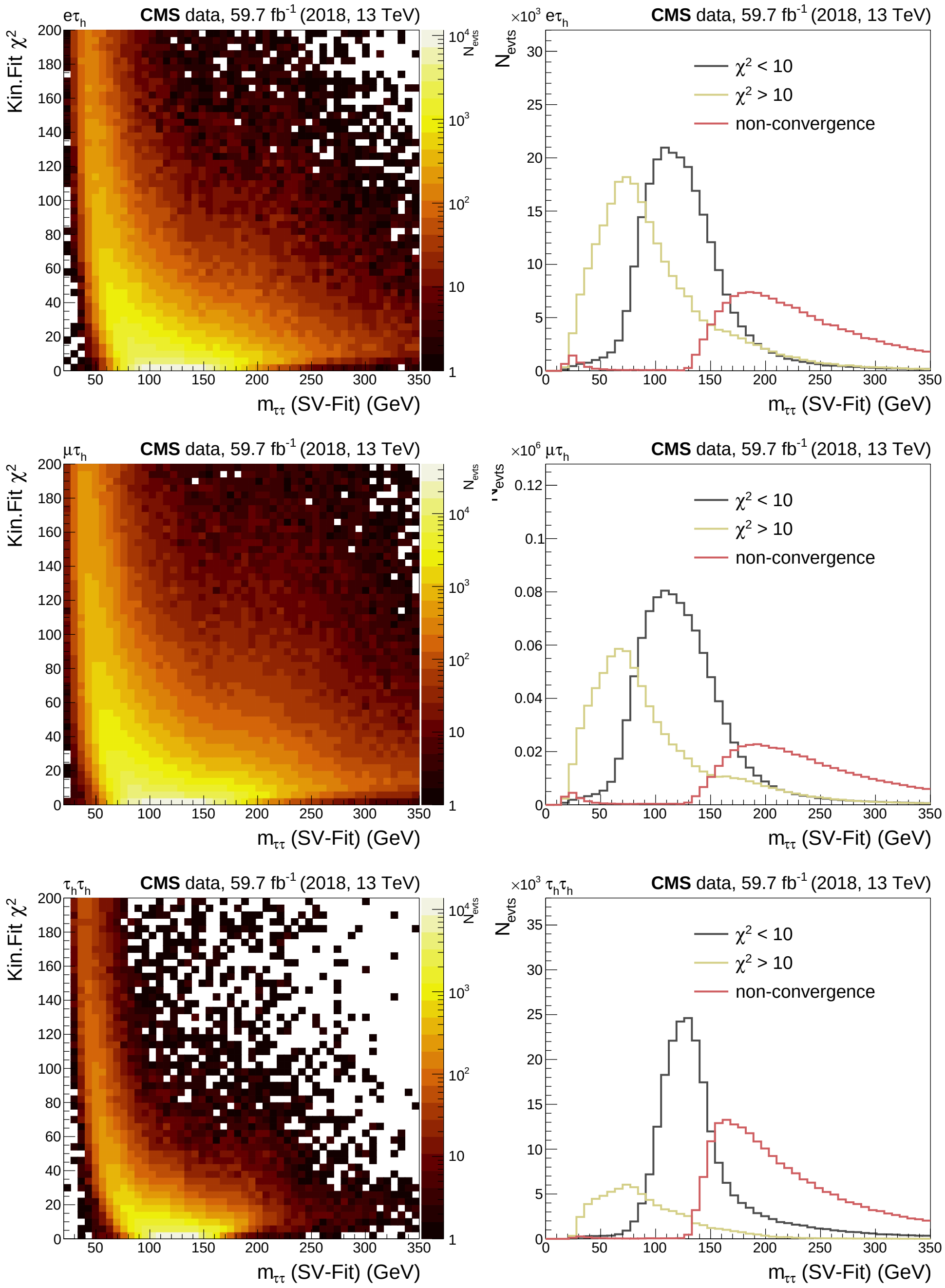


Figure 4.32: Correlation of the reconstructed mass of the $\tau\tau$ system (x-axis) with the quality of the kinematic fit (y-axis), expressed by its χ^2 in the $e\tau_h$ (top row), $\mu\tau_h$ (middle row), and $\tau_h\tau_h$ (bottom row) final states using data collected during the 2018 run period. The clear correlation of low values of χ^2 with $m_{\tau\tau}$ close to the constraint of $m_{\tau\tau} \overset{!}{=} 125\,\text{GeV}$ is visible in the two-dimensional distributions on the left. On the right, the 1D-distributions of $m_{\tau\tau}$ for events with $\chi^2 < 10$ (black line), $\chi^2 > 10$ (yellow line), or nonconverged fits (red line) are shown. The χ^2 of the fit indicates the compatibility of the event with the signal hypothesis and can serve as a highly discriminating variable between signal and background events.

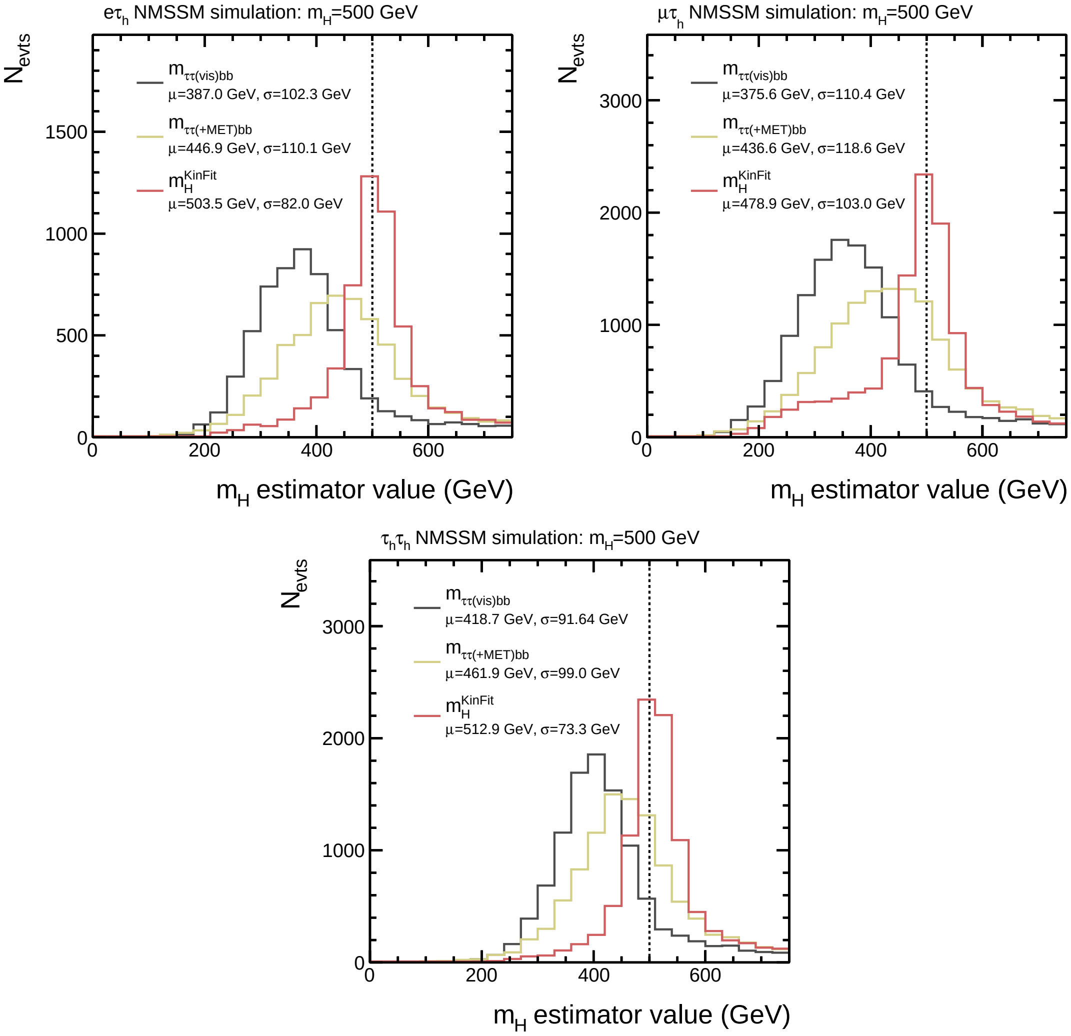


Figure 4.33: Various estimators of the mass of the heavy scalar boson $m(\mathrm{H})$: Mass of the visible di-tau lepton plus di-b-jet system (black), mass of the same system when the $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ is included in the di-tau lepton system (yellow) and mass estimator derived from a kinematic fit (red). Shown are the estimators for a simulated sample of the $\mathrm{H}(500\,\mathrm{GeV}) \rightarrow \mathrm{h}_{\mathrm{SM}}(125\,\mathrm{GeV})\mathrm{h}_{\mathrm{S}}(100\,\mathrm{GeV})$ process with $\mathrm{h}_{\mathrm{SM}} \rightarrow \tau\tau$ and $\mathrm{h}_{\mathrm{S}} \rightarrow \mathrm{bb}$. The mass estimators are compared in the $\mathrm{e}\tau_{\mathrm{h}}$ (top left), $\mu\tau_{\mathrm{h}}$ (top right) and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ (bottom) final states.

CHAPTER 5

Analysis strategy and results

The search for decays of a heavy neutral Higgs boson into two lighter Higgs bosons, $\mathrm{H} \rightarrow \mathrm{h_{SM}h_{S}}$, consist of several steps: First, the selected events are categorized to enrich background and signal processes in separate categories. Second, the observed events as well as the events used for background estimation are binned in a variable with discriminating power between background and signal events. The goal is to confine the signal events to a small number of bins with minimal background contribution. Finally, a statistical model covering the statistical and systematic uncertainties relevant for the analysis is built and a statistical inference is performed. Here, the signal significance is measured or upper limits of the possible production cross section times branching fractions in the case of non-observation of signal are derived.

5.1 Categorization of selected events via neural networks

For the optimal separation of events into exclusive categories, neural networks (NN) are used. These NNs take extensive information of each event into account and have the task of classifying the event into one of five categories. The category is chosen based on the output scores at the five output nodes of the NNs. Each output node translates to a specific physics process, or group of physics processes, in which the categorization for the analysis is desired.

The used networks have a fully connected feed-forward architecture. They consist of 23-28 input nodes, representing the physics inputs to the NN, and five output nodes used for the final categorization of the analysis. Furthermore, two hidden layers with 200 nodes each are used. The setup of the NN is sketched in Figure 5.1.

The hyperbolic tangent

$$\sigma(x) = \tanh(x) \tag{5.1}$$

is used as activation function of the hidden layers. For an output node i, the softmax activation function

$$\sigma_i(\vec{x}) = \frac{e^{x_i}}{\sum_i e^{x_i}} \tag{5.2}$$

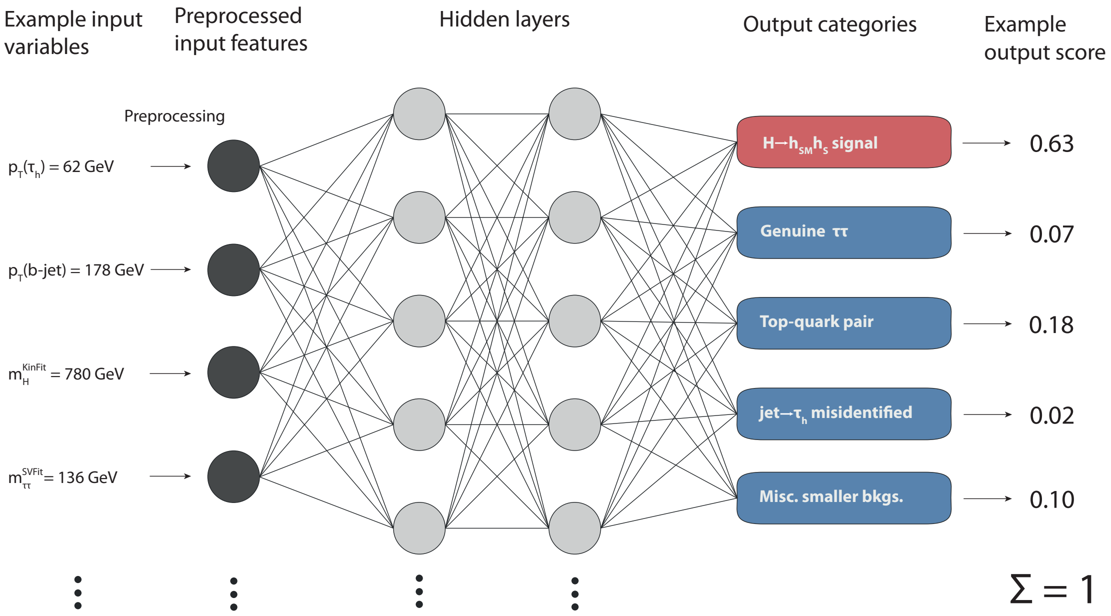


Figure 5.1: Illustration of the multiclass NN categorization. Between 23 and 28 event quantities are used as input and fed into the NN after preprocessing them to the same mean and standard deviation. Two hidden layers with 200 nodes each and an output layer with 5 nodes are used.

is used as activation function. It has the advantage of mapping the outputs of the NN to $[0, 1]$, with the sum of all output nodes $\sum_i \sigma_i(\vec{x}) = 1$. The value of this function for each individual output node can thus be interpreted as a Bayesian probability of the event to belong to the category represented by the output node.

The input variables are preprocessed via a linear transformation to have mean zero and standard deviation one. The loss function, representing the differences of the NN output with the desired output that is chosen for the training is the categorical cross-entropy defined as

$$\text{Loss} = -\sum_{n}^{N}\sum_{i}^{C} t_i^{(n)} \cdot \log(y_i^{(n)}) \tag{5.3}$$

where n labels the current event and N equals the batch size after which the weights of the NN are updated, i the category index and C the number of categories, which equals the number of output nodes. The $t_i^{(n)}$ is the binary indicator of the truth-value category for event n (1 in case category i is the correct one, 0 otherwise), and $y_i^{(n)}$ the model prediction for event n.

It is important to note that the goal of the NN training is not per se to separate signal from background, but to minimize the loss function defined in Equation 5.3. In the equation, no distinction between the signal and background categories is used, so separating between two given background categories is of equal importance to the training as separating between signal and any background category.

The minimization of the loss is achieved via backpropagation for which the Adam optimizer [74] with a learning rate of 10^{-4} is used. The initial weights are set according to the Glorot technique [75], avoiding the problem of vanishing gradients in higher NN layers.

A common problem in the application of machine learning is overtraining the NN on the training dataset, and thus making it incapable of generalization. Two techniques are applied to avoid overtraining the NN. First, L2 regularization is applied, adding a penalty term to the loss function proportional to the squared-sum of all trainable parameters and thus penalizing a large dependence of the classification on individual parameters. The exact proportionality, usually called λ, is a tuneable parameter of the model and for this analysis is chosen as $\lambda = 10^{-5}$.

The second technique is called dropout and consist of randomly skipping nodes in the NN during the gradient update with a predefined probability, which is chosen as 30 %. It avoids assigning a very high importance of the NN output on individual nodes, which can often be a sign of overtraining.

Before training, all events are randomly split in two samples. All training and validation steps described below are performed for each sample independently. In the final application, the resulting models are then applied to the events of the other sample respectively, to enable using the full event statistics for training without applying the model on events which it was trained on.

The two samples are split again, with 75% of events used for training and 25% used for the validation and subsequent selection of the best model. The training is performed on batches of 30 events out of each category, so 150 events in total before the weights are updated. After 1000 gradient steps, the training is validated against the validation sample. If no improvement occurs for 30 of such validations, the training is considered converged and terminated, with the model showing best performance against the validation sample used for the analysis. This convergence usually occurs after $\mathcal{O}(10)$ validation cycles in the $\tau_h\tau_h$ final state and $\mathcal{O}(100)$ validation cycles in the $e\tau_h$ and $\mu\tau_h$ final states.

5.1.1 Grouping of events in physics processes

The output nodes of the NN correspond to specific physics processes. The aim of the multiclassification is not only to separate signal from background events, but also to enrich the main background processes in distinct control categories. The advantage of having a category with a high purity of a specific process lies in the ability to constrain the systematic uncertainties acting on the process during the statistical inference. If e.g. the value of cross section of the $t\bar{t}$ production is estimated too low or high, this will unambiguously appear in the category enriching such events and can thus be constrained.

The categorization of the different background processes defines categories for the three main sources of background events for the analysis, which are genuine $\tau\tau$ events, jet → τ_h

misidentified events and $\mathrm{t\bar{t}}$ decays. A fourth category is created for all minor backgrounds with low cross sections. The categories will in the following be referred to as

1. Genuine $\tau\tau$: Decays of $\mathrm{Z/t\bar{t}/VV} \rightarrow \tau\tau$, described by τ-embedded events in the analysis.

2. Jet $\rightarrow \tau_\mathrm{h}$ misidentified: Events entering the analysis due to misidentification of a quark or gluon induced jet as τ_h, described by the F_F method in the analysis.

3. Top-quark pairs: Decays of $\mathrm{t\bar{t}}$ involving a prompt muon or electron in addition to a genuine tau lepton. These events create the largest source of background events in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states and are a small background in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state. They are described by simulated $\mathrm{t\bar{t}}$ events in the analysis.

4. Miscellaneous smaller backgrounds: The remaining minor backgrounds include events with two vector bosons, or a single Z boson or top-quark not included in (1.), such as $\mathrm{Z} \rightarrow \ell\ell$, $\mathrm{VV} \rightarrow \ell\tau$ ($\ell = \mathrm{e}, \mu$) as well as the production of a single $\mathrm{h_{SM}}$ boson. These processes contribute much less than 10% of the total background in all final states, however especially single $\mathrm{h_{SM}} \rightarrow \tau\tau$ decays can be hard to separate from the signal process.

The training of categories (3) and (4) as well as the signal categories to be discussed below is straight-forward, as simulated events of the specific processes can be used. Similarly, category (1) is trained on τ-embedded events.

In case of the jet $\rightarrow \tau_\mathrm{h}$ misidentified category, the training is not straight-forward, because there are no specific event samples that can be fed into the training as for simulated or embedded events. Therefore, the NN is trained on events with anti-isolated τ_h candidates taken from the application region of the F_F method. The extrapolation factors of the F_F method, referring to the event-by-event probability of the event being selected due to jet $\rightarrow \tau_\mathrm{h}$ misidentification, are applied to weight the events during the training.

For the categorization of signal events, the many different signal mass hypotheses create challenges for the categorization. The target of the categorization is to isolate the signal process $\mathrm{H} \rightarrow \mathrm{h_{SM}}(\tau\tau)\mathrm{h_S}(\mathrm{bb})$ with $m(\mathrm{h_{SM}}) = 125\,\mathrm{GeV}$, $m(\mathrm{H})$ between 240 and 3000 GeV, and $m(\mathrm{h_S})$ between 60 and 2800 GeV. In total, 420 mass points are tested as explained in section 4.5.2. Collecting all signal events in a single category would result in suboptimal results, as the signal could sit anywhere in the mass spectrum and no clear definition of the kinematics of a signal event would be possible.

This could be mitigated by creating multiple signal categories, split after groups of signal masses and training the NN once on all mass points. While this significantly improves the results with respect to a single signal category, the results are still suboptimal especially for masses at the boundary of the chosen categories: If two categories for all events with e.g. $m(\mathrm{H}) = 500\,\mathrm{GeV}$ and $m(\mathrm{H}) = 550\,\mathrm{GeV}$ are used, observed events in which the mass estimator lies around $m(\mathrm{H}) = 525\,\mathrm{GeV}$ would get assigned low scores in both categories and shift the event towards the background-rich low NN score region of the final discriminator.

The ideal way to classify these different mass hypotheses, which are exclusive within the NMSSM, would therefore be to train a separate NN with exactly one of the 420 simulated points, which is then used to classify signal and background events. However, the computing effort of training the NN and creating the final histograms 420 times instead of once makes this option unfeasible. Also, in this case a sufficient number of signal events for each mass point needs to be ensured to arrive at a stable training, whereas a grouping of very similar signal mass hypotheses increases the number of signal events to be used for training.

In this analysis, the results are obtained by training separately for individual mass points, which are however merged towards a reduced scheme. This procedure was optimized to balance the achievable results with the highest feasible computing effort of the analysis. The 420 signal mass points are combined to form 68 groups. The grouping as shown in Figure 5.2 is performed as follows: For all simulated mass values of the heavy bosons $m(\mathrm{H})$ up to a mass of 1000 GeV, a single $m(\mathrm{H})$ value is used per group. For the specific $m(\mathrm{H})$ value, up to four neighboring $m(\mathrm{h_S})$ points are merged in a single group.

For mass points $m(\mathrm{H}) > 1000\,\mathrm{GeV}$, the strategy changes. For these higher masses, the expected number of observed events in the signal categories is very low, as only events in which a record of both tau lepton and b-quark decays with very high energies exists qualify. To avoid splitting these events too finely between categories, all $m(\mathrm{H})$ values above 1000 GeV are combined. For each $m(\mathrm{H})$ point, up to nine $m(\mathrm{h_S})$ points are merged.

A separate NN training is performed for each of the 68 groups exclusively on the mass pairs in the group. The existence of possible signal masses of the other groups is unknown to the NN to avoid confusion of the NN between different signal masses, while the same background events are used to train the background categories.

This results in 68 total trainings to be performed, which is to be multiplied by the three final states. These are treated independently from each other, as the event topologies and composition of background events differs significantly between them.
For the three run periods of the CMS detector, differences exist due to detector updates and changes of the setup of the triggering system used for event selection. For this reason, the run periods are usually treated independently for e.g. the simulation of events. For the training, the run periods are however combined and events from all three run periods enter the NN. In order to maintain the information and a degree of freedom regarding the run period and thus to allow the NN to adapt to their specifics, information about the run period is provided to the NN as three boolean input nodes where only one of them is active depending on the run period.

During the application, an event gets assigned an output score for each of the five categories. The choice of category for the event is then given by the largest of the these scores. Furthermore, this maximal score is saved and later used as a final discriminator for the statistical inference.

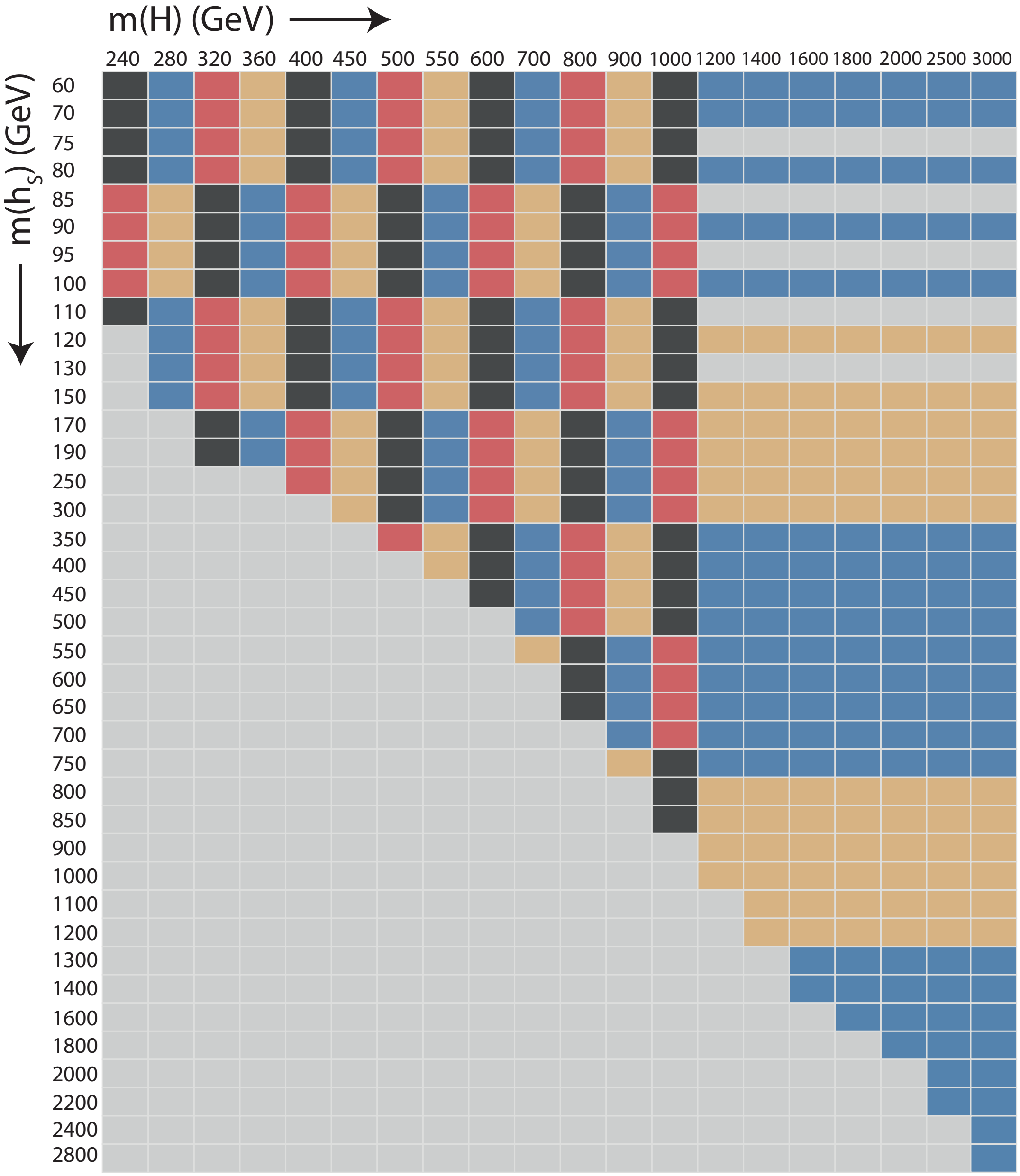


Figure 5.2: Grouping of the signal mass pairs for NN training. Separated areas shown in the same color are grouped and a common NN is trained on these mass pairs simultaneously, while signal mass pairs from other groups are excluded from the training. For $m(\mathrm{H}) \leq 1000\,\mathrm{GeV}$, the grouping consists of single values of $m(\mathrm{H})$ and up to four values of $m(\mathrm{h_S})$. For $m(\mathrm{H}) > 1000\,\mathrm{GeV}$, all $m(\mathrm{H})$ values are merged, and the training is performed depending on $m(\mathrm{h_S})$, for which between five and nine values are also merged. This results in 68 separate regions and thus 68 NN trainings to be performed per final state.

