Contents

Chapter 1

Searches for New Physics in $\tau^+\tau^-$ Final States

The $\tau^+\tau^-$ final states are a powerful tool to search for new physics at collider experiments. As the heaviest lepton, they are sensitive to resonant production of new neutral particles where the couplings have mass hierarchy. They are also sensitive to non-resonant effects from new physics mediators. This chapter will detail the searches for two such areas of new physics: additional Higgs bosons and vector leptoquarks. These searches are split up into three sections:

i) A model independent search for single narrow spin-0 resonance, ϕ, produced via gluon fusion ($gg\phi$) or in association with a bottom quark ($bb\phi$). The SM Higgs boson is treated as a background. The Yukawa couplings that contribute to the gluon fusion loop are set to SM values.

ii) A search for the MSSM Higgs sector, in a number of benchmark scenarios. The benchmark scenarios are defined in Section 1.1.1. The production of SM Higgs boson is also used to constrain the available phase space.

iii) A search for the t-channel exchange of a U_1 vector leptoquark. Two scenarios are taken, based of the best fit to the b anomalies. These scenarios are detailed in Section 1.1.2.

These searches are performed with the full run-2 dataset (138 fb^{-1}) collected by the CMS experiment. The search for additional Higgs bosons had previously been performed with data collected in 2016 (39 fb^{-1}) and results were consistent with the SM background prediction.

1.1 Signal Modelling

1.1.1 Additional Higgs Bosons

Extended Higgs sectors, such as that of the MSSM, can be probed by direct searches for the additional bosons and further precise measurements of the Standard Model Higgs boson. This search for an extended Higgs sector is motivated by Type II 2HDMs, such as the MSSM. In these models $\tan\beta$ enhances couplings of additional Higgs bosons to bottom-like quarks and leptons, whilst top-like couplings are suppressed. This narrows down the most important production modes of the Higgs boson into two categories: Gluon fusion and production in association with a bottom quark. Examples of these are shown in Figure 1.1.

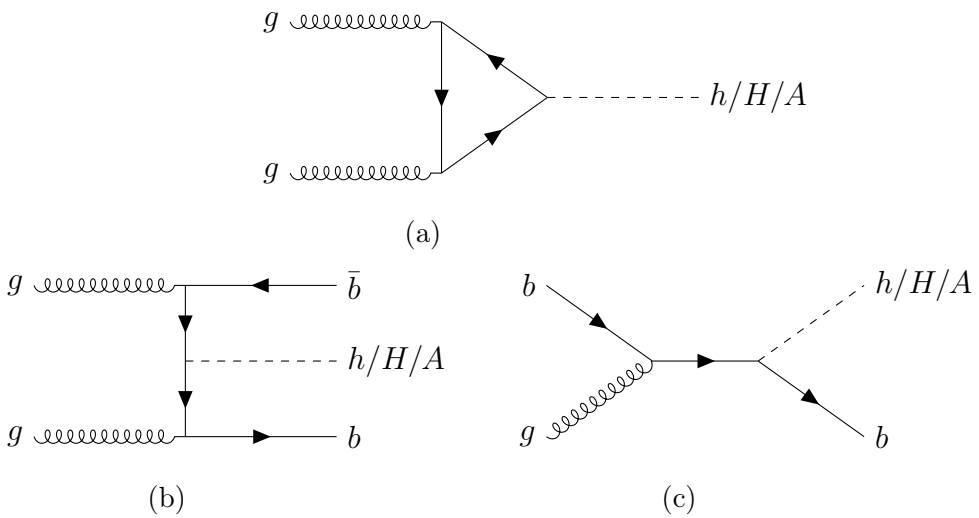


Figure 1.1: Diagram (a) shows the production of neutral Higgs bosons from gluon fusion. The dominant loop contributions to this diagrams are from top-only, bottom-only and top-bottom interference. Diagrams (b) and (c) show production in association with b quarks.

With the $\tan\beta$ enhancement, the decays of additional Higgs bosons to tau leptons and bottom quarks are most likely. Tau leptons are identified with a higher purity than bottom quarks a the CMS detector. It is also easier to separate $\tau^+\tau^-$ from the large QCD multijet background produced from the high energy proton-proton collisions. This hypothesis was tested with the 2016 dataset and although no deviations were observed, the strongest limits on the MSSM phase space was placed by the $\tau^+\tau^-$ final states.

For this analysis, the production of additional Higgs bosons over a mass range of 60 GeV to 3.5 TeV are generated. Gluon fusion is simulated at NLO precision

using the 2HDM implementation of POWHEG 2.0. The kinematic properties are highly dependent on the contributions to the loop, which vary dependent on the specific signal model. To account for the different loop contributions at the NLO plus parton shower prediction, weights based off the p_T spectra are calculated to split the contributions from the t quark only, b quark only, and tb-interference. Once individual templates have been determined for each contribution to the loop, the 2HDM samples can be scaled to the correct specific MSSM scenario prediction by the following formula.

$$\frac{d\sigma_{\text{MSSM}}}{dp_T} = \left(\frac{Y_{t,\text{MSSM}}}{Y_{t,\text{2HDM}}}\right)^2 \frac{d\sigma^t_{\text{2HDM}}}{dp_T}(Q_t) + \left(\frac{Y_{b,\text{MSSM}}}{Y_{b,\text{2HDM}}}\right)^2 \frac{d\sigma^b_{\text{2HDM}}}{dp_T}(Q_b) + \left(\frac{Y_{t,\text{MSSM}}}{Y_{t,\text{2HDM}}}\frac{Y_{b,\text{MSSM}}}{Y_{b,\text{2HDM}}}\right)\left\{\frac{d\sigma^{t+b}_{\text{2HDM}}}{dp_T}(Q_{tb}) - \frac{d\sigma^t_{\text{2HDM}}}{dp_T}(Q_{tb}) - \frac{d\sigma^b_{\text{2HDM}}}{dp_T}(Q_{tb})\right\} \quad (1.1)$$

where Q_i are resummation scales that depend on the mass of the additional Higgs boson. Further contributions from any Supersymmetric partners have been checked and account for less than a few percent and so are neglected. This is also done separately for the scalar and pseudoscalar additional Higgs bosons, as the p_T distributions can differ. The MSSM benchmark scenarios considered are detailed in Ref. []. The scenarios provide the relative Yukawa couplings (to calculate the cross sections) and branching fractions of the MSSM Higgs bosons. An example of the changes to gluon fusion production, in the MSSM M_h^{125} scenario with $m_A = 1600$ GeV and $\tan\beta$ varying is shown in Figure 1.2. The distributions peak at a higher p_T for the top quark loop, therefore at smaller $\tan\beta$, where the top quark contribution is dominant, an additional Higgs boson would be more boosted.

Production in association with bottom quarks is simulated at NLO precision using the corresponding POWHEG 2.0 implementation in the four-flavour scheme. All additional Higgs boson signal generation is performed using the parton distribution function (PDF) NNPDF3.1. Tau lepton decay, parton showering and hadronisation are all modelled with the PYTHIA event generator where the PU profile is matched to data. All events generated are passed through a GEANT4-based simulation of the CMS detector and reconstructed in the same way as data.

The model dependent search for the MSSM also looks to find differences from the observed SM Higgs boson and the predicted MSSM SM-like Higgs boson. In each MSSM benchmark scenario, an uncertainty of ± 3 GeV is given on the prediction for the SM Higgs boson mass. This uncertainty is to reflect the contribution from any

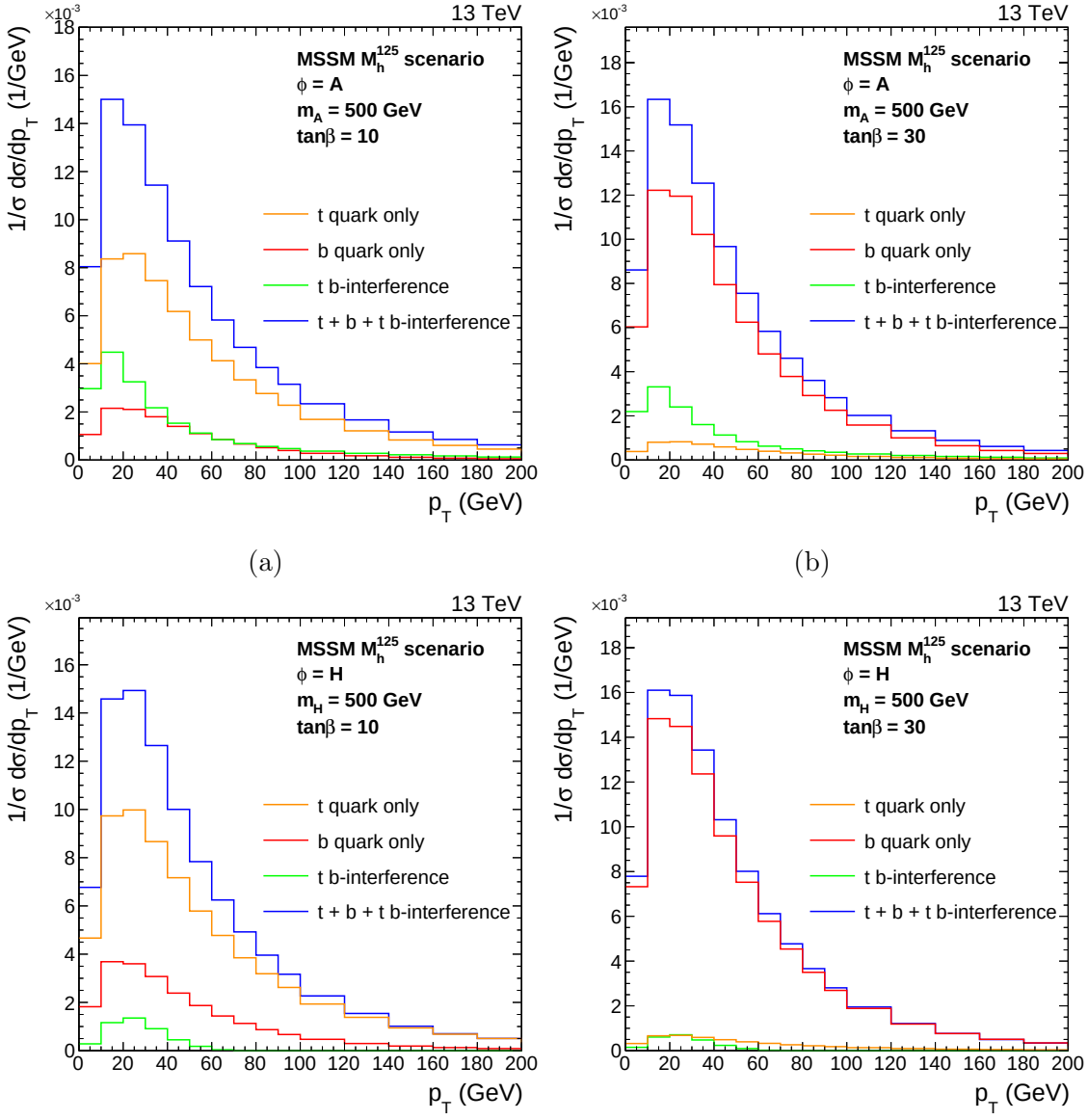


Figure 1.2: p_T density distributions of the A (top) and H (bottom) boson, with contributions to the gluon fusion loop displayed individually and summed. These are shown for $\tan\beta$ values of 10 (left) and 30 (right) where $m_A = 500$ GeV in the MSSM M_h^{125} scenario.

unknown higher-order corrections. The value of the mass is allowed to vary within this window, however the Yukawa couplings are rescaled the observed mass.

1.1.2 Vector Leptoquarks

The best fit in the vector leptoquark phase space to the B anomalies yielded large bottom quark and tau lepton couplings to the U_1 particle. The possible production modes of a $\tau^+\tau^-$ final state are shown in Figure 1.3.

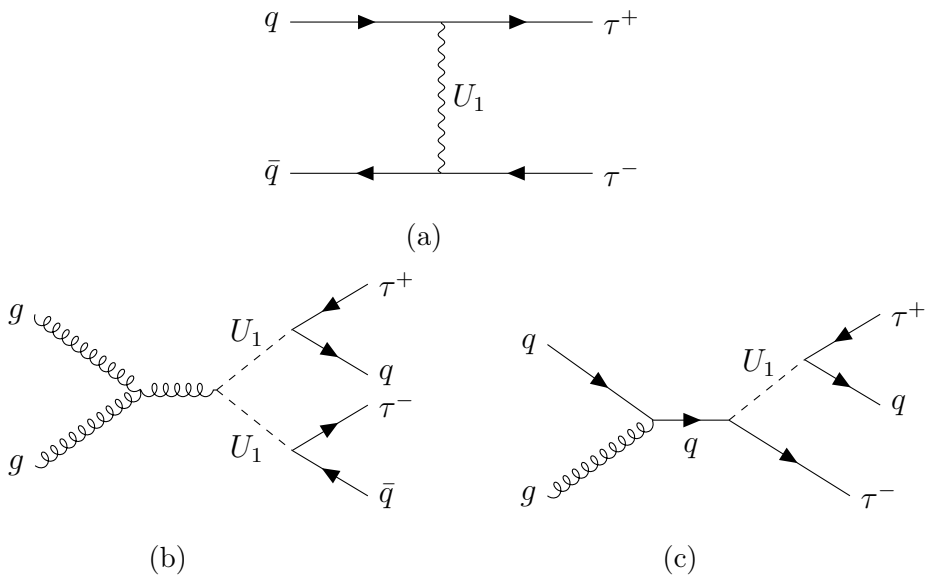


Figure 1.3: Feynman diagrams showing the contribution from U_1 vector leptoquarks to the final state with a pair of oppositely charged tau leptons. Diagram (a) shows t-channel, (b) pair and (c) single production of a vector leptoquark.

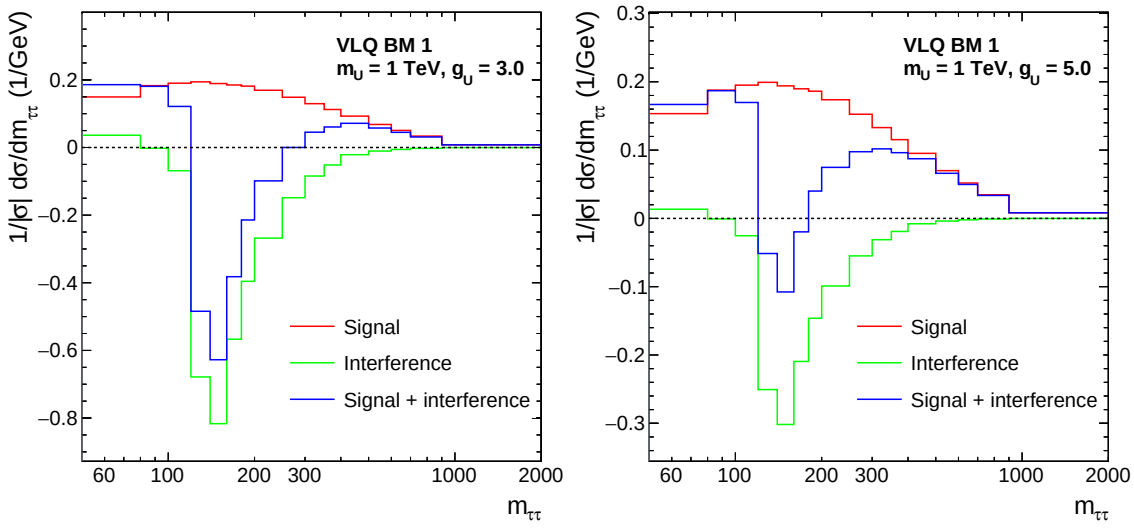
Pair and single production of a vector leptoquark is dependent on its strong coupling, which is highly model dependent. For large mass, m_U, the probability of producing an on-shell U_1 singlet or pair is heavily suppressed due to the momentum of the initial partons. These production processes are not discussed further in this search. Further studies have been used to search for single and pair production at the CMS experiment and no statistically significance derivation was observed.

The t-channel process contain two vertices with a U_1 vector leptoquark, a quark and a tau lepton, and hence the cross section will scale with g_U^4. From the best fit to B anomalies this vertex will be dominated by the b quark and hence the initial state will be mostly from $b\bar{b}$, with sub-dominant contributions from $b\bar{s}$, $s\bar{b}$ and $s\bar{s}$. Although there are no additional b quarks in the final state in the LO process, initial state radiation can lead to additional b quarks in the final state. In this search the two scenarios discussed in Section **??** are considered. The only non negligible

parameter for $\tau^+\tau^-$ final states from the fit in the m_U-g_U phase space is the $\beta_L^{s\tau}$ parameter. This is set to the best fit value.

