Faculteit Wetenschappen
Departement Fysica

Search for the Standard Model Higgs boson produced via vector boson fusion and decaying to bottom quarks with the CMS detector at the Large Hadron Collider

---

Zoektocht naar het Standaard Model Higgs boson geproduceerd via vector boson fusie en vervallend naar bottom quarks met de CMS detector bij de Large Hadron Collider

Proefschrift voorgelegd tot het behalen van de graad van
doctor in de wetenschappen: fysica
aan de Universiteit Antwerpen
te verdedigen door

Sara Alderweireldt

Promotor:
Prof. Dr. Nick van Remortel

Antwerpen, 2016

Doctoraatscommissie:

Promotor:	prof. dr. Nick van Remortel (Universiteit Antwerpen)
Overige leden:	prof. dr. Sandra Van Aert (Universiteit Antwerpen)
	prof. dr. Michiel Wouters (Universiteit Antwerpen)
	dr. Barbara Clerbaux (Université libre de Bruxelles)
	dr. Ivo van Vulpen (Universiteit van Amsterdam)

Search for the Standard Model Higgs boson produced via vector boson fusion and decaying to bottom quarks with the CMS detector at the Large Hadron Collider

Printed in Belgium by drukkerij Provo
Cover art by Nieuwe Mediadienst (UA)

This work was supported by a *Ph.D. fellowship of the Research Foundation – Flanders (FWO)*.
It was carried out at the *Universiteit Antwerpen (UA)* in Antwerp, Belgium.

What I am going to tell you about is what we teach our physics students
in the third or fourth year of graduate school [. . .]
It is my task to convince you not to turn away because you don't understand it.
You see, my physics students don't underunderstand it either.
That is because I don't understand it.
Nobody does.

– from *QED: The Strange Theory of Light and Matter*
by Richard Feynman, whose *Six easy pieces* ignited my interest in physics

And above all, watch with glittering eyes the whole world around you
because the greatest secrets are always hidden in the most unlikely places.
Those who don't believe in magic will never find it.

– Roald Dahl

The universe will express itself as long as somebody will be able to say,
"I read, therefore it writes."

– Italo Calvino, in *If on a Winter's Night a Traveller*

Preface

Each piece, or part, of the whole of nature, is always merely an approximation to the complete truth, or the complete truth so far as we know it. [...] The test of all knowledge is experiment.[1]

Our current best understanding of the particles and interactions observed in nature is encapsulated in the Standard Model of particle physics [1–7]. A multitude of theoretical and experimental contributions over many years, jointly led to its development in the 1960s and 70s, and many more have seen to its validation ever since. The SM includes fermions, half-integer spin particles which constitute all matter, and gauge bosons, integer spin particles which mediate the interactions. It unifies three of the four known forces: electromagnetism, the weak interaction and the strong interaction, but excludes gravity. As a quantum gauge theory, the SM reflects the symmetries observed in nature in its formulation. A consequence thereof is that in its barest form, the SM requires all particles to be massless. To explain the existence of massive particles, it incorporates the Brout-Englert-Higgs (BEH) mechanism [8–11].

The BEH mechanism was proposed independently by several theoretical physicists in 1964, and describes how particles, seemingly required to be massless, can become massive. It postulates a scalar field – the BEH field – and an associated scalar boson – the Higgs boson. This field induces spontaneous symmetry breaking, and the coupling of particles to the field allows them to acquire mass. The Higgs boson mass itself however, is an important parameter in the SM, and cannot be predicted by theory. Instead it has to be determined experimentally.

For decades, ever since its postulation, scientists have actively searched for the Higgs boson. Searches were conducted at the Large Electron-Positron collider at CERN[2] (1989–2000), at the Tevatron at Fermilab (1987–2011), and most recently at the Large Hadron Collider (LHC) at CERN (2009–). The LHC is the current largest and most-powerful particle accelerator, developed and built between 1984 and 2009, with the discovery of the Higgs boson as one of its main physics goals. It is designed to collide protons at a centre-of-mass energy of 14 TeV. Since 2009, the LHC has been recording data at centre-of-mass energies of 7, 8 and 13 TeV. The Higgs boson remained elusive until 2012, when on the 4th of July both the ATLAS and CMS collaborations presented their results announcing the discovery of a scalar particle matching the properties of the SM Higgs boson. Experiments have since continued to verify the properties of the discovered particle, which so far remains fully compatible with the SM Higgs scalar.

[1] R. Feynman, in *Six easy pieces.*
[2] European Organisation for Nuclear Research

Higgs bosons can be produced and observed at collider experiments in a variety of ways. In this thesis, I present a search for the SM Higgs boson, studying its production via vector boson fusion (VBF) and subsequent decay to bottom quarks ($H \to b\bar{b}$). The data used for this search was recorded with the CMS detector at the LHC, in 2012.

This thesis is structured as follows:

The first part provides an introduction to the concepts required to understand the analysis which lends its title to this thesis. The analysis itself is discussed in the second part.

Chapter 1 provides an introduction to the SM of particle physics, including the electromagnetic, the weak, and the strong force, and the fermions and the gauge bosons. The origin of particle masses is explained by the BEH mechanism, introducing the BEH field, the Higgs boson, and the principle of electroweak symmetry breaking.

Chapter 2 focuses on the Higgs boson. We study how the particle can be produced and observed at collider experiments. Also in chapter 2 we review the many theoretical and experimental studies performed since the postulation of the BEH mechanism. Finally we discuss the Higgs boson discovery in 2012.

Chapter 3 concerns the experimental setup relevant for this analysis: the LHC and the CMS detector. In the first part we study the LHC properties and the various detectors active around the LHC. In the second part we focus on the design of the CMS detector, the detector used to record the data for this analysis.

Chapter 4 describes the simulation methods available to obtain theoretical predictions for collider experiments. We examine the workings of event generators, and also discuss detector simulation.

Chapter 5 returns to the characteristics of the CMS detector, and discusses object reconstruction. From the raw data recorded by the detector, physical quantities need to be reconstructed to allow the data to be used in analyses.

The second part of this thesis discusses the VBF $H \to b\bar{b}$ analysis, which can be logically divided into three major steps.

Chapter 6 is a general introduction to the studied process. We look at the characteristics of the VBF $H \to b\bar{b}$ signal, and identify background processes presenting a similar signature in the detector. We then discuss the analysis strategy, and the data used to perform the analysis.

Chapter 7 covers the event selection, both online during the experiment, and offline afterwards. We discuss the design and optimisation of a dedicated online selection method in 2011, and study how it performed during data taking in 2012. Subsequently, efficiency corrections can be

derived, and the knowledge about the performance of the online selection method can be used to define an optimal offline selection method.

Chapter 8 studies the maximisation of the signal versus background discrimination, using multivariate methods. Before defining the discriminant, three signal properties that can aid in the discrimination are analysed in detail. Afterwards, we perform data validation.

Chapter 9 involves the fit of the bottom quark pair invariant mass spectrum, aiming to detect a possible Higgs boson signal on top of the background contributions. The fit procedure is first validated by performing a fit for of the known Z boson resonance in the same mass spectrum, and then used to search for a Higgs boson resonance.

Chapter 10 presents the results of the VBF $\mathrm{H} \to \mathrm{b\bar{b}}$ analysis. We study the signal strength obtained in the fit and the statistical significance of the results. We also derive upper limits on the cross section of the process. Finally we perform a combination of the VBF $\mathrm{H} \to \mathrm{b\bar{b}}$ result with other $\mathrm{H} \to \mathrm{b\bar{b}}$ results, leading to an increased significance.

Chapter 11 provides a summary and brief outlook.

The physics analysis results presented in this thesis have been published in the following paper, public CMS note, and conference contributions:

- CMS collaboration, *Search for the Standard Model Higgs boson produced through vector boson fusion and decaying to $b\bar{b}$*, Phys.Rev. **D92** (2015) 032008, doi:10.1103/PhysRevD.92.032008, arXiv:1506.01010.
- CMS collaboration, *Higgs to $b\bar{b}$ in the vector boson fusion channel*, CMS-PAS-HIG-13-011 (2013), http://cds.cern.ch/record/1547579.
- S. Alderweireldt, *Study of Higgs bosons decaying to bottom quarks at CMS*, talk at SUSY 2015, 24th August 2015, Lake Tahoe (US).
- S. Alderweireldt, *Search for the Standard Model Higgs boson produced by vector boson fusion decaying to bottom quarks*, poster at Lepton-Photon 2015, 17-22nd August 2015, Ljubljana (SI).
- S. Alderweireldt, *Search for the Standard Model Higgs boson produced by vector boson fusion decaying to bottom quarks*, poster at EPS-HEP 2015, 22-29th July 2015, Vienna (AT).
- S. Alderweireldt, *Vector Boson Fusion leading to Higgs production and subsequent decay in bottom quarks*, PoS EPS-HEP2013 (2013) 128, poster and proceedings for EPS-HEP 2013, 18-24th July 2013, Stockholm (SE).

Contents

Part One

Chapter 1

The Standard Model

The Standard Model (SM) of particle physics [1–7] was first developed almost 50 years ago and is still the most successful and generally accepted theory describing the properties and interactions of the fundamental components of matter. It uses a quantum field theory (QFT)[1] approach, describing particles as excitations of underlying physical quantum fields and identifying (local) gauge symmetries which leave the Lagrangian invariant. It unifies the electromagnetic, weak and strong interactions and incorporates the Brout-Englert-Higgs (BEH)[2] mechanism [8–11] which gives rise to massive particles. The fourth known interaction, gravity, is not included in the SM, as there is no fitting QFT description of it available and it cannot be added to the model trivially.

This chapter discusses the different components of the theory[3]. Section 1.1 starts with a very brief introduction to gauge symmetry. Section 1.2 describes the particle content of the SM. The electroweak part of the theory is covered by the Glashow-Weinberg-Salam (GWS) model (section 1.3) and the BEH mechanism (section 1.4), while the strong part of the theory is covered by quantum chromodynamics (QCD) (section 1.5). Details about the properties of the Higgs boson will follow in chapter 2. Further discussions on these topics can be found in [12, 13].

1.1 Gauge symmetry

In a QFT, each particle is defined as the excitation of a field $\psi(x)$, in four-dimensional space-time. The dynamics of the particles are described by a Lagrangian density (in the same way as is done in classical mechanics), adhering to certain symmetries. Imposing a symmetry means requiring invariance of the Lagrangian under a symmetry transformation; for a global symmetry, a transformation that is the same in all points of space-time, and for a local symmetry, a transformation that is dependent on the point in space-time. Mathematically, a symmetry is characterised by a symmetry group.

In a gauge theory, the Lagrangian is invariant under local gauge transformations. The Lagrangian contains redundant degrees of freedom, i.e. degrees of freedom not corresponding to changes in the physical state. There are different possible gauges (i.e. configurations of the degrees of

[1] An introduction to QFT and gauge theory is available in [12].

[2] The Brout-Englert-Higgs denomination will be used when referring to the BEH mechanism and the therein postulated field; when referring the particle(s) accompanying the field, the short name Higgs boson will be adopted, sometimes abbreviated as H.

[3] Note that throughout this entire thesis, natural units will be used ($\hbar = c = 1$), unless specified differently.

freedom), and transformations can be made between them. The invariance of the Lagrangian under such gauge transformations, is guaranteed by the inclusion of gauge terms (matching the gauge group or symmetry group) in this Lagrangian. The gauge terms are vector fields, which correspond to the fundamental forces or interactions between the particles in the theory. The strength of the forces is given by the gauge coupling constants. The excitations of the gauge fields, the gauge particles, act as force carriers.