5.1.2 Selection of neural network input features

For the event classification, the NN receives a set of input features for each event. The variables expected to hold the strongest discriminating power over the presence of the process are the quantities related to the tau leptons and b-jets as these are the decay products of the signal process. Furthermore, quantities related to additional jets in the event are included. These mainly can discriminate between background sources as well as allow to infer about presence of a strongly-boosted particle in the event, which is usually accompanied by a recoiling jet in opposite polar direction.

Especially the mass estimators of the $\tau\tau$, bb and bb$\tau\tau$ systems, as well as the quality of the kinematic fit discussed in section 4.6 are of high importance for the multivariate analysis. The signal events are expected to show a peaking structure in all three systems, in contrast to background events, and a very good compatibility with the NMSSM signal hypothesis of two peaking mass systems tested by the kinematic fit. These four quantities are shown in Figure 5.3 for the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state.

The NN is trained on 23 event variables in the $\mu\tau_\mathrm{h}$ and $\mathrm{e}\tau_\mathrm{h}$ final states, and 28 event variables in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state. The choice of variables is based on their importance ranking using a Taylor expansion of the NN response to the individual input nodes [76], to be discussed in more detail below. The different variable selection in the different channels is chosen due to the different background composition, which in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state is dominated by the misidentification of light quark of gluon jets as b-jet. Additional variables containing information regarding the b-jet identification are thus included in this final state. The variable selection is summarized in Table 5.1.

The importance of individual input variables can be tested by decomposing the NN function into a Taylor expansion of the output function of the NN with respect to the input feature as described in [77]. The Taylor coefficients derived by the expansion correlate with the sensitivity of the NN output to the input feature. The ranking of both the marginal values of the input variables as well as the pairwise correlations between input variables are shown in Figure 5.4 for the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state. The figure of merit to determine the ranking is the mean of the absolute Taylor coefficient values $\langle t_i \rangle$ as defined in [77]. The ranking is shown for all five output categories of the NN. As different physics processes are subsumed in the different categories, also the ranking can differ between them.

For all categories, the leading input features are second-order features of the input space, i.e. correlations between variables. A notable example for such a second-order feature is the correlation between the visible mass of the bb$\tau\tau$ system and its mass estimator of the kinematic fit, $m^\mathrm{vis}_{\tau\tau+\mathrm{bb}}$ vs. $m^\mathrm{KinFit}_\mathrm{H}$ which is the leading feature of the signal category, confirming the physics intuition that an invariant mass close to the signal expectation in both systems is a strong indicator for a signal event. An example for a first-order feature with a high ranking is the p_T of the leading τ_h candidate of the event in the classification of jet → τ_h events. These events usually have a significantly softer p_T spectrum than events in which the τ_h candidate is from a genuine tau lepton decay by a heavy resonance.

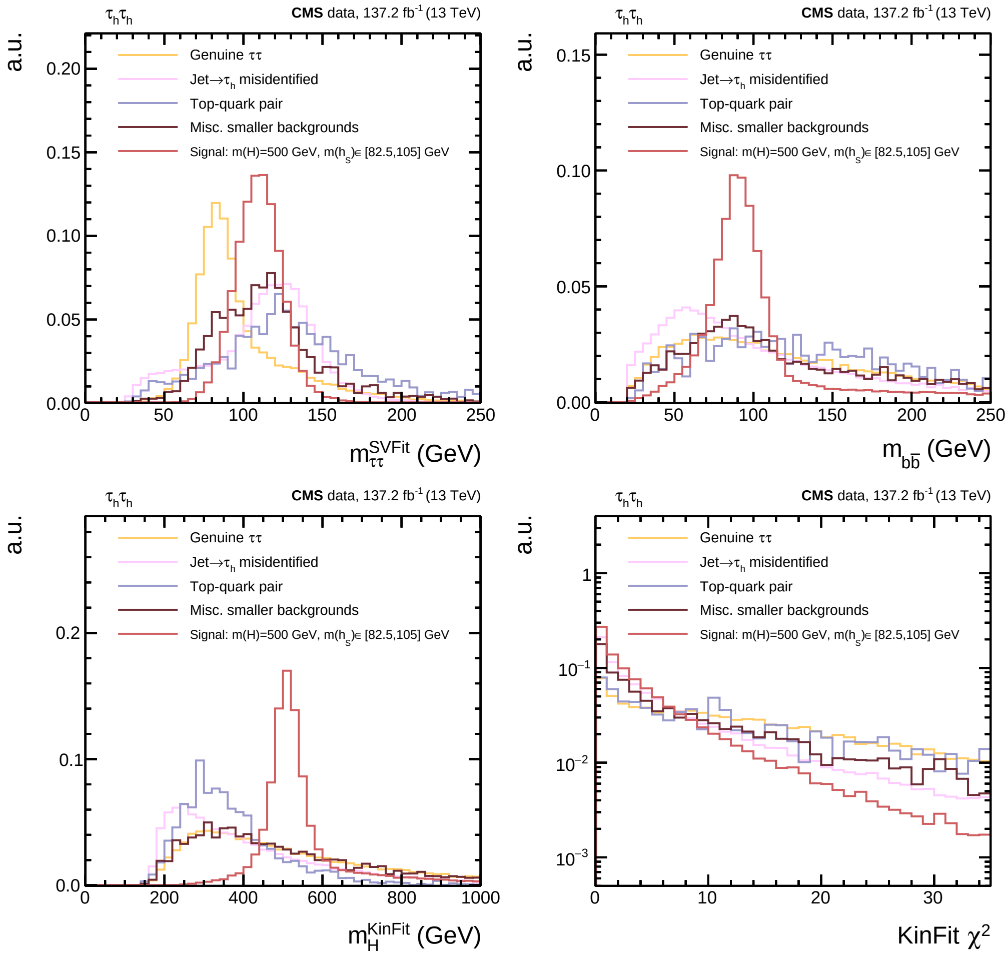


Figure 5.3: Distributions of events used to train the respective background and signal categories in the $\tau_h\tau_h$ final state. Shown are four out of 28 training variables with especially high discriminating power between signal and background: The mass of the $\tau\tau$ system (top left) as estimated by the SVFit algorithm, the reconstructed mass of the di-b-jet system (top right), the mass estimator for $b\bar{b}\tau\tau$ derived by the kinematic fit (bottom left), and the quality of the kinematic fit expressed by its χ^2 value (bottom right). For the estimation of background categories, τ-embedded events are used to describe genuine $\tau\tau$ events (yellow line), the F_F method for jet $\rightarrow \tau_h$ misidentified events (pink line), simulated $t\bar{t}$ events for top-quark pair decays (purple line), and simulated diboson and $h_{SM} \rightarrow \tau\tau$ events make up the miscellaneous smaller backgrounds (brown line). One of 68 signal mass groups is shown in red, in which four similar NMSSM mass pairs in the given mass range are merged. All distributions are scaled to unity to allow their comparison independent of the cross section of the process.

Table 5.1: Selection of the 28 event variables used to train the NN for event classification into signal and background categories. Variables marked by a † are only used in the $\tau_h\tau_h$ final state.

Label	Description
pt_1	p_T of the muon, electron or p_T-leading τ_h
pt_2	p_T of the τ_h (p_T-subleading in $\tau_h\tau_h$)
m_vis	Visible mass of the $\tau\tau$ system
ptvis	Visible p_T of the $\tau\tau$ system
m_sv_puppi	SVFit mass of the $\tau\tau$ system
nbtag	Number of b-tagged jets
bpt_1	p_T of p_T-leading b-tagged jet
bpt_2	p_T of p_T-subleading b-tagged jet
mbb	Invariant mass of the two b-tagged jets
ptbb	p_T of the two b-tagged jets
njets	Number of non-b-tagged jets
jpt_1	p_T of the p_T-leading non-b-tagged jet
jpt_2	p_T of the p_T-subleading non-b-tagged jet
jdeta	$\Delta\eta$ between the two p_T-leading non-b-tagged jets
mjj	Invariant mass of the two p_T-leading non-b-tagged jets
dijetpt	p_T of the two p_T-leading non-b-tagged jets
m_ttvisbb	Invariant mass of the visible $\tau\tau$+bb system
kinfit_mH	m(H) estimator derived by the kinematic fit
kinfit_mh2	Discrete $m(\mathrm{h_S})$-value selected for the minimal χ^2-value of the kinematic fit
kinfit_chi2	Minimal χ^2-value of the kinematic fit
2016	True if the event was recorded in the 2016 run period, false otherwise
2017	True if the event was recorded in the 2017 run period, false otherwise
2018	True if the event was recorded in the 2018 run period, false otherwise
bm_1†	Mass of the p_T-leading b-jet
bm_2†	Mass of the p_T-subleading b-jet
bcsv_1†	b-jet discriminator score of the p_T-leading b-jet
bcsv_2†	b-jet discriminator score of the p_T-subleading b-jet
jetCSV†	In case of only one jet passing the medium b-discriminator working point, b-jet discriminator score of non-b-tagged jet used for the bb system

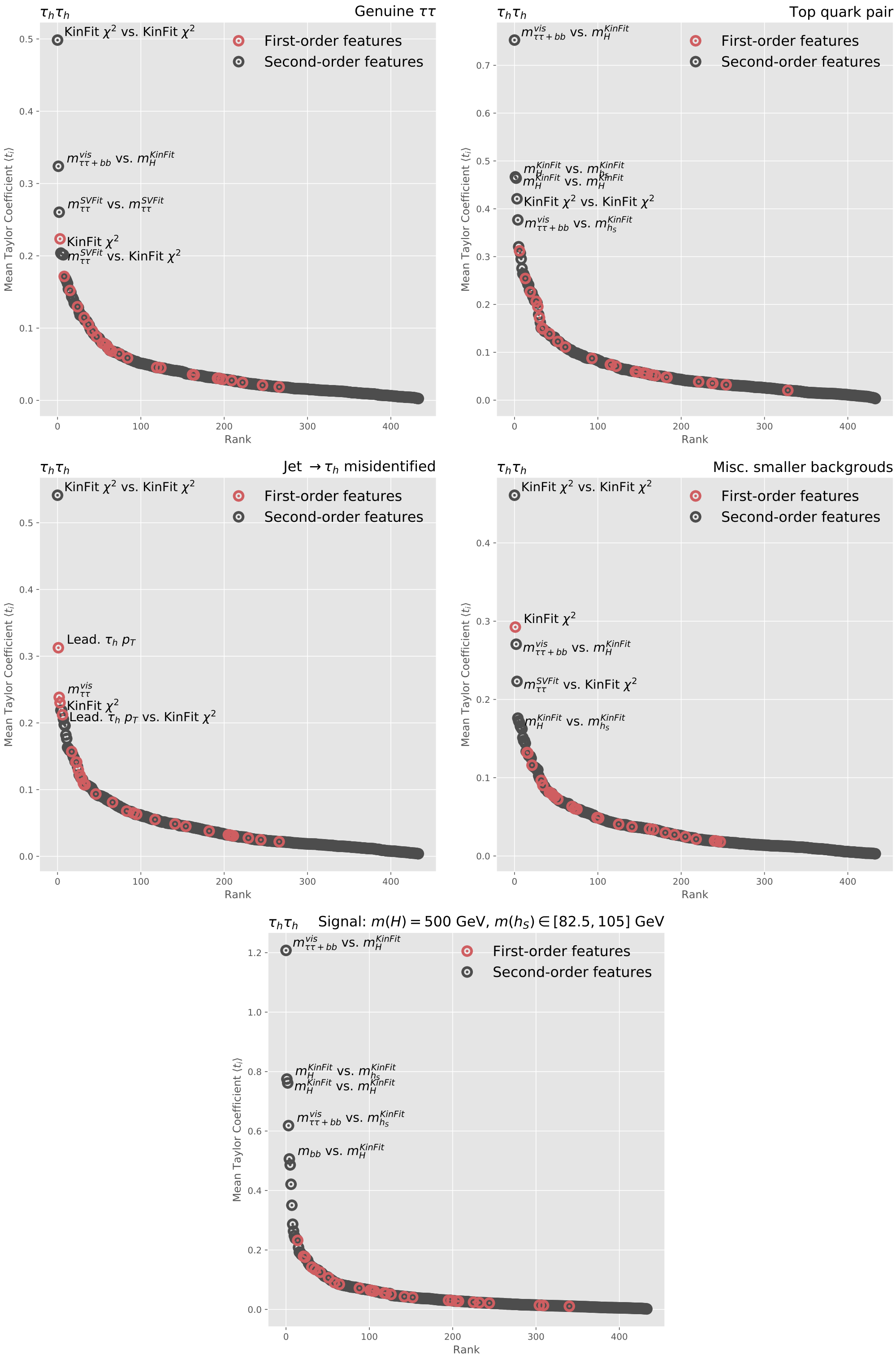


Figure 5.4: Importance ranking of the NN input features used to discriminate events between four background processes (top four) and the NMSSM signal process (bottom) for masses $m(\mathrm{H}) = 500\,\mathrm{GeV}$ and $m(\mathrm{h_S}) \in [82.5, 105]\,\mathrm{GeV}$ in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state. The mean absolute values of the Taylor coefficients are used to determine the ranking. The first-order features of the variables are shown in red and second-order features in black circles. The five leading features for the respective category are labeled in the figures. Second-order features containing information regarding the quality of the kinematic fit (KinFit χ^2), the mass estimations of the $\tau\tau\mathrm{b}\bar{\mathrm{b}}$ systems via the kinematic fit ($m_\mathrm{H}^\mathrm{KinFit}$) or its visible invariant mass ($m_{\tau\tau\mathrm{bb}}^\mathrm{vis}$), as well as the mass of the $\tau\tau$ system ($m_{\tau\tau}^\mathrm{SVFit}$) are usually among the highest-ranking input features.

The high ranking of a self-correlation of a variable, e.g. KinFit χ^2 vs. KinFit χ^2 which is the highest-ranked feature in three background categories, is to be understood as the NN taking not the marginal value, but rather the peaking structure of the variable into account. This can be very useful in cases such as the KinFit χ^2, which shows a strong peak towards zero in the signal categories as opposed to the background categories, as was seen in Figure 5.3.

The magnitude of the coefficients drops exponentially in each category with the increasing rank of the input features. While specific examples with high Taylor coefficients have been highlighted above, the power of the NN multi-classification is derived by the combination of several hundred input features.

5.1.3 Validation of the NN input using goodness-of-fit testing

The selected variables are subject to goodness-of-fit tests to validate their description by the background model. A goodness-of-fit test is a way of quantifying how well the observation is described by the model, taking into account all statistical and systematic sources of uncertainty. The saturated model test [78] is used as primary test for the validation of NN input features. It is a likelihood-based generalization of the χ^2-test, which is in its usual form defined as

$$\chi^2 = \sum_i \frac{(d_i - f_i)^2}{\sigma_i^2} \tag{5.4}$$

where d_i denotes the ith observed datapoint with standard deviation σ_i, and f_i the model prediction for this datapoint. A likelihood function is derived for the final statistical model. The function used for the analysis as well as the goodness-of-fit tests will be discussed in the following section and shown in Equation 5.10. In the following, the saturated model goodness-of-fit test will be explained using a simple model containing only statistical uncertainties. In this case the likelihood is given by

$$\mathcal{L} = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(d_i - f_i)^2}{2\sigma_i^2}\right) . \tag{5.5}$$

This likelihood is then evaluated with respect to a hypothesis in which the data fit the model exactly, i.e. $f_i = d_i$ at every measured value. This is called the saturated model and yields in this case

$$\mathcal{L}_{\text{saturated}} = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} . \tag{5.6}$$

The ratio of equations 5.5 and 5.6 is then used to compute the likelihood ratio λ as

$$\lambda = \frac{\mathcal{L}}{\mathcal{L}_{\text{saturated}}} = \prod_i \exp\left(\frac{-(d_i - f_i)^2}{2\sigma_i^2}\right) . \tag{5.7}$$

The goodness-of-fit test statistic q_{obs} is then given by

$$q_{\mathrm{obs}} = -2\ln\lambda \tag{5.8}$$

For the simple case used to describe the saturated model, the equality of $q_{\mathrm{obs}} = \chi^2$ becomes apparent from equations 5.8 and 5.7. This equality is not given in case of a more complex likelihood function containing systematic uncertainties, as is used for this analysis.

As the saturated model maximizes the likelihood, the likelihood ratio λ is always ≤ 1, and the test statistic is always positive. Lower values of the test statistic refer to closer agreement of the measurement and the model.

One has to keep in mind that the saturated model goodness-of-fit test ignores information about the direction of the deviations as well as their order and is thus weak in inferring about correlated deviations of the observed data with respect to the model, as well as data that is consistently below or above the expectation.

For each distribution of the input features to the NN, the test statistic is computed once for the observed data yielding q_{obs}. In this analysis, the value of q_{obs} cannot be assumed to follow a χ^2-distribution as the likelihood contains systematic uncertainties and its degrees of freedom are not well-defined due the unknown correlations of these systematic uncertainties in the model.

The p-value of the test statistic is thus computed in a Monte Carlo approach. Pseudo-datasets ("toys") are randomly generated based on the uncertainty of all nuisances parameters in the model. For each test, 500 toys are generated and their test statistic is computed. The comparison of these toys with the observed value of the test statistic allows the computation of a p-value, which is in this case defined as the fraction of toys with equal or larger test statistic value as the measurement.

$$p = \frac{N_{\mathrm{toys}}(q_{\mathrm{toy}} \geq q_{\mathrm{obs}})}{N_{\mathrm{toys}}} \tag{5.9}$$

The p-value ranges from zero to one, with values close to zero hinting at large deviations of the observed data from the model.

Each variable distribution is tested after the event selection, calculating the p-value of the observed data distribution under the assumption of the given model with simulation and data-driven methods and including the full uncertainty model of the analysis.

No signal estimation is used during the goodness-of-fit testing. As the goodness-of-fit test is performed on binned data, a binning is automatically chosen depending on the tested variable. For one-dimensional distributions, bin edges are chosen for ten bins, such that each bin is equally populated by the observed data. This is illustrated in Figure 5.5. Here, the estimator of the heavy scalar boson mass H derived via the kinematic fit, $m_{\mathrm{H}}^{\mathrm{KinFit}}$ is shown binned in ten equipopulated bins as described above. The observed data thus has the same yield in each bin. In case of a variable with a peaking structure such as $m_{\mathrm{H}}^{\mathrm{KinFit}}$, the binning is very heterogenous with bins in the peak region below 500 GeV spanning a range of around 30 GeV, and the last bin covering the complete tail of the distribution spanning a range of 3000 GeV.

The resulting value of the test statistic of both the observed data and the distribution of the test statistic of the 500 toys can also be found in Figure 5.5. The p-value is then derived from this distribution using Equation 5.9.

A good modeling of the one-dimensional distributions does not ensure a good modeling of the NN output, as the NN takes more than the first-order input features into account. A NN can utilize the input features by considering higher-order features between variables. For this reason, also the two-dimensional distributions of every variable pair are validated by a goodness-of-fit test.

An equipopulated binning is again chosen for each variable in five bins. For each variable and bin, the data is restricted to this bin and all other variables are distributed again in five bins over their full spectrum. This results in a distribution of $5 \times 5 = 25$ bins for each variable pair. The distribution is then subject to a goodness-of-fit test using the saturated model. The number of possible combinations of N input variables are $\frac{1}{2}(N^2 - N)$, i.e. 190 in the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states ($N = 20$) and 300 in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state ($N = 25$). Multiplied by the three years of CMS data-taking, for which the goodness-of-fit tests are performed independently, this results in 195 one-dimensional goodness-of-fit tests, and 2040 two-dimensional tests. Note that three input variables, the boolean input nodes representing the run period are not subject to goodness-of-fit testing.

It is important to note that the utilization of the input space by the NN is not restricted to two dimensions and also features of higher order can be used. The number of goodness-of-fit tests to be performed when including even higher dimensions would however pose an unfeasibly large computing effort, and the tests are thus restricted to one and two dimensions.

Thresholds to the p-values are defined below which the input features are subject to further scrutiny to exclude systematic mis-modelings which are not covered by the uncertainty model. The threshold for one-dimensional test is chosen at $p = 0.05$, the threshold for two-dimensional tests at $p = 0.005$. Due to the approximately uniform distribution of the p-values, about ten p-values of both the one-dimensional and the two-dimensional test are expected to fail the threshold due to statistical fluctuations alone. It is therefore important not to exclude variables only based on a low p-value. In case a systematic mis-modeling of the variable cannot be excluded, i.e. if the data shows a visible trend compared to the prediction, if the value of the observed test statistic is much larger than all toys, or if the same variable shows low p-value in many run periods and final states, measures are taken to improve the modeling. Alternatively, if the impact of considering the variable on the final results is small, the variable is excluded from the NN training.

In this regard, the five variables only included in the $\tau_\mathrm{h}\tau_\mathrm{h}$, but not the $\mu\tau_\mathrm{h}$ and $\mathrm{e}\tau_\mathrm{h}$ do not only have a low impact on these final states as discussed above, also tensions in the modeling related to the $\mathrm{t\bar{t}}$ simulation appear during the goodness-of-fit testing. In the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, the vast majority of $\mathrm{t\bar{t}}$ events are taken from data-driven methods, i.e. τ-embedded events and the F_F method, which model the respective distributions well.

The results of the one-dimensional goodness-of-fit tests can be found in Figures A.8-A.16. The results of the two-dimensional tests are found in Figures A.17-A.25. In the figures, the labels as defined in Table 5.1 are used. The summary of all tests can be found in Figure A.26. The goodness-of-fit test results show good compatibility with the expected distribution of the p-values, indicating that the systematic model is able sufficiently to describe the input space within the systematic uncertainties. The distribution is expected to be approximately uniform, however also deviations from uniformity can occur in case strong correlations between the variables exist. Cases with low p-values are investigated and, except the five excluded variables mentioned above, the p-values are attributed to statistical fluctuations.

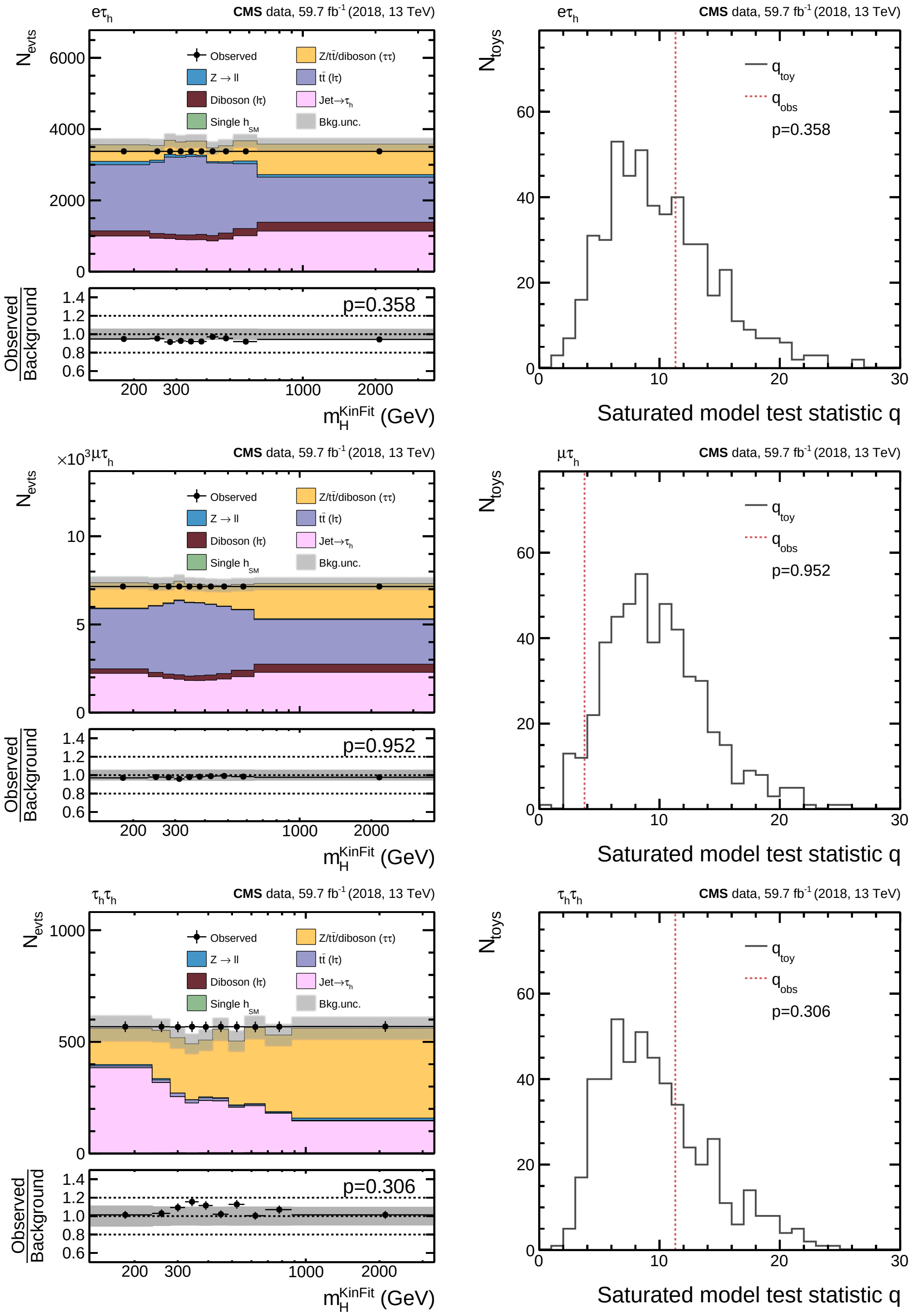


Figure 5.5: Distributions of $m_{\mathrm{H}}^{\mathrm{KinFit}}$ in the equipopulated binning used for the saturated model goodness-of-fit test using data collected in 2018 (left), and resulting distributions of the saturated model test statistic (right) of 500 toys (q_{toy}) in relation to the observed test statistic (q_{obs}) in the eτ_{h} (top row), $\mu\tau_{\mathrm{h}}$ (middle row), and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ (bottom row) final states. The value of q_{obs} is indicated by a red dotted line. The resulting p-value is defined as the fraction of toys with equal or greater test statistic.