The signal process of the U_1 t-channel exchange is simulated in the five-flavour scheme (5FS) at LO precision using the MadGraph5_aMC@NLO event generator, v2.6.5. Events are generated with one or fewer outgoing partons from the matrix element and the MLM prescription is then used for matching, with a scale set to 40 GeV. Negligible dependence of the U_1 decay width (Λ) is observed, for simulation this is chosen to approximately match the value predicted by the B anomaly fit. Samples with a mass between 1 and 5 TeV at $g_U = 1$ are generated.

The interference between the U_1 signal and $Z/\gamma^* \to \tau\tau$ production was checked. A large destructive affect is observed, with the magnitude dependent on g_U. To account for this, separate samples are produced for this interference, generated in the same way as the t-channel exchange. The interference samples are then split into two with a di-tau mass split in order to have a sufficient number of events in the high di-tau mass regions. The cross section of these interference samples scale with g_U^2. Examples of the generator level di-tau mass distributions are shown in Figure 1.4.



(a) (b)

Figure 1.4: The generator level $m_{\tau\tau}$ density distributions of the t-channel vector leptoquark signal and the interference with Drell-Yan. This is shown in the VLQ BM 1 scenario for a leptoquark of mass 1 TeV for coupling strengths of $g_U = 3$ (a) and $g_U = 5$ (b).

The t-channel signal produces a broad distribution in $m_{\tau\tau}$ due to its non-resonant nature. The interference is mostly a destructive effect (except for at small $m_{\tau\tau}$), with the yield becoming less negative at higher $m_{\tau\tau}$. The interference peaks negatively between 100 and 200 GeV and in this region the combined yield can be negative. Due to the difference in scaling of the two effects, at small g_U the interference is more dominant than the signal and hence the yield of the combined result is reduced.

1.2 Event Selection

The possible decays of two tau leptons and their branching fractions, where the tau decay is grouped into three categories e, μ and τ_h as defined in Section **??**, are shown in Table 1.1. For this search the four largest branching fraction channels used: $\tau_h\tau_h$, $e\tau_h$, $\mu\tau_h$ and $e\mu$. This accounts for approximately 94% of di-tau events. The two same lepton channels are neglected due to small branching ratio and the dominating $Z \to ee$ and $Z \to \mu\mu$ backgrounds.

Channel	Branching Fraction
$\tau_h\tau_h$	42.0%
$e\tau_h$	23.1%
$\mu\tau_h$	22.6%
$e\mu$	6.2%
ee	3.2%
$\mu\mu$	3.0%

Table 1.1: Branching fractions of the decays of two tau leptons.

1.2.1 Trigger Requirements

In the four final state pairs a number of different online trigger requirements are needed. In the $\tau_h\tau_h$ channel, two possible triggers are available: the double-τ_h and single-τ_h triggers. The single-τ_h trigger has a high p_T threshold at 120 (180) GeV for events recorded in 2016 (2017-2018), whilst the double-τ_h has a p_T threshold at 40 GeV. Therefore, the double-τ_h trigger is used individually where the τ_h has p_T is below the single-τ_h threshold and the union of single-τ_h and double-τ_h triggered events are taken above the threshold.

In the $e\tau_h$ and $\mu\tau_h$ channels, there are three possible triggers available: the single-e/μ, single-τ_h and the e/μ-τ_h cross-trigger. The cross-trigger is used for events where the light lepton has p_T between the thresholds for the cross-trigger and single-e/μ shown in Table 1.2. The light lepton used in the cross-trigger is required to be in the central barrel of the detector within $|\eta| < 2.1$. Above these light lepton p_T thresholds the single-e/μ trigger is used, where here it is required that the τ_h has $p_T > 30$ GeV. At τ_h p_T above the single-τ_h thresholds, the single-τ_h trigger is used in combination with the single-e/μ trigger.

Year/ Trigger	$e\tau_h$ cross-trigger	single-e	$\mu\tau_h$ cross-trigger	single-μ
2016	23	26	20	23
2017	25	28	20	25
2018	25	33	21	25

Table 1.2: Lower trigger light lepton thesholds p_T in GeV for the $e\tau_h$ and $\mu\tau_h$ channels.

In the $e\mu$ channel, there are three possible triggers available: the single-e, single-μ and the e-μ cross-trigger. However, only the cross-trigger is used in this analysis, due to the larger efficiencies of correctly selecting light leptons. The e and μ are required to have $p_T > 15$ GeV and $|\eta| < 2.4$.

1.2.2 Offline Requirements

All offline selections stated are in addition the object selection discussed in Section **??**. In this analysis, hadronic tau candidates are required to pass the `Medium` $D_{\text{jet}}^{\text{WP}}$. D_e^{WP} and D_μ^{WP} are dependent on the channel. The `VVLoose`, `Tight`, `VVLoose` D_e^{WP} and the `VLoose`, `VLoose` and `Tight` D_μ^{WP} are used in the $\tau_h\tau_h$, $e\tau_h$ and $\mu\tau_h$ channels respectively. The tighter working point for the same light lepton discrimination as tagged in the event is used to remove light leptons faking hadronic taus from the $Z \to ll$ process. The light lepton isolation requirement is $I_{\text{rel}}^{e/\mu} < 0.15$ except for in the $e\mu$ channel where the muon is required to have $I_{\text{rel}}^{\mu} < 0.2$.

The selected τ lepton decay candidates are required to have opposite charge and to be separated by more than $\Delta R > 0.5$ in all channels except $e\mu$ where $\Delta R > 0.3$. In events where the numbers of an object in the event is greater than the required number of objects in the $\tau\tau$ decay channel, the objects are sorted by the maximum $D_{\text{jet}}^{\text{score}}$ if τ_h or minimum I_{rel} if a light lepton and the leading objects are chosen. In order to maintain orthogonality between channels, events with additional light leptons passing looser selections than the nominal requirements, are rejected from the selection. The looser selections help to suppress the $Z \to ll$ background process further.

In the $e\tau_h$ and $\mu\tau_h$ channels, a cut is placed at 70 GeV on the transverse mass between the light lepton $\vec{p}_T$ and the missing $\vec{p}_T$, where the transverse momentum is

defined as,

$$m_{\mathrm{T}}(\vec{p}_T^{\,i}, \vec{p}_T^{\,j}) = \sqrt{2p_T^i p_T^j (1 - \cos\Delta\phi)}, \tag{1.2}$$

where $\Delta\phi$ to the azimuthal angle between $\vec{p}_T^{\,i}$ and $\vec{p}_T^{\,j}$. The variable is used to remove W + jets background events, where a jet fakes a hadronic tau and the MET and light lepton from the W decay are aligned and hence the event has a large $m_T(\vec{p}_T^{\,e/\mu}, \vec{p}_T^{\,\mathrm{MET}})$. In the $e\mu$ channel an additional cut is placed on a variable named D_ζ, which is defined as,

$$D_\zeta = p_\zeta^{\mathrm{miss}} - 0.85 p_\zeta^{\mathrm{vis}}; \qquad p_\zeta^{\mathrm{miss}} = \vec{p}_T^{\,\mathrm{miss}} \cdot \hat{\zeta}; \qquad p_\zeta^{\mathrm{vis}} = (\vec{p}_T^{\,e} + \vec{p}_T^{\,\mu}) \cdot \hat{\zeta} \tag{1.3}$$

where $\vec{p}_T^{\,e/\mu}$ corresponds to the transverse momentum vector of the electron or muon and $\hat{\zeta}$ to the bisectional direction between the electron and the muon in the transverse plane [**?**]. A diagram of the inputs is shown Figure 1.5.

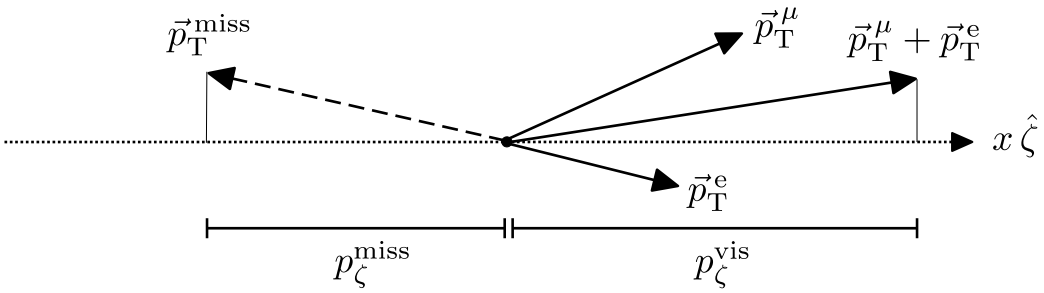


Figure 1.5: Diagram of inputs to the D_ζ variable.

The linear combination is optimised for genuine di-tau events to peak around $D_\zeta = 0$ GeV. It is motivated by the expectation that in di-tau decays from a resonance, the visible and missing (from tau neutrinos) momentums are roughly aligned and of similar magnitudes. In W + jets and $t\bar{t}$ events the directions of the visible and missing products are expected to be more randomly distributed and lead to a non-peaking D_ζ. Therefore only events with $D_\zeta >$ -35 GeV are considered for signal events. No b tagged events with this cut are vetoed and b tagged events with this cut are used for a $t\bar{t}$ control region and discussed further in Section 1.4.

1.3 Search Optimisation

The optimisation of the signal extraction depends on which of the three scenarios, set out at beginning of this section, is searched for. The components of the optimisation are named the high mass, low mass and SM Higgs optimisation procedures. For the model independent search (i) the high or low mass optimisation procedures are used depending on whether the mass of the resonance is greater or less than 250 GeV. The search for the MSSM Higgs sector (ii) uses the high mass or the SM Higgs optimisation procedures depending on whether the reconstructed di-tau mass [] is greater or less than 250 GeV. Finally, the search for vector leptoquarks (iii) uses only the high mass optimisation procedure. The procedures are discussed in detail below.

The high mass optimisation procedure follows what was done in Ref. []. Firstly each category is split into categories with no b tagged and with one or more b tagged events. This firstly helps target the additional Higgs boson production modes gluon fusion and b associated production respectively. Secondly, the initial state radiation of a t-channel vector leptoquark signal dominated by initial states of b quarks, can lead to additional b jets in the final state. The reduced backgrounds in b tagged events allows for a more sensitive vector leptoquark search in this category.

The $e\tau_h$ and $\mu\tau_h$ channels are further subdivided into categories depending on the transverse mass between the light lepton and missing transverse momentum vectors as defined in Equation 1.2. The corresponding categories are defined as:

- `Tight-`m_{T}: $m_{\mathrm{T}}(\vec{p}_T^{\,e/\mu}, \vec{p}_T^{\,\mathrm{miss}}) < 40$ GeV;
- `Loose-`m_{T}: $40 \leq m_{\mathrm{T}}(\vec{p}_T^{\,e/\mu}, \vec{p}_T^{\,\mathrm{miss}}) < 70$ GeV.

The majority of the signal events fall within the `Tight-`m_{T} sub-category. The `Loose-`m_{T} category is used improve the signal acceptance for resonant masses of $m_\phi > 700$ GeV.

In the $e\mu$ channel, is also subdivided into three signal categories based of the cuts on the variable D_ζ as defined in Equation 1.3. The three categories are defined as:

- `Low-`D_ζ: $-35 \leq D_\zeta < -10$ GeV;
- `Medium-`D_ζ: $-10 \leq D_\zeta[\mathrm{GeV}] < 30$ GeV;
- `High-`D_ζ: $D_\zeta[\mathrm{GeV}] \geq 30$ GeV.

By design, the majority of signal events are located in the `Medium-`D_ζ sub-category. The Low and High-D_ζ categories are used to catch the tail of the signal distributions. A schematic of all high mass optimisation categories are shown in Figure 1.6.

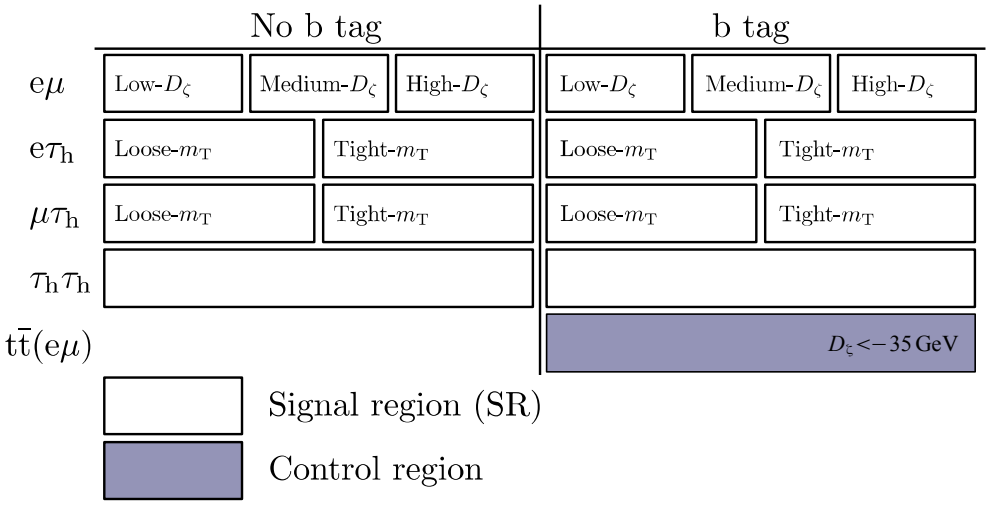


Figure 1.6: Overview of the categories used for the extraction of the signal in the high mass optimisation procedure.

Once all category divisions have been applied, events are drawn in histograms based off a discriminating variable. The discriminating variable used in this analysis is m_T^{tot} and is defined below.

$$m_T^{\text{tot}} = \sqrt{m_T(\vec{p}_T^{\,\tau_1}, \vec{p}_T^{\,\text{miss}})^2 + m_T(\vec{p}_T^{\,\tau_2}, \vec{p}_T^{\,\text{miss}})^2 + m_T(\vec{p}_T^{\,\tau_1}, \vec{p}_T^{\,\tau_2})^2}, \tag{1.4}$$

where τ_1 and τ_2 refer to the visible products of the two tau leptons decays. This variable provides excellent discriminating power between higher mass resonant signals compared to other non-peaking backgrounds whilst still maintaining some separation between signal masses. It is also excellent at separating the high mass non-resonant di-tau signatures where a di-tau mass is unphysical for the signal, due to the use of the p_Ts and in the variable. For the t-channel signal with the mediator has high mass, no significant mass separation is expected in any variable.

The low mass optimisation procedure loosely follows the high mass procedure with a few key difference. Firstly categories that are only sensitive to high mass signals are dropped. This includes the Low-D_ζ and Loose-m_T categories. Each no b tag subcategory is further divided into four bins of reconstructed di-tau visible p_T with bin edges: 0,50,100,200 and ∞. This is not done in the b tag subcategories due to

the lack of statistics in this region. A schematic of the categories used in the low mass optimisation procedure is shown in Figure 1.7.