The SM is a gauge QFT, with a Lagrangian defined according to the fundamental local gauge symmetries observed in nature.

1.2 Elementary particles

Fermions, elementary spin-$\frac{1}{2}$ particles, are the building blocks of matter. At present we know twelve different fermions which can be classified into two groups, six quarks and six leptons, and each group can then again be subdivided into three generations of increasing mass. The leptons interact only through electromagnetic and weak forces, while the quarks will also interact strongly. Furthermore, the leptons are observed as free particles, while the strong force confines the quarks within hadrons, bound states of two or three quarks. Note also that the second and third generations of fermions are not stable, they decay into lighter particles, which implies that ordinary matter will consist of first generation particles only. The fermion classification is given in Table 1.1.

Table 1.1: Classification of the fermions.

	generations			Q	interact through
	1st	2nd	3rd		
leptons	e	μ	τ	-1	electroweak forces
	ν_e	ν_μ	ν_τ	0	
quarks	u	c	t	$+\frac{2}{3}$	all forces
	d	s	b	$-\frac{1}{3}$	

The gauge bosons then, are spin-1 particles serving as force carriers for the interactions. The electromagnetic gauge boson, the photon, is massless and the interaction it represents is long-ranged. The weak bosons however, turn out to be massive, and the weak interaction works at very small scale. In a simple gauge theory these masses would not appear, and it is through the BEH mechanism that they can be explained. The strong interaction finally is short-ranged, and its force carrier, the gluon, is again massless. The gauge boson classification is given in Table 1.2.

Table 1.2: Classification of the gauge bosons.

	force	mass (GeV)	range (fm)	coupling constant[†]
photon (γ)	electromagnetic	$m_\gamma = 0$	∞	$e = \sqrt{4\pi\alpha} \sim 0.3$
W^+, W^-	weak, charged	$m_W = 80.4$	0.001	$g = \frac{8 m_W^2 G_F}{\sqrt{2}} \sim 0.4$
Z^0	weak, neutral	$m_Z = 91.2$	0.001	$g' = g \tan\theta_W \sim 0.2$
gluons (g)	strong	$m_g = 0$	1	$g_s = \sqrt{4\pi\alpha_s} \sim 1$

[†] with electromagnetic finestructure constant α, Fermi constant G_F, Weinberg mixing angle θ_W, and the analogue of the electromagnetic finestructure constant for the strong force α_S.

1.3 Electroweak theory

As an introduction we discuss quantum electrodynamics (QED), the gauge theory describing the electromagnetic interaction. The gauge group for the electromagnetic interaction is $\mathrm{U(1)_{EM}}$, with electric charge Q as the generator. The free fermion Lagrangian is given by:

$$\mathcal{L} = \bar{\psi}\left(i\partial\!\!\!/_\mu - m\right)\psi, \tag{1.1}$$

with spinor field ψ representing the fermion (mass m, charge Q), adjoint spinor field $\bar{\psi} = \psi^\dagger\gamma^0$, and index μ running over the four space-time coordinates 0-3. Derivative $\partial\!\!\!/_\mu$ is written in the Feynman slash notation: $\partial\!\!\!/_\mu = \gamma^\mu\partial_\mu$, with γ^μ the Dirac gamma matrices.

To make this Lagrangian work for QED, requirements for local $\mathrm{U(1)_{EM}}$ gauge invariance should be satisfied. Using the regular derivative, the invariance fails. Take for example the following local gauge transformation:

$$\psi(x) \to e^{i\alpha(x)}\psi(x), \tag{1.2}$$

then:

$$\mathcal{L}' = \bar{\psi}\, i\gamma^\mu\left(\partial_\mu + i\left(\partial_\mu\alpha\right)\right)\psi - m\bar{\psi}\psi \neq \mathcal{L}. \tag{1.3}$$

To restore the invariance, a modified derivative is introduced (covariant derivative D_μ), including a vector field A_μ with specific transformation properties, matching $\mathrm{U(1)_{EM}}$:

$$\begin{aligned} D_\mu &= \partial_\mu - ieA_\mu \\ A_\mu &\to A_\mu + \frac{1}{e}\partial_\mu\alpha. \end{aligned} \tag{1.4}$$

Applying this to the Lagrangian (Eq. 1.1) leads to:

$$\begin{aligned} \mathcal{L} &= \bar{\psi}\left(iD\!\!\!/_\mu - m\right)\psi \\ &= \bar{\psi}\, i\gamma^\mu\left(\partial_\mu - ieA_\mu\right)\psi - m\bar{\psi}\psi \\ &= \bar{\psi}\left(i\partial\!\!\!/_\mu - m\right)\psi + e\bar{\psi}\gamma^\mu A_\mu\psi, \end{aligned} \tag{1.5}$$

which is invariant under transformation 1.2.

For completeness, also a kinetic term for vector field A_μ needs to be added:

$$-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \qquad \text{with field strength tensor} \quad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{1.6}$$

Field A_μ can be interpreted as the photon field, the photon being the gauge boson of $\mathrm{U(1)_{EM}}$. It couples to particles represented by spin-$\frac{1}{2}$ field ψ (cf. $e\bar{\psi}\gamma^\mu\psi A_\mu$), and the strength of the coupling is given in terms of the coupling constant, electron charge e. Again due to gauge invariance, the presence of a mass term is prohibited, and the mediating vector boson has to be massless. In other words, the electromagnetic interaction of two particles is governed by the exchange of a massless vector boson, the photon.

The full QED Lagrangian is given in Eq. 1.7.

$$\mathcal{L}_{EM} = \bar{\psi}\left(i\partial\!\!\!/_\mu - m\right)\psi + e\bar{\psi}\gamma^\mu A_\mu\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \tag{1.7}$$

Moving on to electroweak theory, unifying the electromagnetic and weak interactions, one can follow the same method as used for QED in the introduction. The relevant gauge group as used in the SM is now $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$. The components of the weak isospin, T_1, T_2 & T_3, are the generators of $\mathrm{SU(2)_L}$, while the weak hypercharge, Y, is the generator of $\mathrm{U(1)_Y}$. Note that $\mathrm{U(1)_{EM}}$ and $\mathrm{U(1)_Y}$ are not the same, they are different copies of U(1). The electromagnetic charge can however be related to the isospin and the hypercharge as $Q = T_3 + \frac{Y}{2}$, linking both groups. Imposing local gauge invariance leads to the introduction of four vector fields, $W_\mu^{1,2,3}$ & B_μ with four matching massless gauge bosons. Through the process of spontaneous symmetry breaking (SSB) and the BEH mechanism, discussed in section 1.4, they will mix into three massive and one massless boson: the W^+ and W^-, the Z^0, and the γ. Coupling strengths are given in terms of coupling constants g and g'. The covariant derivative for this case is:

$$D_\mu = \partial_\mu - igT_aW_\mu^a - ig'\frac{Y}{2}B_\mu. \tag{1.8}$$

The fermion classification can now be revisited in terms of symmetry group $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$. Groups are made based on how particles transform when subjected to weak interactions. Since electroweak theory is chiral, it is necessary to distinguish between left-handed (opposite spin & direction of motion) and right-handed (parallel spin & direction of motion) particles. Only left-handed fermions will interact weakly, they exist in doublets of opposite T_3, the third component of the isospin. Right-handed ones don't interact and exist in singlets of zero isospin. There are two types of doublets: $(\nu\ e_L)$ and $(u_L\ d_L)$, and three types of singlets: (u_R), (d_R) and (e_R). The right-handed neutrinos, which would form the fourth singlet, do not exist within the framework of the SM. Hypercharges can be determined through the relation $Q = T_3 + \frac{Y}{2}$. This leads to Table 1.3.

Fully parallel to the QED Lagrangian given in Eq. 1.7, the kinetic part of the electroweak Lagrangian is given by:

$$\mathcal{L}_{EW} = \bar{\chi}_{jL}\, i\not{D}_\mu\, \chi^j_L + \bar{\xi}_{jR}\, i\gamma^\mu \left(\partial_\mu - ig'\frac{Y}{2}B_\mu\right)\xi^j_R - \frac{1}{4}W^a_{\mu\nu}W^{a\,\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{1.9}$$

differentiating between doublets (χ_L) and singlets (ξ_R) for the fermion fields, using vector fields W^a_μ and B_μ, and the covariant derivative of Eq. 1.8. Interaction terms for fermions and gauge bosons are included, as well as the self-interaction terms for the gauge fields. No mass terms are present however, neither for the fermions, nor for the gauge bosons. A solution for this will be introduced in section 1.4, in the form of the BEH mechanism.

1.4 The Brout-Englert-Higgs mechanism

As stated in section 1.3, without an extensions of the Lagrangian as given in Eq. 1.9, the gauge bosons of the electroweak interactions would be massless. To make them massive, symmetry needs to be broken. This is where the BEH field comes in. It is described by an $\mathrm{SU(2)_L}$ doublet of complex fields, with hypercharge $Y = 1$:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{1.10}$$

A kinetic and a potential term for the BEH field can be added to the Lagrangian, aiming at electroweak symmetry breaking:

$$\mathcal{L}_{Higgs} = |D_\mu\phi|^2 - V\left(\phi^\dagger\phi\right). \tag{1.11}$$

Table 1.3: Classification of the fermions, revisited.

		multiplet	generations 1st	2nd	3rd	Q	T	T_3	Y
left-handed	leptons	doublet	ν_e	ν_μ	ν_τ	0	$\frac{1}{2}$	$+\frac{1}{2}$	-1
			e_L	μ_L	τ_L	-1	$\frac{1}{2}$	$-\frac{1}{2}$	-1
	quarks	doublet	u_L	c_L	t_L	$+\frac{2}{3}$	$\frac{1}{2}$	$+\frac{1}{2}$	$+\frac{1}{3}$
			d_L	s_L	b_L	$-\frac{1}{3}$	$\frac{1}{2}$	$-\frac{1}{2}$	$+\frac{1}{3}$
right-handed	leptons	singlet	e_R	μ_R	τ_R	-1	0	0	-2
	quarks	singlet	u_R	c_L	t_R	$+\frac{2}{3}$	0	0	$+\frac{4}{3}$
		singlet	d_R	s_R	b_R	$-\frac{1}{3}$	0	0	$-\frac{2}{3}$

This does however not break the local gauge invariance under $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ yet, since on one hand a covariant derivative is used which transforms correctly under symmetry operations, and on the other hand the potential term is only dependent on $|\phi|$. It is through SSB, the hiding of a symmetry, that massive gauge bosons and fermions are acquired in the SM.

The potential term in Eq. 1.11 can be expanded as:

$$V\left(\phi^\dagger\phi\right) = \mu^2\phi^\dagger\phi + \lambda\left(\phi^\dagger\phi\right)^2. \tag{1.12}$$

The different options for the μ^2 parameter are what makes this interesting. With $\lambda > 0$, the case where it is positive would simply describe a massive scalar boson with mass term $\mu^2\phi^\dagger\phi$, and not fix any of the previously mentioned problems. If it were negative however, it initially appears as though the mass term has the wrong sign, and the potential has an infinite number of minima:

$$\begin{aligned} & \frac{\partial V}{\partial|\phi|} = \left(\mu^2 + 2\lambda\phi^\dagger\phi\right)\cdot\frac{\partial\phi^\dagger\phi}{\partial|\phi|} = 0 \\ \Leftrightarrow \quad & \phi^\dagger\phi = \frac{1}{2}\left(\phi_1^2+\phi_2^2+\phi_3^2+\phi_4^2\right) = -\frac{\mu^2}{2\lambda} = \frac{v^2}{2} \\ \Leftrightarrow \quad & \phi = \frac{v}{\sqrt{2}}\,e^{i\theta}, \end{aligned} \tag{1.13}$$

with v the vacuum expectation value.