5.2 Statistical inference and uncertainty model

After the event classification five output categories are obtained, of which four are background control categories, and one is the signal category. The categories are filled with the events which have their maximal NN score in this category. This score simultaneously serves as final discriminator of the analysis, with the minimal value being the inverse of the number of categories, so 0.2 in the case of five categories, and the maximum value being one. The histograms used for statistical inference are derived by binning the scores in bins of width 0.05. If less than 10 events are contained in a bin, or if the combined background uncertainty of the bin is larger than 90% of its content, bins are merged with neighboring bins, starting from high NN score values going towards lower ones.

As higher scores correspond to increased probabilities that the respective event corresponds to the target process of the category, the upper bins of a category are expected to be very pure in the respective target process. In case of background categories, these serve as control regions for independent background sources, to be used by the final fit to constrain the uncertainties on the background. In the signal categories, the upper bins are expected to have the highest signal-over-background ratio and drive the measurement of the signal significance or exclusion limit.

All histograms enter a combined binned likelihood function of the form

$$\mathcal{L}(d\,|\,\mu\cdot s(\theta)+b(\theta)) = \prod_{i\in\text{bins}} \mathcal{P}(d_i\,|\,\mu\cdot s_i(\theta)+b_i(\theta)) \times \prod_{j\in\text{nuis}} \mathcal{C}(\hat{\theta}_j\,|\,\theta_j)\ . \tag{5.10}$$

The function derives the likelihood of observing the measurement d, taking into account the Poisson-distributed ($\mathcal{P}$) value of observed events in each bin d_i given the model prediction for this bin for the signal s_i with signal strength modifier μ and the background b_i. The prediction of s and b depend on the values of the nuisance parameters θ. Each individual nuisance parameter θ_j is following a probability density function $\mathcal{C}$, with $\hat{\theta}_j$ labelling the estimate of θ_j used to derive s and b.

This function is maximized to find the best-fit estimate of the signal strength modifier μ, which can be scaled during the fit without penalty to the likelihood.

The nuisance parameters reflect all systematic uncertainties influencing the signal and background estimation. A crucial part of the statistical inference is the correct estimation of these systematic uncertainties. These will be discussed in detail in the following.

5.2.1 Uncertainties on the NMSSM signal simulation

Sources of uncertainty during the signal event simulation are propagated to the signal estimation as systematic uncertainties. The uncertainties appear from the limited precision of the matrix-element calculation of the hard interaction vertex resulting in the $\mathrm{gg} \to \mathrm{H} \to \mathrm{h}_{\mathrm{SM}}\mathrm{h}_{\mathrm{S}}$ process and are split into two sources:

1. An uncertainty on the parton density function (PDF) used to simulate the hard proton-proton interaction and

2. an uncertainty describing the impact of the chosen re-normalization and factorization scales at which the process is calculated.

NNPDF [62] is used as estimate of the underlying PDF of the protons. The final PDF is derived from a set of 100 individual PDF fits in which all input parameters are varyied according to their uncertainty. The mean of the 100 fits is used as the nominal function used for the simulation in `MadGraph5_aMC@NLO` [60]. The impact of using any of the 100 fits is stored as an additional set of weights on an event-by-event basis. Using these weights results in 100 slightly different signal estimations. The standard deviation σ of the 100 functions with respect to the nominal one is used as final PDF uncertainty.

In an uncorrelated uncertainty, the choice of renormalization scaled μ_R and factorization scale μ_F is varied and the impact of the variation on the final signal estimation evaluated. The scales are dynamically chosen on an event-by-event basis in `MadGraph5_aMC@NLO` by clustering the external states until the system is reduced to a $2 \to 2$ topology. The squared transverse mass of this system is used as scale choice.

$$\mu_R = \mu_F = m_{\mathrm{T}}^2 \tag{5.11}$$

As $\pm 1\sigma$ uncertainty estimation, the scale choice is divided and multiplied by a factor of two and the event simulation repeated with this scale choice.

Both uncertainties are shown in Figure 5.6. The PDF uncertainty introduces mainly a flat acceptance effect of around 18%, whereas the scale uncertainties also introduce a shape effect especially for lower momenta of the $\mathrm{h_{SM}}$ and $\mathrm{h_S}$ bosons.

In addition to these uncertainties on the simulation of the hard interaction, all uncertainties referring to the detector simulation, such as uncertainties on the jet energy and resolution or the efficiency of the lepton and b-jet identification which will be discussed below are also applied for the signal simulation and fully correlated with the simulated samples used for the background estimation.

5.2.2 Common uncertainties in τ-embedded and simulated events

All uncertainties referring to the reconstruction of the electron, muon, or τ_h in the event affect both the τ-embedded events, in which only the tau lepton decays are simulated, as well as the fully simulated events. To reflect both the shared effects of the simulated detector response as well as the observed data used to derive the corrections, the uncertainties are correlated by 50% between τ-embedded and fully simulated events.

The specific sources of uncertainty shared between the two are listed below:

- Electron and muon identification: Global 2% normalization uncertainties are introduced, as the p_T dependence of the efficiencies is low.
- Electron and muon triggering: As for the identification, a 2% normalization uncertainty is introduced, which is uncorrelated across the type of trigger (e/μ+τ_h or single e/μ) used to select the event. It can introduce a shape effect in the final

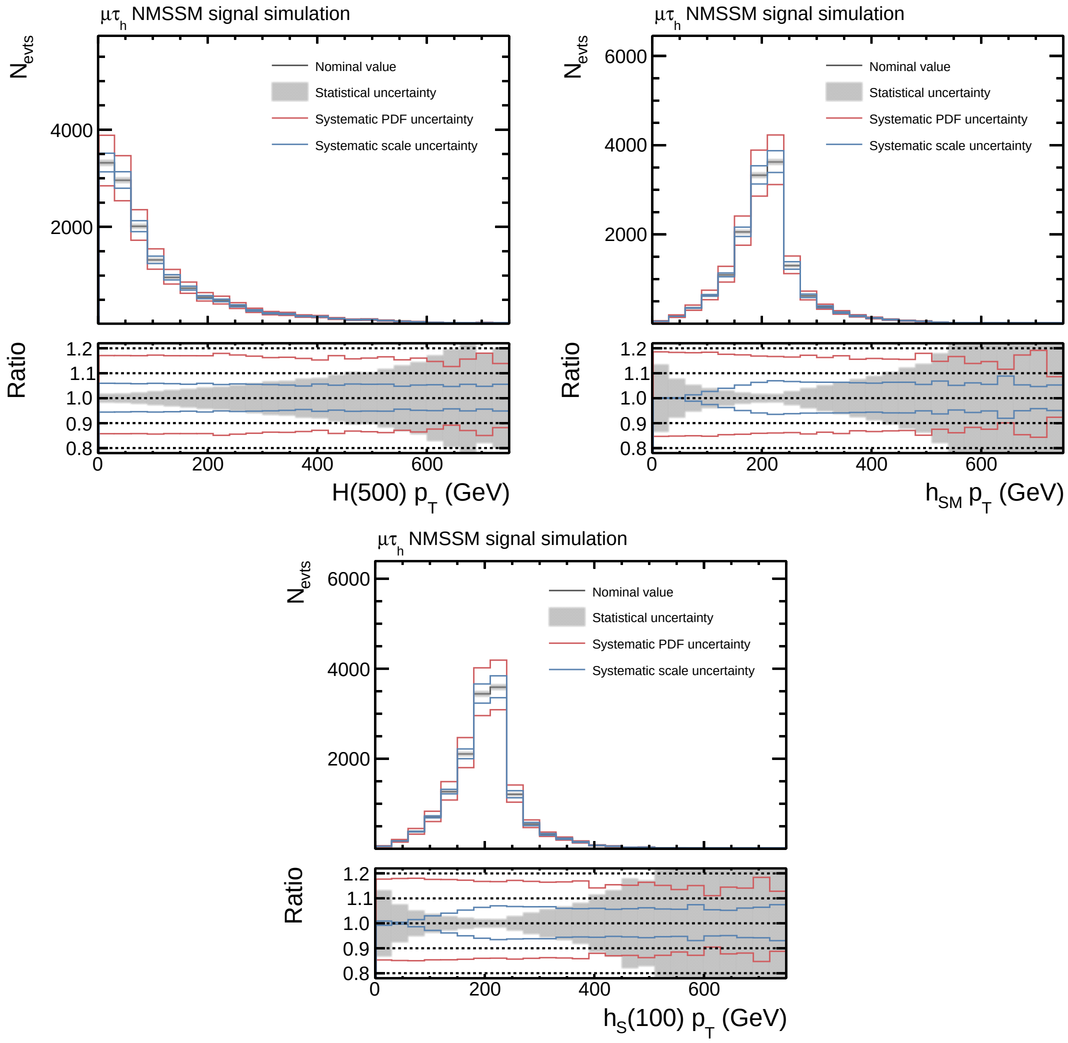


Figure 5.6: The two sources of systematic uncertainties on the NMSSM signal simulation, shown for signal simulation after event selection in the $\mu\tau_{\mathrm{h}}$ final state for the 2018 run period. The simulation of the process gg → H(500 GeV) → $\mathrm{h_{SM}}$(125 GeV)$\mathrm{h_S}$(100 GeV) is shown. The impact of the uncertainties is derived as a function of the simulated values of the p_{T} of the three Higgs bosons: The heavy boson H (top left) and its decay products $\mathrm{h_{SM}}$ (top right) and $\mathrm{h_S}$ (bottom). The nominal value of the signal estimation is shown as black line. The systematic uncertainty on the PDF used is shown as red band, the uncertainty on the factorization scale as blue band. For comparison, the statistical uncertainty of the sample is indicated by a grey band.

analysis as the efficiency of different triggers impact the p_{T} spectrum of the leptons in different regimes.

- Electron energy scale: Here, the uncertainties are treated uncorrelated between τ-embedded and simulated events. The uncertainties for simulated events are based on the event-by-event uncertainty of the electron energy resolution corrections discussed in section 4.5.4. In τ-embedded events, the uncertainties given in Table 4.6 are propagated to the analysis.

- τ_h identification: In the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ final states, the $p_\mathrm{T}^{\tau_\mathrm{h}}$ dependence of the correction factors is reflected in the uncertainty, which is also split into the respective p_T regions and can thus change both the yield and the shape of the $p_\mathrm{T}^{\tau_\mathrm{h}}$ spectrum. In the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, the correction factors binned in the decay mode are used, which is also reflected in the uncertainties.
- τ_h triggering: The uncertainty of the fit over $p_\mathrm{T}^{\tau_\mathrm{h}}$ which is applied to the measured efficiency values is propagated to the uncertainty model.
- τ_h energy scale: The uncertainties derived from the maximum likelihood fits as given in Tables 4.7 and 4.8 are propagated to the final analysis as systematic uncertainties.

5.2.3 Uncertainties specific to τ-embedded events

For the τ-embedding method, a global 4% normalization uncertainty is used, reflecting uncertainties on the measured efficiency of the di-muon selection in the data. Furthermore, an uncertainty is added taking into account the decays of $\mathrm{t\bar{t}} \rightarrow \tau\mu + X \rightarrow \mu\mu + X'$ events which are selected as prompt di-muon events during the event selection. Even though their contribution is very small as was discussed in section 4.3.1, they can impact the final shape of the distribution for an analysis enriching the fraction of $\mathrm{t\bar{t}}$ events in the τ-embedded sample by requiring a b-jet, as is the case for this analysis. The number and shape of $\mathrm{t\bar{t}}$ events contained in the embedded event sample is estimated using $\mathrm{t\bar{t}}$ simulation, in which the simulated information is used to select only $\mathrm{t\bar{t}} \rightarrow \tau\tau + X$ events. Of this distribution, 10% is added and subtracted to the τ-embedded distribution as $\pm 1\sigma$ confidence intervals. This makes the magnitude of the uncertainty proportional to the estimated fraction of $\mathrm{t\bar{t}}$ events in the τ-embedded event sample.

In Figure 5.7, the impact of this uncertainty on the final discriminator is shown in the $\mu\tau_\mathrm{h}$ final state. The uncertainty is expected to be 10% of the fraction of $\mathrm{t\bar{t}} \rightarrow \tau\tau + X$ events in a specific bin. In the genuine $\tau\tau$ category, mainly $\mathrm{Z} \rightarrow \tau\tau$ events are collected and the uncertainty is negligible. In the top-quark pair category however, the $\mathrm{t\bar{t}} \rightarrow \tau\tau + X$ event contribution of τ-embedded events is enriched, indicated by the uncertainty being around 10% especially for higher NN scores. If the uncertainty reaches 10%, the fraction of $\mathrm{t\bar{t}} \rightarrow \tau\tau + X$ events is estimated to be 100%. The uncertainty can even exceed 10% if the $\mathrm{t\bar{t}}$ simulation estimates more events in this bin than the τ-embedding method does.

5.2.4 Uncertainties specific to simulated events

In fully simulated events, additional uncertainties enter, taking into account that in these events also additional (b-)jets in the event as well as the kinematics of the heavy resonance bosons or top-quarks are simulated and can be different from the data.

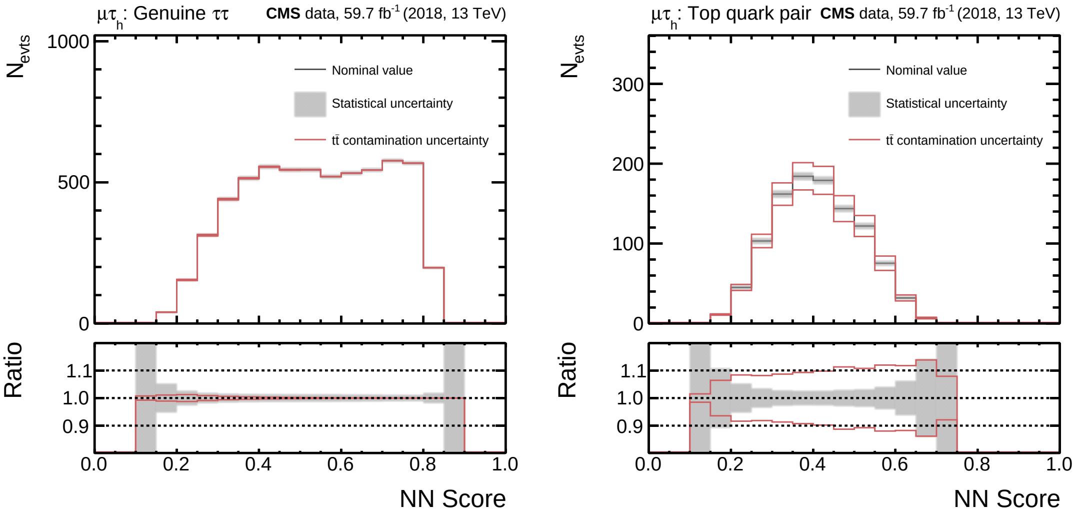


Figure 5.7: The distribution of τ-embedded events is shown as a black line in two NN background categories in the $\mu\tau_\mathrm{h}$ final state of the 2018 run period: The genuine $\tau\tau$ category (left), enriching Z/$\mathrm{t\bar{t}}$/VV$\rightarrow \tau\tau$ events, and the top-quark pair category (right), enriching $\mathrm{t\bar{t}}(\ell\tau)$ events. The NN separates the Z $\rightarrow \tau\tau$ from the $\mathrm{t\bar{t}} \rightarrow \tau\tau + X$ events in the embedded event sample, with the former being assigned to the left plot, and the latter to the right plot. The $\mathrm{t\bar{t}}$ contamination uncertainty, shown as a red band, is estimated from simulated events and is proportional to the fraction of $\mathrm{t\bar{t}} \rightarrow \tau\tau + X$ events in this bin. The uncertainty plays a role mainly in the top-quark pair category, where this fraction is close to 100% for large NN scores. The statistical uncertainty of the distribution is indicated by a grey band for comparison.

- Jet energy scale: The complex reconstruction and calibration of jets recorded at the CMS detector was discussed in detail in section 3.2.5. During the calibration and finally correction of the jet energies in simulated events, also a set of 27 uncertainty sources for the 2016 run period, and 28 sources for the 2017 and 2018 run periods are derived. The sources are grouped in a reduced scheme by merging strongly-correlated uncertainties to obtain eleven nuisance parameters to be used in the final analysis. These refer to e.g. the statistical limitations of the measurements used for calibration, a time-dependence of the energy measurements in the data due to aging of the detector which does not happen in simulated events, or non-closure corrections introduced to describe remnant differences between simulation and the data.

- Jet energy resolution: Not only the central value of the measured energy is different between data and simulation, also the resolution of the energy measurement is usually narrower in simulated than in observed jets. The correction of this effect is accompanied by an additional systematic uncertainty.

- $\vec{p}_\mathrm{T}^\mathrm{miss}$ and recoil uncertainties: Depending on the physics process described by the simulation, two separate uncertainties on the measured value of $\vec{p}_\mathrm{T}^\mathrm{miss}$ are taken into account. For processes which do not contain a single heavy resonance decay, i.e. $\mathrm{t\bar{t}}$

and diboson events, no additional correction of the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ other than a recalculation after application of the jet and lepton energy corrections is applied. The uncertainty on the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ estimate is in this case derived by the propagation of the respective uncertainties of the corrections to the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ as well as by the amount of unclustered energy in the event, i.e. energy in the electromagnetic and hadronic calorimeters which could not be assigned to specific particle candidates and might be attributed to detector noise.

For events that do contain a heavy resonance decay such as Z, H or $\mathrm{h_{SM}}$, the recoil corrections described in section 4.5.4 are applied. The level of confidence in this correction is calculated using the hadronic recoil

$$\vec{H}_{\mathrm{T}} = -(\vec{p}_{\mathrm{T}}^{\,\mathrm{Z/H/h_{SM}}} - \vec{p}_{\mathrm{T}}^{\,\nu} + p_{\mathrm{T}}^{\mathrm{miss}}) \tag{5.12}$$

of the event. This recoil is split in a parallel and orthogonal component with respect to the vector $\vec{p}_{\mathrm{T}}^{\,\mathrm{Z/H/h_{SM}}}$, and the two components shifted up and down by $0.3 - 5.8\%$ with respect to their previous mean. The exact value of the shift is derived from the $\mathrm{Z} \to \mu\mu$ control region in which also the initial correction is derived, and depends on the $|\vec{p}_{\mathrm{T}}^{\,\mathrm{Z/H/h_{SM}}}|$ as well as the number of jets in the event. The shifted parallel and orthogonal components are independently used to re-derive $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$. The resulting values are used as two uncorrelated nuisance parameters reflecting the confidence interval of the calibration.

- Top-quark pair p_{T} spectrum: In simulated events, the spectrum of the top-quark pair p_{T} is corrected as discussed in section 4.5.4. As confidence intervals, the shape of the $\mathrm{t\bar{t}}$ simulation is used in which the correction is not applied at all, or applied twice.

- Efficiency of b-jet identification: The centrally measured correction factors discussed in section 4.5.4 contain also the p_{T} and η dependent uncertainty of the respective measurements, and are applied in the analysis.

- Prefiring: The correction of the missing prefiring issue in simulated events as discussed in section 4.5.4 is taken into account by propagating the confidence intervals of the prefiring probability used in the correction to the final analysis.

- Luminosity: Simulated events are scaled according the the recorded luminosity of the CMS detector during a specific run period. This luminosity is measured for the three run periods in references [70, 79, 80] and known to a precision of 2.5% (2016, 2018) or 2.3% (2017). This is propagated to a set of nuisance parameters, impacting the normalization of all simulated processes, taking into account both uncorrelated and correlated sources between the three run periods.

- Cross section of simulated processes: Next to the luminosity, simulated events are also scaled according the cross section measured for the described physics process. These uncertainties are 6% for the $\mathrm{t\bar{t}}$ process, 5% for the VV and single-top processes, 2% for Z boson production and between 1.3% and 3.9% for single $\mathrm{h_{SM}}$ production, depending on its production mechanism. All cross section uncertainties are fully correlated across all three run periods.

5.2.5 Uncertainties specific to the F_F method

The individual F_F^i (i = QCD,W+jets,$\mathrm{t\bar{t}}$) values of the F_F method described in section 4.4 as well as their corrections are subject to statistical fluctuations in the respective determination regions (DR^i) of the method. The statistical uncertainties are saved during the measurement and applied on the final estimation of jet → τ_h misidentified events.

For this, the uncertainties on the parameters of the fit applied to the measured values are derived. They are parametrized into two main contribution, the first representing the y-intercept of the fit and thus affecting only the overall normalization of the jet → τ_h estimate, and the other reflecting only the slope of the fit. This results in two independent nuisance parameters per measurement, of which one can scale the overall event yield of jet → τ_h misidentified events, and the other the $p_\mathrm{T}^{\tau_\mathrm{h}}$ dependence of the measurement. The same procedure is performed in all final states. In the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, where also higher-order polynomials are used for the fit, the fit function is approximated to a linear dependency for the derivation of uncertainties. The two nuisance parameters are added independently for the three measurements F_F^i and the two or three jet multiplicity regions.

For the non-closure corrections and the corrections for the extrapolation from the DR to the signal region (SR), also the statistical uncertainties are propagated to the analysis using two independent nuisance parameters. As the corrections are not fitted, but rather smoothed by a Gaussian kernel smoothing, the parametrization is performed by calculating the envelope of the smoothed curve in a Monte Carlo approach, in which the measured points are shifted within their uncertainties 100 times and the smoothing repeated. The $\pm 1\sigma$ intervals of the envelope are then used to extract the normalization and slope parameter as a function of the variable which was used to derive to correction, e.g. the muon or electron p_T, or m_vis. These uncertainties are treated uncorrelated from the uncertainties described above.

In addition to these statistical uncertainties propagated to the jet → τ_h estimation, also systematic uncertainties of the F_F method are reflected in the uncertainty model. Each non-closure as well as extrapolation correction is reflected in an uncertainty taking into account the magnitude of the correction. The idea behind this method is that the confidence in the prediction by F_F depends not only on the statistical precision of its measurement, but also on the validity of the method itself. If e.g. the difference between the DR and SR is found to be very large and needs a significant correction, a larger uncertainty is applied respectively. This is done by adding a nuisance parameter proportional to the magnitude of the each correction, giving the final fit the freedom to strengthen or weaken the applied correction. Depending on the correction, these nuisance parameters can impact the shape and the normalization of the estimation simultaneously.

The impact of these uncertainties on the final jet → τ_h estimation are shown in Figure 5.8, where the two uncertainty parameters reflecting the statistical precision of the $F_\mathrm{F}^\mathrm{QCD}$

measurement as well as the uncertainty proportional to the magnitude of the corrections are shown. While the uncertainty due to the statistical precision of the measurement is around $1-5\%$, the uncertainty on the corrections can be up to 10% if large deviations between the DR and the SR are found.