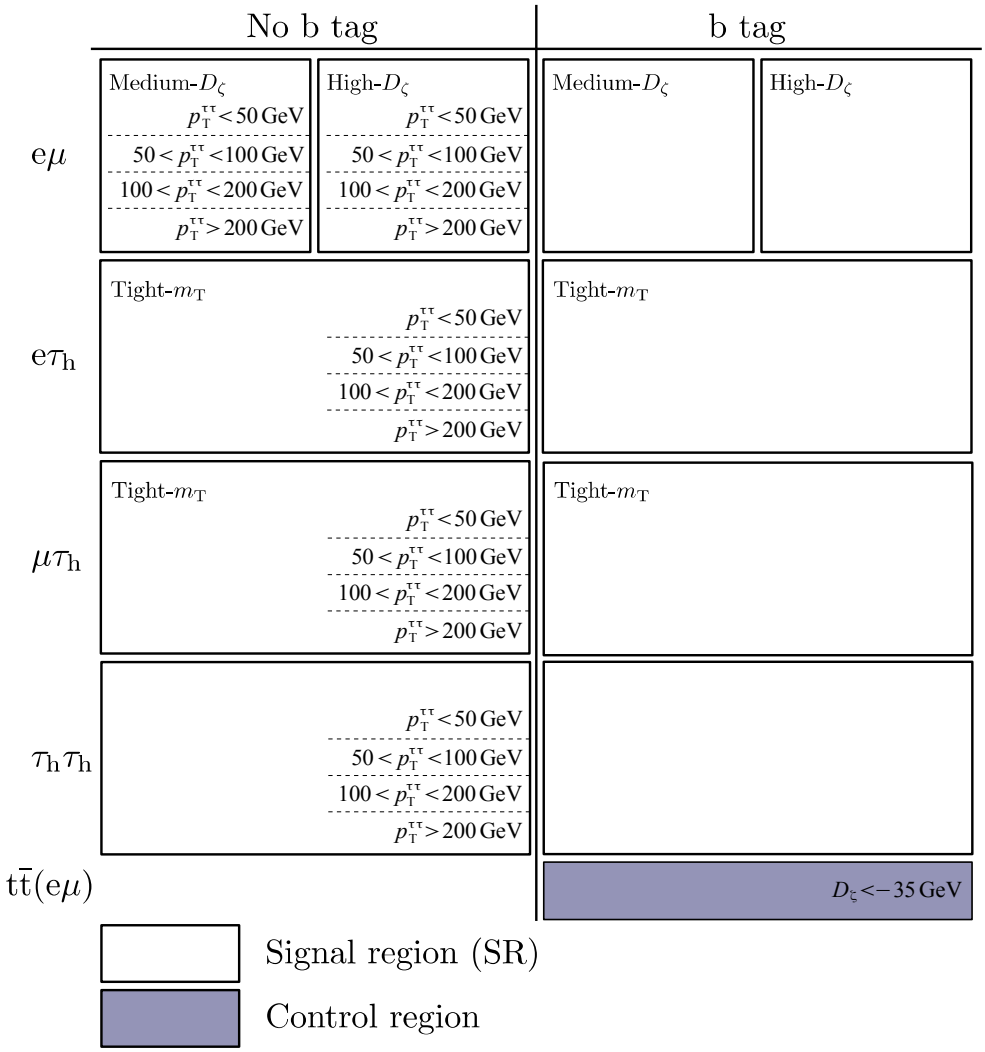


Figure 1.7: Overview of the categories used for the extraction of the signal in the low mass optimisation procedure.

The final difference with the high mass optimisation procedure is the discriminator used, here the reconstructed di-tau mass is used. This helps to separate signal events from the Z boson peak in this region. Examples of the signal mass separations in the $\tau_h\tau_h$ channel with the low and high mass discriminators are shown in Figure **??**

FIGURES OF DISTRIBUTIONS VLQ AND INTERFERENCE GGH AND BBH heavy GGH AND BBH light

Finally the SM Higgs optimisation procedure is taken from the CMS SM $H \to \tau\tau$

analysis. This was previously used for simplified template cross section measurements. This uses a neural-network-based (NN) to obtain the most precise estimates from data of the SM Higgs produced via gluon fusion, vector boson fusion or vector boson associated production. The NN based analysis introduces 26 categories, 8 of which are optimised to pull the out the Higgs boson signal. Although the NN is trained specifically to target events with an SM-like Higgs boson, signal events with differing masses can also enter the NN categories.

1.4 Background Modelling Overview

The analysis considers several backgrounds including Drell-Yan, $t\bar{t}$, W+jets, QCD, di-boson, single-top, and electroweak W and Z bosons production. These are split into a five categories:

i) Events containing only genuine tau leptons.

ii) Events with a jet misidentified as a hadronic tau (jet$\rightarrow \tau_h$) in the $e\tau_h$, $\mu\tau_h$ or $\tau_h\tau_h$ channels.

iii) Events with jets faking both light leptons (jet$\rightarrow l$) in the $e\mu$ channel.

iv) Events from $t\bar{t}$ with a prompt light lepton (e or μ not from a τ decay) and the other object (if there are not two prompt light lepton) is from a genuine tau leptons.

v) Other events. This is a small contribution and hence why it is grouped.

- Non $t\bar{t}$ events with a prompt light lepton (e or μ not from a τ decay) and the other object (if there are not two prompt light lepton) is from a genuine tau leptons.
- Events with a light lepton faking a hadronic tau and the other object (if there are not two light leptons faking a hadronic tau) are reconstructed as prompt light lepton or from genuine tau leptons.
- Events with a jet faking a light lepton and the other object is from genuine tau leptons in the $e\tau_h$, $\mu\tau_h$ or $\tau_h\tau_h$ channels.
- Events with one jet faking a light lepton and the other object from a prompt light lepton in the $e\mu$ channel.

Backgrounds from (i) consists of largely $Z/\gamma^* \rightarrow \tau\tau$ events but there are also smaller contributions from other processes. This background is modelled by a data-simulation hybrid method called the embedding method and this is described in detail in Section 1.6. Group (ii) is dominated by QCD, W + jets and $t\bar{t}$ events with a jet$\rightarrow \tau_h$ misidentification. This is modelled from data by the fake factor method (F_F) and is explained in Section 1.7. Group (iii) is modelled from data to describe QCD multijet contribution to the background in the $e\mu$ channel. The method to obtain this background is described in Section 1.5. The data driven background estimations for (i), (ii) and (iii) contribute >98% of all expected background events in the $\tau_h\tau_h$ channel, >90% in $e\tau_h$ and $\mu\tau_h$ channels and >50% in the $e\mu$ channel.

The final groups, (iv) and (v), are modelled with MC. The $t\bar{t}$ process is separated due to its large contribution to the phase space where a b jet is required.

The W + jets and $Z \to ll$ processes are simulated at leading order (LO) using the MadGraph5_aMC@NLO 2.2.2 (2.4.2) event generator [?, ?] for the simulation of the data taken in 2016 (2017–2018). To increase the number of simulated events in regions of high signal purity, supplementary samples are generated with up to four outgoing partons in the hard interaction. For diboson production, MadGraph5_aMC@NLO is used at next-to-LO (NLO) precision. In each case, the FxFx [?] (MLM [?]) prescription is used to match the NLO (LO) matrix element calculation with the parton shower model. For $t\bar{t}$ [?] and (t-channel) single top quark production [?], samples are generated at NLO precision using POWHEG 2.0 [?, ?, ?, ?]. The POWHEG version 1.0 at NLO precision is used for single top quark production in association with a W boson (tw channel) [?].
When compared with data, W + jets, $Z \to ll$, $t\bar{t}$, and single top quark events in the tW channel are normalised to their cross sections at next-to-NLO (NNLO) precision [?, ?, ?]. Single top quark (t-channel) and diboson events are normalized to their cross sections at NLO precision or higher [?, ?, ?].

1.5 QCD Estimation in the $e\mu$ Channel

The QCD model in the $e\mu$ channel, that attempts to model events where two jets are misidentified as an election muon pair, is taken from data with same sign election muon pair with a transfer factor (F_T). The transfer factor determines differences from the same sign to opposite sign region is calculated from a sideband region with an anti-isolated muon ($0.2 < I^{\mu}_{\text{rel}} < 0.5$). F_T is initially parameterised by the ΔR between the election and muon, and the number of jets in the event, however additional dependencies on the electron and muon p_T enter via a correction.
Good agreement is observed in events with no b jets, for the discriminating variables discussed in Section 1.10, when applying F_T onto same sign events compared to opposite sign events where both regions have an anti-isolated muon. However in events with b jets, an additional correction is needed. This is determined to be 0.75 (differs very slightly between data taking years). As this correction is large, it is validated by switching the light lepton anti-isolation, so that the election is required to have $0.15 < I^{e}_{\text{rel}} < 0.5$. Also events where both light leptons are anti-isolated are looked at. The correction for b tagged events is equivalent in all three regions, and a global average of the three is taken for the final correction.

To understand the physical reason for the large difference in no b tag and b tag events in same sign and opposite pairs, studies were performed on simulated samples. It was observed that the electron muon pair is usually produced from pairs of heavy quarks, $pp \to b\bar{b}$ ($c\bar{c}$). If the two jets are initiated from the heavy quarks there is a large bias towards opposite sign jets due to the opposite signs of the quark anti-quark pair. However if one of the heavy quarks is tagged as a b jet, another object has to be the jet initiator (a radiated gluon for example) and there is therefore no charge preference in the pair. As F_T is originally fit inclusively in numbers of b jets and the 0 bin is dominant, the correction over predicts the opposite sign to same sign ratio and so a large correction is needed as observed.

1.6 Embedding Method

The background for genuine di-τ lepton pairs is modelled via the embedding method. This is a hybrid method that utilises both data and MC techniques to produce high statistic samples, where the bulk of the event comes from data. This minimises both the chance of MC fluctuations and the size of the uncertainties. The background is dominated by $Z \to \tau\tau$ decay however there will be smaller contributions from $t\bar{t}$ and di-boson processes.

The algorithm first selects $\mu\mu$ events from data. The selection is chosen to naturally target the pure $Z \to \mu\mu$ region but still be loose enough to catch events from other processes, so not to introduce a bias on the Z boson mass. Events are required to pass the DoubleMuon trigger with minimum requirements on the invariant mass of the two muons ($m_{\mu\mu}$) and the the p_T of the leading and trailing muon. Also required at the trigger level is a loose association of the track to the PV and a loose isolation in the tracker. Offline objects matched to the trigger muons, are then required to have standard d_z and η selections and originate from a global muon-track, as defined in Section **??**. The muon pair are required to have opposite charge and have $m_{\mu\mu} > 20$ GeV. The fraction of processes within this selection is tested with MC background samples and a QCD model from same sign muon pairs with an extrapolation factor. Approximately 97% of selected events are expected to come from $Z \to \mu\mu$ events with smaller contributions from $Z \to \tau\tau$ ($\tau \to \mu$), di-boson, $t\bar{t}$ and QCD. The di-boson and $t\bar{t}$ relative contributions are greater at higher $m_{\mu\mu}$ and in events with tagged b jets whilst the QCD contribution is largest at lower $m_{\mu\mu}$. The events selected are biased by detector acceptances. Therefore, corrections on the reconstruction and identification efficiencies are performed in muon η and p_T using the "tag-and-probe".

Next, all energy deposits in the detector from the selected muons are removed. This involves removing the hits on global-muon track in the tracker, hits in the muons systems and clusters in the calorimeters that intercept the muon trajectory. Once completed, the selected muons and its kinematic properties are replaced with a tau lepton. To account for the difference in mass between the muon and tau, the muons are boosted into the center-of-mass frame of the di-muon system and then this 4-vector is taken for the tau but boosted back into the laboratory frame. The event simulation is performed from the PV. The tau lepton decay is then simulated with PYTHIA and separate samples are produced for differ $\tau\tau$ decay channels. Only the decay of the tau leptons are then processed through the detector simulation and the

remainder of the $\mu\mu$ event is added back. A schematic of the process is shown in Figure 1.8.

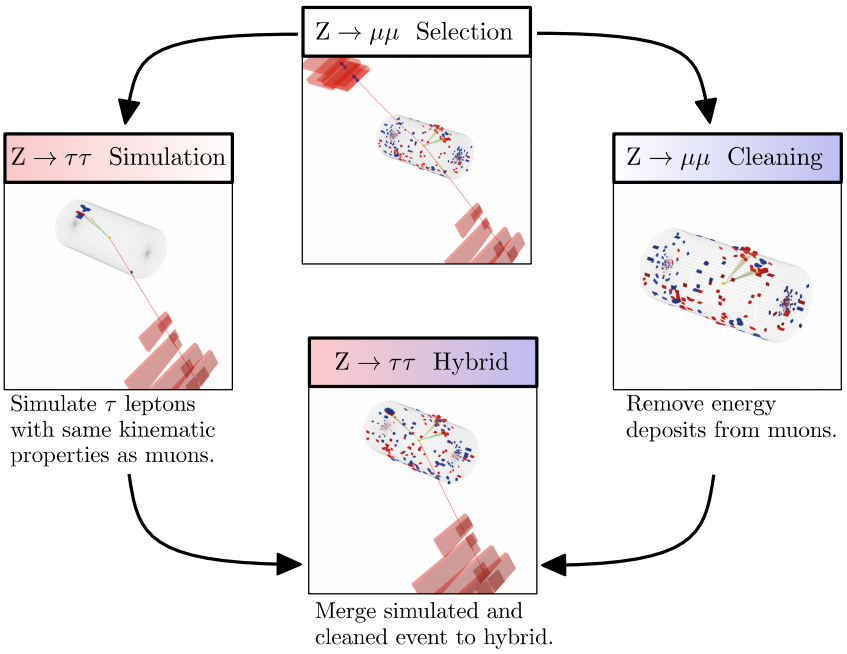


Figure 1.8: Schematic of the embedding method to model genuine di-tau backgrounds from di-muon events in data.

The embedding method is validated on dedicated samples, where the muons from data are replaced by simulated muons instead of taus.

1.7 Fake Factor Method

Backgrounds in which a jet fakes a τ_h can be difficult to model using MC due to the poor description of the jet $\rightarrow \tau_h$ fake rate in simulation. In addition, the small probability of a jet being misidentified as a τ_h necessitates the production of high statistics MC samples at a significant computational expense. These shortcomings motivate the use of data-driven estimates for these processes. One such procedure is the fake factor (F_F) method.

The F_F method utilises regions in the data to model the jet $\rightarrow \tau_h$ background. Firstly, the determination regions, which are jet $\rightarrow \tau_h$ enriched control regions orthogonal to the signal region. It is used to calculate F_F by taking the ratio of number of jet fake events that pass the nominal hadronic tau ID requirement ($N(\texttt{Nominal})$), to the number of jet fake events that fail the nominal hadronic tau ID but pass a looser alternative hadronic tau ID requirement ($N(\texttt{Alternative \&\& !Nominal})$), as shown in Equation 1.5.

$$F_F = \frac{N(\texttt{Nominal})}{N(\texttt{Alternative \&\& !Nominal})}. \tag{1.5}$$

In the remaining text this numerator and denominator are referred to as the pass and fail regions. The derivation of this ratio is done differentially with respect to key parameters that differ in the two regions. Once F_F have been derived it is common to calculate corrections in other sideband regions (a region orthogonal to the signal region) and combine F_F measured from different processes. Finally, the F_F are applied to the application region (AR). This is defined as the SR but with the criteria that the jet fakes fail the nominal hadronic tau ID but pass the looser alternative tau ID requirement. This now models the background from jet $\rightarrow \tau_h$ events in SR.

The following Sections 1.7.1–1.7.4 detail the complexities of how this method is applied to this analysis. For these searches the nominal hadronc tau ID used is the `Medium` $D_{\text{jet}\text{WP}}$ and the alternative hadronic tau ID used is the `VVVLoose` $D_{\text{jet}\text{WP}}$.

1.7.1 Determination Regions

The fake factors are measured separately in each year of data taking period (2016, 2017, 2018), in each channel containing hadronic taus ($e\tau_h$, $\mu\tau_h$, $\tau_h\tau_h$) and in enriched regions of dominant processes that contribute jet $\rightarrow \tau_h$ events. In the $e\tau_h$ and $\mu\tau_h$ channels F_F are measured for three processes: QCD, W + Jets and $t\bar{t}$. In the $\tau_h\tau_h$ channel F_F are measured only for the dominant QCD process. The QCD

process is assumed to produce two jet fakes and so the fake factors is chosen to be calculated from leading p_T hadronic tau candidate only. Section 1.7.4 discusses how single jet fake events in the $\tau_h\tau_h$ channel are modelled.

Each separate measurement region is split into three sideband regions based off two cuts that surround the signal region. These regions are named the `Determination Region` (C), `Alternative Determination Region` (D) and `Correction Region` (B) and are schematically shown in Figure 1.9.

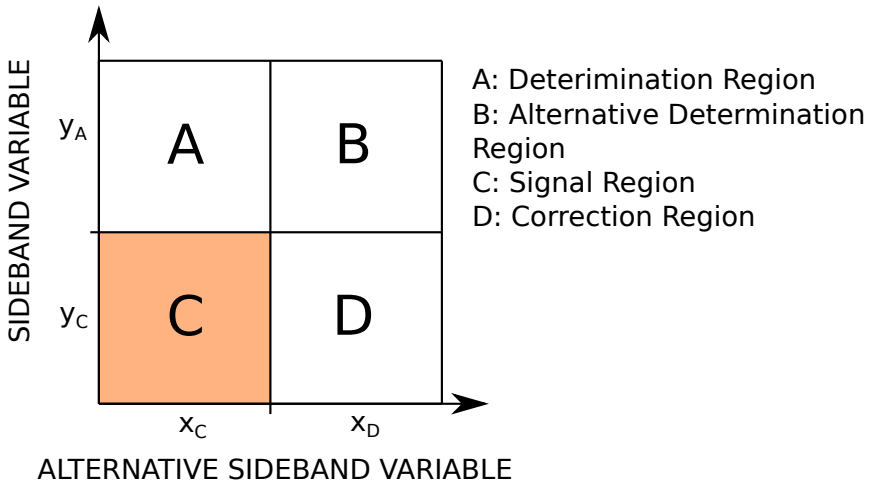


Figure 1.9: Schematic of the regions used for fake factor derivation.