This is the typical Mexican hat potential as shown in Fig. 1.1. Now, without losing generality, the minimum can be fixed at a given point, and the field can be expanded around this point (which will have vacuum expectation value v). For example:

$$\phi_1^2 = \phi_2^2 = \phi_4^2 = 0 \quad , \quad \phi_3^2 = v^2 \tag{1.14}$$

and

$$\phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}. \tag{1.15}$$

It is now possible to have a closer look at $\mathcal{L}_{Higgs}$, given in Eq. 1.11. Substituting $\phi(x)$, it is obvious that there will be three sets of terms: related to the scalar Higgs boson only, related to the coupling between the Higgs boson and the electroweak vector bosons, and related to the electroweak bosons only. It can be expected to find the sought for vector boson mass terms in this last set.

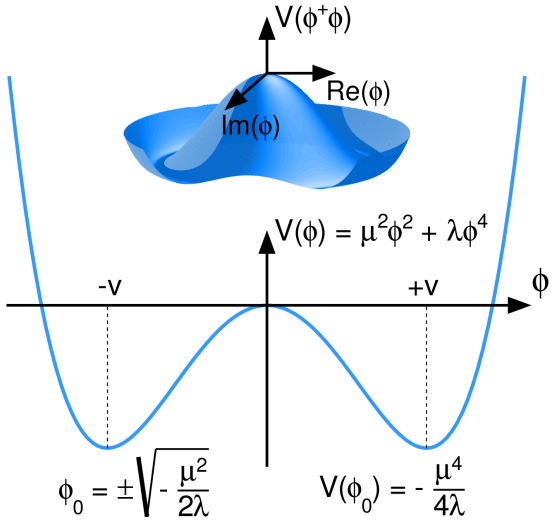


Figure 1.1: Graphical representation of the $V(\phi^\dagger\phi)$ potential of Eq. 1.12, typically called a Mexican hat potential.

Taking only the relevant terms of the Lagrangian[4]:

$$\mathcal{L}_{m_V} = \frac{1}{2}\left|\left(g\frac{t_a}{2}W_\mu^a + \frac{1}{2}g'B_\mu\right)\begin{pmatrix}0\\ v\end{pmatrix}\right|^2 \tag{1.16}$$

$$= \frac{1}{2}\begin{pmatrix}0 & v\end{pmatrix}\begin{pmatrix}\frac{1}{2}\left(gW_\mu^3 + g'B_\mu\right) & \frac{g}{2}\left(W_\mu^1 - iW_\mu^2\right)\\ \frac{g}{2}\left(W_\mu^1 + iW_\mu^2\right) & \frac{1}{2} - \left(gW_\mu^3 - g'B_\mu\right)\end{pmatrix} \cdot$$
$$\begin{pmatrix}\frac{1}{2}\left(gW_\mu^3 + g'B_\mu\right) & \frac{g}{2}\left(W_\mu^1 - iW_\mu^2\right)\\ \frac{g}{2}\left(W_\mu^1 + iW_\mu^2\right) & -\frac{1}{2}\left(gW_\mu^3 - g'B_\mu\right)\end{pmatrix}\begin{pmatrix}0\\ v\end{pmatrix}$$

$$= \frac{1}{8}\begin{pmatrix}g\left(W_\mu^1 + iW_\mu^2\right)v & -\left(gW_\mu^3 - g'B_\mu\right)v\end{pmatrix}\begin{pmatrix}g\left(W_\mu^1 - iW_\mu^2\right)v\\ -\left(gW_\mu^3 - g'B_\mu\right)v\end{pmatrix}$$

$$= \frac{v^2}{8}\Big(g^2\left(W_\mu^1 + iW_\mu^2\right)\left(W_\mu^1 - iW_\mu^2\right) + \left(gW_\mu^3 - g'B_\mu\right)\left(gW_\mu^3 - g'B_\mu\right)\Big)$$

$$= \frac{g^2v^2}{4}\,W_\mu^+W^{-\mu} + \frac{\left(g^2+g'^2\right)v^2}{8}\,Z_\mu^0Z^{0\mu} \tag{1.17}$$

$$= m_W^2\,W^+W^- + \frac{1}{2}m_Z^2\,Z^2 + \frac{1}{2}m_A^2\,A^2. \tag{1.18}$$

[4] $t_aW_\mu^a$ has been expanded using $\mathrm{SU(2)_L}$ generating matrices t_1, t_2 and t_3:
$t_1 = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$, $t_2 = \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}$, $t_3 = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$.

Equation 1.17 is written in terms of new physical fields W^+, W^-, Z^0 and A^0, linear combinations of $W^{1,2,3}$ and B:

$$\begin{aligned} W_\mu^- &= \frac{1}{\sqrt{2}}\left(W_\mu^1 - iW_\mu^2\right) \\ W_\mu^+ &= \frac{1}{\sqrt{2}}\left(W_\mu^1 + iW_\mu^2\right) \\ Z_\mu^0 &= \frac{1}{\sqrt{g^2+g'^2}}\left(gW_\mu^3 - g'B_\mu\right) \\ A_\mu^0 &= \frac{1}{\sqrt{g^2+g'^2}}\left(g'W_\mu^3 + gB_\mu\right). \end{aligned} \tag{1.19}$$

The vector boson masses finally follow from comparing Eqs. 1.17 and 1.18. The lack of a term for the photon, A, is expected since the $\mathrm{U}(1)_{\mathrm{EM}}$ symmetry remains intact after SSB, and thus the associated boson has to be massless. One finds:

$$m_W = \frac{1}{2}vg\,, \quad m_Z = \frac{1}{2}v\sqrt{g^2+g'^2}\,, \quad m_A = 0. \tag{1.20}$$

Expanding further the potential term of $\mathcal{L}_{Higgs}$, as given in Eq. 1.12, taking only the term of $\mathcal{O}(h^2)$, one can identify the Higgs boson mass:

$$\begin{aligned} \frac{1}{2}m_h^2\,\bar{h}h &= -\mu^2\,\bar{h}h = \lambda\, v^2\,\bar{h}h \\ \Rightarrow m_h &= \sqrt{-2\mu^2} = \sqrt{2\lambda\, v^2}. \end{aligned} \tag{1.21}$$

Note that due to free parameter λ the Higgs boson mass cannot be calculated from theory only, it needs to be assessed experimentally.

Next the covariant derivative in Eq. 1.8 can be rewritten in terms of the physical fields:

$$\begin{aligned} D_\mu = \partial_\mu - ig\,\frac{1}{\sqrt{2}}\left(W_\mu^+T^+ + W_\mu^-T^-\right) - i\frac{Z_\mu^0}{\sqrt{g^2+g'^2}}\left(g^2T^3 - g'^2Y\right) \\ -i\frac{A_\mu^0 gg'}{\sqrt{g^2+g'^2}}\left(T^3+Y\right). \end{aligned} \tag{1.22}$$

Since the coupling strength for the photon was defined earlier as e, it is possible to identify:

$$e = \frac{gg'}{\sqrt{g^2+g'^2}}. \tag{1.23}$$

Additionally, the mixing from $\left(W_\mu^3, B_\mu\right)$ to $\left(Z_\mu^0, A_\mu^0\right)$, described in Eq. 1.19, can also be defined in terms of the Weinberg weak mixing angle, θ_W:

$$\begin{pmatrix} Z_\mu^0 \\ A_\mu^0 \end{pmatrix} = \begin{pmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} W_\mu^3 \\ B_\mu \end{pmatrix}. \tag{1.24}$$

Using again Eq. 1.19 this entails:

$$\begin{aligned} \cos\theta_W &= \frac{g}{\sqrt{g^2+g'^2}} \\ \sin\theta_W &= \frac{g'}{\sqrt{g^2+g'^2}} \quad , \end{aligned} \tag{1.25}$$

and

$$\begin{aligned} e = g\sin\theta_W &= g'\cos\theta_W \\ m_W &= m_Z\cos\theta_W. \end{aligned} \tag{1.26}$$

This concludes the initial aim of this section: having a unified description for the electroweak interaction, and in particular, explaining the masses of its vector bosons.

There are however a few remaining terms in the electroweak part of the SM Lagrangian which have not yet been discussed here. Next to the masses of the vector bosons, also the cubic and quartic couplings of the Higgs boson to vector bosons can be derived from the $\mathcal{L}_{Higgs}$. Making the expansion parallel to Eq. 1.16, looking at the terms of order $\mathcal{O}(vh)$ and $\mathcal{O}(h^2)$, one finds:

$$\begin{aligned} \mathcal{L}_{hVV} &= \frac{g^2 v}{2}\, h\, W^+W^- + \frac{(g^2+g'^2)\, v}{4}\, h\, Z^0_\mu Z^{0\mu} = g_{hWW}\, h\, W^+W^- + g_{hZZ}\, h\, Z^0_\mu Z^{0\mu} \\ \mathcal{L}_{hhVV} &= \frac{g^2}{4}\, h^2\, W^+W^- + \frac{(g^2+g'^2)}{8}\, h^2\, Z^0_\mu Z^{0\mu} = g_{hhWW}\, h^2\, W^+W^- + g_{hhZZ}\, h^2\, Z^0_\mu Z^{0\mu}. \end{aligned} \tag{1.27}$$

The potential term of $\mathcal{L}_{Higgs}$ (Eq. 1.12) additionally includes the cubic and quartic Higgs boson self-coupling terms:

$$\begin{aligned} \mathcal{L}_{hhh} &= \lambda v h^3 \\ \mathcal{L}_{hhhh} &= \frac{\lambda}{4} h^4, \end{aligned} \tag{1.28}$$

while the Higgs boson self-energy is again part of the kinetic term of $\mathcal{L}_{Higgs}$:

$$\mathcal{L}_{\partial_\mu h} = \frac{1}{2}\left(\partial_\mu h\right)\left(\partial^\mu h\right). \tag{1.29}$$

The last set of terms to be added describes the fermion masses and the Higgs boson to fermion couplings. Consider the following Lagrangian:

$$\mathcal{L}_{m_f,hff} = -G_f\left(\bar{\chi}_L\,\phi\,\xi_R + \bar{\chi}_R\,\phi_c\,\xi_L)\right), \tag{1.30}$$

with G_f the arbitrary strength of the Yukawa coupling to fermions, and (χ_L) the left-handed doublets and (ξ_R) the right-handed singlets for the fermion fields. Expanding once more as in

Eq. 1.16, first for the leptons, one finds:

$$\begin{aligned}\mathcal{L}_{m_l,hll} &= -\frac{G_e}{\sqrt{2}}\left(\bar{e}_L e_R + \bar{e}_R e_L\right) v - \frac{G_e}{\sqrt{2}}\left(\bar{e}_L e_R + \bar{e}_R e_L\right) h \\ &= -m_e\,\bar{e}e - g_{hee}\,\bar{e}e\,h. \end{aligned} \tag{1.31}$$

The electron mass can be identified as $m_e = G_e v/\sqrt{2}$ (with G_e the Yukawa coupling for the electron), and the coupling strength as $g_{hee} = G_e/\sqrt{2}$. Since both are dependent on the arbitrary Yukawa coupling, they cannot be predicted. Note also that while the coupling strength of the Higgs boson to gauge bosons g_{hVV} (cf. Eq. 1.27) is proportional to the square of the gauge boson masses, the coupling strength of the Higgs boson to fermions is directly proportional to the fermion masses.