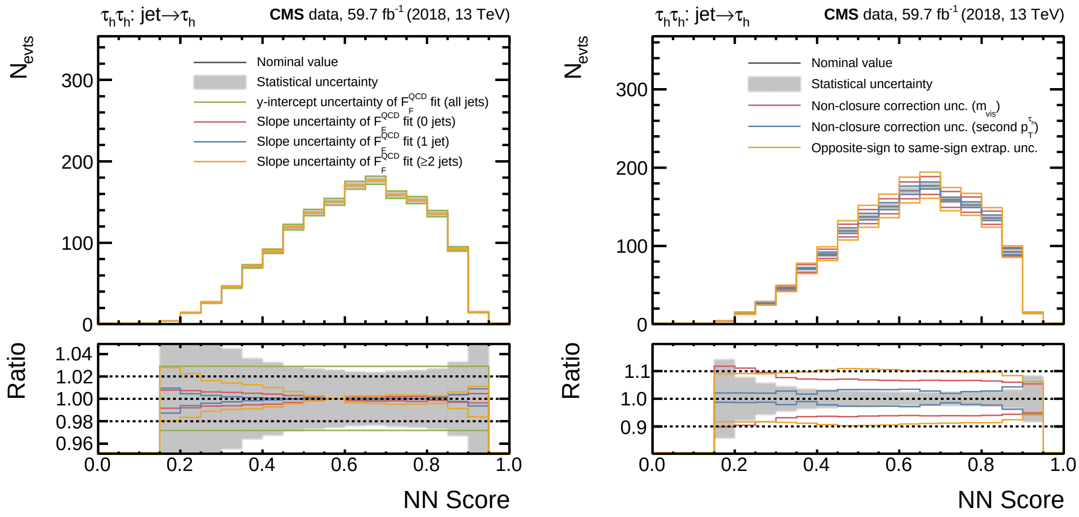


Figure 5.8: The distribution of the jet → τ_{h} estimation is shown for the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state of the 2018 run period as a black line in the jet→ τ_{h} NN background category designed for these events. On the left, the impact of the statistical uncertainty of the fit result of the $F_{\mathrm{F}}^{\mathrm{QCD}}$ measurement is shown. This uncertainty is split in two parameters, of which the parameter reflecting the uncertainty on the slope of the fit is shown for events with no jet (red line), one jet (blue line), and at least two jets (orange line). In this final state, the additional normalization uncertainty of 2.9%, shown as green line, enters from the uncertainty on the y-intercept of the fit. On the right, the systematic uncertainties proportional to the magnitude of the non-closure (red and blue line) and extrapolation corrections (orange line) applied on the $F_{\mathrm{F}}^{\mathrm{QCD}}$ are shown. These systematic uncertainties usually have a much larger impact than the statistical uncertainty of the measurement itself, as can be seen in the different subplot ranges of the two plots. In the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state especially the extrapolation from the determination region using same-sign events to the opposite-sign signal region, requires a significant correction and thus uncertainty.

An additional source of uncertainty concerns the subtraction of processes other than QCD multijet or W+jets events in the QCD and W+jets DR respectively. As has been discussed in section 4.4, these are subtracted from the data using simulated or τ-embedded events. If these are modeled incorrectly, the measurement of the $F_{\mathrm{F}}^{\mathrm{QCD/W+jets}}$ can be biased. The combined shape of the events to be removed is therefore scaled by $\pm 7\%$ and the measurement repeated. The impact of this variation on the $F_{\mathrm{F}}^{\mathrm{QCD/W+jets}}$ is used as uncertainty.

Finally, an uncertainty is added reflecting the estimation of the fractions in the application region. The individual contributions are again scaled by $\pm 7\%$, with the remaining fractions increased or decreased by the same factor relative to their previous size such that the sum of all contributions stays constant, with the resulting differences to the jet → τ_h estimation used as systematic uncertainties.

5.2.6 Uncertainties on the statistical precision of background estimation methods

In all three methods of background estimation, the estimation is performed using event samples which themselves contain only a limited number of events. This number is intrinsically limited by the computation for the specific process in the case of simulation-based estimation, by the available di-muon events in the data in the case of τ-embedded events, or by the number of events in the application region used for the F_F method.

This creates a systematic uncertainty purely statistical in nature, which is thus uncorrelated in each individual category and bin of the final discriminator. The uncertainty is described using the Barlow-Beeston approach [81], in which a single Gaussian nuisance parameter is added for each bin. In this analysis, in which the signal significance is expected to be limited to a few bins with low expected background, these uncertainties can often be among the leading sources of systematic uncertainties.

5.3 Results of the search

For each final state, single NNs are trained based on one of the 68 mass groupings described above. This results in 68 different categorizations, which are however highly correlated as background-like events often end up in the same categories independent of the training, and the signal categories from which the sensitivity is derived are composed of only a small subset of the available events. This is shown in Figure 5.9.

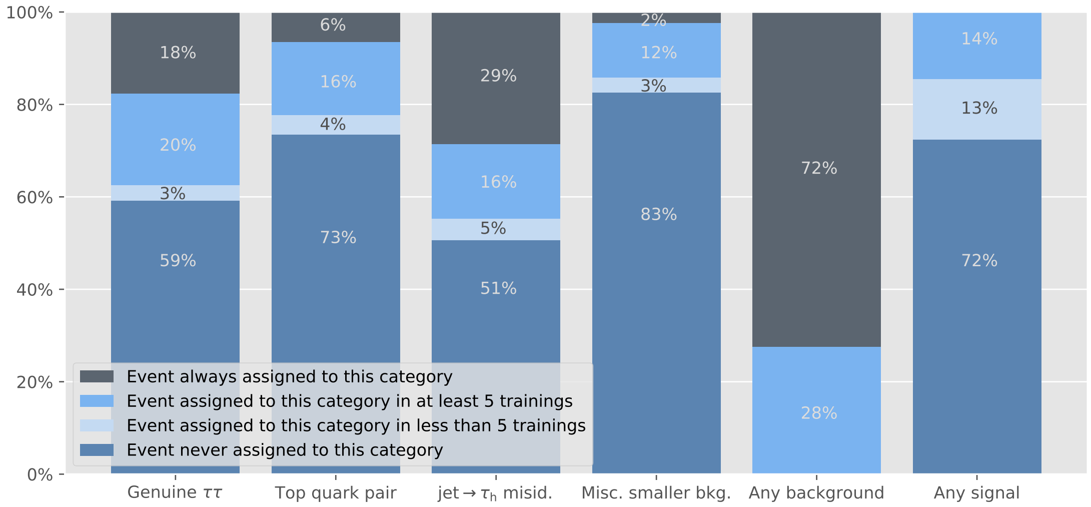


Figure 5.9: Consistency of the event classification across all 68 trainings used for the classification, depending on the masses of the bosons involved in the NMSSM signal process. The numbers are shown for the categorized data events in the $\tau_h\tau_h$ final state of the 2018 run period. The majority (72%) of events are always classified as background, independent of the target mass of the signal process used in the training. For the specific background categories, the effect is strongest for the jet→ τ_h category, in which 29% of events always end up in this category. As the signal mass hypotheses lead to very different event signatures, no event is assigned to the signal category for all trainings. The 28% (which is the rounded sum of 14% and 13% in the right-most bar of the figure) of events which are however classified as signal at least once usually end up in multiple signal categories. Due to this, the measurements derived for the 68 categorizations are highly correlated, as they use mostly the same data events.

The distributions of the NN output score in the four background categories of the three final states utilized for the analysis are shown in Figures 5.10-5.12, after the maximum-likelihood fit using the likelihood function discussed in Equation 5.10 is performed. In the background categories, excellent modeling of the observation is achieved using the simulation- or data-driven background estimation methods.

An example for the distribution of the data in the signal categories, where the NMSSM signal is expected to appear, is shown independently for the three final states in Figure 5.13 for the categorization given by a training optimized for masses of the additional Higgs bosons of $m(\mathrm{H}) \approx 500\,\mathrm{GeV}$, and $m(\mathrm{h_S}) \in [110, 160]\,\mathrm{GeV}$. Here and for all other tested mass hypotheses, no significant excess over the background model has been observed.

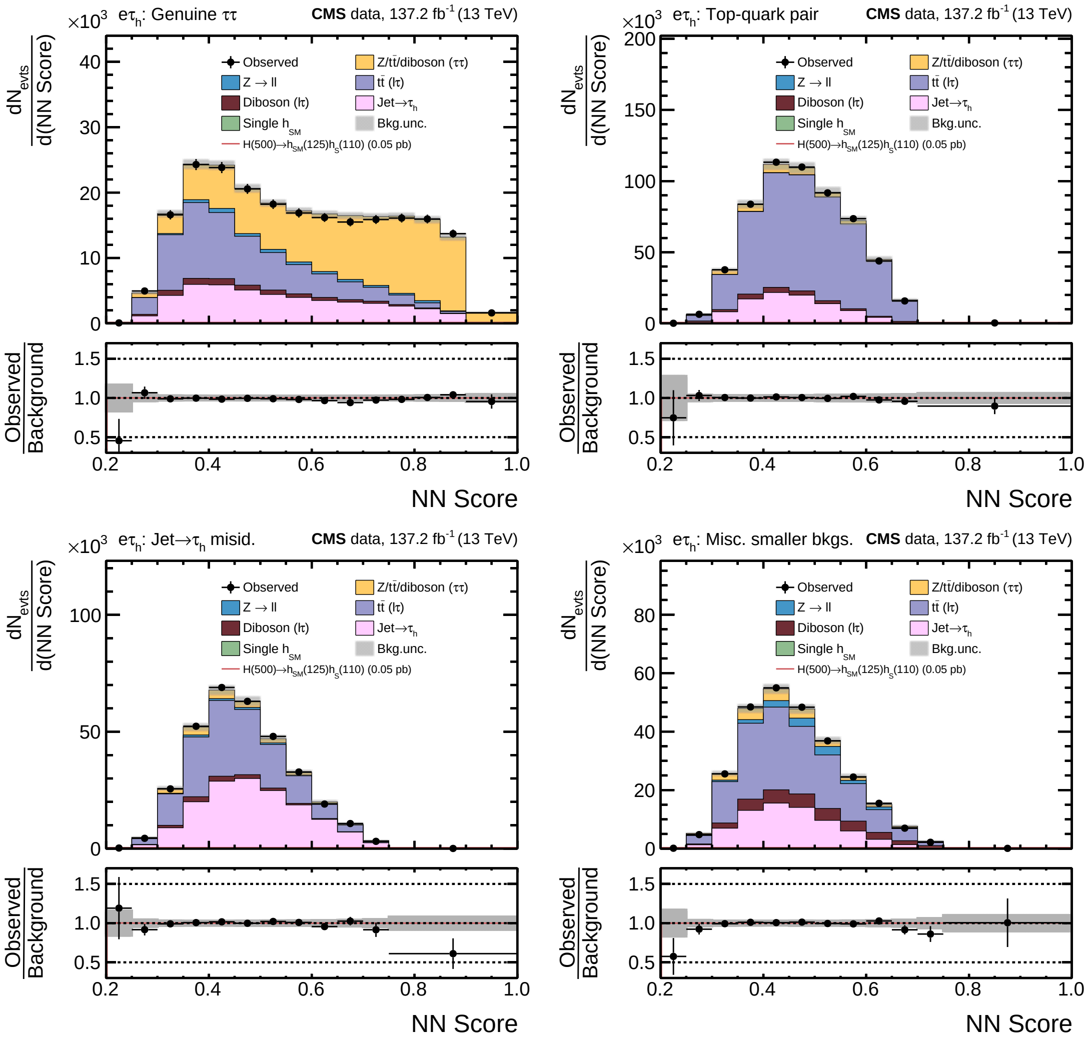


Figure 5.10: Background categories in the $e\tau_h$ final state after the fit to the data is performed: Genuine $\tau\tau$ category (top left), top-quark pairs (top right), jet → τ_h misidentified (bottom left), and miscellaneous smaller backgrounds (bottom right). The shown background categories given for a training in which $m(\mathrm{H}) = 500\,\mathrm{GeV}$ and $m(\mathrm{h_S}) \in [110, 160]\,\mathrm{GeV}$.

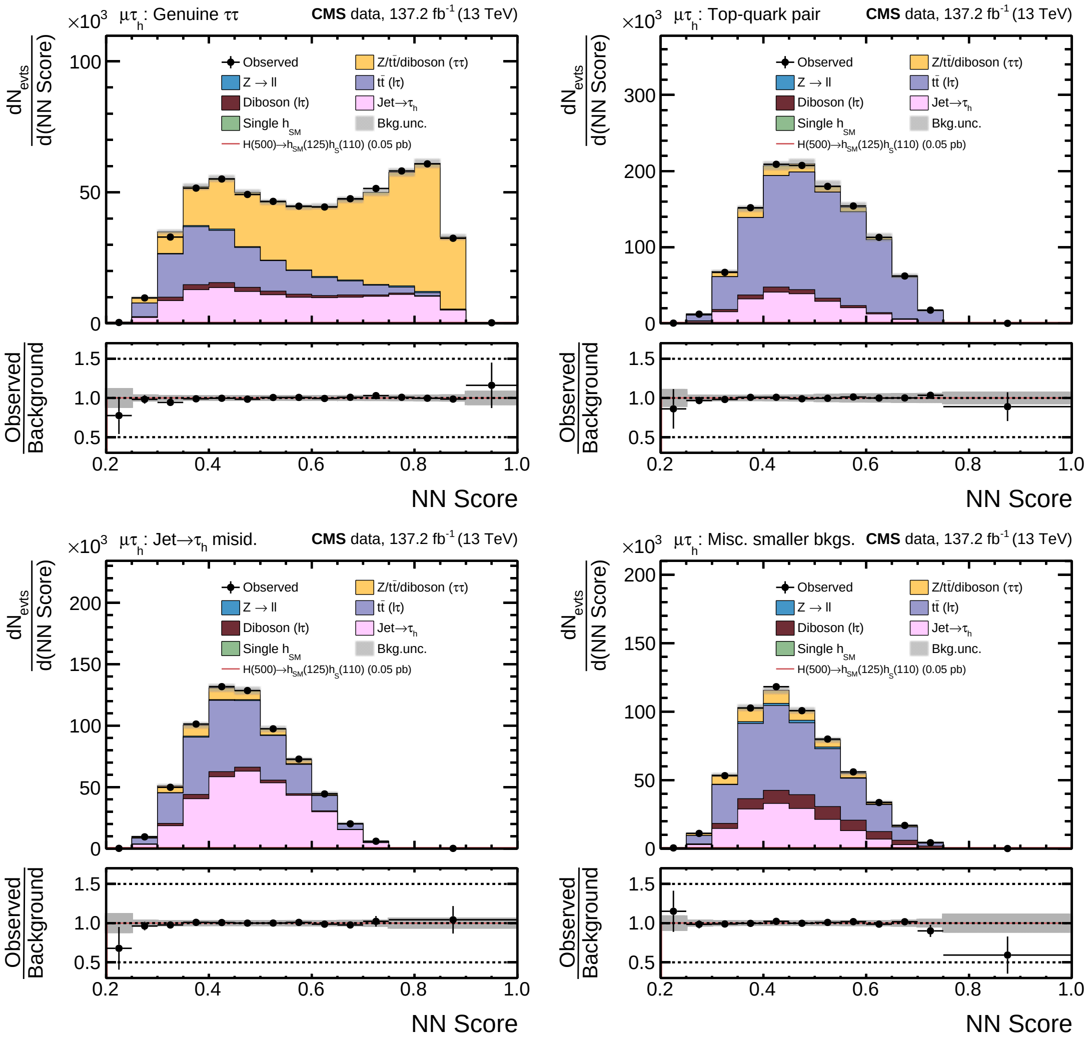


Figure 5.11: Background categories in the $\mu\tau_{\text{h}}$ final state after the fit to the data is performed: Genuine $\tau\tau$ category (top left), top-quark pairs (top right), jet $\rightarrow \tau_{\text{h}}$ misidentified (bottom left), and miscellaneous smaller backgrounds (bottom right). The shown background categories given for a training in which $m(\text{H}) = 500\,\text{GeV}$ and $m(\text{h}_{\text{S}}) \in [110, 160]\,\text{GeV}$.

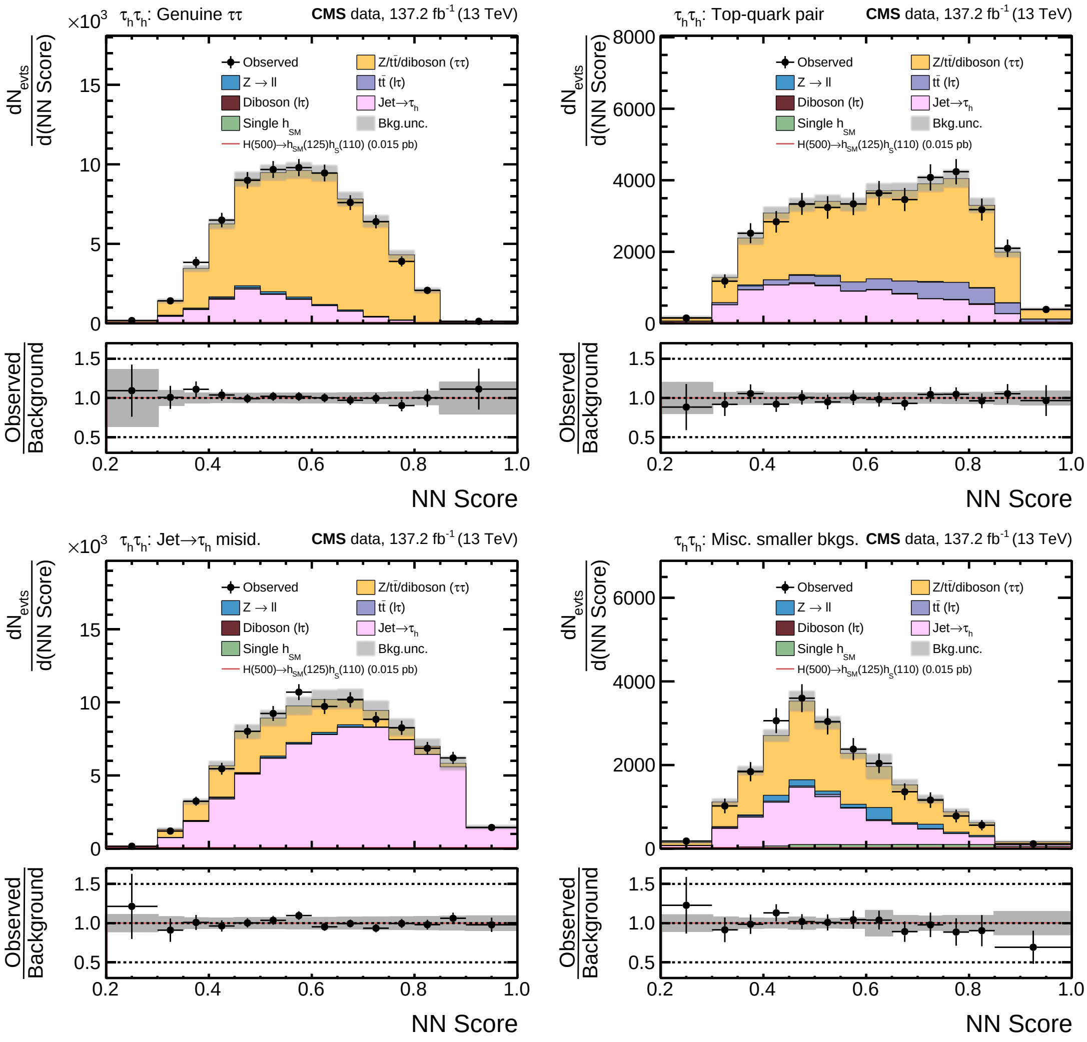


Figure 5.12: Background categories in the $\tau_h\tau_h$ final state after the fit to the data is performed: Genuine $\tau\tau$ category (top left), top-quark pairs (top right), jet → τ_h misidentified (bottom left), and miscellaneous smaller backgrounds (bottom right). The shown background categories given for a training in which $m(\mathrm{H}) = 500\,\mathrm{GeV}$ and $m(\mathrm{h_S}) \in [110, 160]\,\mathrm{GeV}$.

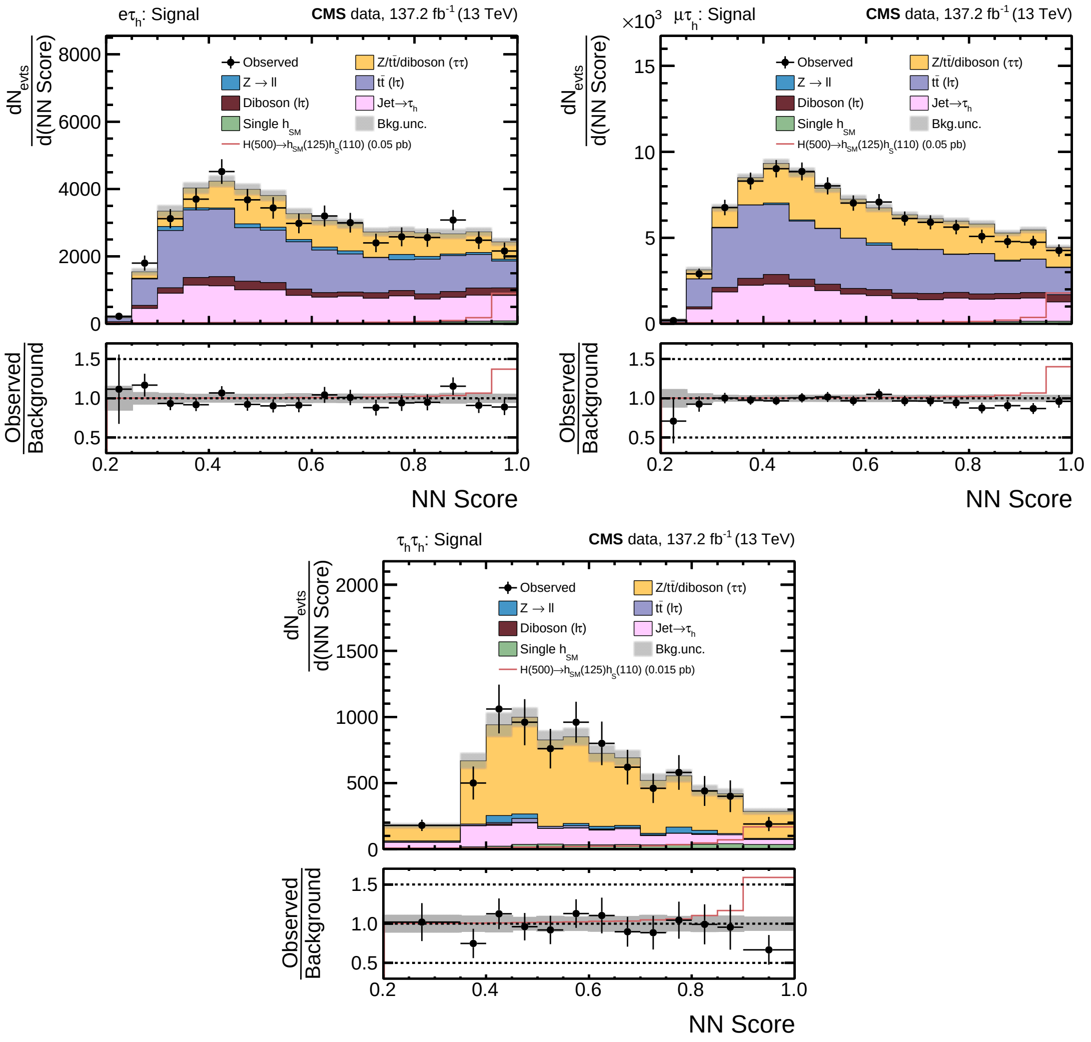


Figure 5.13: Signal categories in the $\mathrm{e}\tau_{\mathrm{h}}$ (top left), $\mu\tau_{\mathrm{h}}$ (top right) and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ (bottom) final states after the fit to the data is performed. The shown signal category is designed to enrich signal events with $m(\mathrm{H}) = 500\,\mathrm{GeV}$ and $m(\mathrm{h_S}) \in [110, 160]\,\mathrm{GeV}$ and is used to set the upper exclusion limits for these signal hypotheses. An exemplary signal mass hypothesis of $m(\mathrm{H}) = 500\,\mathrm{GeV}$ and $m(\mathrm{h_S}) = 110\,\mathrm{GeV}$ is shown by a red line with a cross section times branching fractions of $0.015 - 0.050\,\mathrm{pb}$ for illustration purposes. The best-fit of the signal cross section times branching fractions for this mass point is $-0.009 \pm 0.006\,\mathrm{pb}$ with an upper 95% confidence level limit of $0.005\,\mathrm{pb}$.

In the absence of an excess, results will be given as model-independent 95% confidence level upper limits on the cross section of the expected signal process, multiplied by its branching fraction into the studied final state involving b-quarks and tau leptons. For the computation of the upper limits, the $\mathrm{CL_s}$ method [82] is used. A $\mathrm{CL_s}$ value is calculated, defined as the ratio of the p-values of the signal hypothesis with signal strength modifier μ, p_μ, and the p-value of the null hypothesis p_0, given by the formula

$$\mathrm{CL_s} = \frac{p_\mu}{1-p_0} = \frac{\int_{q_\mathrm{obs}}^{\infty} f(q|\mu)\,dq}{\int_{q_\mathrm{obs}}^{\infty} f(q|0)\,dq} \leq \alpha = 0.05 \tag{5.13}$$

in which the test statistic q is defined as $q = -2\ln\mathcal{L}$ using the likelihood function defined in equation 5.10, and q_obs is the observed value of the test statistic in the data. The value α is defined such that $1-\alpha$ represents the confidence level of the exclusion, which is set to 95% for these results. The $\mathrm{CL_s}$ method has large advantages over using the p-value of the signal hypothesis p_μ alone to exclude the signal hypothesis, as it increases the effective p-value in cases where the two distributions become close and the analysis loses sensitivity to exclude the signal process. This can be the case in this analysis as the signal strength modifier μ can become arbitrarily small. Using only p_μ to exclude these cases in which the distributions are similar would result in excluding a given signal hypothesis with a probability of 5%, without having experimental sensitivity. A toy example of how a signal hypothesis excluded by the standard p-value approach is not necessarily excluded by using the $\mathrm{CL_s}$ method is shown using a graphical representation of the method in Figure 5.14.