Region A is used to measure and fit fake factors. Region B is an alternative region used to measure and fit fake factor to account for the difference in fake factors between A and C. These alternative fake factors are applied to the fail region in D and corrections are calculated comparing it to the pass region in D. The total fake factor per measurement region is calculated as the fake factors derived in region A multiplied by the correction calculated from region B to D.

The selection for x_C, x_D, y_C and y_A, as defined in Figure 1.9, in each separate measurement region are shown below. These are chosen to balance the number of events and the purity of each background in the region.

i) $\tau_h\tau_h$ QCD
   y_C: The τ_h candidates are required to have the opposite sign.
   y_A: The τ_h candidates are required to have the same sign.
   x_C: The subleading tau passes the `Medium` $D_{\text{jet}}^{\text{WP}}$.
   x_D: The subleading tau fails the `VVLoose` $D_{\text{jet}}^{\text{WP}}$ but passes the `VVVLoose` $D_{\text{jet}}^{\text{WP}}$.

ii) $e\tau_h$ and $\mu\tau_h$ QCD

y_C: The e/μ and τ_h candidates are required to have the opposite sign.

y_A: The e/μ and τ_h candidates are required to have the same sign and the e/μ to have $I_{\text{rel}} > 0.05$.

x_C: The e/μ candidate is required to have $I_{\text{rel}} < 0.15$.

x_D: The e/μ candidate is required to have $0.25 < I_{\text{rel}} < 0.5$.

iii) $e\tau_h$ and $\mu\tau_h$ W + Jets

y_C: The m_T between the e/μ and the MET < 70 GeV.

y_A: The m_T between the e/μ and the MET > 70 GeV and no b jets in the event.

x_C: Data.

x_D: W + Jets MC.

iv) $e\tau_h$ and $\mu\tau_h$ $t\bar{t}$

y_C: Data.

y_A: MC ($t\bar{t}$ in B and W + Jets D).

x_C: $m_T < 70$ GeV.

x_D: $m_T > 70$ GeV and no b jets.

In the $\mu\tau_h$ and $e\tau_h$ channels QCD and W + Jets jet fake events are in general the most significant and contribute with approximately equal weights. $t\bar{t}$ inclusively is small but becomes more significant when searching for events with a b jet. The additional $I_{\text{rel}} > 0.05$ requirement in these channels for QCD is to reduce processes producing genuine leptons and the $N_{\text{b-jets}} = 0$ requirement for W + Jets is to reduce $t\bar{t}$ contamination. It is not possible to define a DR that is sufficiently pure in $t\bar{t}$ events to make a reasonable measurement of $F_F^{t\bar{t}}$ from data. Therefore $F_F^{t\bar{t}}$ are derived from MC. A comparison of the $F_F^{\text{W+jets}}$ measured in data and MC shows only $\sim$ 10–20% differences in the fake rates in data and MC. This observation coupled with the fact that the $t\bar{t}$ contribution is small compared to the other processes means that any bias introduced by using $F_F^{t\bar{t}}$ measured in MC is small compared to the uncertainties on the fake factors, discussed in Section **??**.

1.7.2 Parametrisation

The raw F_F^i take into account dependencies on N_{jets} via the analysis tailed variable $N_{\text{pre b-jets}}$, the p_T of the τ_h candidate ($p_T^{\tau_h}$) and the p_T of the jet matched in ΔR to

the τ_h (p_T^{jet}). $N_{\text{pre b-jets}}$ is defined to map the dependence of F_F^i on N_{jets} and describe the categorising variable $N_{\text{b-jets}}$ well. Although not local to the tau, it helps control other dependencies on the constituents of the event. It is the number of jets in the event with $|\eta| < 2.4$ and $p_T > 20$. These are the same η and p_T thresholds required for a b-jet. The data is split into two bins of $N_{\text{pre b-jets}}$, equal to 0 and greater than 0. It is then further split by the ratio of p_T^{jet} to $p_T^{\tau_h}$. An example of the dependence of these two transverse momentums on the fake factor is shown in Figure 1.10.

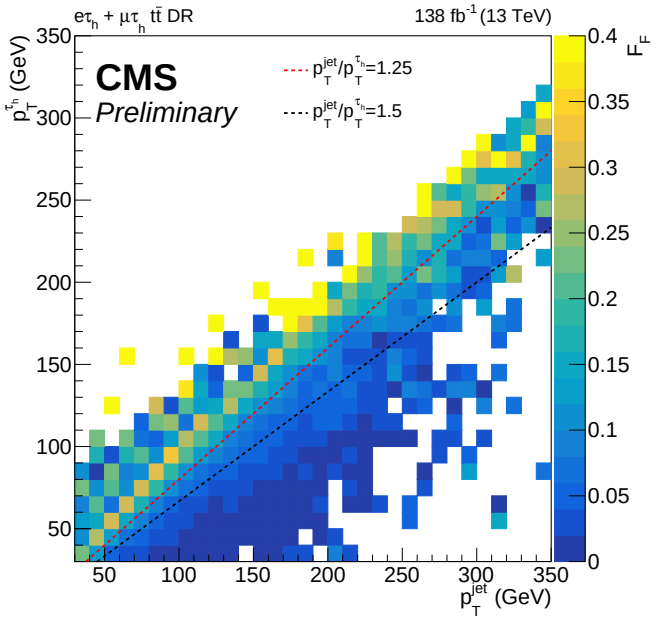


Figure 1.10: A 2D heat map of the fake factors determined from $t\bar{t}$ MC for the full run-2 dataset in the combined $e\tau_h$ and $\mu\tau_h$ channels. This is shown with respect to the hadronic tau p_T and the p_T of the jet matched to the hadronic tau. The ratio of jet to hadronic tau p_T categorisation used is shown split by the dashed lines.

It is motivated by the observation that the fake factor is largest when the p_T^{jet} and $p_T^{\tau_h}$ are closest. The physical motivation for this is when they are close, the hadronic tau candidate is likely to be isolated from any other hadronic activity and so more likely to be identified as a tau. However, when p_T^{jet} is larger than $p_T^{\tau_h}$, the candidate is likely surrounded by other hadronic activity and so more likely to be a jet fake. When p_T^{jet} is less than $p_T^{\tau_h}$, charge pions are likely not close enough to the PV to be clustered into the jet and so the event is more likely to be classified as a jet fake. This will then lead to the fake factor dependence as seen in Figure 1.10.

For all divisions of the phase space, dependence on the $p_T^{\tau_h}$ is fit using the superposition of a Landau and a zeroth order polynomial in the low-p_T region. The fake factors are seen to rise sharply at high-p_T. This increase happens in either the bin

$140 < p_T^{\tau_h} < 200$ GeV or $p_T^{\tau_h} > 200$ GeV. To map this effect, binned values are taken based off the algorithm shown in Figure 1.11 and the fit is used below the minimum bin.

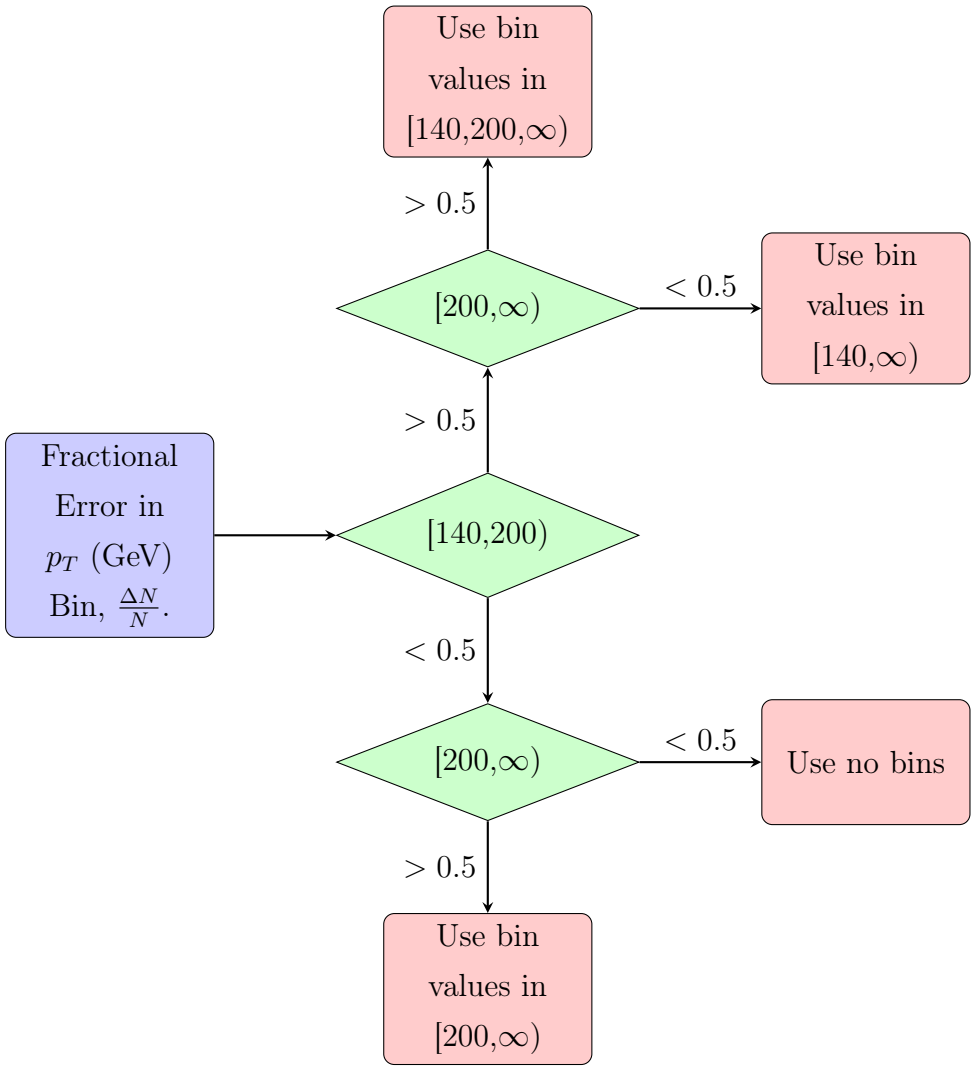
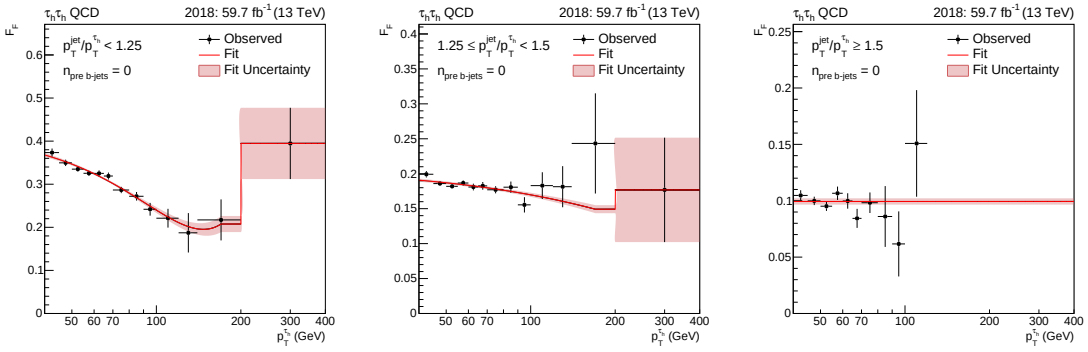


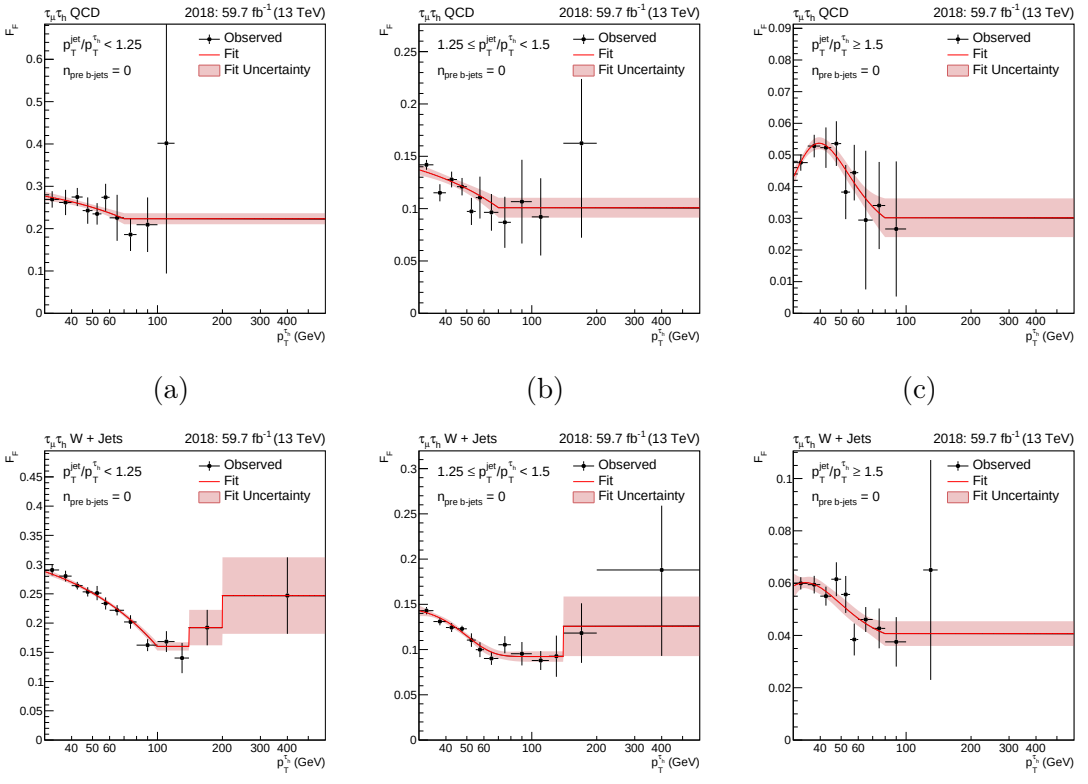
Figure 1.11: Flow chart of the algorithm used to determine where binned values are taken instead of the fit. The blue box represents the input, the green diamonds represent the decisions and the red boxes represent the outputs.

The Landau and zeroth order polynomial fits are flattened at $p_T^{\tau_h}$ values where there is no significant downwards shift or at the final bin. Fake factor fits with respect to $p_T^{\tau_h}$ are shown in Figures 1.12-1.13. The fake factors are highest in the lowest $p_T^{\text{jet}}/p_T^{\tau_h}$ bin and lowest in the highest bin as expected. Otherwise the fake factors fall with p_T in each category until the thresholds used for the high p_T binning algorithm.



(a) (b) (c)

Figure 1.12: Fake factor fits in $\tau_h\tau_h$ channel for the QCD $N_{\text{pre b jets}} = 0$ category with 2018 data. The three jet p_T to hadronic tau p_T categories are shown.



(a) (b) (c)

(d) (e) (f)

Figure 1.13: Fake factor fits in $\mu\tau_h$ channel for the QCD and W + Jets $N_{\text{pre b jets}} = 0$ category with 2018 data. The three jet p_T to hadronic tau p_T categories are shown for each process.

1.7.3 Corrections

In the $\tau_h\tau_h$ channel, the measured F_F^{QCD} are then corrected to account for non-closures in other variable in the `Determination Region`. The only significance non-closures are observed for E_T^{miss} related variables and are largest for events with $N_{\text{pre b-jets}} = 0$. Closure corrections are performed for the variable ΔR in bins of $N_{\text{b-jets}}$. In the $\mu\tau_h$ and $e\tau_h$ channels, the measured F_F^{QCD} and $F_F^{\text{W+jets}}$ are corrected for non-closures observed in the E_T^{miss} variables and $p_T^{e/\mu}$ distributions. A study was performed to determine the nature of these non-closures and it was found that the cause was due to fake E_T^{miss} arising from mismeasurement of the energies of particles in a jet. A diagram of this effect is shown in Figure 1.14.