The same logic can be followed for the quarks, which leads to:

$$\begin{aligned}\mathcal{L}_{m_q,hqq} &= -G_d\left((\bar{u}\ \bar{d}')_L\,\phi\,d_R + h.c.\right) - G_u\left((\bar{u}\ \bar{d}')_L\,\phi_c\,u_R + h.c.\right) \\ &= -m_d\,\bar{d}d\left(1+\frac{h}{v}\right) - m_u\,\bar{u}u\left(1+\frac{h}{v}\right), \end{aligned} \tag{1.32}$$

where $(\bar{u}\ \bar{d}')_L$ takes into account quark mixing (i.e. mass eigenstates being linear combinations of the flavour eigenstates), and conjugate ϕ_c of the BEH field is used to allow up-type quark masses in a gauge invariant manner. G_d and G_u are the Yukawa couplings for down-type and up-type quarks.

As a final result for the Lagrangian for the electroweak sector of the SM one can then write:

$$\begin{aligned}\mathcal{L}_{EWSB} = {} & \bar{\chi}_{jL}\,i\not{D}_\mu\,\chi_L^j + \bar{\xi}_{jR}\,i\gamma^\mu\left(\partial_\mu - ig'\frac{Y}{2}B_\mu\right)\xi_R^j - \frac{1}{4}W^a_{\mu\nu}W^{a\,\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \\ & + i\gamma^\mu\left|\left(\partial_\mu - igT_aW_\mu^a - ig'\frac{Y}{2}B_\mu\right)\phi\right|^2 - V\left(\phi^\dagger\phi\right) \\ & - G_f\left(\bar{\chi}_L\,\phi\,\xi_R + \bar{\xi}_L\,\phi_c\,\xi_R\right). \end{aligned} \tag{1.33}$$

1.5 Quantum chromodynamics

Quantum chromodynamics (QCD) is the part of the SM that outlines the strong interactions of partons, i.e. quarks and gluons. As mentioned before, when approaching from a gauge theoretical point of view, it is essential to consider which symmetries are in play. From experimental studies of bound quark states it became apparent that an additional quantum number needed to be allocated to quarks, if one wanted to explain bound states with three "identical" quarks, qqq, while still respecting Fermi-Dirac statistics. In this new picture, quarks carry one of three colours: red, green or blue; they are colour charged under gauge group $\mathrm{SU(3)_c}$. Gluons are assigned combinations of a colour and an anti-colour.

The QCD Lagrangian is hence required to be invariant under:

$$\psi(x) \rightarrow e^{i\alpha_s(x)\frac{\lambda_a}{2}}\psi(x), \tag{1.34}$$

where ψ represents the partons, α_s is the coupling constant of the strong interaction and $\lambda_a/2$ are the generators of $SU(3)_c$. This invariance can be achieved by writing a covariant derivative:

$$D_\mu = \partial_\mu - ig_s\frac{\lambda_a}{2}G^a_\mu, \tag{1.35}$$

which introduces eight gauge fields, G^a_μ. The vector bosons of the strong interaction are eight massless gluons, independent through eight different non-zero colour charges. Mathematically, it would be equally possible to use a U(3) gauge group and have nine gluons, including a colour neutral one, but this is experimentally discouraged.

So far, this description develops in a very similar way to QED. The first difference is that the gluons are not colour neutral[5], allowing for gluon self-interaction. In the Lagrangian, this is represented by the presence of a third term in the field-strength tensor:

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + if^{abc}G^b_\mu G^c_\nu, \tag{1.36}$$

compared to the form known from QED (Eq. 1.6). Factors f^{abc} denote the structure constants of $SU(3)_c$.

Combining the above elements one finds:

$$\mathcal{L}_{QCD} = -\frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a + \sum_{flav}\bar{\psi}\left(i\not{D}_\mu - m\right)\psi, \tag{1.37}$$

with gluon colours a and quarks fields ψ, and summing over all quark flavours. Combination of Eqs. 1.33 and 1.37 gives the full definition of the SM Lagrangian.

Due to the presence of gluon self-interaction terms, QCD obtains a few more properties which warrant discussion. Continuing the comparison to QED, there is a profound difference at the level of the coupling constant. Due to renormalisation[6], a measurement of a coupling constant, α, depends on the energy or distance scale, μ, at which it is measured. This can be expressed as:

$$\frac{1}{\alpha(\mu)} = \frac{1}{\alpha(\mu_{ref})} + b\log\left(\frac{\mu^2_{ref}}{\mu^2}\right), \tag{1.38}$$

with b related to the number of fermions, with mass below scale μ, transforming according to the gauge group of the interaction under consideration [14].

[5] as the photon is electromagnetically neutral in QED

[6] In renormalisation, infinities in the theoretical expressions are absorbed into redefinitions of elementary charges, which appear "dressed" in real life experiments [12].

At an energy scale of the order of the Z-boson mass, m_Z, both the electromagnetic and the strong coupling strength are small:

$$\alpha_{em}(m_Z) = \frac{e^2(m_Z^2)}{4\pi} = 0.0078$$
$$\alpha_s(m_Z) = \frac{g_s^2(m_Z^2)}{4\pi} = 0.12, \tag{1.39}$$

which allows perturbative methods to be used in calculations.

When changing the energy scale however, both forces behave in opposite ways. For the electromagnetic force, factor b in Eq. 1.38 is positive. As scale μ increases, also the coupling strength increases. For the strong force on the other hand, this factor is negative, and the coupling strength decreases as the scale increases. QCD exhibits asymptotic freedom: at very high energies – or very short distance – the strong force is effectively weak. A consequence hereof is that perturbative calculations in QCD are invalidated at low energy.

The second typical QCD property is the concept of confinement: the only eigenstates of the QCD Hamiltonian[7] which have finite energy are colour neutral. In other words: quarks must always appear in colour neutral combinations, the most obvious ones being mesons ($q\bar{q}$, colour/anti-colour) and baryons (qqq, fully antisymmetric in colour). It also explains why quarks, carrying colour charge, may not be observed as free particles. Confinement has, albeit very much experimentally supported, not yet been analytically derived from the Lagrangian, and remains, in principle, a hypothesis.

Finally there is the concept of hadronisation, which can be intuitively understood keeping in mind both asymptotic freedom and confinement. As for example the constituents of a quark/anti-quark pair travel apart, the force between them increases. At some point, it will become energetically favourable for a new $q\bar{q}$ pair to be created from the vacuum, and for the constituents to form new colour neutral bound states, satisfying the confinement principle. Hadronisation also has an experimental consequence. Partons scattered or radiated in a collision cannot continue to exist as free particles and will hadronise. Each parton emerging from an interaction will result in a collimated flow of hadrons to be produced, following the direction of motion of this parton; such a collection of particles is called a *jet*. The dynamic properties of the hadrons will resemble those of the parton, a feature called local parton-hadron duality. In detector experiments, specific jet-algorithms are used to identify jets, and to trace back as well as possible the properties of the originating parton.

This chapter introduced the Standard Model of particle physics, which describes the properties and interactions of fundamental particles. It includes electroweak theory and QCD, and explains the masses of particles using spontaneous symmetry breaking and the BEH mechanism. The following chapter 2 will focus on the Higgs boson, more specifically on how it is produced experimentally and how it can be observed. In addition, an overview is given of past searches for this particle.

[7] Mathematical description of a system, cf. Lagrangian.

Chapter 2

The Higgs boson

To study fundamental particles and interactions and test models such as the SM (cf. chapter 1), scientists perform collider experiments at accelerator facilities. The largest and most powerful particle accelerator today is the Large Hadron Collider (LHC) at CERN[1] in Geneva, Switzerland; it will be discussed in chapter 3.

This chapter focuses on the study of one particle in particular, the Higgs boson. Section 2.1 explains how it can be produced and observed in experiments. In the second part of the chapter, an overview is given of indirect and direct searches performed for the Higgs boson in the past (section 2.2), ultimately leading to its discovery in 2012 (section 2.3).

2.1 Higgs boson production mechanisms and decay modes

The SM Higgs boson can be produced at collider experiments in various ways, depending on the energy scale of the collisions[2], as well as on the properties of the colliding particles. At a proton-proton collider such as the LHC (cf. chapter 3), it will happen only very rarely: about once every ten billion collisions (1 in 10^{10}). Four main mechanisms can be identified that provide a significant contribution to the total Higgs boson production cross section, i.e. the rate at which Higgs bosons are generated.

Due to the abundance of gluons in proton–proton collisions, *gluon-gluon fusion (GF)* is the dominant Higgs boson production mechanism throughout the entire mass range studied at the LHC, $m_H = 100 - 600$ GeV. Figure 2.1a shows the Leading Order (LO) Feynman diagram for the process: the Higgs boson is produced from two gluons through a loop, most commonly a top-quark loop.

Second most important is *vector boson fusion (VBF)* production, where two quarks each radiate a vector boson, either a W or a Z boson, and the vector bosons merge to form the Higgs boson. This is illustrated in Fig. 2.1b. Typical for this mode are the two remaining forward quarks in the final state, in addition to the Higgs boson decay products.

Production *in association with a vector boson (VH)*, either a W or a Z boson, is shown in Fig. 2.1c. A quark and an anti-quark annihilate to a vector boson, which subsequently decays into another vector boson of the same type, and a Higgs boson. Note that at a proton–anti-proton collider,

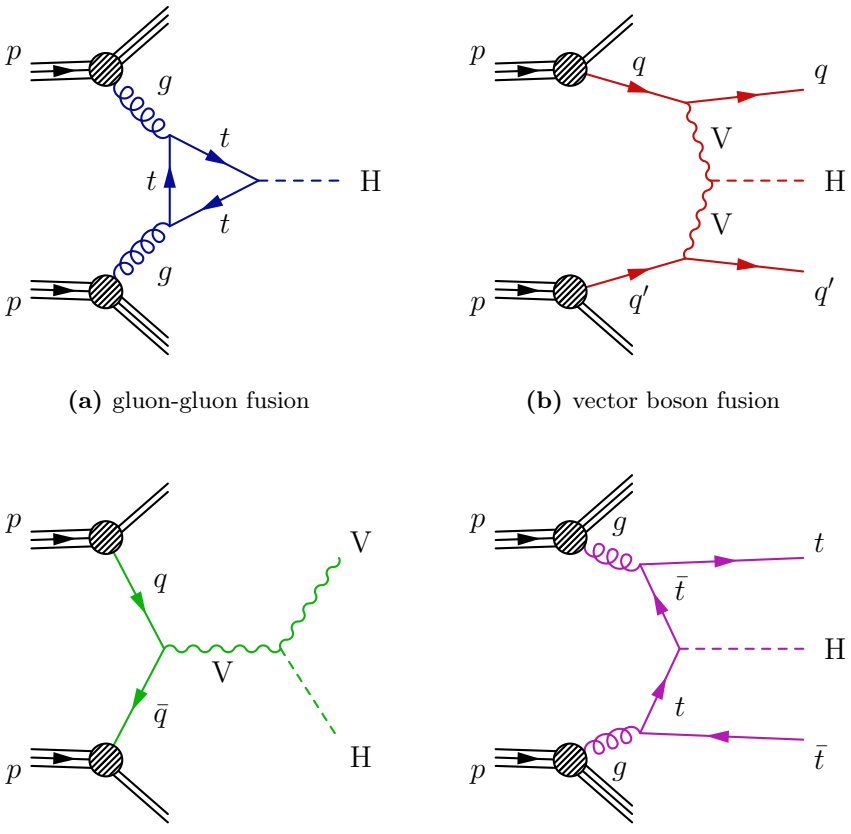
[1] Conseil Européen pour la Recherche Nucléaire or European Organization for Nuclear Research.
[2] Usually expressed in terms of the centre-of-mass energy $\sqrt{s}$ of the collisions.

this production mode would be more important, as the quark and anti-quark could both be valence quarks, whereas at LHC the anti-quark has to be a sea quark.