To derive model-independent exclusion limits on the cross section times branching fraction, the signal strength modifier μ is used to vary the product of the cross section as well as the branching fractions $\mathcal{B}$ as

$$\mu \propto \sigma(\mathrm{gg} \rightarrow \mathrm{H}) \times \mathcal{B}(H \rightarrow \mathrm{h_{SM}}(\tau\tau)\mathrm{h_S}(\mathrm{bb})) \tag{5.14}$$

These limits can be interpreted within the context of the NMSSM, and compared to the not-yet excluded cross section times branching fractions in this model, but also within the context of other beyond-the-SM models predicting such a process. As the analysis is not sensitive to the CP phase of the heavy Higgs bosons, the exclusion limits also cover cases in which the two additional bosons are pseudoscalar, i.e. $\mathrm{A_2} \rightarrow \mathrm{h_{SM}A_1} \rightarrow \tau\tau\mathrm{bb}$.

The observed model-independent limits for all probed mass pairs are shown in Figure 5.15. The comparison of the observed with the expected limits is shown in Figure 5.16. The same points are shown in more detail in Figures A.1-A.5. A signal would manifest itself as a significant upward fluctuation of the observation. The strongest fluctuation of all 420 tested hypotheses is at the level of 2σ for the mass hypothesis $m(\mathrm{H}) = 1000\,\mathrm{GeV}$, $m(\mathrm{h}_S) = 350\,\mathrm{GeV}$. The strongest downward fluctuation is at the level of 2.5σ. The observation is thus statistically consistent with the background-only expectation.

The observed exclusion limits can be compared to the maximally allowed cross section times branching fractions of the process in the context of the NMSSM, taking into

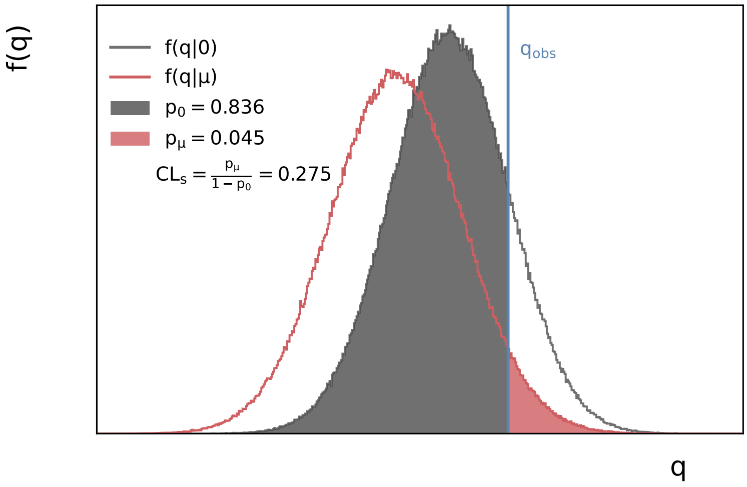


Figure 5.14: Toy example showcasing the $\mathrm{CL_s}$ method used to derive 95% confidence level upper limits on the signal process. The distribution of the test statistic q given by the background only hypothesis $f(q|0)$ is shown as black line, the distribution given by the signal hypothesis with signal modifier μ, $f(q|\mu)$, is shown as red line. The observed value of the test statistic q_{obs} is shown as a vertical blue line. The p-value p_{μ} is shown as a filled red area, divided by the integral of the full distribution. For this toy example the signal hypothesis of μ could be excluded with over 95% confidence if the standard p-value approach were to be used. Instead, this p-value is divided by $1 - p_0$ to derive a $\mathrm{CL_s}$ value of 0.275. In this toy example, the shown signal hypothesis will thus not be excluded anymore by 95% using the $\mathrm{CL_s}$ method.

account all experimental constraints of previous searches. These values are at the level of $10^{-3} - 10^{-2}$ pb depending on the involved masses [83] and are derived via `NMSSMTools 5.5.0` [84] and `NMSSMCALC` [85].

The comparison of the exclusion limits derived in this search to the maximally allowed cross section times branching fractions in the context of the NMSSM is shown in Figure 5.17 in the form of an exclusion contour in the $m(\mathrm{H}) - m(\mathrm{h_S})$−plane. The contour is shown for $m(\mathrm{H}) \in [400, 800]$ GeV. For lower masses of H, a doublet-like H is already largely excluded in the context of the NMSSM, and for higher masses, the cross section times branching fractions drops too low to be experimentally accessible using the available data.

As shown in the figure, the allowed cross section times branching fractions within the NMSSM can be constrained by this search for masses between 400 and 600 GeV. The strongest of such constraints is achieved for the mass hypothesis of $m(\mathrm{H}) = 450$ GeV and light states of $\mathrm{h_S}$ between 60 and 80 GeV, where a 95% confidence upper limit of the cross section times branching fractions of around 0.004 pb is observed, reducing the allowed value of 0.02 pb within the NMSSM by a factor of five.

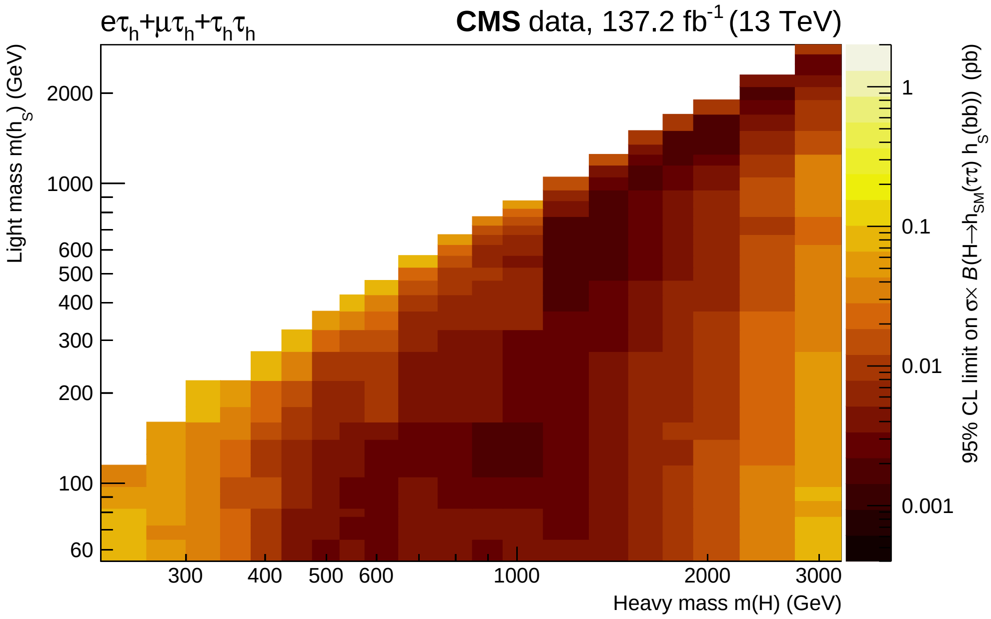


Figure 5.15: Observed 95% confidence level upper limits on $\sigma(\mathrm{gg} \to \mathrm{H}) \times \mathcal{B}(H \to \mathrm{h_{SM}}(\tau\tau)\mathrm{h_S}(\mathrm{bb}))$. The limits are shown in the $m(\mathrm{H}) - m(\mathrm{h_S})$−plane and encoded by a color scale, with darker colors indicated lower exclusion limits. The cross section times branching fractions of the process can be excluded down to $\approx 1\,\mathrm{fb}$.

Also shown in Figure 5.17 is an extrapolation of the sensitivity of the presented analysis. The extrapolation is calculated by estimating the influence of extending the analysis to the final state $\mathrm{h_{SM}} \to \mathrm{bb}$ and $\mathrm{h_S} \to \tau\tau$ as well as to a data set of $300\,\mathrm{fb}^{-1}$. This data set is expected to be available after the Run-III period of the LHC, to be concluded in the year 2024.

The results presented in this search can be compared to the previously published results of a similar search, optimized to the resonant production of two $\mathrm{h_{SM}}$ in the $\mathrm{bb}+\tau\tau$ final state using data of the 2016 run period [86]. The comparison is shown in Figure 5.18. If restricted to the expected limits only as well as to the same data set, and thus only updating the experimental methods such as the improved detection of b-jets and τ_h's as well as the categorization using NNs, this search provides improved exclusion limits by a factor of up to three, without being explicitly tuned to the case in which the mass of $\mathrm{h_S}$ is also 125 GeV.

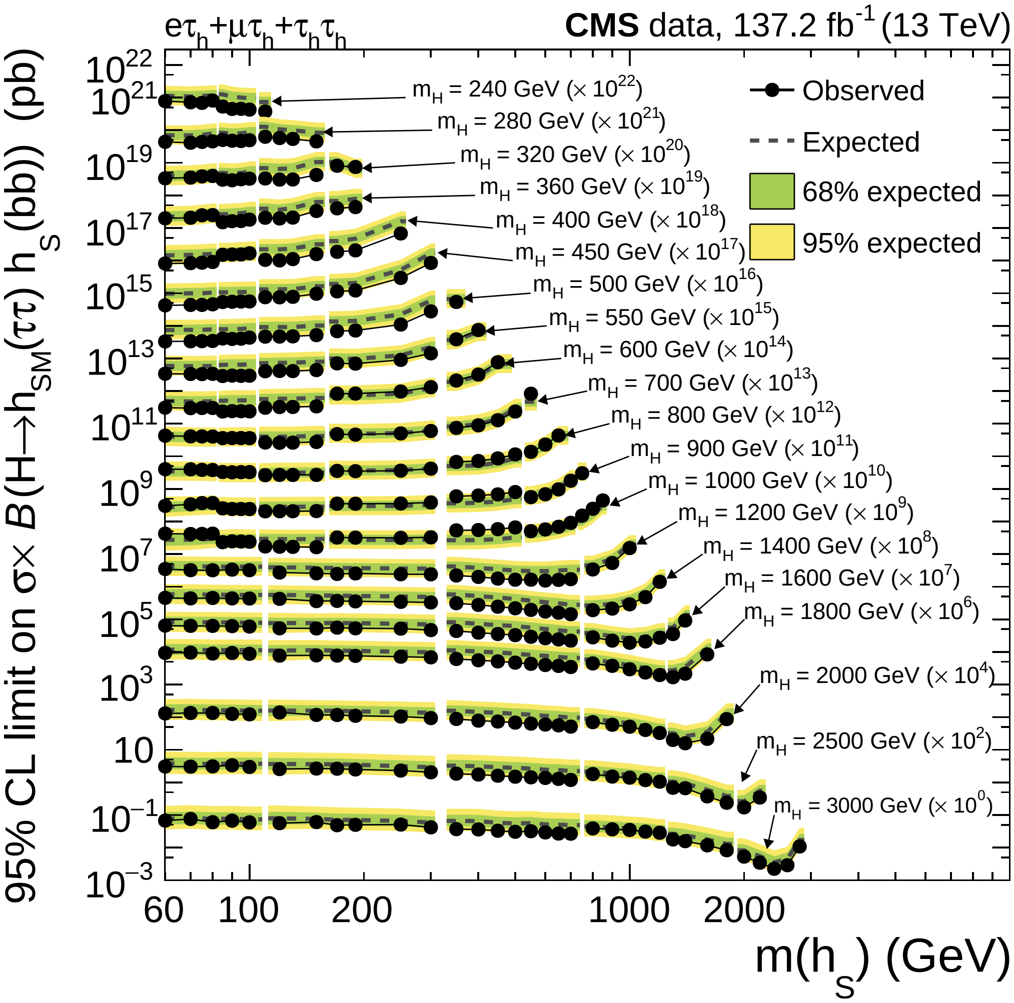


Figure 5.16: Observed and expected 95% confidence level upper limits on $\sigma(\mathrm{gg} \to \mathrm{H}) \times \mathcal{B}(H \to \mathrm{h}_{\mathrm{SM}}(\tau\tau)\mathrm{h}_{\mathrm{S}}(\mathrm{bb}))$. A scaling in orders of ten is used as indicated in the annotations to display the results in a common figure. The expected limits are shown as a dashed line with the 68% and 95% confidence interval of the expectation given by the green and yellow bands. The observation is shown by black points. The correlation of similar probed mass pairs due to the grouping of masses used for the NN training is indicated by the interruptions between the lines. No deviation beyond the 2.5σ level is found for all 420 probed mass pairs.

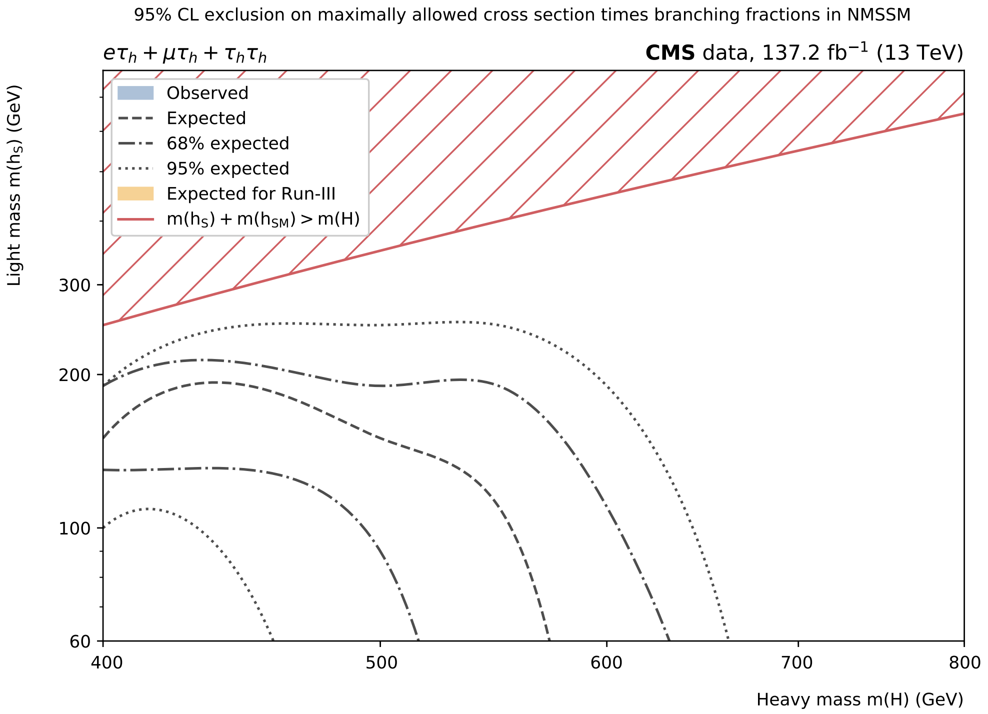


Figure 5.17: Comparison of the 95% confidence level upper limits on the cross section times branching fractions of the NMSSM signal process to the maximally allowed value within the theoretical framework of the NMSSM. The region in which the allowed cross section times branching fractions is expected to be constrained in this search is indicated by the dashed line, with the 68% and 95% uncertainty bands of the expectation given by a dash-dotted and dotted line respectively. The blue region highlights the observed region which is constrained. The yellow region refers to an extrapolation of the analysis to a total of 300 fb^{-1} of data expected to be available after the LHC Run-III as well as the extension of the analysis to the case $\mathrm{h_{SM}} \rightarrow \mathrm{bb}$ and $\mathrm{h_S} \rightarrow \tau\tau$.

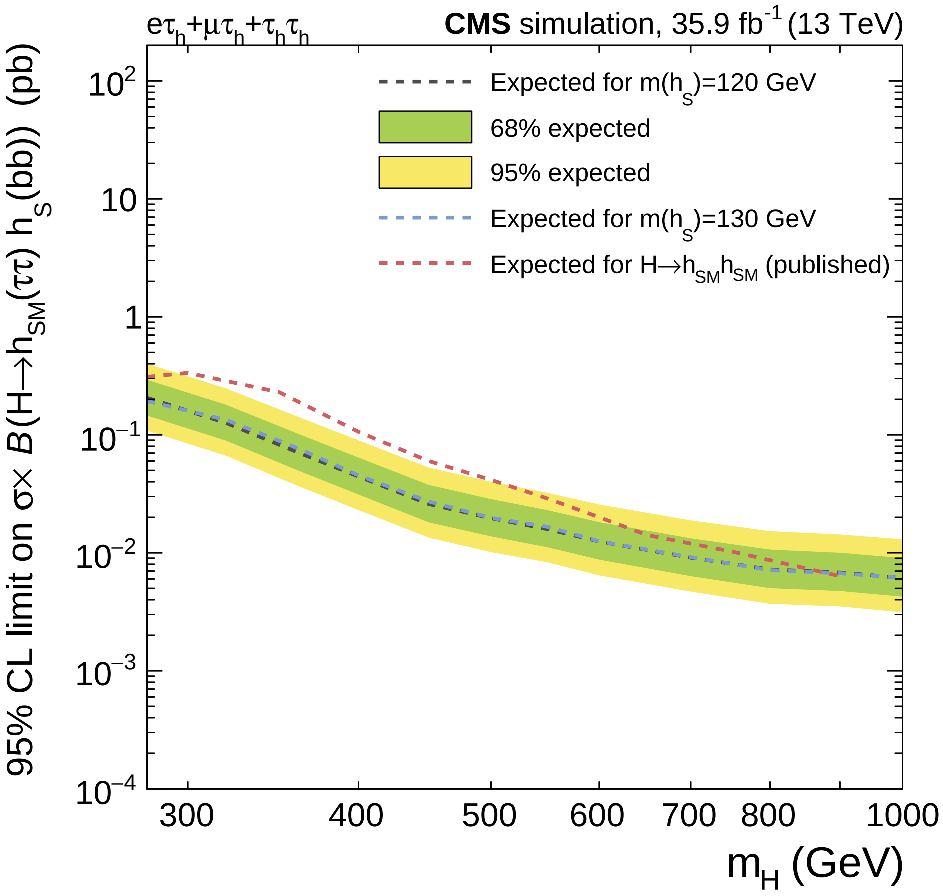


Figure 5.18: Comparison of the expected exclusion limits of this analysis to the analysis published in 2018 by the CMS collaboration searching for $\mathrm{H} \to \mathrm{h}_{\mathrm{SM}}\mathrm{h}_{\mathrm{SM}}$ [86]. For the comparison, the analysis is restricted to the same data set that has been used for the published analysis, which is the $35.9\,\mathrm{fb}^{-1}$ of data collected during the 2016 run period. As the point of $m(\mathrm{h}_{\mathrm{S}}) = m(\mathrm{h}_{\mathrm{SM}}) = 125\,\mathrm{GeV}$ is not explicitly modeled in this analysis, the comparison is given with respect to the two mass points at $m(\mathrm{h}_{\mathrm{S}}) = 120\,\mathrm{GeV}$ or $130\,\mathrm{GeV}$ respectively. The excluded values of the cross section times branching fractions in this analysis are lower than for the previous publication by a factor of up to three, without being explicitly tuned for the case in which $m(\mathrm{h}_{\mathrm{S}}) = 125\,\mathrm{GeV}$. This improvement is enabled by experimental developments such as the improved detection of b-jets or τ_{h}'s in the events.

CHAPTER 6

Conclusions

The goal of this thesis is the development of an analysis to search for decays of a heavy Higgs boson to the discovered Higgs boson and an additional light Higgs boson. The analysis is interpreted in the context of the NMSSM in which the heavy boson is expected to be doublet-like, and the light boson to be singlet-like. With the analysis, an important part of probing supersymmetric extensions of the SM in future run periods of the LHC is introduced. As the search has not been done before, many cornerstones of the analysis are built from scratch in the context of this thesis. This involves the simulation of the signal process in the context of the NMSSM, which was performed over a span of close to one year, requiring over two million hours of processing time.

The challenge of the analysis furthermore lies in the accurate description of the many different background processes which can result in event signatures similar to the signal process. Using data-driven estimation methods to a large extend proved to provide a very robust way of modeling the background, while reducing systematic uncertainties. The use of τ-embedded events for the estimation of decays of the Z boson or top-quark pairs allowed to model the additional b-jets in the event from data, turning many corrections and systematic uncertainties obsolete. The background from events in which jets are misidentified as τ_{h}, which poses the largest source of background in the most sensitive $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state, was estimated specifically for the phase space of this analysis using the F_{F} method, utilizing the fact that events in which no b-jet is identified can be treated as sideband region for the analysis.

Whether unconstrained regions of the NMSSM phase space would be reachable at all with LHC Run-II data in the probed $\tau\tau + \mathrm{bb}$ final state was yet to be seen before the analysis was performed. Even though no signal was observed, the fact that additional exclusion on the Higgs sector of the NMSSM is set for a large range of masses, spanning from $m(\mathrm{H}) \in [400, 600]\,\mathrm{GeV}$ and $m(\mathrm{h_S}) \in [60, 250]\,\mathrm{GeV}$ made this parameter space of the NMSSM experimentally available for the first time. The maximally allowed cross section multiplied by the branching fractions of the $\mathrm{H} \to \mathrm{h_{SM}}(\tau\tau)\mathrm{h_S}(\mathrm{bb})$ process is lowered by a factor of up to five by the results of this thesis.

The improvements that could be achieved with the modern experimental techniques that are utilized for this search becomes apparent in comparison with previously published searches of the H → $h_{SM}h_{SM}$ process in the same final state [86]. The expected results to find such a signature in the $\tau\tau$ + bb final state improved for all mass points by a factor of up to three when using the same data set. These improvements are achieved by using modern identification algorithms such as `DeepTau` and `DeepJet`, which make use of deep learning techniques to discriminate hadronic tau lepton decays or genuine b-jets from light quark or gluon induced jets.

Furthermore, improvements are achieved using machine learning to derive the final discriminator in the analysis instead of manual feature selection to separate signal events from background. Especially for the H → $h_{SM}(\tau\tau)h_{S}$(bb) process, information regarding the signal-like nature of an event is contained in multiple event quantities, such that correlations between these event quantities are often utilized to a higher degree than the marginal values themselves.

The analysis can be used as baseline for similar analyses for future run periods of the LHC, with potential improvements and extensions given in the following. A technical development with the potential to benefit the analysis is the research of the machine learning methods used to discriminate the events in cases where many different signal masses are possible. In this thesis, training a large number of individual neural networks for different mass groups proves to be close to optimal in terms of the resulting sensitivity of the analysis, however this procedure comes at the cost of a significant computing effort.

An obvious physical extension is the simulation of NMSSM signal events in the decay channel H → h_{SM}(bb)$h_{S}(\tau\tau)$. This simulation and subsequent analysis would boost the sensitivity of the analysis by a factor of two.

The methodology and results presented in this thesis have been approved for publication by the CMS collaboration and will published in the name of the CMS Collaboration in the future.

CHAPTER 7

Bibliography

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

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

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

[4] ATLAS Collaboration. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. *Physics Letters B* 716.1 (2012), pp. 1–29. ISSN: 0370-2693.
DOI: https://doi.org/10.1016/j.physletb.2012.08.020.

[5] CMS Collaboration. “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693.
DOI: https://doi.org/10.1016/j.physletb.2012.08.021.

[6] *Standard Model of Elementary Particles.* https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg. Accessed 22.10.2020.

[7] Sheldon L. Glashow. “Partial-symmetries of weak interactions”. *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582.
DOI: https://doi.org/10.1016/0029-5582(61)90469-2.

[8] Steven Weinberg. “A Model of Leptons”. *Phys. Rev. Lett.* 19 (21 Nov. 1967), pp. 1264–1266.
DOI: 10.1103/PhysRevLett.19.1264.

[9] Abdus Salam. “Weak and Electromagnetic Interactions”. *Conf. Proc. C* 680519 (1968), pp. 367–377.
DOI: 10.1142/9789812795915_0034.

[10] John Ellis. “Higgs Physics”. KCL-PH-TH-2013-49. KCL-PH-TH-2013-49. LCTS-2013-36. CERN-PH-TH-2013-315 (Dec. 2013). 52 pages, 45 figures, Lectures presented at the ESHEP 2013 School of High-Energy Physics, to appear as part of the proceedings in a CERN Yellow Report, 117–168. 52 p.
DOI: `10.5170/CERN-2015-004.117`.

[11] Steven Weinberg. “A Model of Leptons”. *Phys. Rev. Lett.* 19 (1967), pp. 1264–1266.
DOI: `10.1103/PhysRevLett.19.1264`.

[12] P. van Nieuwenhuizen. “Supergravity”. *Physics Reports* 68.4 (1981), pp. 189–398.
ISSN: 0370-1573.
DOI: `https://doi.org/10.1016/0370-1573(81)90157-5`.