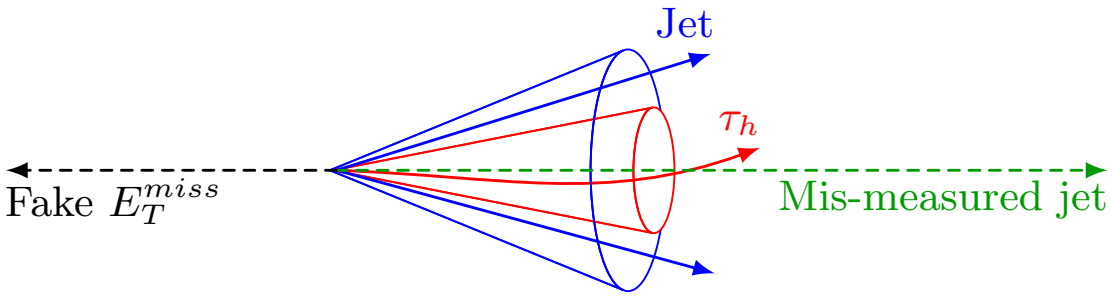


Figure 1.14: Diagram showing how fake E_T^{miss} arises from mismodelling jet constituents in hadronic tau identification.

To correct for this effect, the QCD fake factors are corrected as a function of C_{QCD}, where C_{QCD} is defined as,

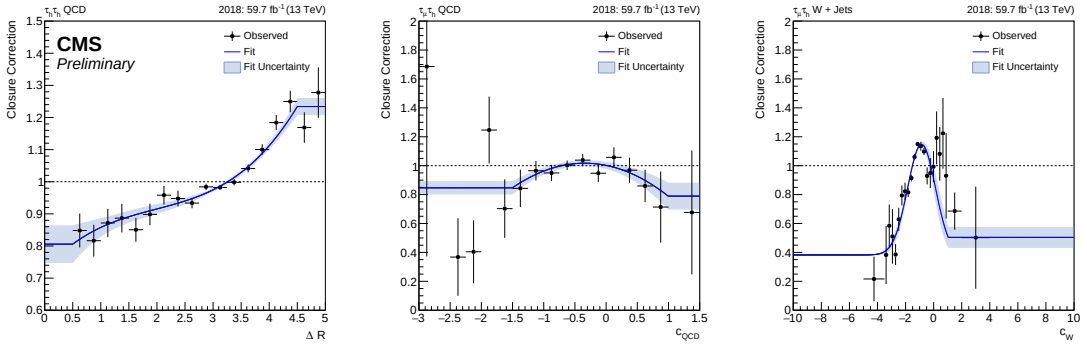
$$C_{\text{QCD}} = \frac{E_T^{\text{miss}} \cos\Delta\phi(\vec{p}_T^{\text{miss}}, \vec{p}_T^{\tau_h})}{p_T^{\tau_h}}. \tag{1.6}$$

where $\Delta\phi(\vec{p}_T^{\text{miss}}, \vec{p}_T^{\tau_h})$ is the separation in the azimuthal angle between the the missing a $\vec{p}_T^{\text{miss}}$ and $\vec{p}_T^{\tau_h}$. The numerator quantifies the missing transverse momentum in the direction of the hadronic tau candidate. Once divided by the τ_h p_T, C_{QCD} is a measure of the fraction of missing to visible hadronic tau transverse momentum aligned with the hadronic tau. For W + jets and $t\bar{t}$ the situation is slightly different due to the presence of genuine missing energy from neutrinos. In this case, the correction variable is modified to approximately subtract the genuine E_T^{miss} from the total. This approximation assumes the neutrino is back-to-back and balanced with the light lepton (which is exactly true for W bosons produce at rest in the transverse

direction). The equation then becomes,

$$C_{\mathrm{W}} = \frac{(E_T^{\mathrm{miss}} + p_T^{e/\mu}) \cos\Delta\phi(\vec{p}_T^{\,\mathrm{miss}} + \vec{p}_T^{\,e/\mu}, \vec{p}_T^{\,\tau_h})}{p_T^{\tau_h}}. \tag{1.7}$$

When either correction variable is separated from 0, a larger quantity of fake E_T^{miss} is expected in the event. In these regions a large correction is needed due to the mis-measured jet energy spectrum shifting the hadronic tau candidate isolation and so shifting the tau identification scores. Examples of these closure corrections are shown in Figure 1.15



(a) (b) (c)

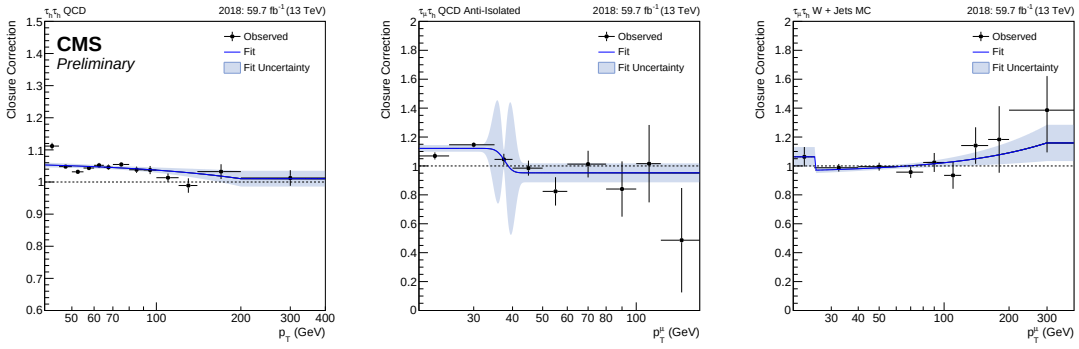
Figure 1.15: Determination region closure correction fits with 2018 data. (a) is the correction parametrised by ΔR in events with $N_{\mathrm{b\ jets}} = 0$ in the $\tau_h\tau_h$ channel. (b) and (c) show the correction for the $\mu\tau_h$ channel parametrised by the specific correction variables defined in Equation 1.6 and 1.7 for QCD and W + jets processes respectively.

After the `Determination Region` is modelled well for all variables of interest, extrapolation corrections from the fake factors derived in B applied to region D are calculated. In the $\tau_h\tau_h$ the correction is parameterised by the p_T of the leading hadronic tau candidate, in the $e\tau_h$ and $\mu\tau_h$ channels it is parameterised by the p_T of the light lepton. Where statistics allow, these corrections are calculated in the high mass optimisation procedure categories. Examples of the extrapolation corrections are shown in Figure 1.16.

1.7.4 Applying Fake Factors

In the $e\tau_h$ and $\mu\tau_h$ channels the F_F^i measured for the different processes are combined into an overall factor, F_F, using

$$F_F = \sum_i f_i \cdot F_F^i, \tag{1.8}$$



(a) (b) (c)

Figure 1.16: Determination region to application region closure correction fits with 2018 data. (a) is the correction moving from same sign to opposite sign tau leptons the parameterised by leading τ_h p_T in events with $N_{\text{b-jets}} = 0$ in the $\tau_h\tau_h$ channel. (b) and (c) show the correction for the $\mu\tau_h$ channel moving from same sign to opposite sign tau leptons and high m_T to low m_T both parameterised by the the muon p_T for QCD and W + jets processes respectively.

where the factor f_i is defined as

$$f_i = \frac{N^i_{\text{AR}}}{\sum_j N^j_{\text{AR}}}, \tag{1.9}$$

which is the fraction of events with a jet $\rightarrow \tau_h$ originating from process i over the total number of jet $\rightarrow \tau_h$ events for all processes in the application region. These fraction of events are estimated with MC, with a QCD model which is extrapolated from same sign tau pairs. It is observed that W + jets is the dominating process in this region, however there are effect from QCD at low m_T and from $t\bar{t}$ in the b tagged categories. These fractions are then multiplied to the relevant corrected fake factor and applied to the fail region in C, with any events which are not jets faking hadronic taus subtracted off with MC.

For the $\tau_h\tau_h$ channel there are two hadronic taus that a jet can fake. For this analysis, the fake factors are only applied the leading hadronic tau candidate failing the tau ID in C. This models all events where the leading hadronic tau candidate is a jet fake. However, this leaves a small fraction of events, where the leading candidate is a genuine tau and the sub-leading candidate is a jet fake. This contribution (mostly from W + jets) is added back with MC.

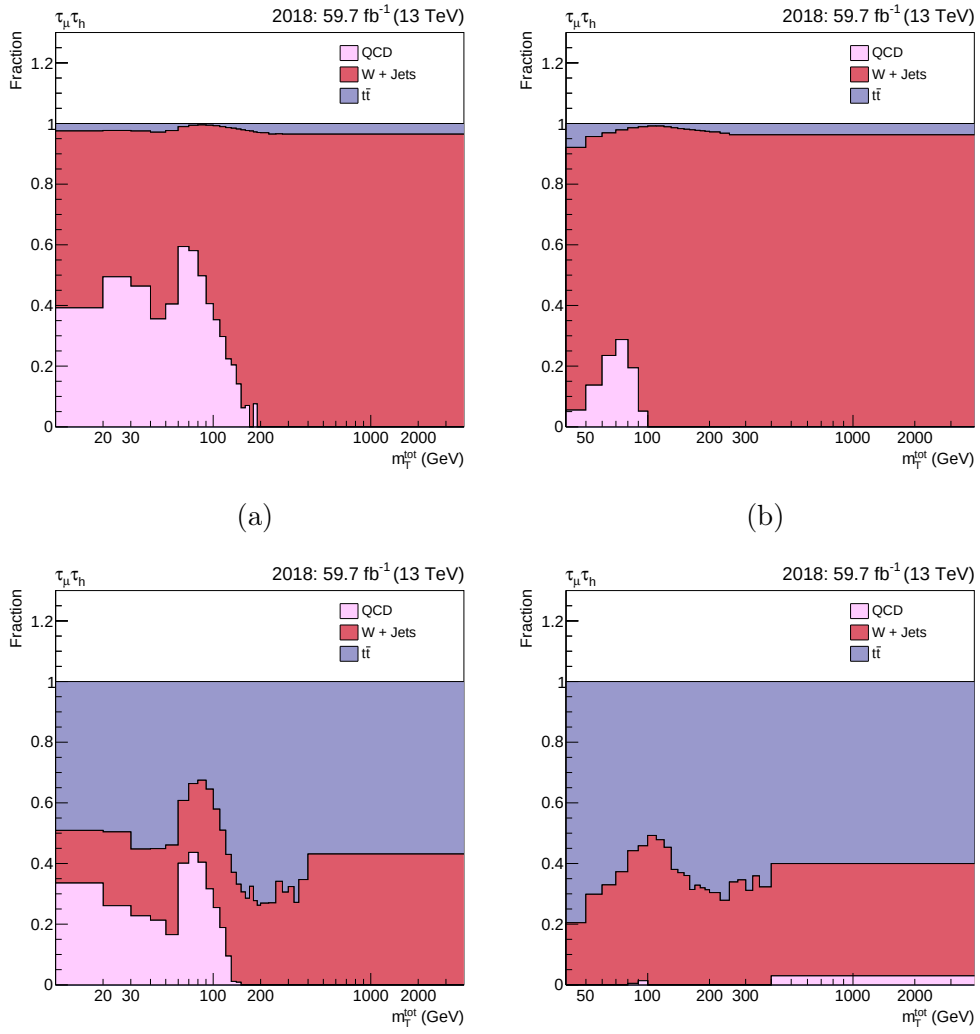


Figure 1.17: The expected application region fractions of the processes in the $\mu\tau_h$ channel. (a) and (b) show the no b tag Tight-m_T and Loose-m_T categories and (c) and (d) show the b tag Tight-m_T and Loose-m_T categories respectively.

1.8 MC Corrections

1.9 Uncertainty Model

The statistical uncertainties are taken into account by the Barlow-Beeston method, described in Ref. []. The systematic model is split into uncertainties based on the online and offline reconstruction of objects and the background and signal modelling. An uncertainty is correlated across channels when it represents a shift on the reconstruction of an object and decorrelated otherwise. It is decorrelated across era of data taking when the shift is derived independently by era. The embedded samples use the same uncertainty scheme as MC but 50% are correlated and 50% uncorrelated with MC uncertainties, because of the shared real data in the measurement.

Hadronic Taus

Uncertainties on the tau triggers are obtained from the fitted scale factors used to derive the corrections for the tau trigger efficiencies. The tau legs of the DoubleTau and lepton+tau cross triggers in different decay mode bins are treated as uncorrelated. For the SingleTau trigger leg, due to limited statistics it is not possible to determine scale factors and uncertainties split by decay mode and therefore a single uncertainty common to all decay modes is applied. The DoubleTau trigger uncertainties are further split the hadronic trigger efficiencies in the p_T regions $<$ 100 GeV and $>$ 100 GeV to allow the fit more freedom to adjust the high p_T regions relative to the low p_T regions. Uncertainties are also applied on the energy scale of the hadronic taus. These uncertainties range between 0.2 and 1.1 %. Finally, an uncertainty on the ID efficiency is placed as a function of p_T in the $e\tau_h$ and $\mu\tau_h$ channels, and of the tau decay mode in the $\tau_h\tau_h$ channels. These vary between 3-9% and is uncorrelated in each variable bin it is derived in. To account for the different anti-lepton discriminator working points, an uncertainty of 3% per tau is applied and treated as uncorrelated between the channels where different $D^{e/\mu}_{\mathrm{WP}}$ are used.

Light leptons

The uncertainty on the trigger efficiencies amounts to 2% per lepton in the $e\tau_h$, $\mu\tau_h$ and $e\mu$ channels. They are basically normalisation uncertainties but implemented as shape uncertainties as they only touch the events triggered by the corresponding cross trigger or single lepton triggers. Uncertainties are also placed on the electron energy scale based off the calibration of ECAL crystals. This information is not reliable for embedding samples and so uncertainties of 0.5-1.25% are placed here. The energy scale variations are negligible and so not included. Another 2% uncertainty

is placed on the ID of any electron or muon in the event.

Jets

Jet energy scale and resolution uncertainties arise from a number of sources. These include limited statistical measurements used for calibration, energy measurement changes due to detector ageing, and bias corrections to address differences between simulation and data. Uncertainty ranges are from sub-percent to (10%). Uncertainties are also placed on the tagging of b jets, which vary from 0–3%.

Leptons misidentified as hadronic taus

Uncertainty shifts are applied for the energy scale of leptons faking hadronic taus parametrised by the p_T of the $l \rightarrow \tau_h$ fake. The magnitude is 1.0% for muons in all eras and for electrons the uncertainties vary between 0.5 and 6.6 %.

Jets misidentified as hadronic taus

The backgrounds with jets misidentified as τ_h are estimated from data with the fake factor method. There are different sources of uncertainty related to this method. The first uncertainties come from subtracting off other background processes with MC to form the determination region. The subtraction is shifted up and down by 10% to determine new weights. Next, statistical uncertainties on all of the fake factor method fits are accounted for, where the binned values are uncorrelated with the rest of the fit. An uncertainty is also placed on the choice of fit function, the shifts are estimated by comparing the fits to a 1st order polynomial fit set to constant above 100 GeV. The final systematic variation are then on the extrapolation corrections by applying the corrections twice and not at all. The size of each systematic uncertainty varies from 0–10%, whilst the statistical element from the fits can be larger in the tails of the distributions.

MET

The MET uncertainties is different dependent on process. For all processes that are not $t\bar{t}$ and diboson, the hadronic recoil response and its resolution are varied within the uncertainties determined during the computation of the recoil corrections. For $t\bar{t}$ and diboson an uncertainty is derived on the energy carried by an unclustered particle []. These uncertainties vary between 0–10%.

Background process specific uncertainties

Uncertainties on the $t\bar{t}$ p_T and DY m_{ll}-p_T reweighting is placed by the applying the correction twice and not at all. An additional uncertainty is placed to cover the $t\bar{t}$ contamination in embedding, where the removed $t\bar{t}$ genuine tau pair is shifted up and down by 10%. Some non-closures are observed in embedded $Z \to \mu\mu$ control samples. Therefore, these non-closures are taken as an additional shape uncertainty as a function of the Z p_T and $m_{\tau\tau}$.

Signal process specific uncertainties

For the ggϕ process, in particular for low mass hypotheses, the variation of `hdamp` parameter of the POWHEG MC generator as well as the μ_R/μ_F scale variations are used to determine the uncertainties on the p_T spectrum of each contribution at NLO QCD to the Higgs boson production via gluon fusion (top, bottom, top-bottom interference). These are also determine from additional samples produced up to the generator level, and applied as event weights dependent on generator level p_T after the parton shower simulation. These uncertainties are included as shape uncertainties as they may affect the shapes of the m_T^{tot} distribution as well as the predicted signal yields.

Luminosity

(1.2 %, 2.3 %, 2.5 %) normalisation luminosity uncertainty is applied to the (2016, 2017, 2018) templates which originate from MC simulation.