Finally the Higgs boson can be produced *in association with a heavy quark pair ($t\bar{t}H$)*, as in Fig. 2.1d. The decay of the top quarks, next to the decay of the Higgs boson, will typically lead to final states with many jets.

The relative importance of each production mechanism can also be seen in Fig. 2.2, which shows the Higgs boson production cross section as a function of its mass, at three centre-of-mass energies, for each of the mechanisms. At $m_H = 125$ GeV, there is roughly an order of magnitude difference between the GF production cross section and that of VBF or VH production, and the contribution of $t\bar{t}H$ production is again an order of magnitude smaller.

Data has been observed experimentally at the LHC for all four production mechanisms, but only the GF process has been shown to exist with sufficient statistical significance.



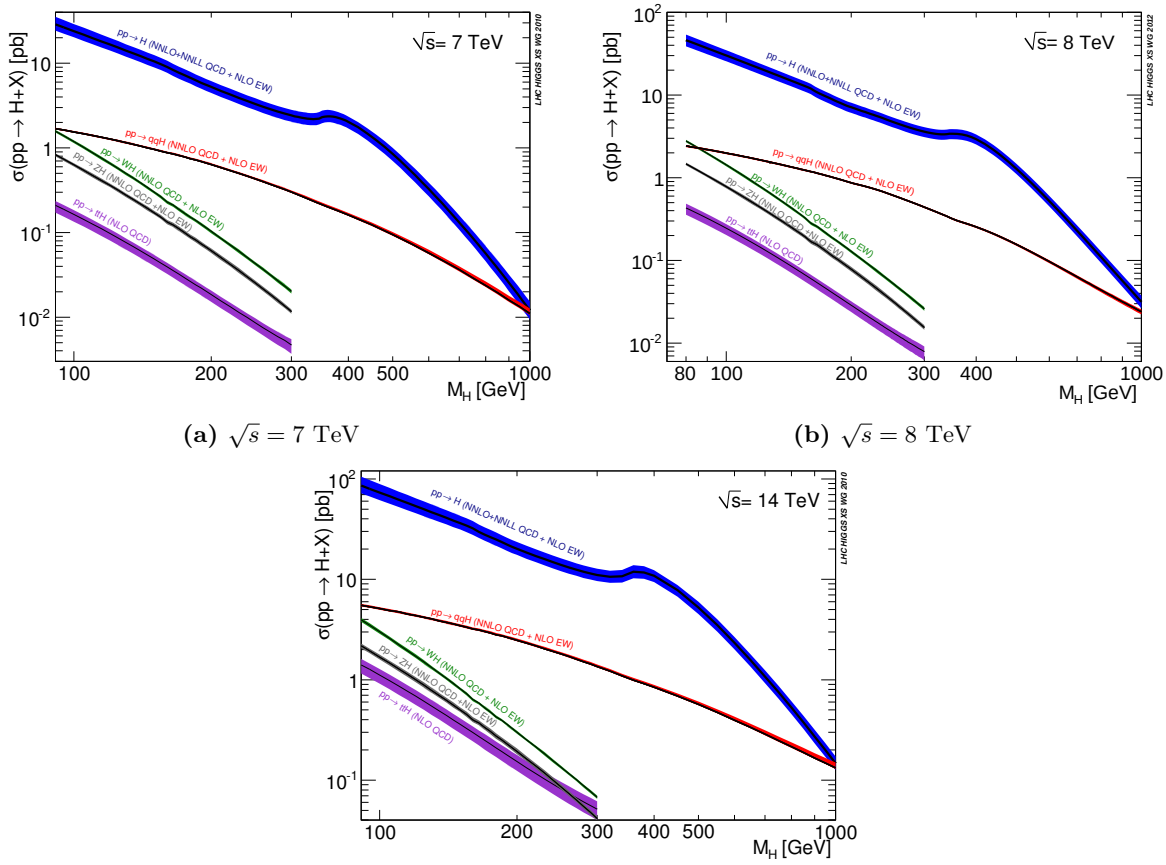
(a) gluon-gluon fusion

(b) vector boson fusion

(c) vector boson associated production

(d) $t\bar{t}$ associated production

Figure 2.1: LO Feynman diagrams for the four most important Higgs boson production mechanisms at the LHC. In order of importance they are (a) gluon-gluon fusion, (b) vector boson fusion, (c) vector boson associated production, and (d) heavy quark associated production.



(a) $\sqrt{s} = 7$ TeV **(b)** $\sqrt{s} = 8$ TeV

(c) $\sqrt{s} = 14$ TeV

Figure 2.2: Higgs boson production cross sections at (a) $\sqrt{s} = 7$ TeV, (b) $\sqrt{s} = 8$ TeV, and (c) $\sqrt{s} = 14$ TeV. The different production mechanisms are shown in blue (GF), red (VBF), green/gray (VH), and purple ($t\bar{t}h$). [15].

The decay width of the SM Higgs boson is predicted to be around $\Gamma_{\mathrm{SM}} = 4$ MeV (for $m_H = 125$ GeV) [14]. As the width of a mass peak is inversely proportional to the lifetime of a particle, $t = \hbar/\Gamma = 1.6 \cdot 10^{-22}$ s, this indicates that the Higgs boson is unstable and will decay very shortly after it has been produced. As a result, direct searches at colliders will always observe the decay products of the Higgs boson, rather than the particle itself.

In the Higgs boson search at LHC, the considered mass range can be divided into three broad regions: low mass ($m_H = 110 - 150$ GeV), medium mass ($150 - 200$ GeV) and high mass ($200 - 1000$ GeV). Different decay channels are relevant in different regions. The branching fractions of the Higgs boson into various products, in function of mass m_H, are shown in Fig. 2.3. Of the channels in this figure, five were used in the analyses that ultimately led to the discovery of the Higgs boson: $H \to \gamma\gamma$, $H \to ZZ$, $H \to W^+W^-$, $H \to \tau^+\tau^-$ and $H \to b\bar{b}$.

In the *low mass region*, decays to two fermions, for example the decays to $b\bar{b}$ or $\tau^+\tau^-$, represent the largest fraction of the width. However, these fermions decay further, in many cases leading to multijet final states, and at hadron colliders any search in channels containing many jets is

rendered difficult by the presence of a large QCD multijet background. Next to the poor S/B, also the mass resolution in these channels is rather low.

More important, at much smaller branching fractions, are the decays to $\gamma\gamma$ and ZZ. While they contribute only a small fraction of the rate, it is possible to obtain a S/B ratio in these channels that is much better than that of most two-fermion decays. Their final states are very clean, and γ and lepton identification guarantee very good mass resolution.

In the *medium and high mass ranges*, decays to a vector boson pair are dominant. The W^+W^- channel is the leading one, but doesn't provide good mass resolution. In the high mass region, it is again the decay to ZZ, with further decay to four leptons, that is most promising.

After the discovery, two fermion channels such as the $b\bar{b}$ and $\tau^+\tau^-$ ones, become more important again. They depend on the coupling of the Higgs boson to fermions, and the strength of this coupling is an important parameter to test the SM hypothesis.

Initially, the VH and $t\bar{t}H$ production modes were used to study the $b\bar{b}$ channel. This thesis presents the complementary analysis of the $H \to b\bar{b}$ decay in the VBF production mode. This analysis is especially relevant, since it provides novel information on two fronts. First, it concerns the coupling of the Higgs boson to down-type fermions, which is currently only badly known. Furthermore, it contributes to the measurement of the VBF mechanism, which has not yet been established with sufficient statistical significance. A more detailed study of this channel will follow in chapter 6.

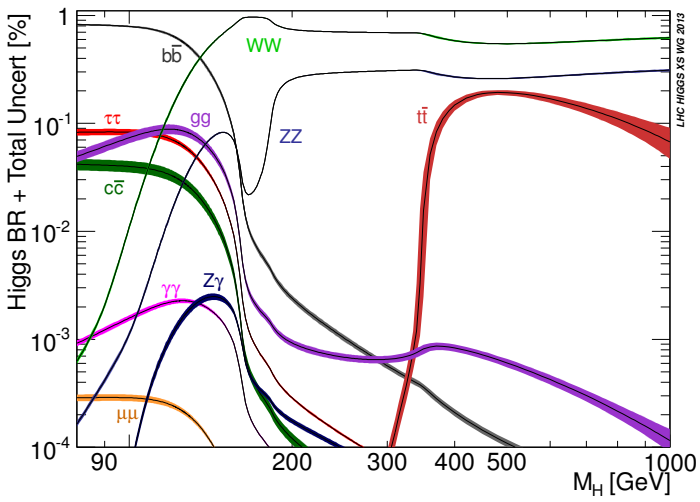


Figure 2.3: Higgs boson branching fractions in function of the Higgs boson mass. Decays to two fermions have the largest branching fractions in the low mass region, while those to two vector bosons are more important in the high mass region. The $\gamma\gamma$ and ZZ decays have a smaller branching fraction overall, but they provide better mass resolution than the other channels. [15].

2.2 Higgs boson searches

As a possible cornerstone of the SM, the Higgs boson has been extensively studied and sought after experimentally ever since the 1970s. Before the discovery at the LHC in 2012, various limits on its mass were set, both from theoretical considerations as well as from experimental results. A few key ones are summarised below.

2.2.1 Limits from theory

A first upper limit for m_H follows from the contribution of Higgs diagrams to the charged vector boson scattering amplitude [16]. While this amplitude does not diverge even when m_H becomes large, there is a problem with perturbative unitarity. If m_H exceeds a certain upper limit, perturbative methods would no longer be valid to describe this interaction. The limit derived for this case is:

$$m_H < \sqrt{\frac{8\pi\sqrt{2}}{3G_F}} \approx 1 \text{ TeV}. \tag{2.1}$$

Another set of limits can be obtained by looking at the renormalisation of parameter $\lambda(\mu_R)$ in the Higgs potential (Eq. 1.12). Important terms in the 1-loop expansion involve the Higgs boson self-coupling (λ) on one hand, and the couplings between the Higgs boson and the top quark (g_t), the W boson (g) and the Z-boson (g, g') on the other hand [16]:

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\ln\mu} = \frac{3}{4\pi^2}\left(\lambda^2 + \lambda\frac{g_t^2}{2} - \frac{g_t^4}{4} + \frac{1}{64}\left[2g^4 + (g^2 + g'^2)^2\right] + \dots\right). \tag{2.2}$$

If the self-coupling parameter is dominant, one can take just the first term of Eq. 2.2 and integrate to find a solution for $\lambda(\mu_R)$:

$$\lambda(\mu_R) = \frac{\lambda(v)}{1 - \frac{3\lambda(v)}{4\pi^2}\cdot\ln\left(\frac{\mu_R}{v}\right)}, \tag{2.3}$$

with the reference scale chosen to be vacuum expectation value v, and μ_R the renormalisation scale. If the self-coupling is too large, this results in a singularity. To avoid this singularity up to a certain physics scale $\mu_R = \Lambda$, $\lambda(v)$ needs to be contained and as a result also $m_H \sim \lambda^2$ has an upper limit. For $\Lambda = \Lambda_{\text{Planck}}$[3] the limit on m_H reads:

$$m_H < \sqrt{\frac{4\pi^2 v^2}{3}\frac{1}{\ln\frac{10^{18}}{v}}} \approx 150 \text{ GeV}. \tag{2.4}$$

[3] $\Lambda_{\text{Planck}} = 10^{18}$ GeV is the Planck scale, the scale at which quantum gravitational effects become relevant.