[13] Edward Witten. “Dynamical breaking of supersymmetry”. *Nuclear Physics B* 188.3 (1981), pp. 513–554. ISSN: 0550-3213.
DOI: `https://doi.org/10.1016/0550-3213(81)90006-7`.

[14] Savas Dimopoulos and Howard Georgi. “Softly broken supersymmetry and SU(5)”. *Nuclear Physics B* 193.1 (1981), pp. 150–162. ISSN: 0550-3213.
DOI: `https://doi.org/10.1016/0550-3213(81)90522-8`.

[15] John Ellis, S. Kelley, and D.V. Nanopoulos. “Probing the desert using gauge coupling unification”. *Physics Letters B* 260.1 (1991), pp. 131–137. ISSN: 0370-2693.
DOI: `https://doi.org/10.1016/0370-2693(91)90980-5`.

[16] Ugo Amaldi, Wim de Boer, and Hermann Furstenau. “Comparison of grand unified theories with electroweak and strong coupling constants measured at LEP”. *Phys. Lett. B* 260 (1991), pp. 447–455.
DOI: `10.1016/0370-2693(91)91641-8`.

[17] Heinz Pagels and Joel R. Primack. “Supersymmetry, Cosmology, and New Physics at Teraelectronvolt Energies”. *Phys. Rev. Lett.* 48 (4 Jan. 1982), pp. 223–226.
DOI: `10.1103/PhysRevLett.48.223`.

[18] D CHUNG et al. “The soft supersymmetry-breaking Lagrangian: theory and applications”. *Physics Reports* 407.1-3 (Feb. 2005), pp. 1–203. ISSN: 0370-1573.
DOI: `10.1016/j.physrep.2004.08.032`.

[19] J.A. Casas and C. Muñoz. “A natural solution to the μ problem”. *Physics Letters B* 306.3 (1993), pp. 288–294. ISSN: 0370-2693.
DOI: `https://doi.org/10.1016/0370-2693(93)90081-R`.

[20] Ulrich Ellwanger, Cyril Hugonie, and Ana M. Teixeira. “The Next-to-Minimal Supersymmetric Standard Model”. *Physics Reports* 496.1-2 (Nov. 2010), pp. 1–77.
ISSN: 0370-1573.
DOI: `10.1016/j.physrep.2010.07.001`.

[21] M. Maniatis. “The next-to-minimal supersymmetric extension of the Standard Model reviewed”. *International Journal of Modern Physics A* 25.18n19 (July 2010), pp. 3505–3602. ISSN: 1793-656X.
DOI: `10.1142/s0217751x10049827`.

[22] S. F. King et al. “Discovery prospects for NMSSM Higgs bosons at the high-energy Large Hadron Collider”. *Physical Review D* 90.9 (Nov. 2014). ISSN: 1550-2368.
DOI: 10.1103/physrevd.90.095014.

[23] CMS Collaboration. “The CMS experiment at the CERN LHC”. *Journal of Instrumentation* 3.08 (Aug. 2008), S08004–S08004.
DOI: 10.1088/1748-0221/3/08/s08004.

[24] Tai Sakuma. “Cutaway diagrams of CMS detector” (May 2019).

[25] Moon Meok Kim et al. “Web-based monitoring tools for Resistive Plate Chambers in the CMS experiment at CERN”. 9 (Oct. 2014).

[26] *CMS Tracker Detector Performance Public Results*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/DPGResultsTRK (accessed 14.10.2020).

[27] CMS Collaboration. “CMS Technical Design Report for the Pixel Detector Upgrade” (Sept. 2012).
DOI: 10.2172/1151650.

[28] CMS Collaboration. “Description and performance of track and primary-vertex reconstruction with the CMS tracker”. *Journal of Instrumentation* 9.10 (Oct. 2014), P10009–P10009.
DOI: 10.1088/1748-0221/9/10/p10009.

[29] CMS Collaboration. “The Phase-2 Upgrade of the CMS Tracker” (June 2017). Ed. by K. Klein.

[30] CMS Collaboration. “The CMS electromagnetic calorimeter project: Technical Design Report”. Technical Design Report CMS. Geneva: CERN, 1997.

[31] CMS Collaboration. “The CMS ECAL performance with examples”. Tech. rep. CMS-CR-2013-430. Geneva: CERN, Nov. 2013.
DOI: 10.1088/1748-0221/9/02/C02008.

[32] CMS Collaboration. “The CMS hadron calorimeter project: Technical Design Report”. Technical Design Report CMS. Geneva: CERN, 1997.

[33] CMS Collaboration. “Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays”. *Journal of Instrumentation* 5.03 (Mar. 2010), T03021–T03021.
DOI: 10.1088/1748-0221/5/03/t03021.

[34] CMS Collaboration. “The CMS muon project: Technical Design Report”. Technical Design Report CMS. Geneva: CERN, 1997.

[35] CMS Collaboration. “CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. CMS trigger and data-acquisition project”. Technical Design Report CMS. Geneva: CERN, 2002.

[36] CMS Collaboration. “Description and performance of track and primary-vertex reconstruction with the CMS tracker”. *JINST* 9.10 (2014), P10009.
DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569 [physics.ins-det].

[37] K. Rose. “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems”. *Proceedings of the IEEE* 86.11 (1998), pp. 2210–2239.

[38] CMS Collaboration. “Particle-flow reconstruction and global event description with the CMS detector”. *Journal of Instrumentation* 12.10 (Oct. 2017), P10003–P10003. ISSN: 1748-0221.
DOI: 10.1088/1748-0221/12/10/p10003.

[39] CMS Collaboration. “Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC” (2020). arXiv: 2012.06888 [hep-ex].

[40] CMS Collaboration. “Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{s} = 8$ TeV”. *JINST* 10.06 (2015), P06005.
DOI: 10.1088/1748-0221/10/06/P06005. arXiv: 1502.02701 [physics.ins-det].

[41] CMS Collaboration. “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$=13 TeV”. *Journal of Instrumentation* 13.06 (June 2018), P06015–P06015. ISSN: 1748-0221.
DOI: 10.1088/1748-0221/13/06/p06015.

[42] Daniele Bertolini et al. “Pileup Per Particle Identification”. *JHEP* 10 (2014), p. 059.
DOI: 10.1007/JHEP10(2014)059. arXiv: 1407.6013 [hep-ph].

[43] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. “The anti-ktjet clustering algorithm”. *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 063–063.
DOI: 10.1088/1126-6708/2008/04/063.

[44] CMS Collaboration. “Jet algorithms performance in 13 TeV data”. Tech. rep. CMS-PAS-JME-16-003. Geneva: CERN, 2017.

[45] CMS Collaboration. “Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV”. *Journal of Instrumentation* 12.02 (Feb. 2017), P02014–P02014.
DOI: 10.1088/1748-0221/12/02/p02014.

[46] CMS Collaboration. “Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s}$=13 TeV”. *Journal of Instrumentation* 13.10 (Oct. 2018), P10005–P10005.
DOI: 10.1088/1748-0221/13/10/p10005.

[47] Konstantin Androsov. “Identification of tau leptons using Deep Learning techniques at CMS”. Tech. rep. CMS-CR-2019-272. Geneva: CERN, Nov. 2019.

[48] CMS Collaboration. “Measurement of $B\bar{B}$ angular correlations based on secondary vertex reconstruction at $\sqrt{s} = 7$ TeV”. *Journal of High Energy Physics* 2011 (Mar. 2011), p. 136.
DOI: 10.1007/JHEP03(2011)136.

[49] CMS Collaboration. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". *Journal of Instrumentation* 13.05 (May 2018), P05011–P05011.
DOI: 10.1088/1748-0221/13/05/p05011.

[50] CMS Collaboration. "Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13TeV with Phase 1 CMS detector" (Nov. 2018).

[51] CMS Collaboration. "Evidence for the Higgs boson decay to a bottom quark-antiquark pair". *Physics Letters B* 780 (2018), pp. 501–532. ISSN: 0370-2693.
DOI: https://doi.org/10.1016/j.physletb.2018.02.050.

[52] M. Tanabashi et al. "Review of Particle Physics". *Phys. Rev. D* 98 (3 Aug. 2018), p. 030001.
DOI: 10.1103/PhysRevD.98.030001.

[53] CMS Collaboration. "An embedding technique to determine $\tau\tau$ backgrounds in proton-proton collision data". *Journal of Instrumentation* 14.06 (2019), P06032–P06032.
DOI: 10.1088/1748-0221/14/06/p06032.

[54] CMS Collaboration. "Measurements of inclusive W and Z cross sections in pp collisions at $\sqrt{s} = 7$ TeV". *Journal of High Energy Physics* 2011.1 (Jan. 2011). ISSN: 1029-8479.
DOI: 10.1007/jhep01(2011)080.

[55] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". *Computer Physics Communications* 191 (2015), pp. 159–177. ISSN: 0010-4655.
DOI: https://doi.org/10.1016/j.cpc.2015.01.024.

[56] CMS Collaboration. "Measurement of the $Z\gamma^* \to \tau\tau$ cross section in pp collisions at $\sqrt{s} = 13$ TeV and validation of τ lepton analysis techniques". *Eur. Phys. J. C* 78.9 (2018), p. 708.
DOI: 10.1140/epjc/s10052-018-6146-9. arXiv: 1801.03535 [hep-ex].

[57] CMS Collaboration. "Measurement of Higgs boson production and decay to the $\tau\tau$ final state". Tech. rep. CMS-PAS-HIG-18-032. Geneva: CERN, 2019.

[58] CMS Collaboration. "Measurement of Higgs boson production in the decay channel with a pair of τ leptons" (Aug. 2020).

[59] CMS Collaboration. "Search for additional neutral MSSM Higgs bosons in the $\tau\tau$ final state in proton-proton collisions at $\sqrt{s} = 13$ TeV". *JHEP* 09.arXiv:1803.06553. CMS-HIG-17-020-003 (Mar. 2018), 007. 56 p.
DOI: 10.1007/JHEP09(2018)007.

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

[61] Simone Alioli et al. “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX”. *JHEP* 06 (2010), p. 043. DOI: `10.1007/JHEP06(2010)043`. arXiv: `1002.2581 [hep-ph]`.

[62] Richard D. Ball et al. “Parton distributions for the LHC Run II”. *JHEP* 04 (2015), p. 040. DOI: `10.1007/JHEP04(2015)040`. arXiv: `1410.8849 [hep-ph]`.

[63] Richard D. Ball et al. “Parton distributions from high-precision collider data - NNPDF Collaboration”. *Eur. Phys. J. C* 77.10 (2017), p. 663. DOI: `10.1140/epjc/s10052-017-5199-5`.

[64] S. Agostinelli et al. “Geant4—a simulation toolkit”. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: `https://doi.org/10.1016/S0168-9002(03)01368-8`.

[65] Eric Conte et al. “Investigating light NMSSM pseudoscalar states with boosted ditau tagging”. *JHEP* 05 (2016), p. 100. DOI: `10.1007/JHEP05(2016)100`. arXiv: `1604.05394 [hep-ph]`.

[66] *BwForCluster NEMO*. `https://wiki.bwhpc.de/e/Category:BwForCluster_NEMO` (site accessed 29.09.2020).

[67] *ForHLR II*. `https://www.scc.kit.edu/dienste/forhlr.php` (site accessed 29.09.2020).

[68] 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”. *Phys. Rev. D* 95 (9 May 2017), p. 092001. DOI: `10.1103/PhysRevD.95.092001`.

[69] CMS Collaboration. “Measurements of $t\bar{t}$ differential cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV using events containing two leptons”. *JHEP* 02.arXiv:1811.06625. CMS-TOP-17-014-003 (Nov. 2018), 149. 103 p. DOI: `10.1007/JHEP02(2019)149`.

[70] CMS Collaboration. “CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV”. Tech. rep. CMS-PAS-LUM-18-002. Geneva: CERN, 2019.

[71] Malte Hoffmann et al. “HHKinFit - a kinematic fitting package to fit heavy Higgs decays” (July 2014). CMS-AN-2014-163.

[72] 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→ $\tau\tau$”. *Physics Letters B* 755 (2016), pp. 217–244. ISSN: 0370-2693. DOI: `https://doi.org/10.1016/j.physletb.2016.01.056`.

[73] Lorenzo Bianchini et al. "Reconstruction of the Higgs mass in $H \to \tau\tau$ Events by Dynamical Likelihood techniques". *J. Phys. Conf. Ser.* 513 (2014). Ed. by D. L. Groep and D. Bonacorsi, p. 022035.
DOI: 10.1088/1742-6596/513/2/022035.

[74] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". *CoRR* abs/1412.6980 (2015).

[75] Y. Bengio and X. Glorot. "Understanding the difficulty of training deep feed forward neural networks". *International Conference on Artificial Intelligence and Statistics* (Jan. 2010), pp. 249–256.

[76] S. Jörger. "Summary of NMSSM benchmark lines". https://publish.etp.kit.edu/record/21950/. MA thesis. 2020.

[77] Stefan Wunsch et al. "Identifying the Relevant Dependencies of the Neural Network Response on Characteristics of the Input Space". *Computing and Software for Big Science* 2 (Mar. 2018).
DOI: 10.1007/s41781-018-0012-1.

[78] Robert D. Cousins. "Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models, with Application to Histograms". http://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf. 2013.

[79] CMS Collaboration. "CMS Luminosity Measurements for the 2016 Data Taking Period". Tech. rep. CMS-PAS-LUM-17-001. Geneva: CERN, 2017.

[80] CMS Collaboration. "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV". Tech. rep. CMS-PAS-LUM-17-004. Geneva: CERN, 2018.

[81] Roger Barlow and Christine Beeston. "Fitting using finite Monte Carlo samples". *Computer Physics Communications* 77.2 (1993), pp. 219–228. ISSN: 0010-4655.
DOI: https://doi.org/10.1016/0010-4655(93)90005-W.

[82] A L Read. "Presentation of search results: The CLs technique". *Journal of Physics G: Nuclear and Particle Physics* 28.10 (Sept. 2002), pp. 2693–2704.
DOI: 10.1088/0954-3899/28/10/313.

[83] NMSSM subgroup of the LHC Higgs cross section working group. *Studies of the usage of neural networks in particle physics analyses.* https://twiki.cern.ch/twiki/bin/view/LHCPhysics/NMSSMBenchmarksMarch2020. 2020.

[84] Ulrich Ellwanger and Cyril Hugonie. "NMHDECAY 2.1: An updated program for sparticle masses, Higgs masses, couplings and decay widths in the NMSSM". *Computer Physics Communications* 175.4 (Aug. 2006), pp. 290–303. ISSN: 0010-4655.
DOI: 10.1016/j.cpc.2006.04.004.

[85] J. Baglio et al. "NMSSMCALC: A program package for the calculation of loop-corrected Higgs boson masses and decay widths in the (complex) NMSSM". *Computer Physics Communications* 185.12 (Dec. 2014), pp. 3372–3391. ISSN: 0010-4655.
DOI: 10.1016/j.cpc.2014.08.005.

[86] CMS Collaboration. “Search for Higgs boson pair production in events with two bottom quarks and two tau leptons in proton–proton collisions at s=13TeV”. *Physics Letters B* 778 (Mar. 2018), pp. 101–127. ISSN: 0370-2693.
DOI: `10.1016/j.physletb.2018.01.001`.

APPENDIX A

Appendix

Results of the search for all probed mass points

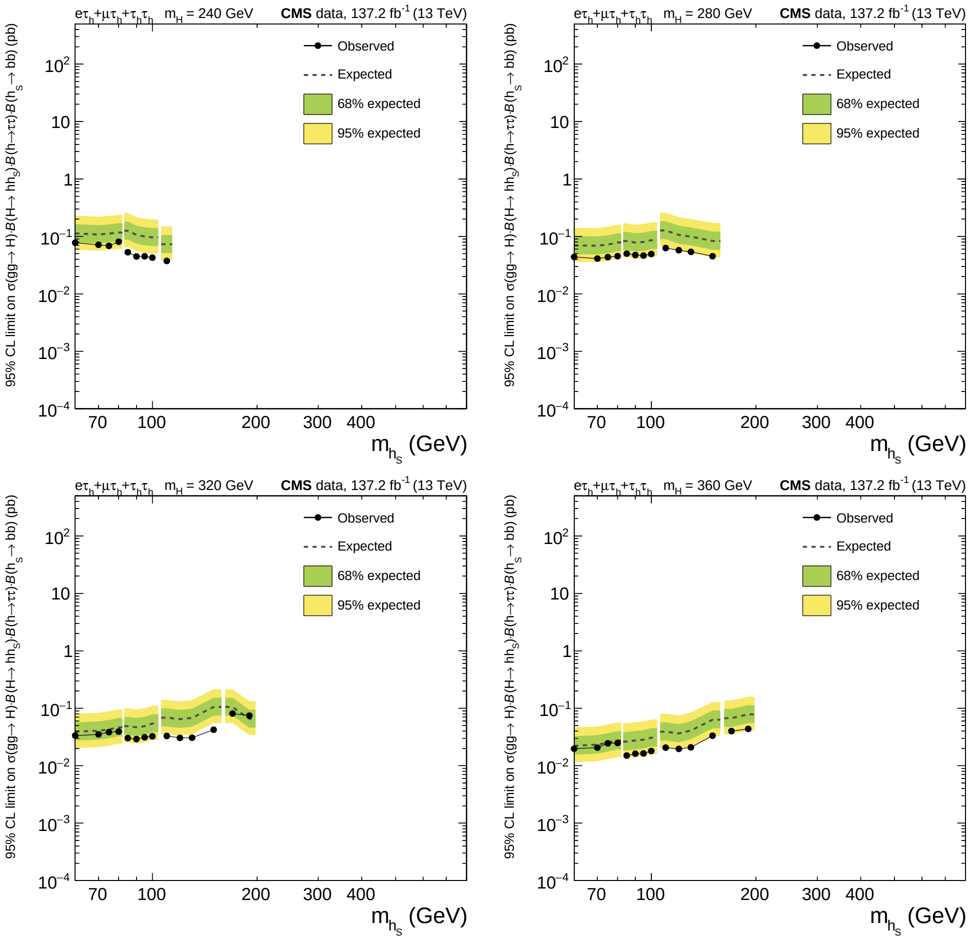


Figure A.1: Expected and observed 95% confidence level upper limits on the NMSSM signal process.

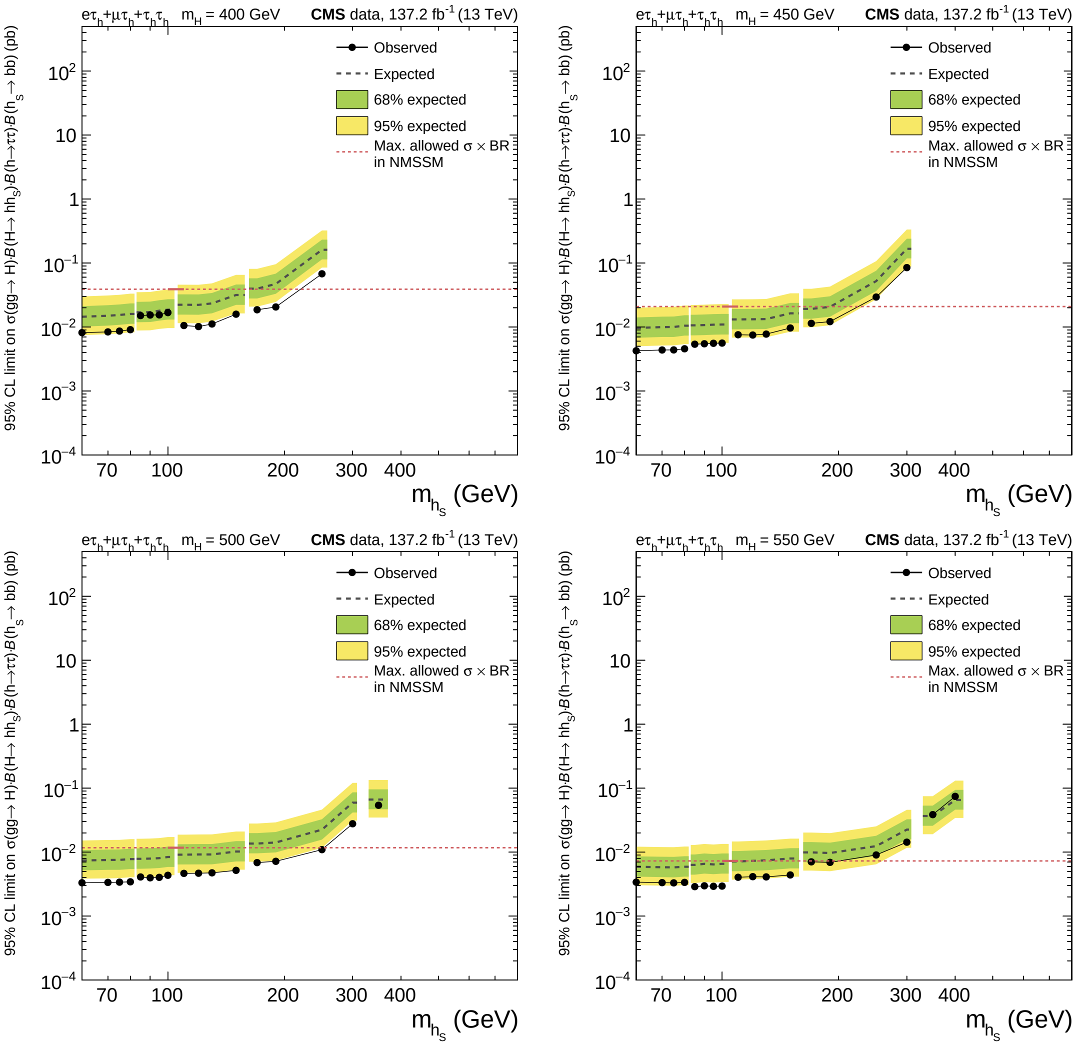


Figure A.2: Expected and observed 95% confidence level upper limits on the NMSSM signal process.

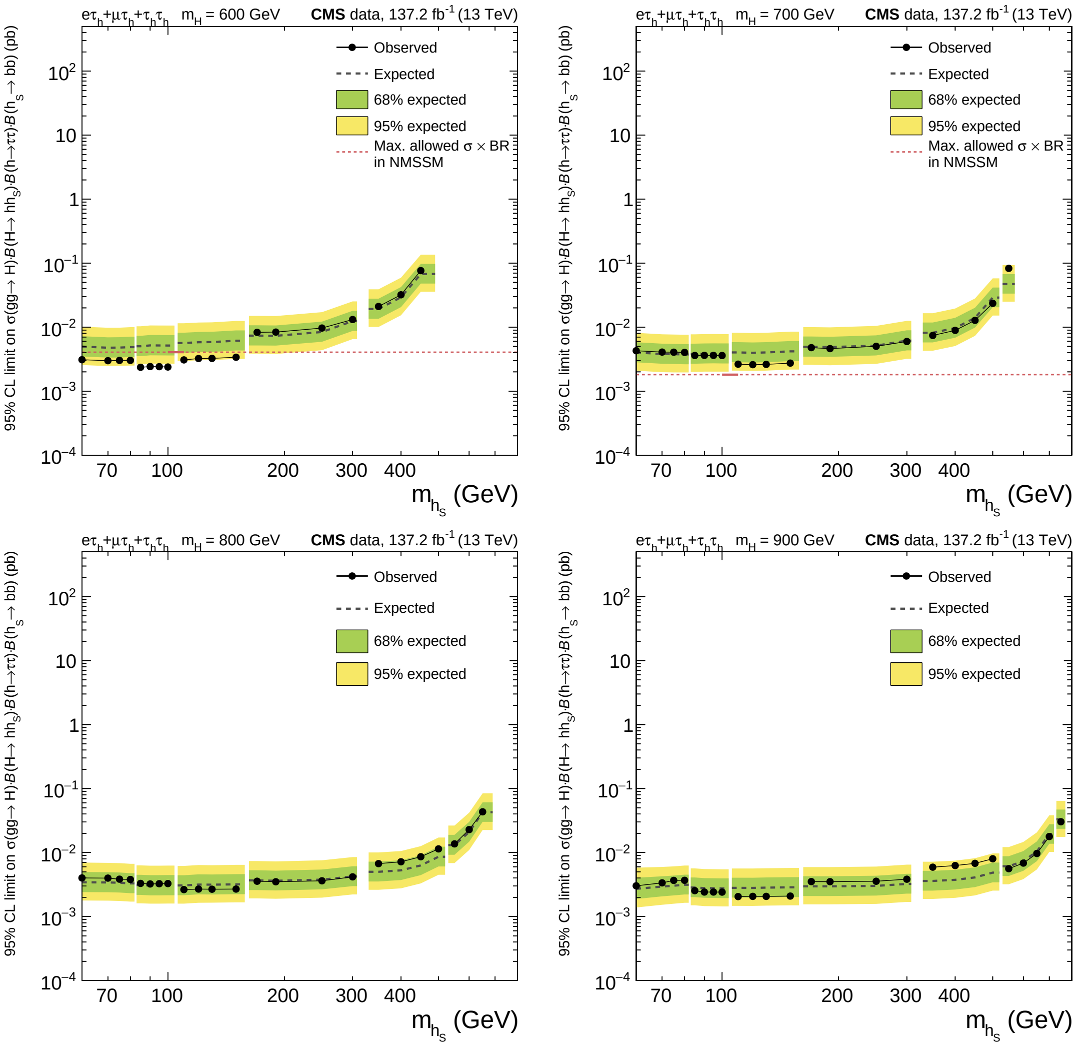


Figure A.3: Expected and observed 95% confidence level upper limits on the NMSSM signal process.