Prefiring

Upper and lower bounds are taken from the efficiency maps provided by the L1 DPG and propagated on all MC samples as shape uncertainty for 2016 and 2017. The size of the uncertainty as the weight itself depends on the event topology. In general the uncertainty is at the order of 1%.

1.10 Signal Extraction

A simultaneous binned maximum likelihood fit over all analysis categories is used to extract the results. The likelihood take the form,

$$\mathcal{L}(\text{data} \mid \mu, \theta) = \prod_{i}^{N_i} \text{Poisson}\Big(n_i \mid \sum_{j}^{N_j} g_j(\mu_{ij}) \cdot s_{ij}(\theta) + \sum_{k}^{N_k} b_{ik}(\theta)\Big) \cdot p(\hat{\theta} \mid \theta), \quad (1.10)$$

where i loops through all histogram bins and analysis categories. j and k loop over all signal and background processes of the hypothesis being fit. n_i, s_i and b_i are the data observed, signal and background expectation respectively in each bin. θ represents the set of nuisance parameters (corresponding to the systematic uncertainties as detailed in Section 1.9) that parametrise the signal and background modelling. μ are rate parameters and $g(\mu)$ are scaling functions that scale to a signal to a signal hypothesis. The form of the Poisson probabilities are,

$$\text{Poisson}(n \mid x) = \frac{x^n e^{-x}}{n!}. \quad (1.11)$$

Finally, $p(\hat{\theta} \mid \theta)$ represents the probability density function (pdf) of each nuisance parameter (θ) with respect to the initial value of the parameter ($\hat{\theta}$).

These come in two forms, the first is for uncertainties that only affect the normalisation of the process and are modelled by log-normal pdfs. The second is for uncertainties that affect the shape of the distribution, these are assigned Gaussian pdfs. The $\pm 1\sigma$ shifts for each shape variations are derived and vertical morphing [] is used to interpolate and extrapolate within and outside the shifts. Both pdfs are depend on the mean (μ) and standard deviations (σ) and the functional forms are shown in Table 1.3.

Gaussian	Log-normal
$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$f(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)$

Table 1.3: pdfs used for nuisance parameters.

The following subsections discuss the results of many such fits. The key fits to understand the results are the background-only fit and the signal-plus-background fits. It is important to mention here that the SM Higgs boson processes are treated

as a background. The background-only fit is performed with N_j (number of signal processes) set to 0. For all signal-plus-background fits, the fit is done with respect to a single mass hypothesis, however within this mass hypothesis can be a number of signal processes. The model independent resonance search has separate $gg\phi$ and $bb\phi$ signal modes and so two rate parameters $\mu_{gg\phi}$ and $\mu_{bb\phi}$ are needed. When the samples are originally scaled to the cross section times branching ratio ($\sigma \times B(\phi \to \tau\tau)$) of 1 pb and $g(\mu) = \mu$ for both processes, $\mu_{gg\phi}$ and $\mu_{bb\phi}$ represent the $\sigma \times B(\phi \to \tau\tau)$ with units of pb. To avoid negative signal strengths, μ will only be taken to be positive. Also used in the following subsections, is a signal-plus-background channel/category compatibility fit. In this fit the signal processes and rate parameters are further split in each channel or category utilising index i in Eq. 1.10. This is used to determine the compatibility of the results in different decay channels and analysis categories. In this case, μ is allowed to take negative values to help fully understand the fits to data in each channel or category.

The vector leptoquark search has two signal modes; the t-channel interaction and the interference with Drell-Yan however as this is a model dependent interpretation of these results both these rate parameters scale together. The scaling functions differ between the two processes with $g_{\text{t-channel}}(\mu) = \mu^4$ and $g_{\text{interference}}(\mu) = \mu^2$ to mimic how the cross sections of each process scales. When the initial samples are scaled to cross section at $g_U = 1$, μ corresponds to the coupling g_U.

For the MSSM interpretation of the results, there are three Higgs bosons to consider in the signal model (h, H and A) produced via both gluon fusion and in association with b quarks. The gluon fusion samples are also split into the separate loop contributors, so the kinematic properties can be properly scaled to MSSM prediction, as described in Section 1.1.1. The SM-like Higgs boson is considered in the MSSM signal model to monitor differences in the observed Higgs boson prediction between the MSSM and the SM. In each benchmark scenario chosen, the signal prediction depends only on m_A and $\tan\beta$ and the scaling to cross section is shown in Eq. 1.1. As the potential scaling functions for MSSM interpretations are not necessarily smooth one-to-one mappings, the likelihood is tested for individual points on the m_A-$\tan\beta$ parameter space. At each point, the MSSM Higgs bosons are scaled to the theory predicted cross section times branching ratio. To test the MSSM hypothesis over the SM hypothesis, the single rate parameter μ is used and only allowed to take values of 1 (MSSM) and 0 (SM) with $g(\mu) = \mu$. As the SM Higgs boson is added to the background modelling and the MSSM prediction of the observed Higgs boson is

added to the signal model when $\mu = 1$, the SM Higgs boson prediction must then be subtracted from the signal model.

The confidence intervals in the best fit results are given by the $-2\Delta \ln \mathcal{L}$, where $\Delta \ln \mathcal{L}$ is the difference between $\ln \mathcal{L}$ of the best fit model and the test value of μ. The 68% and 95% confidence regions with two degrees of freedom (as in the model independent resonant search) are determined by $-2\Delta \ln \mathcal{L} = 2.28$ and 5.99 respectively.

Upper limits are placed using the modified frequentest approach [] with profile likelihood ratio used for the test statistic defined below.

$$q_\mu = -2 \ln \left(\frac{\mathcal{L}(\text{data}|\mu, \hat{\theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta}_{\hat{\mu}})} \right), 0 \leq \hat{\mu} \leq \mu, \tag{1.12}$$

where $\hat{\mu}$ and $\hat{\theta}_{\hat{\mu}}$ are the best fit values of μ and θ. $\hat{\theta}_\mu$ are the values of θ that is maximised by the likelihood for a tested value of μ. The bounds on $\hat{\mu}$ are to ensure a positive signal strength with a one-sided confidence interval. The probability of $q_\mu \geq q_\mu^{\text{obs}}$ is,

$$\text{CL}(\mu) = \int_{q_\mu^{\text{obs}}}^{\infty} f(q_\mu \mid \mu, \theta_\mu^{\text{obs}}), \tag{1.13}$$

where $f(q_\mu \mid \mu, \theta_\mu^{\text{obs}})$ is the pdf of q_μ. CL_b and CL_{s+b} are then defined by the relevant background-only and signal-plus-background fits. CL_s is defined as the ratio of CL_{s+b} and CL_b and then upper limits are placed at the confidence level of $1 - \text{CL}_s$. The $f(q_\mu \mid \mu, \theta_\mu^{\text{obs}})$ are determined using the asymptotic approximation [] and results are cross-checked and deemed consistent with toy MC datasets.

If a deviation from the background expectation is observed, the size of the deviation is quantified by a significance. To begin quantifying the excess, a p-value is determined, μ is replaced with 0 in the test statistic, to test rejection of the background-only hypothesis in favour of the signal-plus-background hypothesis. The p-value, p_0 is then,

$$p_0 = \int_{q_0^{\text{obs}}}^{\infty} f(q_0 \mid 0, \theta_\mu^{\text{obs}}). \tag{1.14}$$

p_0 is uniformly distributed between 0 and 1 for the background-only hypothesis and so the probability and significance of rejecting the background-only hypothesis can be found.

1.11 Postfit Plots

Figures 1.18 and 1.19 show the unblinded distributions in the most sensitive analysis categories. For simplicity, the $e\tau_h$ and $\mu\tau_h$ channels have been combined. Figure 1.18 shows the distributions of the $m_{\tau\tau}$ discriminator in the no b tag low mass optimisation categories. A signal-plus-background fit for a model independent gluon fusion resonant mass hypothesis of 100 GeV is shown and the changes in the background modelling when using a background-only fit is displayed in the ratio. Figure 1.19 shows the distributions of the m_T^{tot} discriminator in the high mass optimisation categories. A background-only fit is shown for the stacked background and best fit signal hypotheses for the model independent resonances ggϕ and bbϕ 1.2 TeV and VLQ (BM 1) 1 TeV mass points from separate signal-plus-background fits are displayed.

In the low mass optimisation categories, a small excess of events is observed on the Z boson peak in the no b tag categories and reasonable agreement is observed in the b tag categories. The excess of events are distributed in $m_{\tau\tau}$ between 80 and 120 GeV. A signal-plus-background hypothesis is best fit with a 100 GeV ggϕ with a cross section times branching ratio of 5.8 pb. In this same fit the bbϕ process is constrained by the b tag categories to give a signal yield of 0. A background-only fit is also performed on the data, it is observed that this can only partly explain the differences observed between background and data. Even after a background-only fit there is still an small excess of data events over the Z boson peak.

In the high mass optimisation categories, another small excess is observed in high m_T^{tot} bins, particularly in the most sensitive no b tag categories. This excess is best fit by a model independent gluon fusion resonant mass at 1.2 TeV with a cross section times branching ratio of 3.1 fb. There are no considerable differences observed in background modelling between signal-plus-background and background-only fits. This is as the uncertainties in these bins are more statistically dominated and the majority of the systematic uncertainties are constrained in the bulk of the distribution. Good agreement is observed in the rest of the distribution. There is a very small deviation in the b tag categories, but as this can also be explained by a ggϕ signal, the bbϕ signal is heavily constrained and so largely does not contribute to the signal-plus-background fit of the excess. Similar to the bbϕ signal, the VLQ BM 1 signal is constrained by the results in the no b tag categories, leading to a non-zero best fit signal strength, but cannot explain the excess in the no b tag categories.

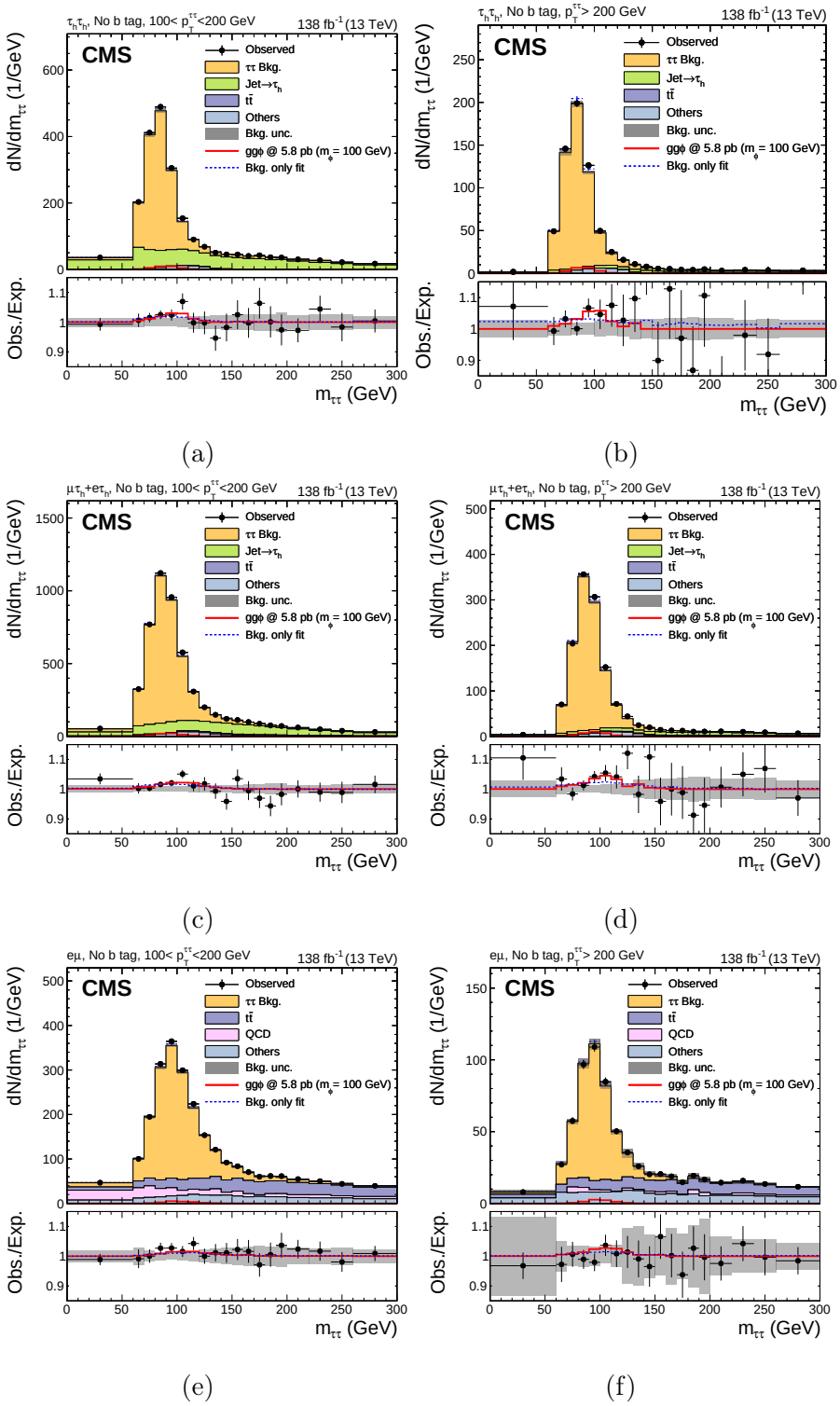


Figure 1.18: Distributions of $m_{\tau\tau}$ in the no b tag second highest (left) and highest (right) p_T category for the $\tau_h\tau_h$ (top), the combined $e\tau_h$ and $\mu\tau_h$ (middle) and the $e\mu$ (bottom) channels. The solid histograms show the stacked background predictions after a signal plus background fit to the data. The best fit gluon fusion signal for $m_\phi = 100$ GeV is shown by the red line. Also shown by a blue dashed line on the bottom pad is the ratio of the background predictions for the background only fit to the signal plus background fit