Alternatively, if the self-coupling parameter is small, the equation becomes [17]:

$$\lambda(\mu_R) = \lambda(v) + \frac{3}{16\pi^2}\frac{1}{v^4}\left(2m_W^4 + m_Z^4 - 4m_t^4\right) \cdot \ln\left(\frac{\mu_R}{v}\right), \tag{2.5}$$

using the $\mathcal{O}(\lambda^0)$ terms in Eq. 2.2 and $g_i = 2\,m_i/v$. To ensure the stability of the vacuum up to Λ, the requirement is $\lambda\,(\Lambda) > 0$. Evaluating again at the Planck scale and transforming to a limit on m_H:

$$m_H > \sqrt{\frac{3}{16\pi^2}\frac{1}{v^2}\left(2m_W^4 + m_Z^4 - 4m_t^4\right) \cdot \ln\left(\frac{10^{18}}{v}\right)} \approx 95 \text{ GeV}. \tag{2.6}$$

These upper and lower bounds are often referred to as the triviality and vacuum stability bounds. A higher order calculation can be found in [18] and the results are shown in Fig. 2.4. If m_H were to exceed either bound at scale Λ, this would indicate that at that scale SM predictions cease to be valid and New Physics sets in. If no Higgs boson would have been found in the range m_H = 95-150 GeV, this would have meant that New Physics was required at a scale $\Lambda < \Lambda_{Planck}$. As will be discussed in the next subsection, over the years experimental searches have narrowed down the possible values of m_H. Taking into account only pre-LHC results, the possibility remains for the SM to be valid all the way up to the Planck scale. With the discovery of the Higgs boson at LHC and the determination of its mass at 125 GeV [19, 20], it appears we sit right at the edge of stability/metastability and more precise measurements will be required to decide the fate of the SM and possible extensions.

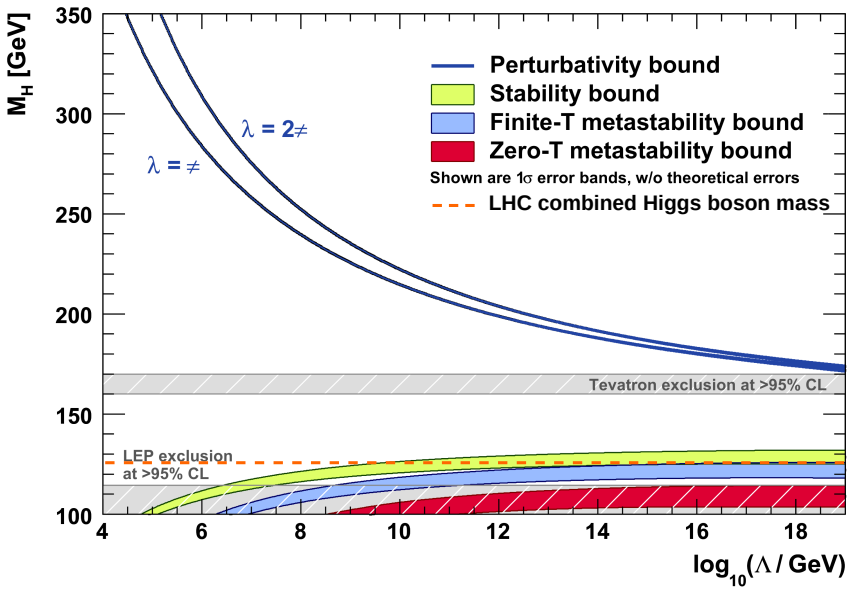


Figure 2.4: Various Higgs boson mass bounds in function of scale Λ. The triviality bound is shown in blue, while the green band indicates the vacuum stability bound. Less strict metastability bounds are shown by the light blue (finite temperature) and red (zero temperature) bands. The latest LHC combined Higgs boson mass result has been added on top in orange. [18, 21].

2.2.2 Limits from experimental searches

The experimental searches for the Higgs boson can be divided into two categories: direct and indirect searches. Three colliders performed dedicated Higgs boson searches: the Large Electron–Positron collider (LEP) at CERN, the Tevatron collider at Fermilab and the current LHC at CERN. A fourth one was proposed, the Superconducting Super Collider, but the project was cancelled in 1993.

A Direct searches

LEP The first contributor to the experimental Higgs boson searches was LEP [22–24]. This electron–positron (e^+e^-) collider was active between 1989 and 2000, and was conceived to perform precision measurements of the Z and W bosons discovered in the 1980s. It initially operated at centre-of-mass energies, $\sqrt{s}$, of around 90 and 161 GeV, the energies required for the production of one Z and two W bosons respectively. From 1998 on however, the beam energies were increased first to 95 GeV and finally to 103 GeV, with the aim of pushing closer towards a possible discovery of the Higgs boson.

The leading Higgs boson production mode at LEP was associated production with a neutral vector boson , while towards the end of the accessible energy range, also contributions from VBF production were expected . With energies below 115 GeV, the dominant decay mode was $H \to b\bar{b}$, with the Z boson decaying either hadronically or leptonically.

In the final dataset consisting of around 2.46 fb^{-1}, no statistically significant excess of events consistent with a Higgs boson signal could be observed on top of the known SM background. As a result a 95% Confidence Level (CL) lower limit for the Higgs boson mass was set at $m_H > 114.4$ GeV [24].

Tevatron Also the Tevatron proton–anti-proton ($p\bar{p}$) collider [22, 25, 26], active between 1983 and 2011, included Higgs boson searches in its research programme. The collider operated at a centre-of-mass energy of up to $\sqrt{s} = 1.96$ TeV, and is best known for its discovery of the top quark in 1995. During its Run II (2001-2011), it conducted searches in the mass range of 100-200 GeV, having the highest chance of discovery for Higgs boson masses between roughly 145 and 190 GeV.

The dominant production mechanism was associated production with a vector boson, followed by GF production and VBF production. The most important decay channels were $H \to b\bar{b}$ for Higgs boson masses below 125 GeV, and $H \to W^+W^-$ for those above 125 GeV.

The final dataset of around 10 fb^{-1} allowed to exclude at 95% CL the presence of a SM Higgs boson in two mass ranges, at the very start and the higher end of the search window: $m_H = 100 - 103$ and $147 - 180$ GeV. In the intermediate mass range of $m_H = 110 - 140$ GeV an

excess of events consistent with the presence of a SM Higgs boson was observed, with a significance of about 2.5 standard deviations [26].

LHC Most recently, the search for the Higgs boson was continued at the LHC, using the ATLAS[4] and CMS[5] detectors. A more detailed description of the LHC and the CMS detector is provided in chapter 3. In 2011, with the LHC operating at a centre-of-mass energy of $\sqrt{s} = 7$ TeV, both ATLAS and CMS focussed on the range $m_H = 110 - 600$ GeV. As discussed in section 2.1, four production mechanisms are in play at the LHC: GF, VBF, VH and $t\bar{t}H$ production.

The ATLAS collaboration [27] considered Higgs boson decays to $\gamma\gamma$, ZZ and W^+W^-. With datasets comprising up to 4.9 fb^{-1}, they excluded three mass ranges at 95% CL: $m_H = 112.9 - 115.5$ GeV, $131 - 238$ GeV and $251 - 466$ GeV. In the intermediate mass range of $m_H = 114 - 146$ GeV, they observed an excess of events consistent with the presence of a SM Higgs boson, with a significance of 2.5 standard deviations.

At the same time, the CMS collaboration [28] studied Higgs boson decays to $\gamma\gamma$, $b\bar{b}$, $\tau^+\tau^-$, W^+W^- and ZZ. Based on a 4.8 fb^{-1} dataset, they published one large 95% CL exclusion range: $m_H = 127 - 600$ GeV, and reported the presence of an excess in the $m_H = 110 - 145$ GeV mass range, consistent with the presence of a SM Higgs boson with a significance of 2.1 standard deviations.

The results of direct searches up until just before the Higgs boson discovery in June 2012 are summarised in Fig. 2.5. In the low mass range the LEP and Tevatron exclusion is confirmed and extended by ATLAS, almost all the way up to $m_H = 122$ GeV, except for a small sliver around 118 GeV. The higher mass Tevatron exclusion range is also confirmed by ATLAS and CMS, as well as extended down to 127 GeV and up to 600 GeV. The unexcluded region between 122 and 127 GeV becomes the most interesting one, with clear 2.5 sigma excesses reported in that vicinity by both ATLAS, CMS and Tevatron.

B Indirect searches

In the SM, quantum loop effects ensure that fundamental parameters, e.g. the mass of the W boson, are dependent on other fundamental parameters, e.g. the top mass or the Higgs boson mass. If a parameter can be measured with high enough precision, the measurement becomes sensitive to the loop effects, and the influencing parameters can be constrained. Using multidimensional fits, it is possible to test the SM and to use precision measurements of well-known parameters to derive information about lesser known ones or even unmeasured ones. Even before the discovery of the top quark at Tevatron, electroweak precision measurements

[4] A Toroidal LHC ApparatuS
[5] Compact Muon Solenoid

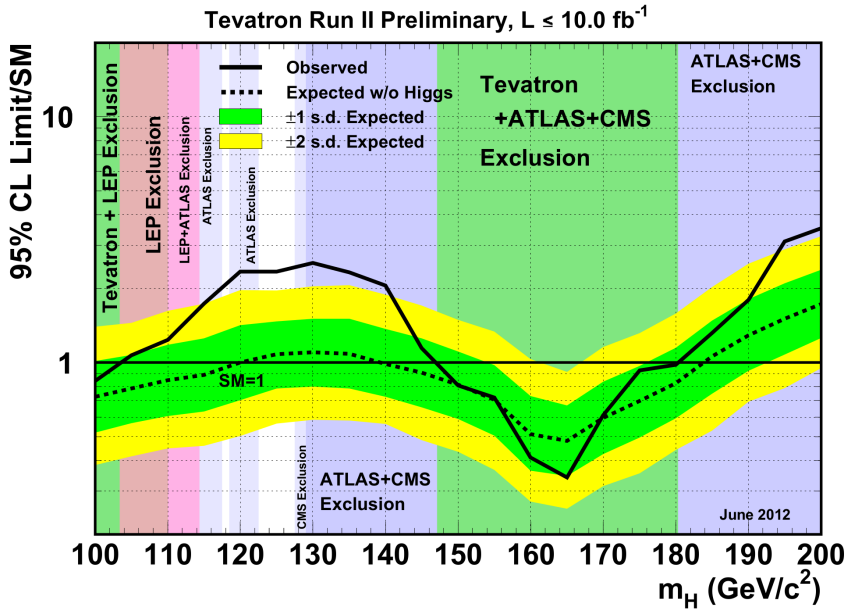


Figure 2.5: Summary of the direct searches up until just before the Higgs boson discovery in July 2012. The observed and expected Tevatron results are shown, with 1σ and 2σ SM expected bands in green and yellow. Exclusion ranges from different experiments are overlaid as vertical bands. [26].

from LEP and other experiments allowed to constrain its mass. In the same way, indirect limits were set on the mass of the Higgs boson before its discovery at LHC.