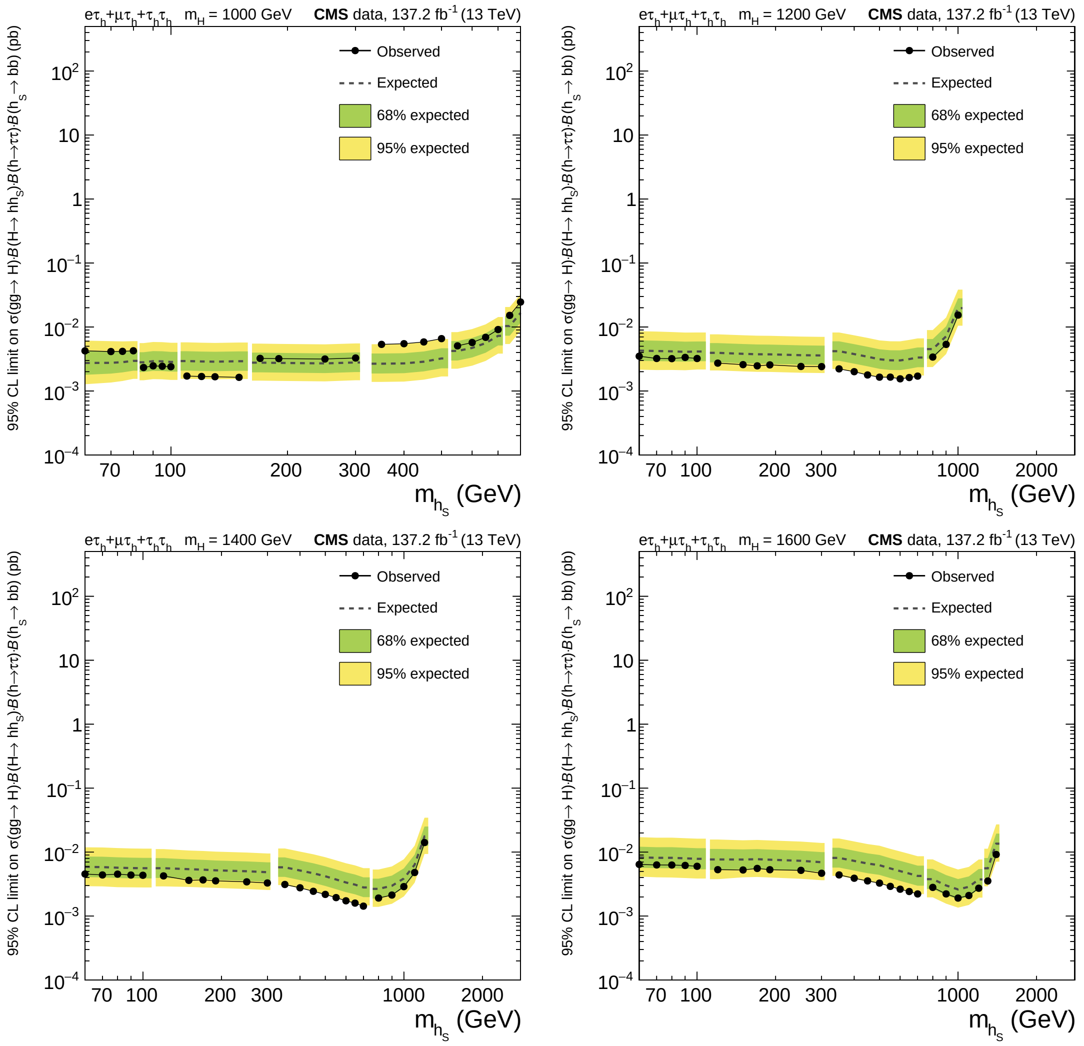


Figure A.4: Expected and observed 95% confidence level upper limits on the NMSSM signal process.

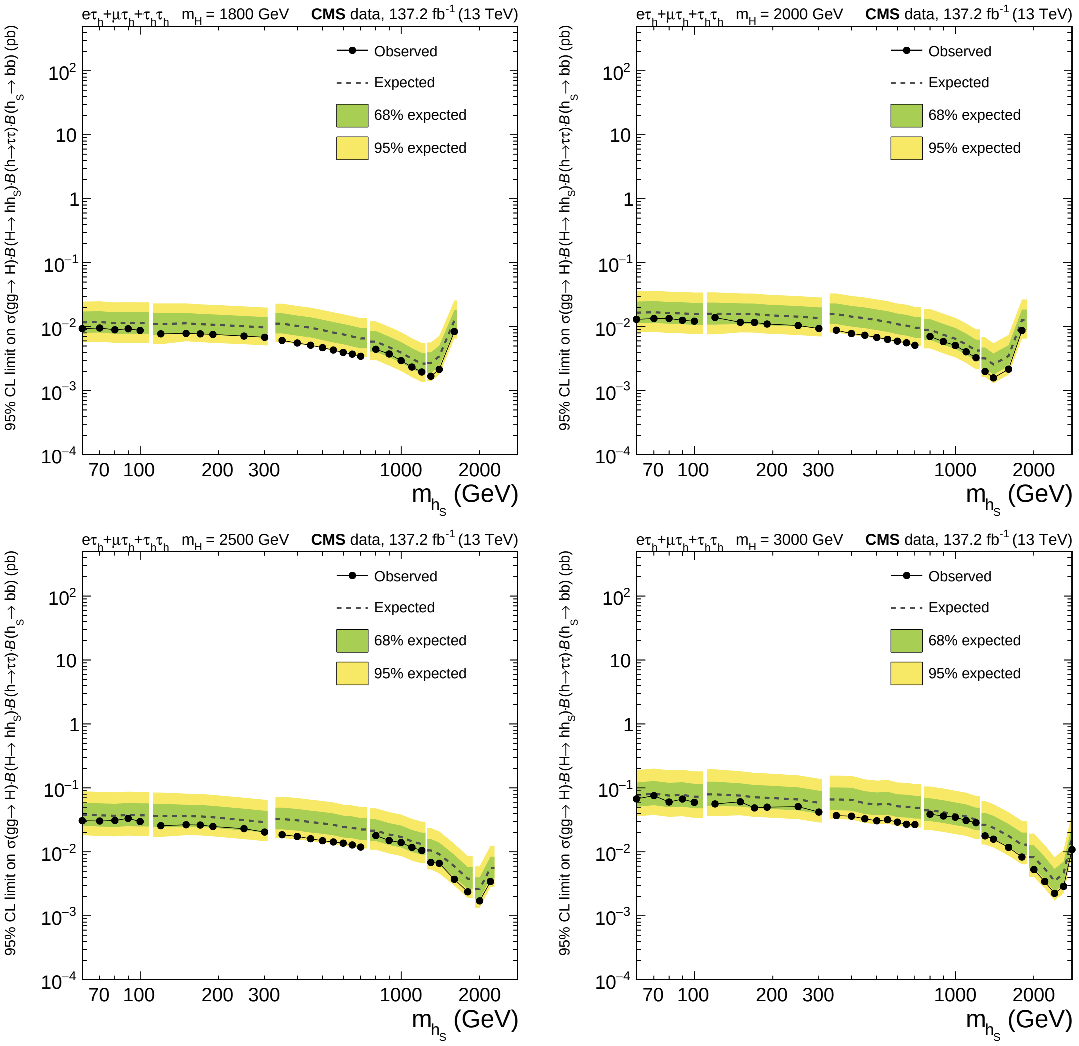


Figure A.5: Expected and observed 95% confidence level upper limits on the NMSSM signal process.

Validation of the NMSSM signal simulation

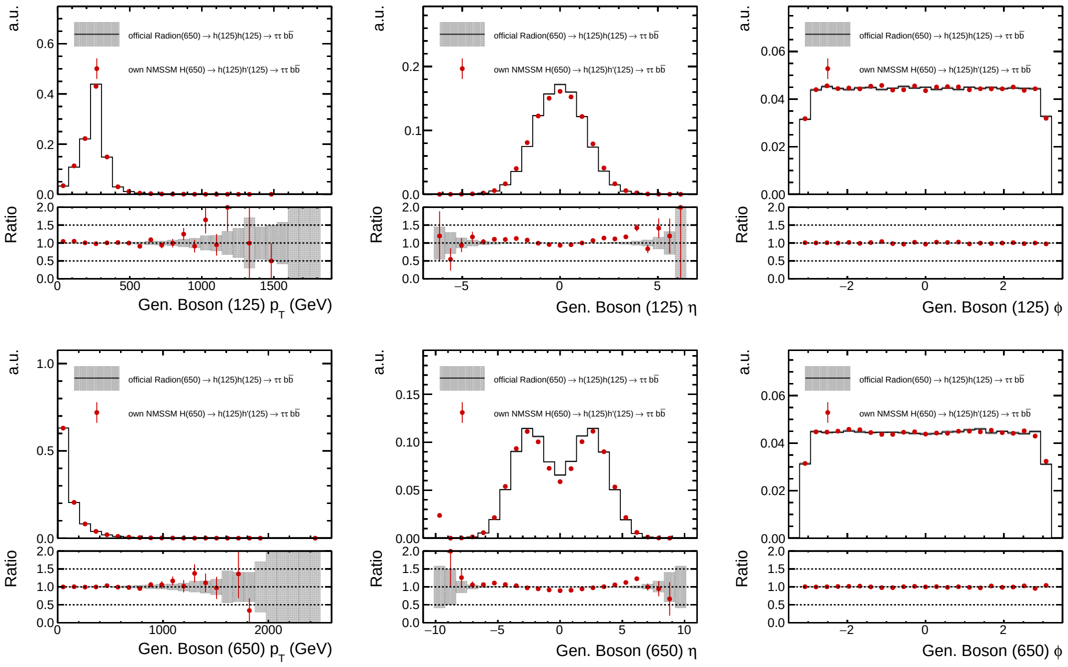


Figure A.6: Comparison of signal samples (red) with official CMS production (black). Shown are the simulated values of the p_T, η and ϕ of the light scalar boson (top) and the heavy scalar boson (bottom). The red error bars and black band represent the statistical uncertainties of the two samples.

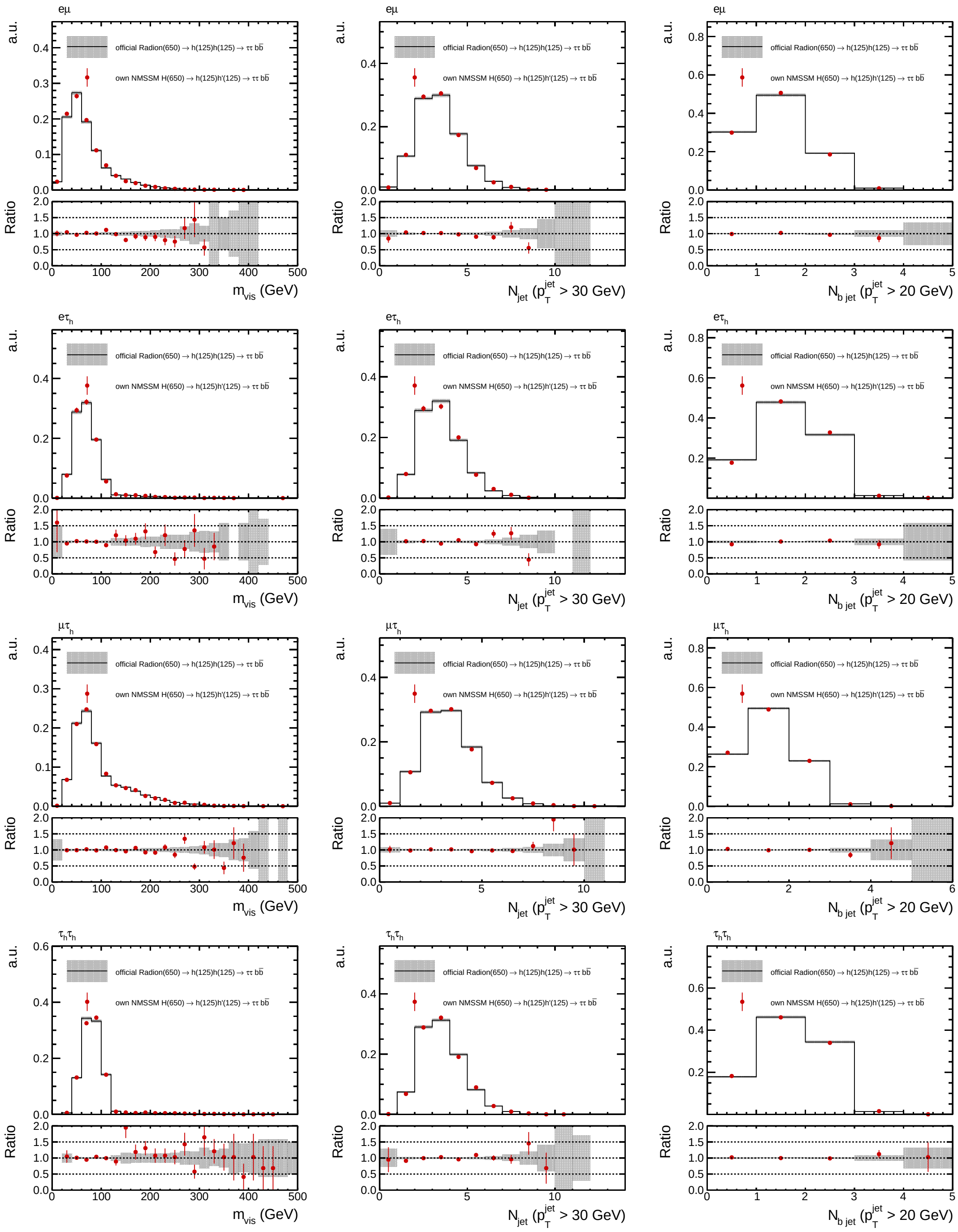


Figure A.7: Comparison of privately produced signal samples (red) with official CMS production (black). Shown are the reconstruction-level quantities of the visible di-tau mass (left), the number of reconstructed jets (middle) and the number of reconstructed b-tagged jets (right) for the $e\mu$, eτ_{h}, $\mu\tau_{\text{h}}$ and $\tau_{\text{h}}\tau_{\text{h}}$ final states of the di-tau system (top to bottom). The red error bars and black band represent the statistical uncertainties of the two samples.

One- and two-dimensional p-values of the goodness-of-fit tests performed on the NN input variables

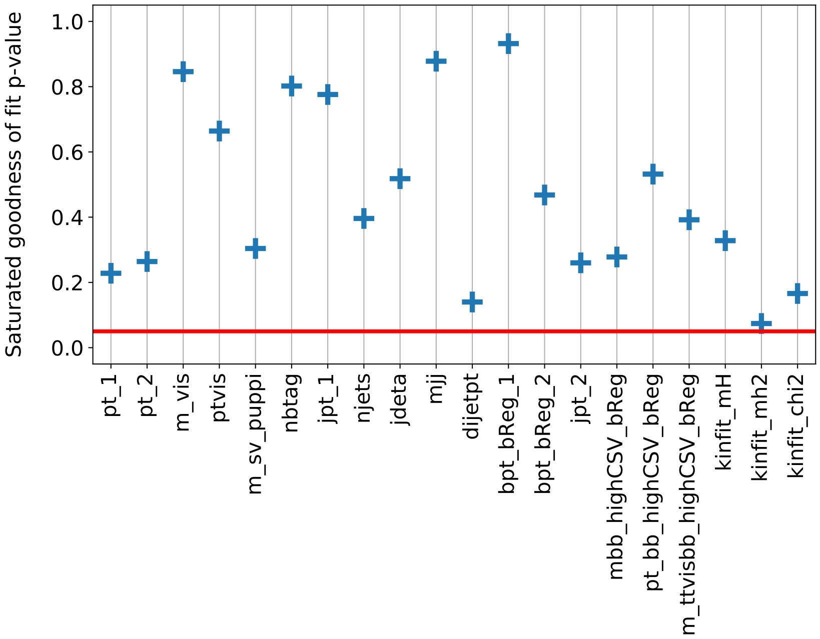


Figure A.8: Results for the 1D goodness-of-fit tests for the input variables used for the NN classification in the $e\tau_h$ final state using 2016 data.

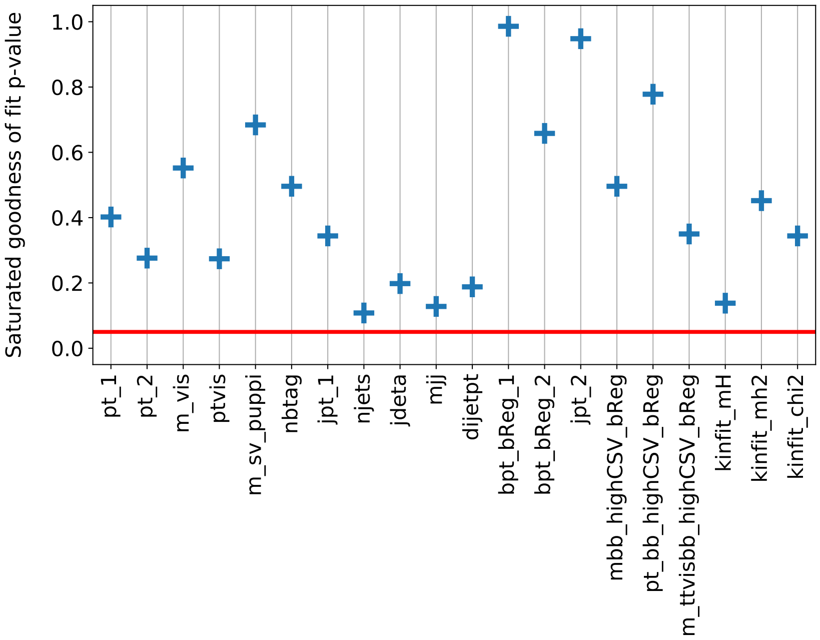


Figure A.9: Results for the 1D goodness-of-fit tests for the input variables used for the NN classification in the $\mu\tau_{\mathrm{h}}$ final state using 2016 data.

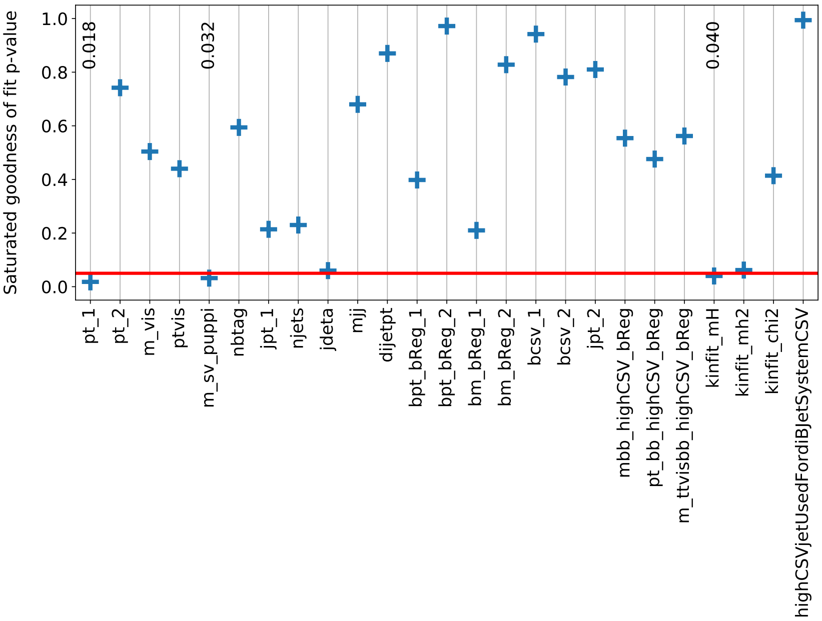


Figure A.10: Results for the 1D goodness-of-fit tests for the input variables used for the NN classification in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state using 2016 data.

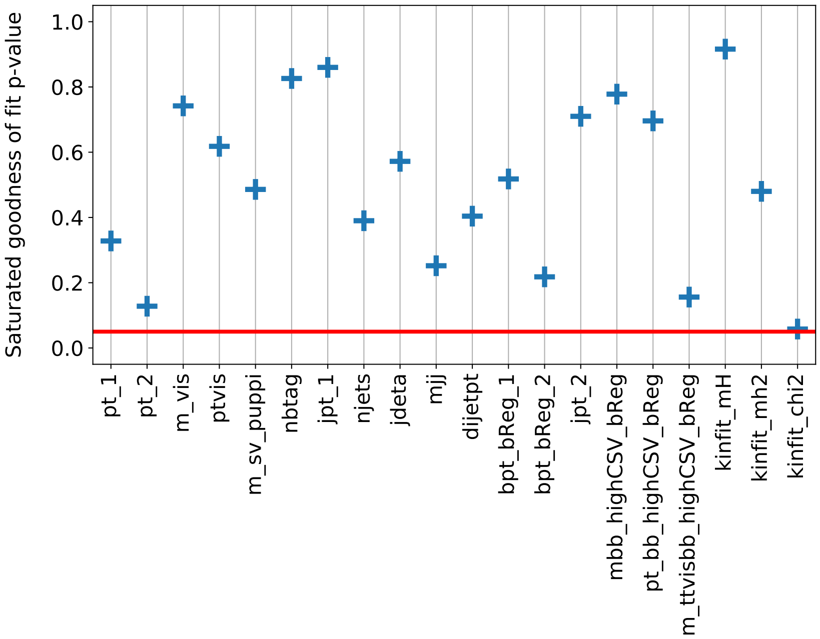


Figure A.11: Results for the 1D goodness-of-fit tests for the input variables used for the NN classification in the $e\tau_{\mathrm{h}}$ final state using 2017 data.

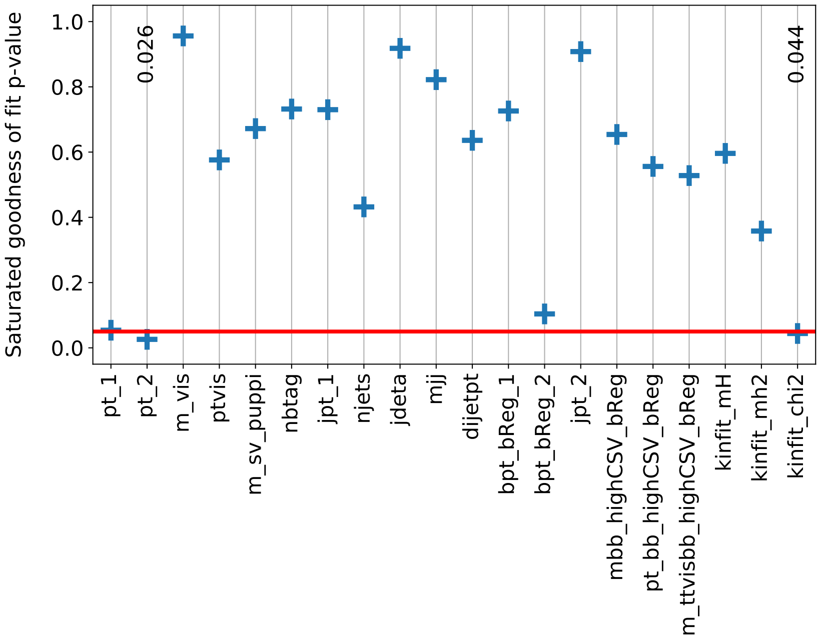


Figure A.12: Results for the 1D goodness-of-fit tests for the input variables used for the NN classification in the $\mu\tau_{\mathrm{h}}$ final state using 2017 data.

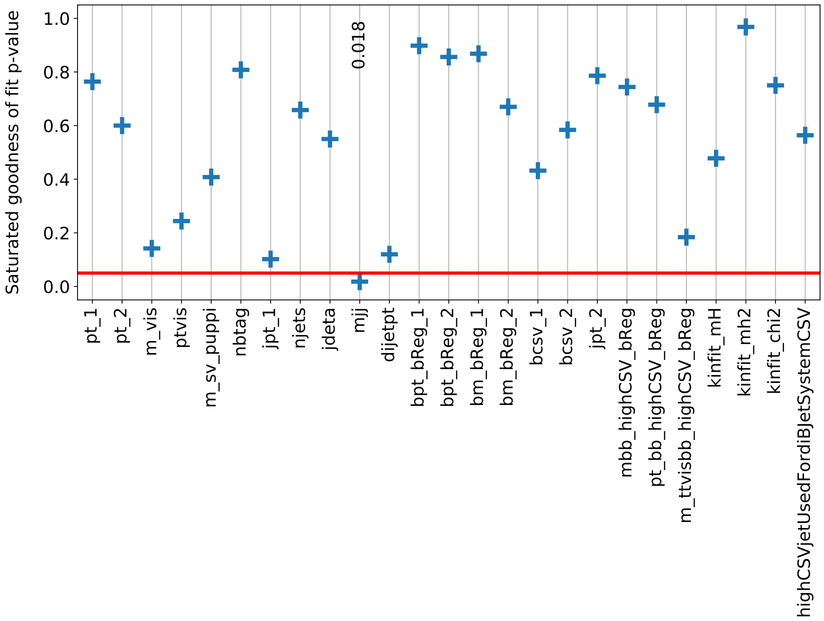


Figure A.13: Results for the 1D goodness-of-fit tests for the input variables used for the NN classification in the $\tau_h\tau_h$ final state using 2017 data.