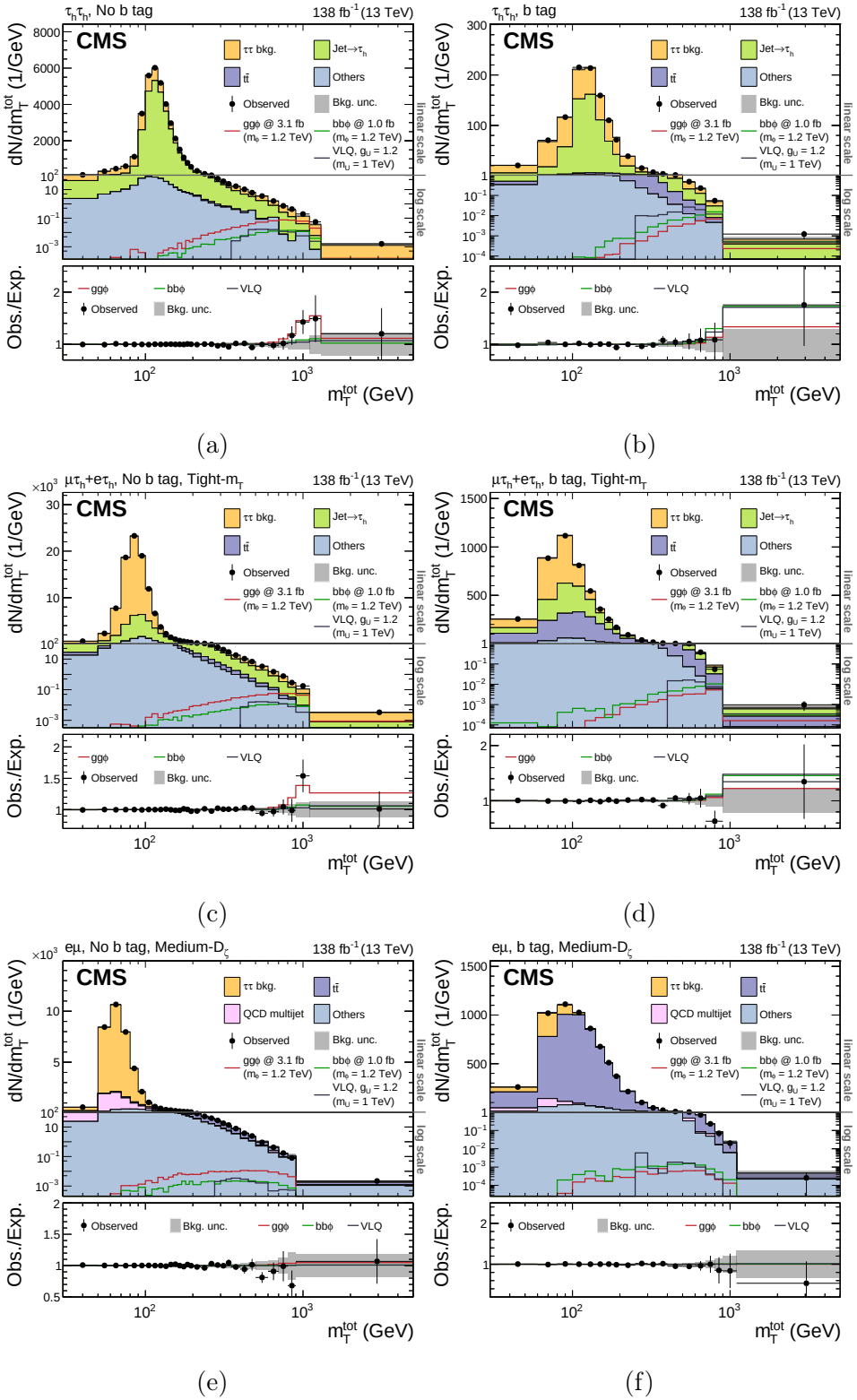
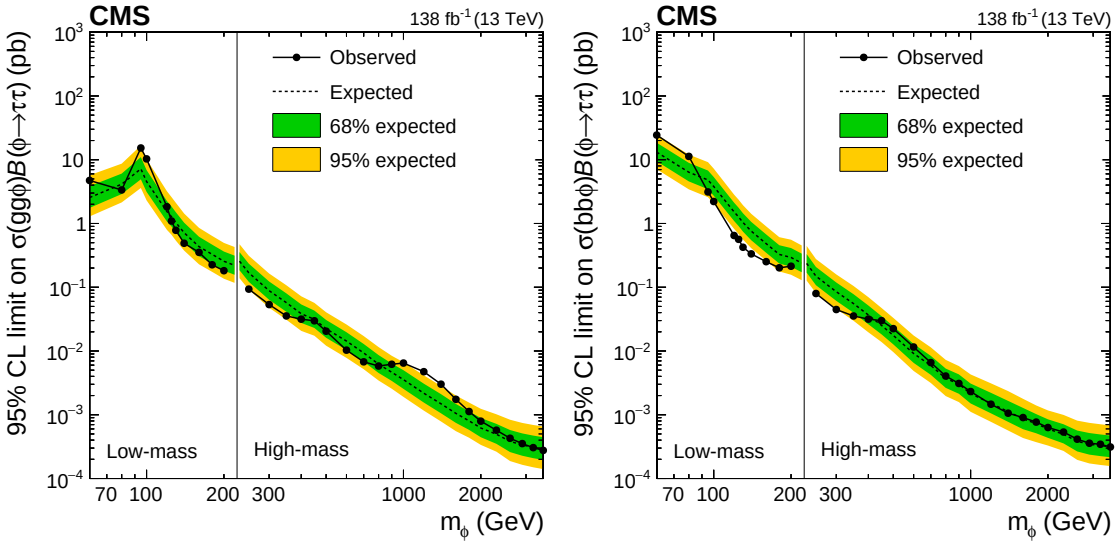


Figure 1.19: Distributions of m_T^{tot} in the $\tau_h\tau_h$ no b tag (a) and b tag (b) categories, the combined $e\tau_h$ and $\mu\tau_h$ no b tag (c) and b tag (d) Tight-m_T categories and the $e\mu$ no b tag (e) and b tag (f) Medium-D_ζ categories. The solid histograms show the stacked background predictions after a background only fit to the data. The best fit gluon fusion signal for $m_\phi = 1.2$ TeV is shown by the red line, b associated production and U_1 signals are also shown for illustrative purposes.

1.12 Model Independent Results

1.12.1 Limits

95% CL limits are set on the assumption of absence of a signal for the search for a ggϕ or bbϕ resonance and shown in Figure 1.20. In each case, the other process is allowed to float freely in the fit. The excesses observed in the postfit distributions act to weaken the observed limit compared to the expected limit at 100 GeV and 1.2 TeV, as more data was observed than expected. For ggϕ production the expected limits flatten under 100 GeV, due to difficulty of separating signal from the Z boson at this mass. Both sets of limits vary from $\mathcal{O}$(10 pb) at 60 GeV to 0.3 fb at 3.5 TeV.

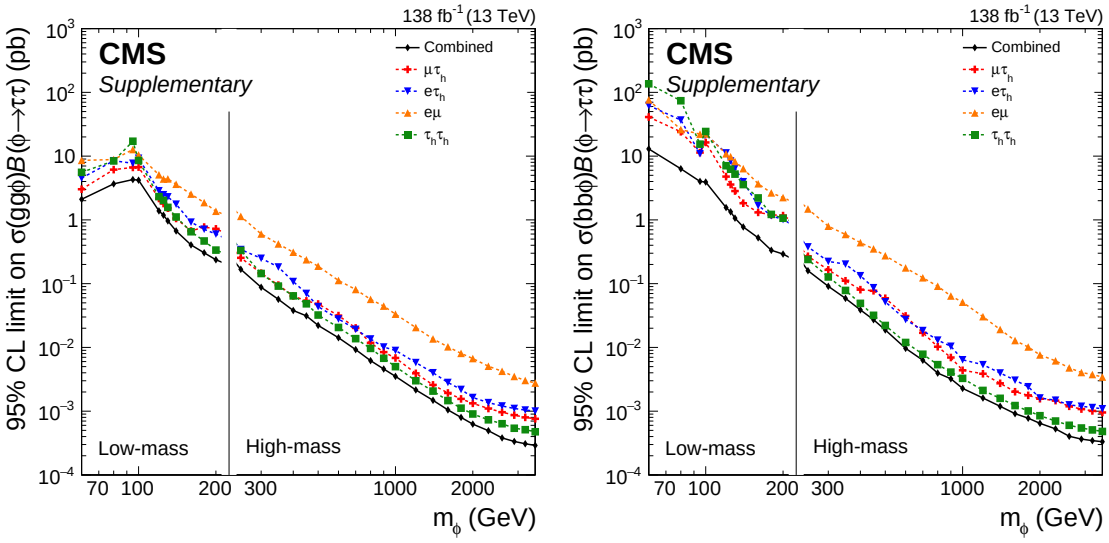


(a) (b)

Figure 1.20: Expected (dashed line) and observed (solid line and dots) 95% CL upper limits on the product of the cross sections and branching fraction for the decay into τ leptons for (a) ggϕ and (b) bbϕ production in a mass range of 60 $\leq m_\phi \leq$ 3500 GeV. The dark green and bright yellow bands indicate the central 68% and 95% intervals for the expected exclusion limit.

95% expected limits are drawn on the fit to each di-tau decay channel individually and are shown in Figure 1.21. This gives a measure of the sensitivity of each channel. In the high mass optimisation categories, the combined limit is heavily dominated by the $\tau_h\tau_h$ channel. This is mostly driven purely by branching fraction, as all channels in this mass range have similar signal separation ability. In the high mass optimisation categories, the combined limit is more a contribution of all channels. In the $\tau_h\tau_h$ channel in this region, the QCD multijet background is the largest fraction of any non $Z \to \tau\tau$ backgrounds in all channels and so the limit for this channel is weakened and the other channels contribute to the combined limit more. The high

DoubleTau trigger p_T thresholds (chosen because of the QCD multijet background) also lowers the signal acceptance in the $\tau_h\tau_h$ channel.



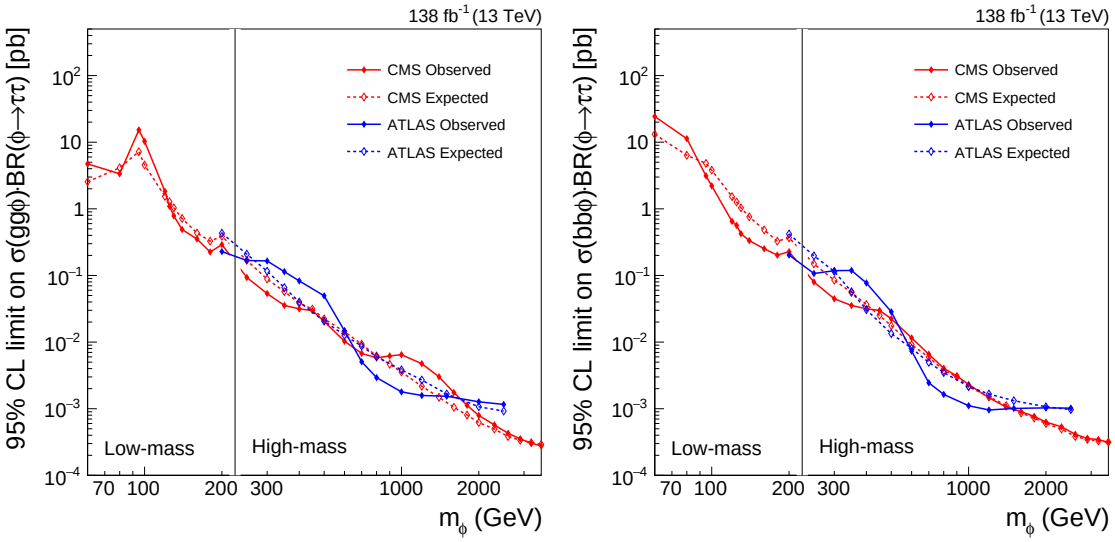
(a) (b)

Figure 1.21: Comparison of the expected 95% CL upper limits on the product of the cross sections and branching fraction for the decay into τ leptons for (left) ggϕ and (right) bbϕ production, split by the $\tau\tau$ decay products fit individually.

A comparison of the limits are also made with the ATLAS experiment and in particular the results presented in Ref. [?]. This ATLAS search looks for the same signal but over a smaller mass range, from 200 GeV to 2.5 TeV. Plots showing the comparison of the expected and observed limits for ggϕ and bbϕ are shown in Figure 1.22. The expected limits from the CMS and ATLAS results are roughly compatible over the shared mass range, except at high mass where the extra statistics from the $Z \to \tau\tau$ samples compared to MC allow for lower background uncertainties and hence a stronger limit. The ATLAS result observed no excess of events compatible with ggϕ signal at 1.2 TeV, in fact a small deficit was observed. Also, ATLAS observed local excesses at 400 GeV of 2.2σ for ggϕ and 2.7σ for bbϕ. None of these excesses are consistent between the ATLAS and CMS results. The ATLAS search does not stretch to the mass of the low mass CMS excess and so cannot be used a cross-check for this.

1.12.2 Significance and Compatibility

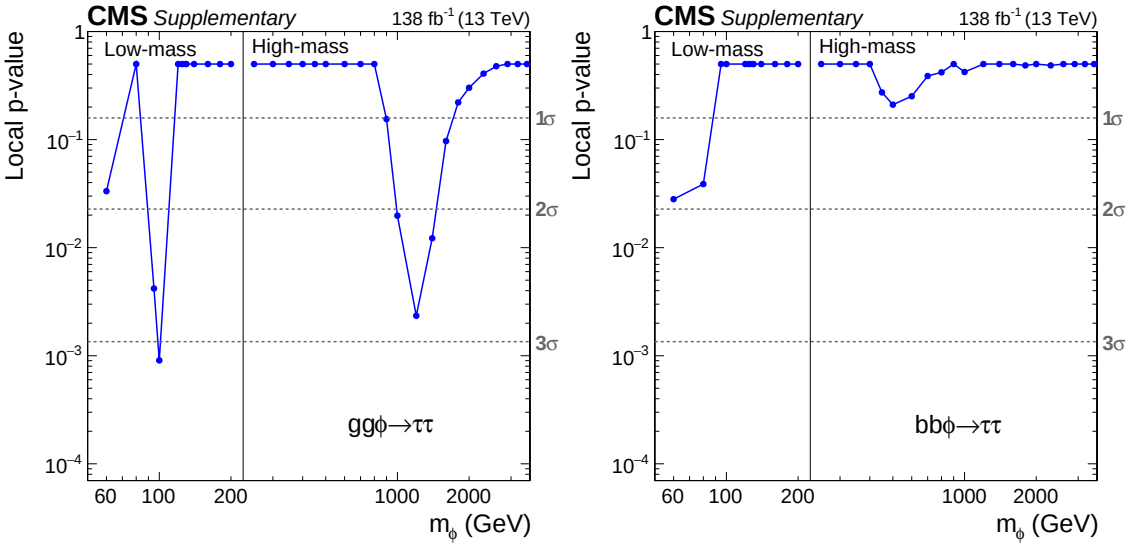
The p-values and significances at each model independent signal hypothesis are calculated as described in Section 1.10 and shown in Figure 1.23. Identical to the model independent limits, the ggϕ or bbϕ process is allowed to float freely if not



(a) (b)

Figure 1.22: Comparison of the expected 95% CL upper limits on the product of the cross sections and branching fraction for the decay into τ leptons for (left) ggϕ and (right) bbϕ production, split by the CMS result detailed in this thesis and the ATLAS result from Ref. [?].

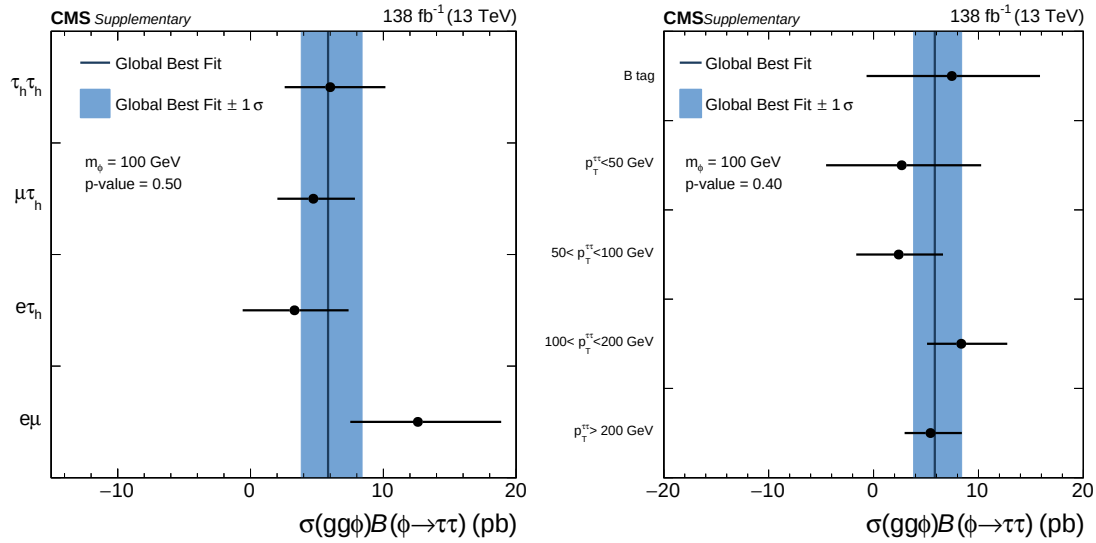
the parameter of interest. The excesses for the ggϕ process peak at 100 GeV and 1.2 TeV and quantify to a local (global) significance of 3.1σ (2.7σ) and 2.8σ (2.2σ). There are also excesses at neighbouring mass points (particularly at high mass), however this is consistent with the mass resolution of the fitted templates for the central values. No deviations beyond 2σ are observed for bbϕ production.



(a) (b)

Figure 1.23: Local p-value and significance of a ggϕ (left) and bbϕ (right) signal as a function of m_ϕ.

As many different decay channels and categories are used to extract these significances, the signal strength is studied in each channel and category. This is done via compatibility fits as described in Section 1.10, where the signal strength parameter in each channel/category is decoupled. No statistically significant differences are observed in the best fit signal strength in any decay channel or category fit and p-values between each channel or category fit are always above 0.05. Figure 1.24 shows the results of the compatibility fits in the low mass optimisation categories split by di-tau decay channels and p_T bins fit and Figure 1.25 shows the compatibility fits in the high mass optimisation categories split by di-tau decay channels. The low mass signal strengths are no more dominant in any p_T region than another. In both low and high mass cases, the signal strengths are consistent across di-tau decay channels. There is a small shift in the high mass $e\mu$ categories to a negative signal strength, these categories have little to no sensitivity to this signal in comparison to others and a small deficit is observed in data, resulting in fits for a negative signal strength with a large uncertainty.



(a) (b)

Figure 1.24: Compatibility plots of the low mass excess split into analysis channels (a) and categories (b). In each case the fitted signal strength is decoupled in the bin shown on the plot.