One example of such a multidimensional fit is the global electroweak fit [29–34], which implements two different approaches: the standard fit and the complete fit. The standard fit includes all available electroweak precision data, but no direct experimental constraints on the Higgs boson mass. The complete fit also includes the latter information. Figure 2.6 shows the evolution of the constraint on the Higgs boson mass as more data is added to the fit. Figure 2.7 shows the standard fit just before the Higgs boson discovery. Note that the uncertainty on the fitted value of m_H in the standard fit remains quite large, even in the later results. This is natural as the dependency on m_H is only logarithmic, and thus the constraint from the fit is reasonably weak. An additional factor is the fact that the precision of several important input parameters, for example the top mass (quadratic dependence), is not yet optimal.

The best fitted value of m_H from the standard (complete) fit, excluding data from the LHC, is 84^{+30}_{-23} (120^{+17}_{-5}) GeV. The latest result from the standard fit, including LHC data, is 93^{+25}_{-21} GeV, which is compatible with the latest direct Higgs boson mass measurement to approximately 1.5 standard deviations. It constitutes as such a weak test of the SM.

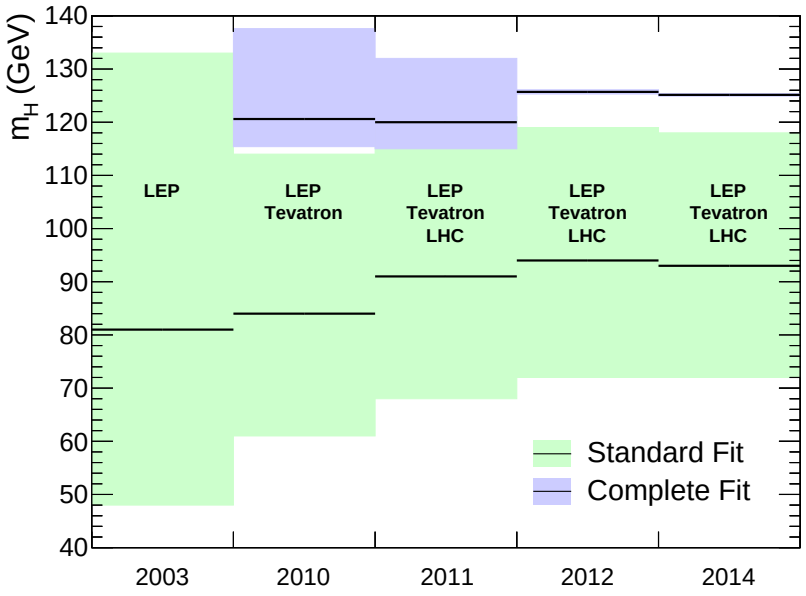


Figure 2.6: Overview of the indirect determination of the Higgs boson mass. Results of the standard (complete) electroweak fit are shown in green (blue). The black lines indicate the value, while the coloured bands cover the $\pm 1\sigma$ intervals.

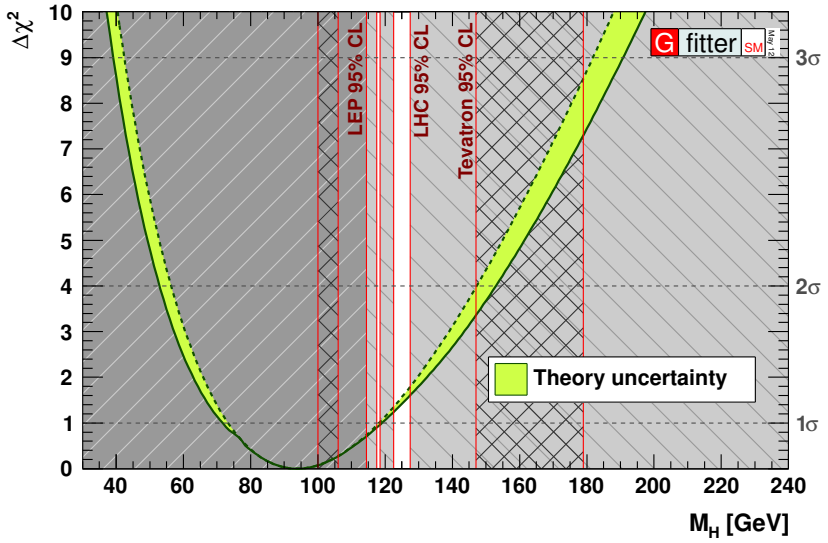


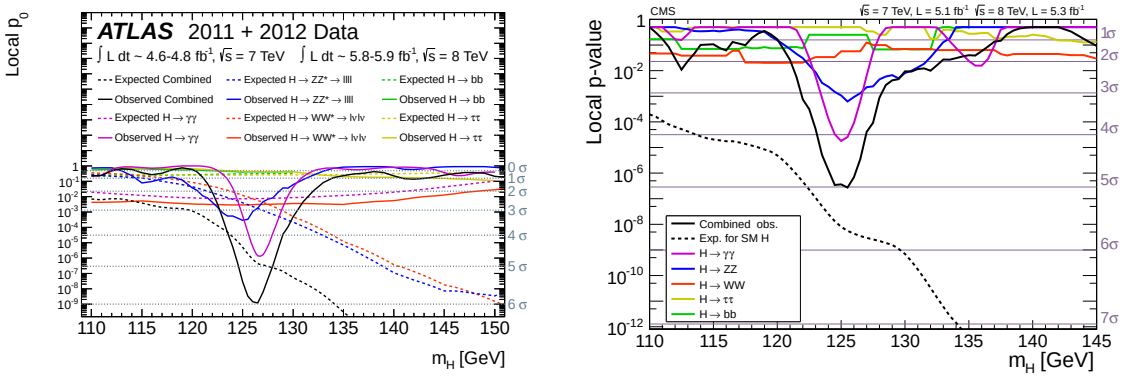
Figure 2.7: Result of the standard electroweak fit just before the Higgs boson discovery in July 2012. Exclusion regions of LEP, Tevatron and LHC are overlaid. [32, 35]

2.3 Higgs boson discovery

The search for the SM Higgs boson culminated in the summer of 2012. Having analysed about 10 fb^{-1} of data, taken at centre-of-mass energies of $\sqrt{s} = 7$ and 8 TeV in 2011 and early 2012, both the ATLAS and CMS collaborations independently confirmed the observation of a new particle with a mass near 125 GeV, compatible with the properties of the SM Higgs boson.

Five decay channels were used in the analyses: $H \to \gamma\gamma$, $H \to ZZ$, $H \to W^+W^-$, $H \to \tau^+\tau^-$ and $H \to b\bar{b}$. In the case of CMS, 7 and 8 TeV data was analysed for all of these channels; ATLAS however, included 7 TeV results only for the $\tau^+\tau^-$ and $b\bar{b}$ channels, and 7 and 8 TeV results for $\gamma\gamma$, ZZ and W^+W^- channels. As noted in section 2.1, the $\gamma\gamma$ and ZZ channels provide high mass resolution, while the other three contribute to the sensitivity of the search, but not so much to the determination of the mass.

Figure 2.8 shows the observed and expected local probabilities for the background to appear as signal-like as the observation, for the different channels as well as for the combination, for both the ATLAS and CMS experiments. An observation of a previously unknown state is generally declared a discovery once it reaches a significance of 5σ or over. The largest observed local significances were 5.9σ (ATLAS) and 5.0σ (CMS), for expected values of 4.9σ and 5.8σ respectively.

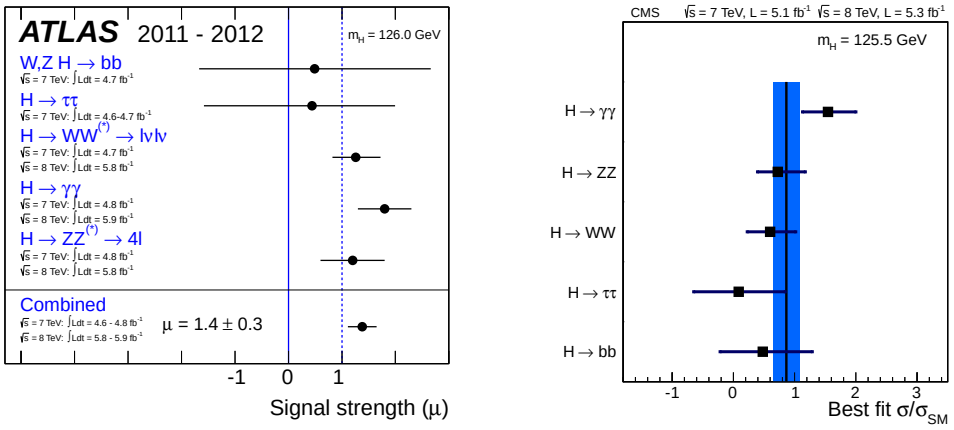


(a) ATLAS, July 2012 **(b)** CMS, July 2012

Figure 2.8: Local probability (as of July 2012) for the observation to be explained by background-only fluctuations, shown for the different channels as well as for the combination, (a) for the ATLAS experiment and (b) for the CMS experiment. *Solid lines* show the observed local p-values. *Dashed lines* show the median expected values under the hypothesis of a Higgs boson at mass m_H. *Horizontal lines* indicate the p-values corresponding to significances of 0-7σ. [19, 20].

The mass of the newly observed particle was measured by both experiments using the $\gamma\gamma$ and ZZ high resolution channels:

$$m_H = 126.0 \pm 0.4\ (\text{stat.}) \pm 0.4\ (\text{syst.})\ \text{GeV} \qquad \text{(ATLAS, July 2012)}$$
$$m_H = 125.3 \pm 0.4\ (\text{stat.}) \pm 0.5\ (\text{syst.})\ \text{GeV} \qquad \text{(CMS, July 2012)}$$



(a) ATLAS, July 2012 **(b)** CMS, July 2012

Figure 2.9: Best-fit values (as of July 2012) of $\mu = \sigma/\sigma_{\mathrm{SM}}$, the production cross section times the relevant branching fractions, relative to the expected SM value, (a) for the ATLAS experiment and (b) for the CMS experiment. In (b) the vertical line and band indicate the combined result. [19, 20].

For each channel and the combination, also the best-fit signal strength with respect to the SM expectation $\mu = \sigma/\sigma_{\mathrm{SM}}$ was determined. The results are shown in Fig. 2.9.

Since the initial observation, various Higgs boson search results and property measurements have been published, including both individual channels and combinations, and covering the entire LHC Run I dataset of about 25 fb^{-1}. Three major updates recently came in the form of a new ATLAS-CMS combination [21], and the ATLAS [36] and CMS [37] Run I legacy results. Figure 2.10 shows the results of the LHC combined Run I mass measurement, based on the high precision $\gamma\gamma$ and ZZ decay channels of both ATLAS and CMS. The mass derived from the full combination is $\mathrm{m_H} = 125.09 \pm 0.21\ (\mathrm{stat.}) \pm 0.11\ (\mathrm{syst.})\ \mathrm{GeV}$.