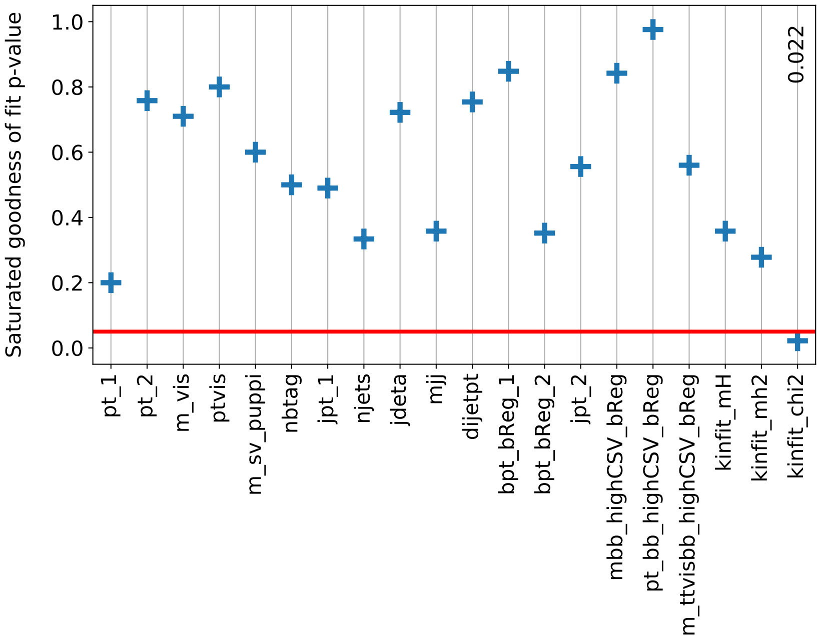


Figure A.14: Results for the 1D goodness-of-fit tests for the input variables used for the NN classification in the $e\tau_h$ final state using 2018 data.

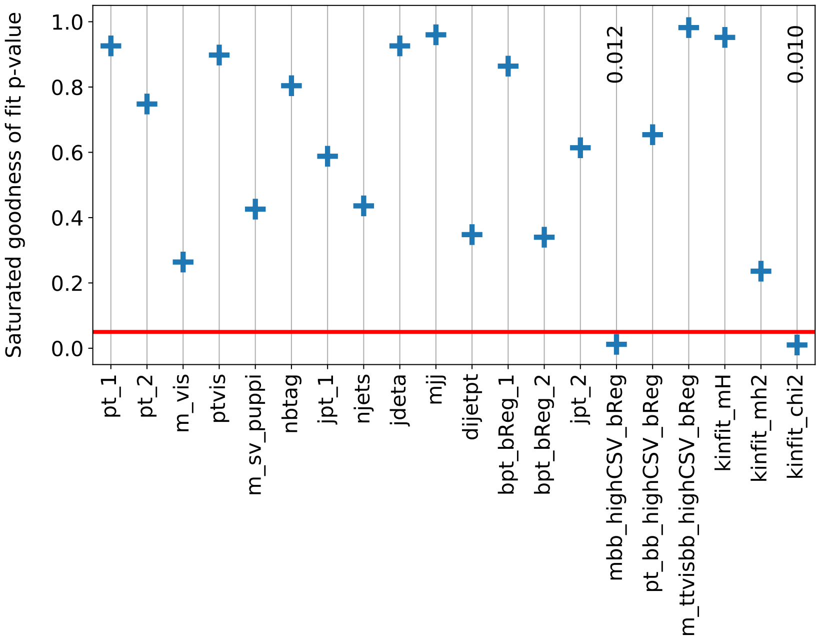


Figure A.15: Results for the 1D goodness-of-fit tests for the input variables used for the NN classification in the $\mu\tau_{\mathrm{h}}$ final state using 2018 data.

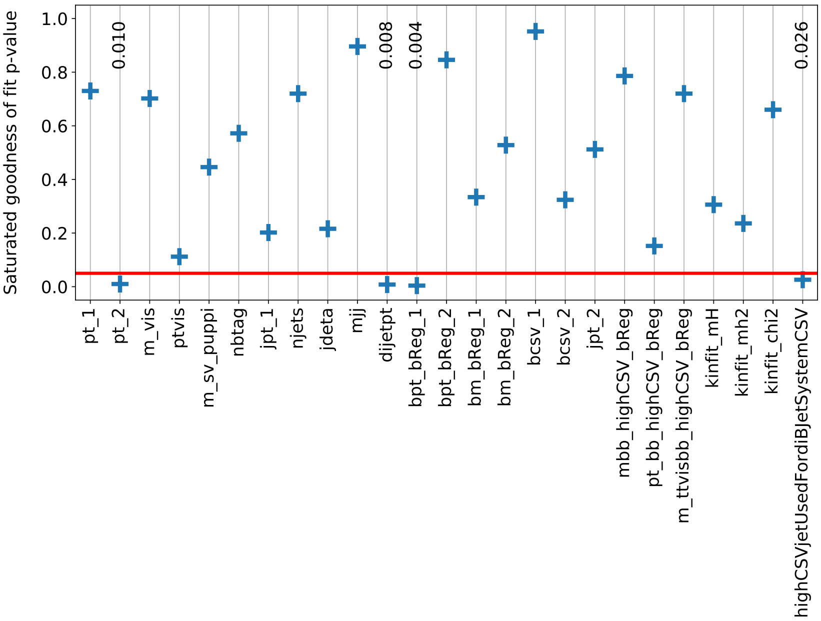


Figure A.16: Results for the 1D goodness-of-fit tests for the input variables used for the NN classification in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state using 2018 data.

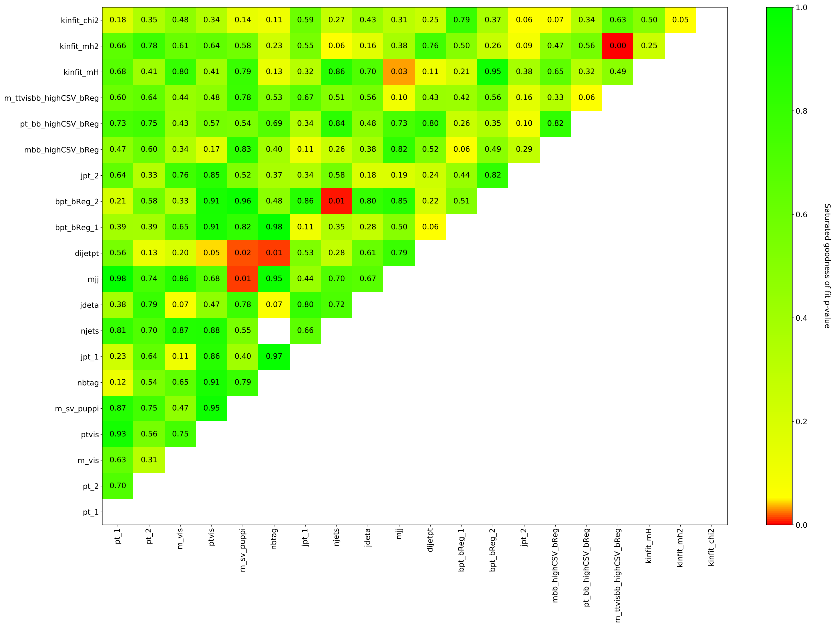


Figure A.17: Results for the 2D goodness-of-fit tests for the input variables used for the NN classification in the $e\tau_\mathrm{h}$ final state using 2016 data.

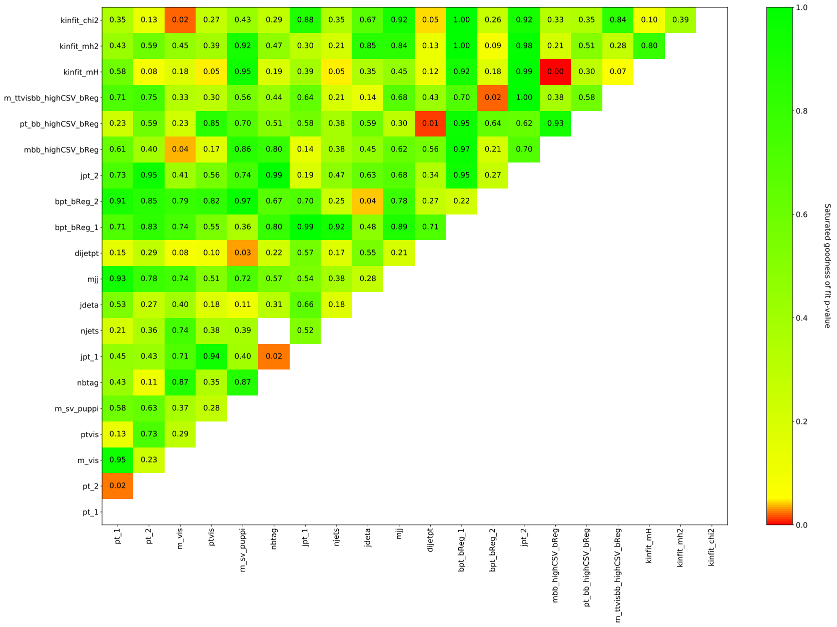


Figure A.18: Results for the 2D goodness-of-fit tests for the input variables used for the NN classification in the $\mu\tau_\mathrm{h}$ final state using 2016 data.

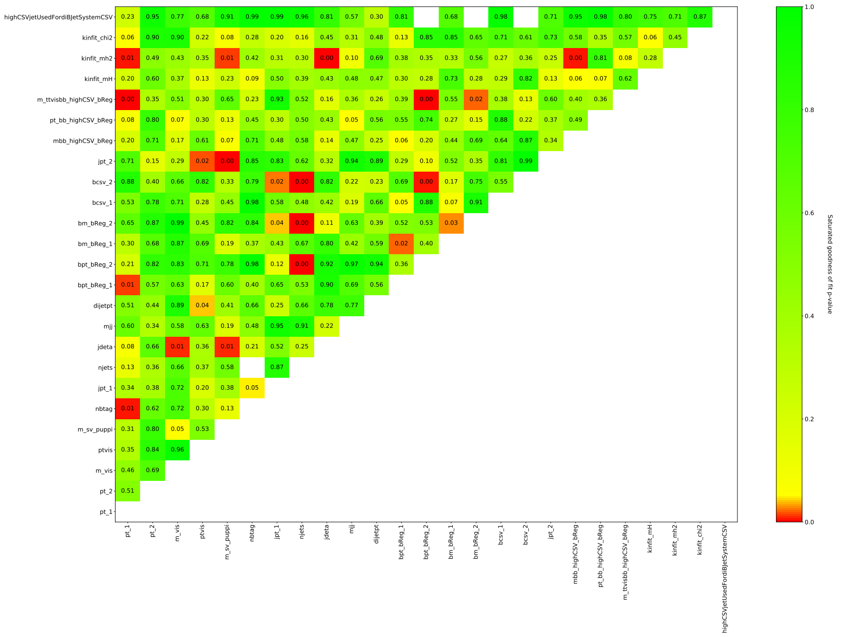


Figure A.19: Results for the 2D goodness-of-fit tests for the input variables used for the NN classification in the $\tau_h\tau_h$ final state using 2016 data.

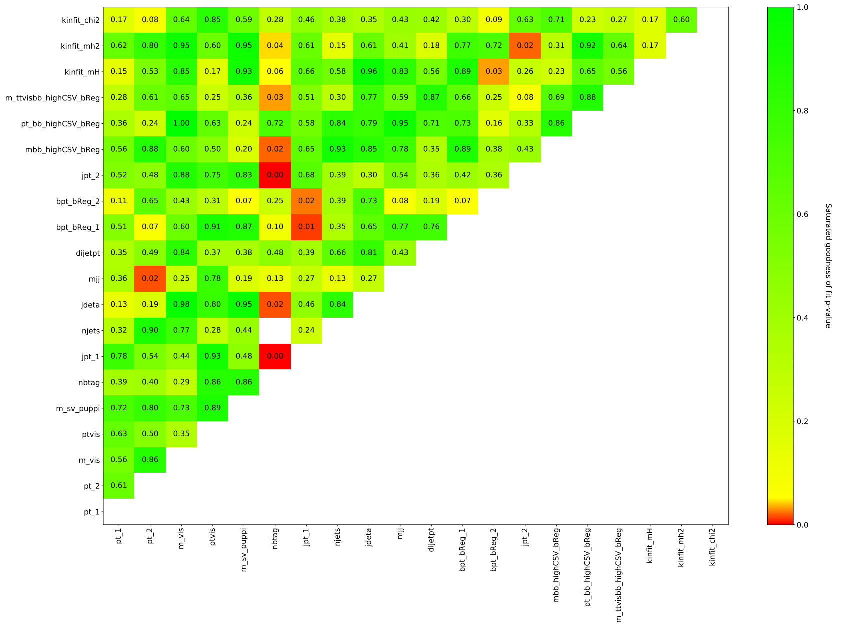


Figure A.20: Results for the 2D goodness-of-fit tests for the input variables used for the NN classification in the $e\tau_h$ final state using 2017 data.

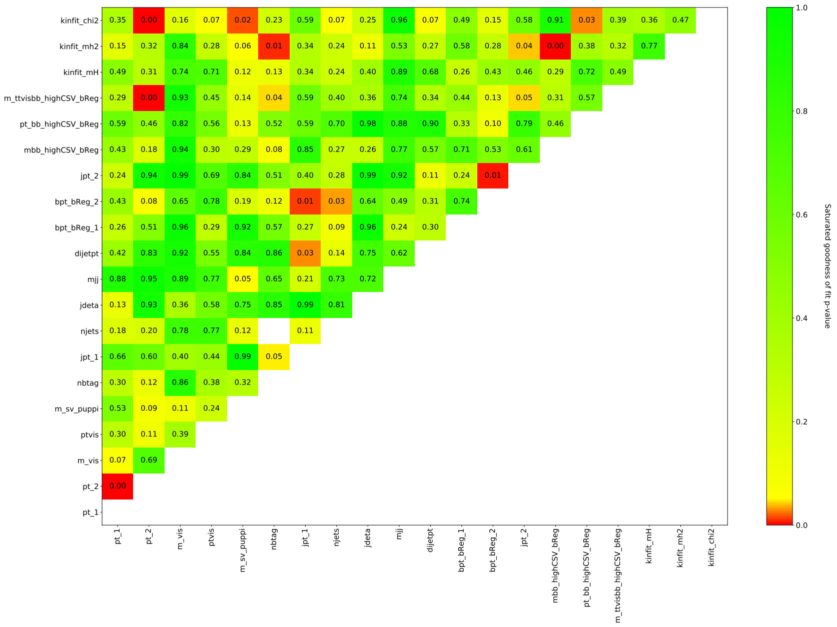


Figure A.21: Results for the 2D goodness-of-fit tests for the input variables used for the NN classification in the $\mu\tau_{\mathrm{h}}$ final state using 2017 data.

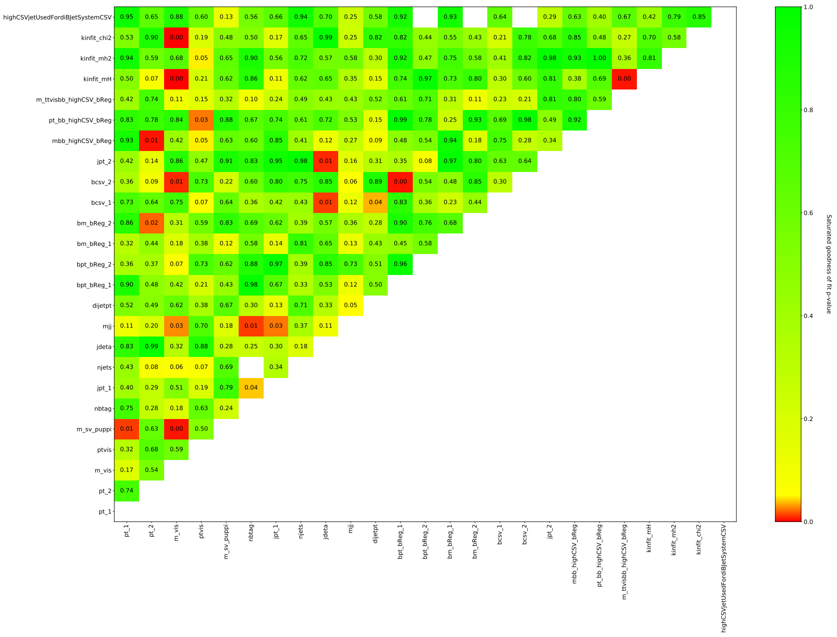


Figure A.22: Results for the 2D goodness-of-fit tests for the input variables used for the NN classification in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state using 2017 data.

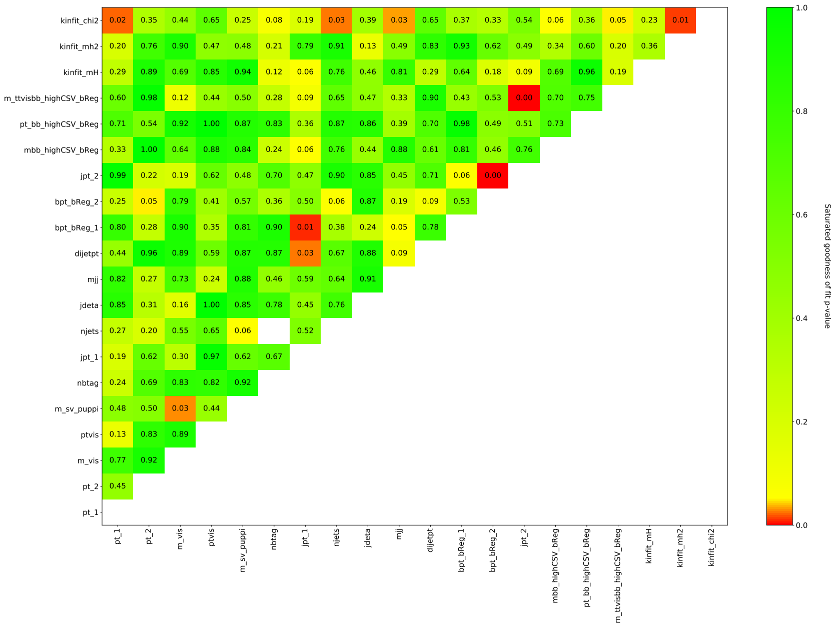


Figure A.23: Results for the 2D goodness-of-fit tests for the input variables used for the NN classification in the $e\tau_{\mathrm{h}}$ final state using 2018 data.

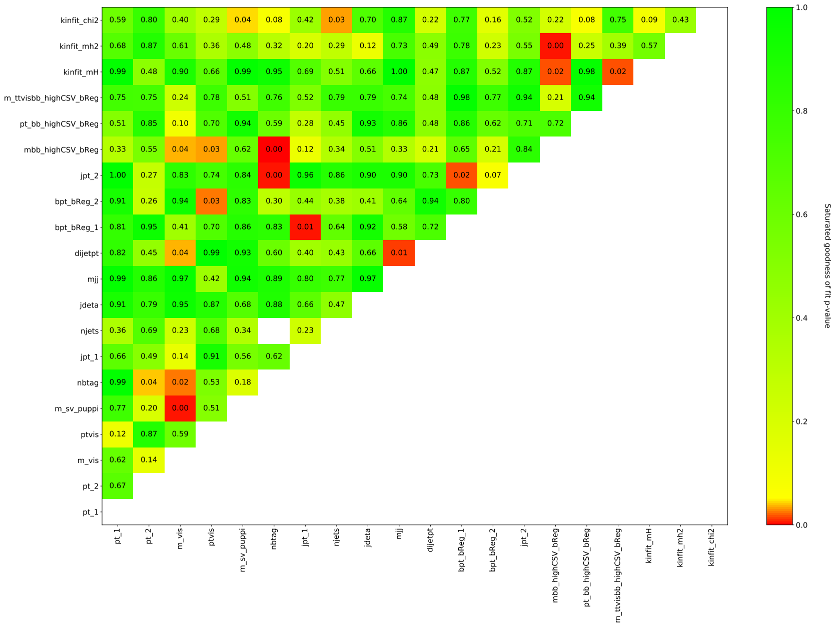


Figure A.24: Results for the 2D goodness-of-fit tests for the input variables used for the NN classification in the $\mu\tau_{\mathrm{h}}$ final state using 2018 data.

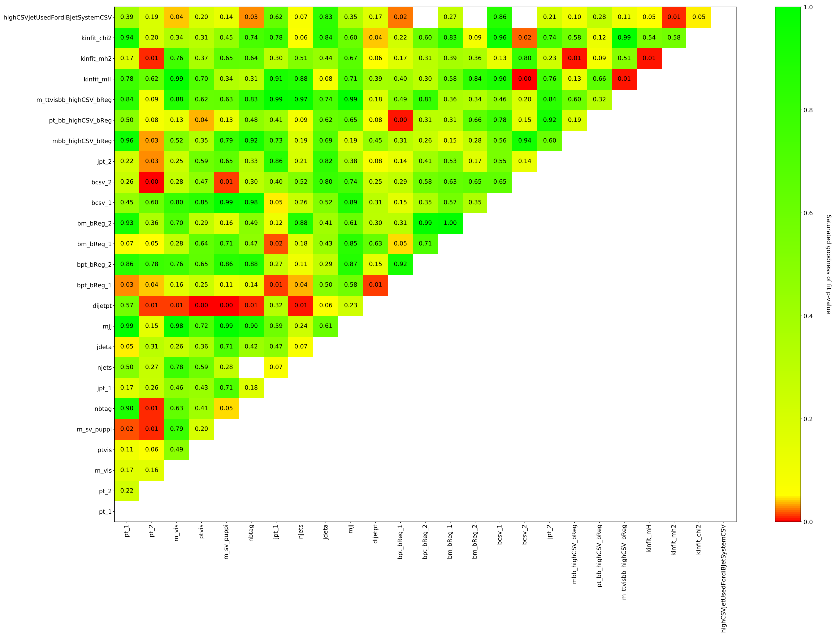


Figure A.25: Results for the 2D goodness-of-fit tests for the input variables used for the NN classification in the $\tau_h\tau_h$ final state using 2018 data.

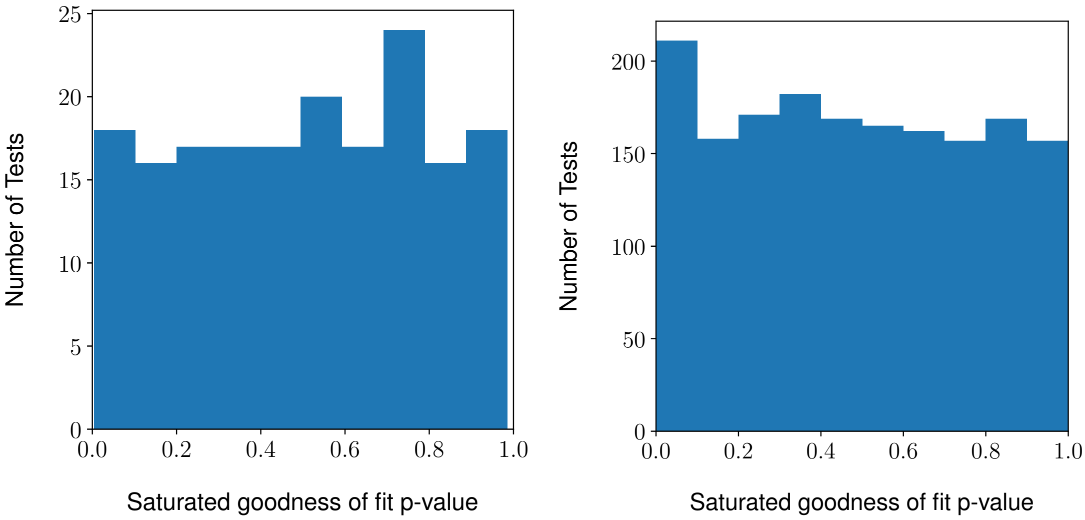


Figure A.26: Summary of all one-dimensional (left) and two-dimensional (right) p-values over the three final states and three run eras. The compatibility of the shown distributions with a uniform distribution is p=0.969 (1D) and p=0.119 (2D). The distributions are expected to be only approximately uniform due to the correlations between the tested variables.

Acknowledgements

I am extremely grateful to Günter Quast and Roger Wolf for warmly welcoming me into the group and giving me the trust and guidance which allowed me to pursue the research topics presented in this thesis, as well as providing me with the opportunity to attend many international schools and conferences as well as a research stay at CERN.

I would like to thank my colleagues Sebastian Brommer, Maximilian Burkart, Artur Gottmann, Sebastian Wozniewski and Stefan Wunsch for a fun and open work environment in which everybody could ask for help when needed and receive it.

Furthermore I am grateful for the support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 39/963-1 FUGG (bwForCluster NEMO). As highlighted in my thesis, the simulation used for the results relied heavily on the resources provided by the initiative and would not have been possible without it. I would like to thank the state of Baden-Württemberg again for providing me with financial support in the context of the Landesgraduiertenförderung during the first two years of my PhD.

In this context I would also like to thank everyone at our institute who cared for and enabled us to efficiently use these and many other computing resources.

Finally, I wish to thank my family for their love, support and encouragement throughout my life.