1.12.3 2D Likelihood Scans

As the model independent search looks for 2 signal modes at each mass point, the results for both processes happening simultaneously are studied. This is done in the form of 2D likelihood scans. The best fit cross section times branching fractions of

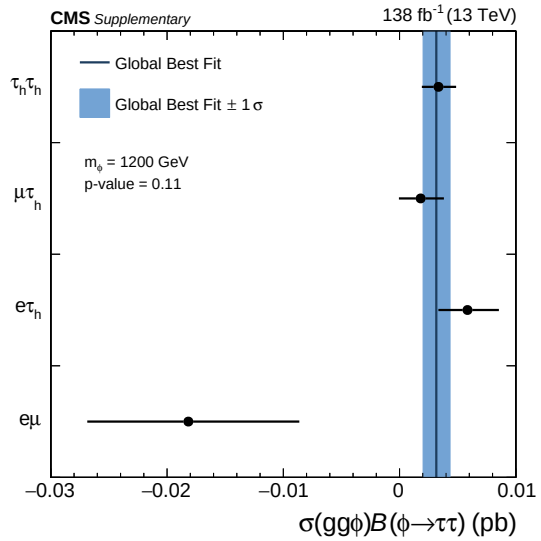


Figure 1.25: Compatibility plots of the high mass excess split by analysis channels. In each case the fitted signal strength is decoupled in each channel.

each process and the 95% and 68% confidence intervals are shown for a number of different mass scenarios in Figure 1.26. The SM prediction in all plots is at (0,0). These results highlight how the excesses at 100 GeV and 1.2 TeV are dominated in the phase space in which ggϕ and not bbϕ signals are allowed. In the 60 GeV example, there are smaller deviations in both ggϕ and bbϕ and so the SM background is over 2σ away. Otherwise, signal strengths are completely compatible with the background expectation.

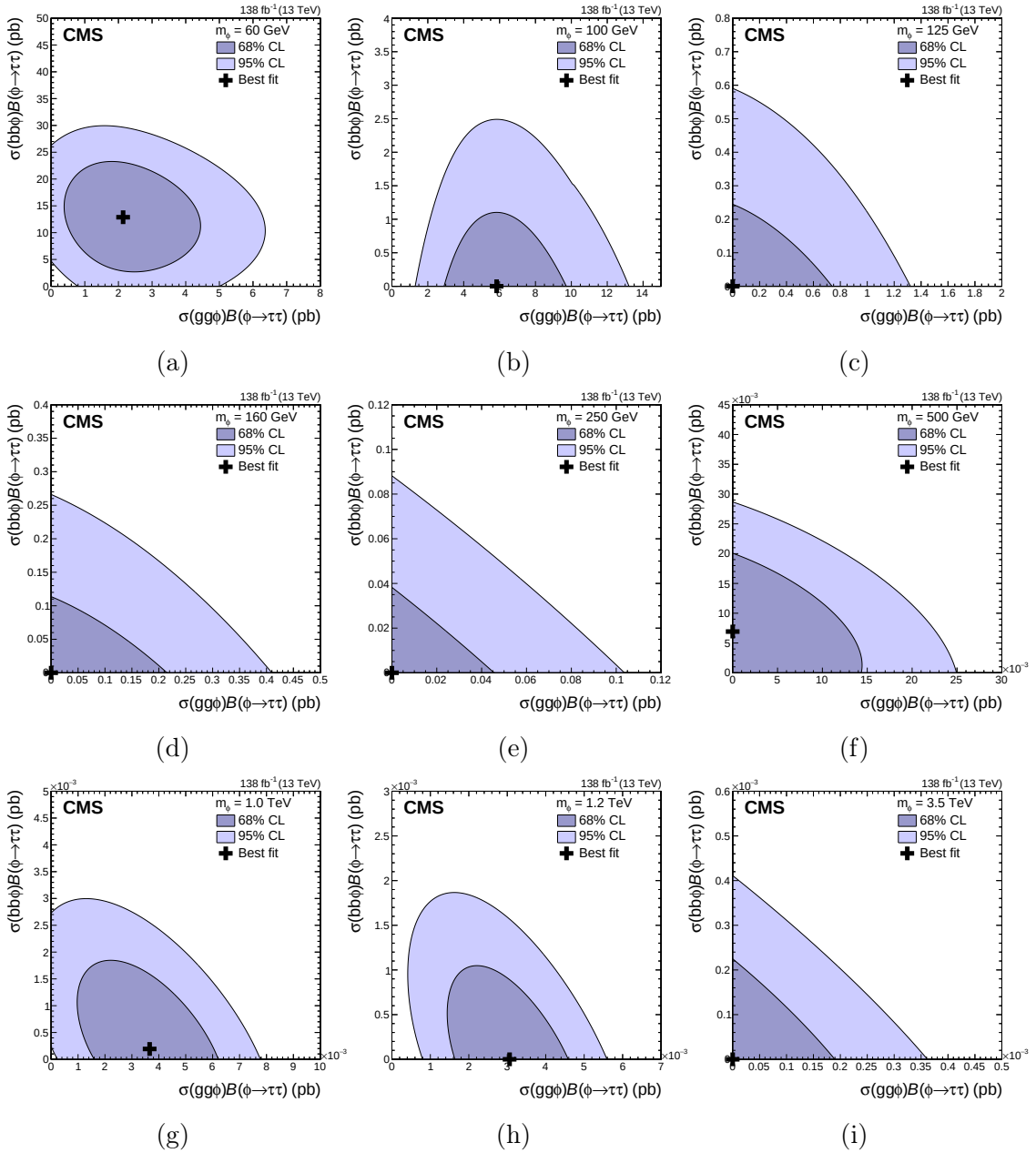


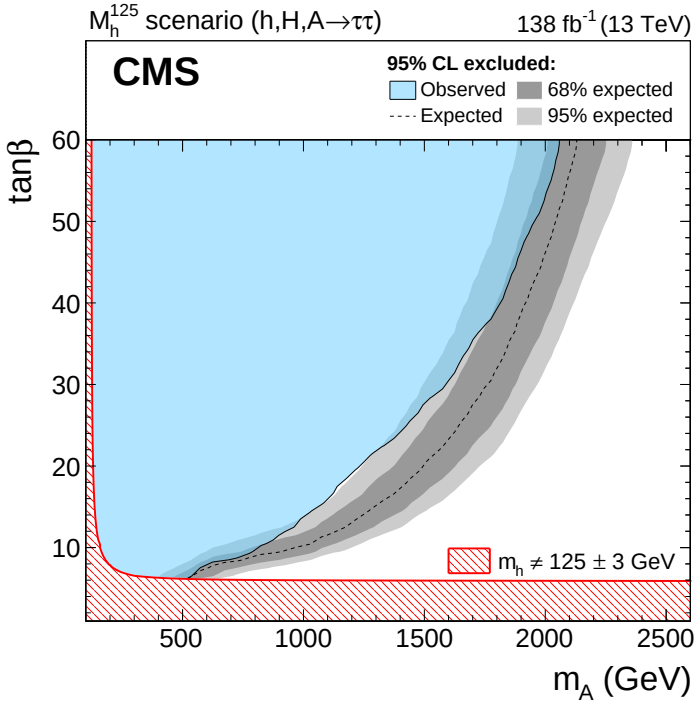
Figure 1.26: Maximum likelihood scans, including 68% and 95% CL contours obtained from the signal likelihood for the model-independent search. The scans are shown for selected values of m_ϕ between 60 GeV and 3.5 TeV.

1.13 Model Dependent Limits

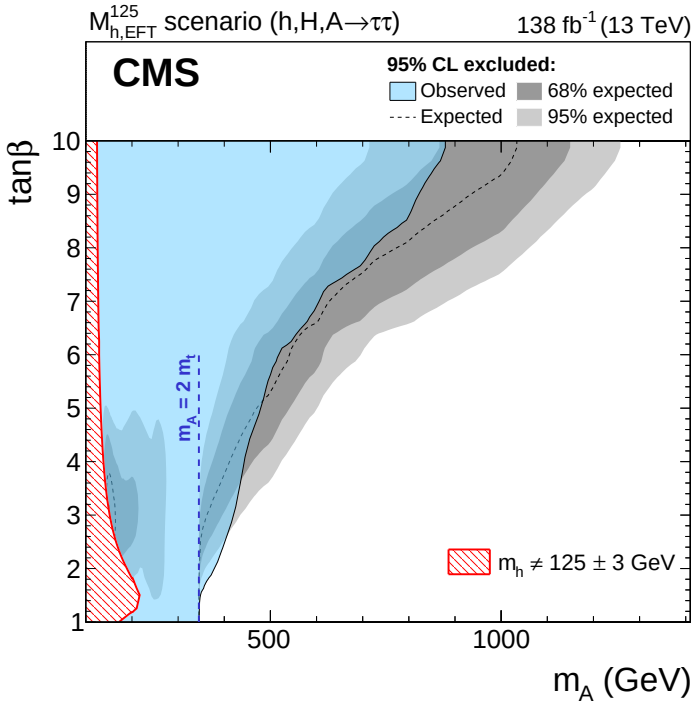
The exclusion contours for two benchmark scenarios of the MSSM, M_h^{125} and $M_{h,EFT}^{125}$, are presented in Figure 1.27. The red hatched regions denote areas where m_h is inconsistent with the observed SM Higgs boson mass within a ± 3 GeV boundary. For low values of $\tan\beta$, higher values of the additional SUSY particle masses, denoted as m_{SUSY}, are needed to explain a mass of approximately 125 GeV for the Higgs boson. In the M_h^{125} scenario, m_{SUSY} is fixed, and the predicted value of m_h is below 122 GeV. In contrast, the $M_{h,EFT}^{125}$ scenario adjusts m_{SUSY} to satisfy the required value of m_h for each point in $(m_A,\ \tan\beta)$ individually, accounting for the logarithmic corrections associated with the large values of m_{SUSY} using an effective field theory approach. The red hatched region in Figure 1.27 (b) indicates that the required values of m_{SUSY} exceed the GUT scale at very low values of m_A in this scenario. The Higgs boson masses, mixing angle α, and effective Yukawa couplings were calculated using FeynHiggs, and branching fractions for the decay into tau leptons and other final states were obtained from a combination of the FeynHiggs and HDECAY, following the prescriptions in Refs.[?, ?, ?], for the scenarios described in Ref.[?].

For the $M_{h,EFT}^{125}$ scenario, the sensitivity sharply drops at $m_A = 2m_t$ due to a drop in the branching fractions for the decay of A and H into tau leptons, when the A and H decays into two on-shell top quarks becomes kinematically accessible. Both scenarios are excluded at 95% CL for $\lesssim 350$ GeV. For $\lesssim 250$ GeV, most of the ggH/A events do not enter the no b tag categories due to the $m_{\tau\tau} > 250$ GeV requirement. In this parameter space, the sensitivity to the MSSM is driven by the measurements of the observed Higgs boson, even though H and A still contribute to the categories here. The sensitivity to the H and A enters mainly via the bbϕ signal in the b tag categories, especially for increasing values of $\tan\beta$.

Other MSSM scenarios are tested and detailed in []. One scenario of note is the M_H^{125} scenario, which is the equivalent scenario to the M_h^{125} but with the observed Higgs boson being the heavier CP-even Higgs boson. Despite the local excess at a resonant mass of 100 GeV, this scenario is entirely excluded by the search. This is mostly due to the sensitivity of the b tag categories to b associated production. The local excess observed at 1.2 TeV is hard to rectify within these MSSM benchmark scenarios. The lack of any excess in the b tag categories strictly constrains the b associated production cross section times branching fraction. It is not possible within these scenarios to predict the excess of gluon fusion events within the constraints placed on b associated production.



(a)

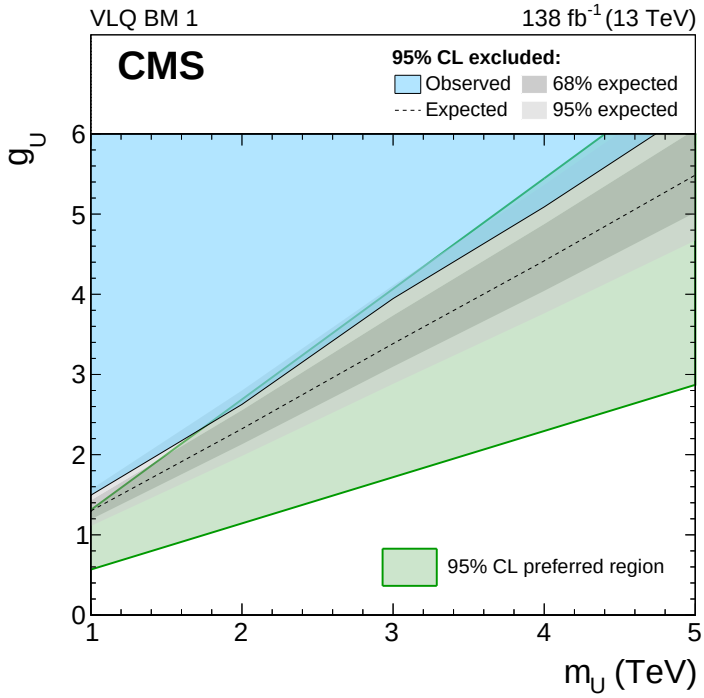


(b)

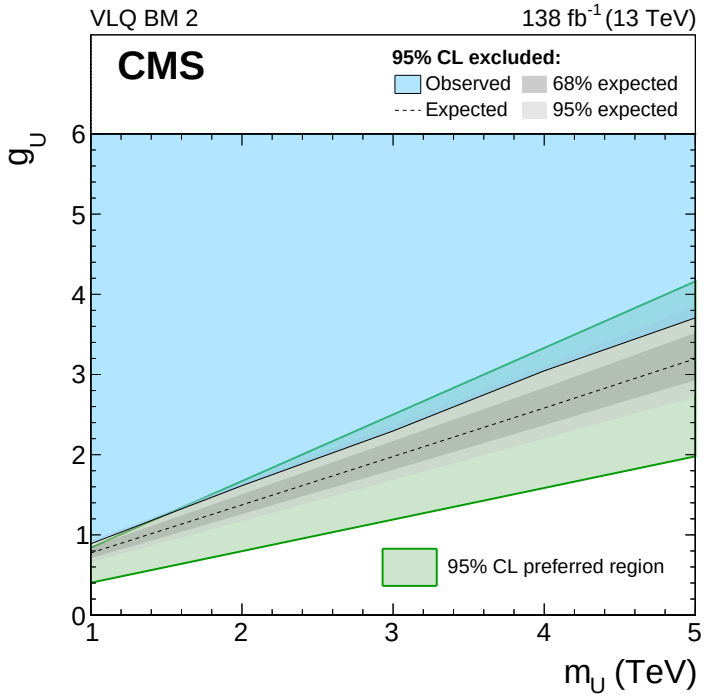
Figure 1.27: Expected and observed 95% CL exclusion contours in the MSSM M_h^{125} (a) and $M_{h,EFT}^{125}$ (b) scenarios. The exclusion limit only on background expectation is shown as a dashed black line, the dark and bright grey bands show the 68% and 95% intervals of the expected exclusion and the observed exclusion contour is shown by the blue area. The parameter space where deviates by more then ±3 GeV from the observed SM Higgs boson mass it shown by a red hatched area.

Upper limits of 95% confidence level for VLQ BM 1 and 2 are shown Figure 1.28. These are drawn with respect to the leptoquark mass (m_U) and coupling (g_U). The limit on g_U decreases as m_U increases, with values of g_U ranging from 1.3 to 5.2 in VLQ BM 1 and 0.8 to 3.2 in VLQ BM 2. VLQ BM 2 has stronger exclusion limits than VLQ BM 1 due to additional right-handed couplings of the leptoquark with a bottom quark and a tau lepton. The observed limits fall within the central 95% intervals of the expected limits when no signal is present. The expected limits are also within the 95% confidence interval of the best fit results reported by Ref.[**?**], indicating that the search is capable of detecting a part of the parameter space that can explain the anomalies observed in b physics.

Similarly to the MSSM scenarios, the local excess at 1.2 TeV is not consistent with a VLQ BM 1 or 2 vector leptoquark. Again this is due to lack of signal in the b tag categories, where the reduction in backgrounds makes the t-channel signal with initial state radiation the dominant search option.



(a)



(b)

Figure 1.28: Expected and observed 95% CL upper limits on in the VLQ BM 1 (a) and 2 (b) scenarios, in a mass range of $1 < m_U < 5$ TeV. The exclusion limit only on background expectation is shown as a dashed black line, the dark and bright grey bands show the 68% and 95% intervals of the expected exclusion and the observed exclusion contour is shown by the blue area. The 95% confidence interval for the preferred region from the global fit presented in Ref. [**?**] is also shown by the green shaded area.