The Higgs boson was studied in six individual channels by both experiments: $\gamma\gamma$, ZZ, W^+W^-, $\tau^+\tau^-$, $b\bar{b}$ and $\mu^+\mu^-$. As can be seen in Table 2.1, the observation is confirmed in each of the first three channels with a significance of over 5σ; in the remaining channels the significance is lower. Figure 2.11 additionally reports the best-fit signal strength for each of the channels and the combination.

Complementary to the mass and signal strength measurements, also spin-parity studies have been performed for the observed particle, and limits have been set on the width of the resonance. The observed limit on the width is $\Gamma_{\mathrm{H}} < 46$ MeV, for an expected limit of $\gamma_{\mathrm{H}} < 73$ MeV [38]. The SM theory expectation is $\Gamma_{SM} = 4$ MeV [14], which is, for now, far beyond the obtainable instrumental precision. The spin-parity studies [39, 40], mostly in the ZZ, W^+W^- and $\gamma\gamma$ channels, exclude a wide range of spin-1 and spin-2 states at 99% CL or higher. Figure 2.12 shows the outcome of testing various spin-parity models against the SM hypothesis. Assuming that the Higgs boson has quantum numbers $J^P = 0^{++}$, the first steps have been taken in testing the SM compatibility of various coupling strengths.

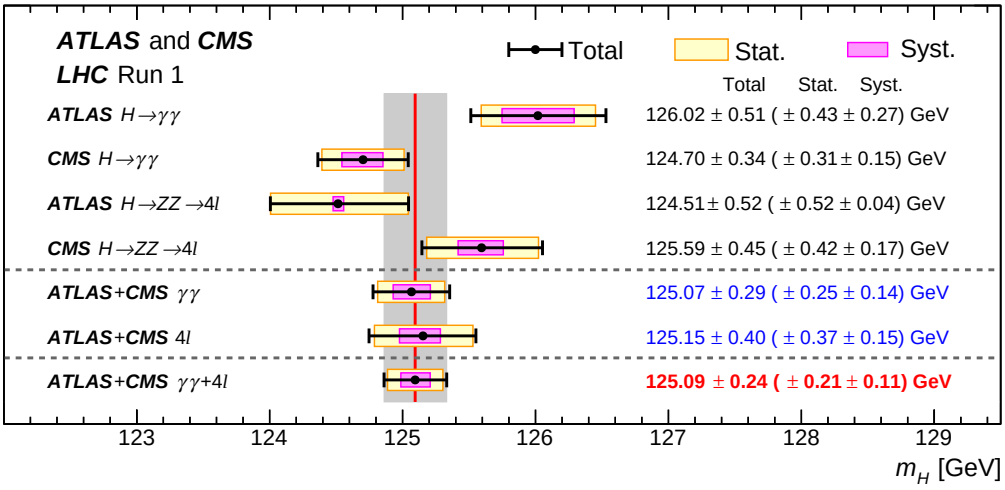
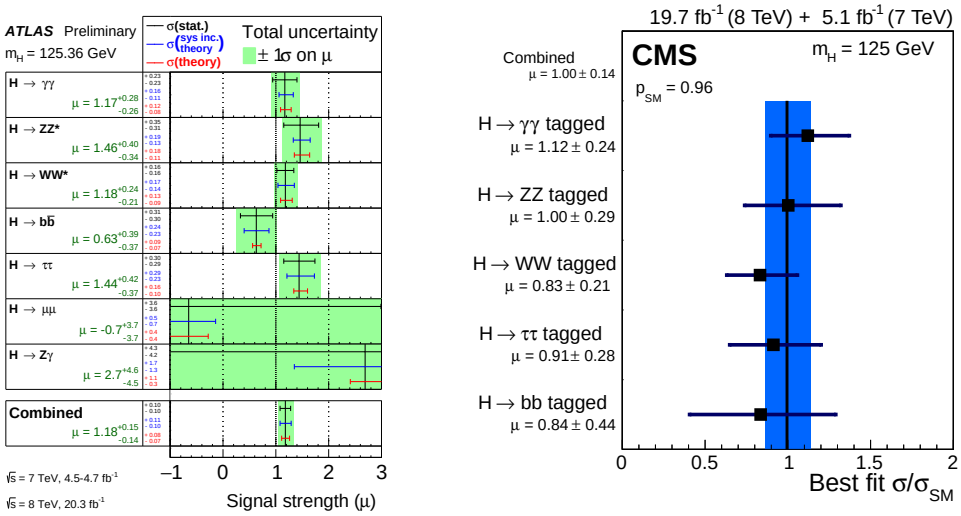


Figure 2.10: Summary of mass measurements (as of March 2015) in the $\gamma\gamma$ and ZZ decay channels of the ATLAS and CMS experiments, and their combinations. The *horizontal bars* indicate the individual measurements and their $\pm 1\sigma$ uncertainties, split up into statistical (*yellow*) and systematic (*pink*). The *vertical red line* and *grey band* repeat the result of the full combination in the bottom line. [21]

Table 2.1: Observed and expected local significances per channel (as of March 2015), assuming m_H = 125.3 GeV (ATLAS) and m_H = 125.0 GeV (CMS). [36, 37]

Channel	**Significance (σ)**			
	ATLAS		**CMS**	
	Observed	Expected	Observed	Expected
$H \to \gamma\gamma$	5.2	4.6	5.6	5.3
$H \to ZZ$	8.1	6.2	6.5	6.3
$H \to W^+W^-$	6.5	5.9	4.7	5.4
$H \to \tau^+\tau^-$	4.5	3.4	3.8	3.9
$H \to b\bar{b}$	1.4	2.6	2.0	2.6
$H \to \mu^+\mu^-$	$\mu < 7.0$	$\mu < 7.2$	< 0.1	0.4

So far, all observations indicate full compatibility with the expectations for a SM Higgs boson. More data will be required to improve the precision of various property measurements, and to either further support the SM nature of the boson, or to disprove it. The newly started Run II at $\sqrt{s} = 13 - 14$ TeV, which is expected to deliver about four times the luminosity of Run I over the next few years, will guarantee the continuity of this research.



(a) ATLAS **(b)** CMS

Figure 2.11: Best-fit signal strength $\mu = \sigma/\sigma_{\mathrm{SM}}$ for various decay channels and their combination, for (a) ATLAS and (b) CMS. In (a) the *vertical black lines* indicate the values and the *green bands* give the overall uncertainties. In (b) the results in the individual channels are given by the *square markers* and *horizontal bars*, while the combination is illustrated by the *vertical line and band*. [36, 37].

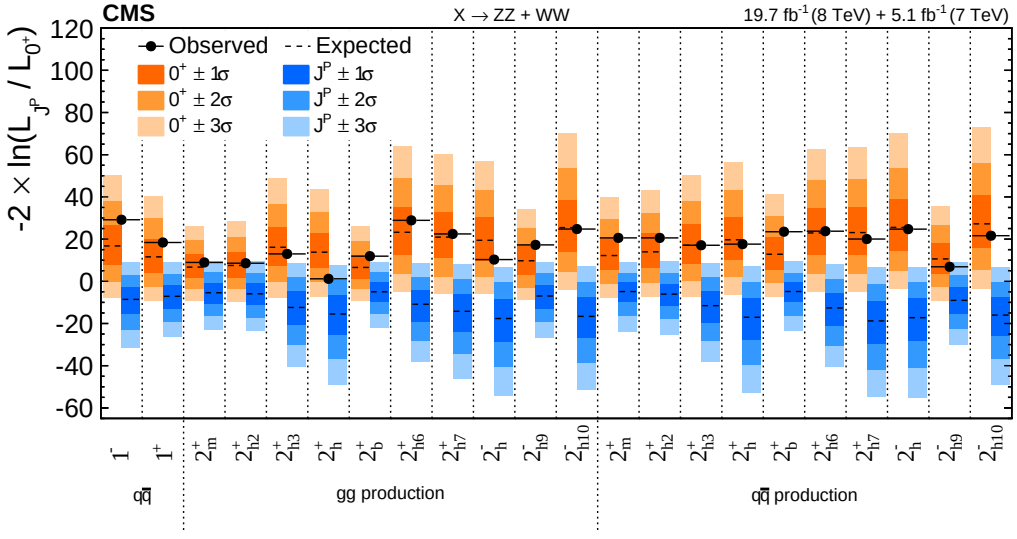


Figure 2.12: Test statistic $-2\Delta \ln L_{J_P}/L_{0_+}$, testing various spin-1 and spin-2 models against the SM Higgs boson hypothesis. Median expected values are shown for the SM Higgs boson (*orange*) and for the alternatives (*blue*). The observation is shown in *black*. [39].

This chapter started with an overview of the possible production modes and decay channels of the SM Higgs boson at collider experiments. It furthermore discussed the indirect and direct searches performed for this particle in the past, starting shortly after its postulation in the 1970s and culminating with its discovery in 2012. Data taking is still ongoing at the LHC, the world's current most powerful particle accelerator, where scientists continue to study the properties of the observed particle and its interactions. In chapter 3, we go into some more detail about the LHC, and introduce the CMS detector, one of the main detectors operating at the LHC.

This chapter also first mentioned the VBF $H \to b\bar{b}$ process, the analysis of which is the main topic of my PhD. All elements of the analysis are discussed in full detail in chapters 6 to 10, contained in part two of this thesis.

Chapter 3

Experimental setup

The VBF $H \to b\bar{b}$ search described in this thesis is based on proton–proton collision data recorded in 2012 at the LHC using the CMS detector, at a centre-of-mass energy $\sqrt{s} = 8$ TeV. The first section of this chapter is dedicated to the LHC, today's leading particle accelerator. Section 3.2 focuses on the CMS detector, one of the four main detectors at the LHC.

3.1 The Large Hadron Collider

The LHC at CERN, located underground at the French-Swiss border near Geneva, is a particle accelerator designed to collide protons or heavy nuclei, at centre-of-mass energies of up to $\sqrt{s} = 14$ TeV. Conceived in the 1980s, the project went through an extensive R&D stage, was approved in 1994, and finally completed in 2007. First beams were circulated in September 2008, and first physics collisions were recorded in 2009.

3.1.1 Accelerator complex

The circular accelerator [41–43], built in the old tunnel of the dismantled LEP collider, uses radiofrequency cavities to accelerate two proton beams, and superconducting dipoles, quadrupoles and higher order magnets to curve and focus them as they travel in opposite directions through the LHC ring. Due to space constraints, the choice was made to use a two-in-one magnet design: both beams run through the same magnets, each in their own separate, contained beam pipe. Several thousand of these magnets are set up along eight straight and eight curved segments of the LHC, altogether forming a rough circle with a circumference of about 27 km. The entire system is cooled to a temperature less than 2 K using superfluid helium-4.

Protons enter the LHC ring after a series of pre-acceleration steps. The sequence starts at linear accelerator Linac 2, where protons are obtained from hydrogen gas and accelerated to 50 MeV before being injected into the Proton Synchrotron Booster. This first circular accelerator in the sequence steps up the energy of the protons to 1.4 GeV. The particles then enter the Proton Synchrotron (PS), which brings them to an energy of 25 GeV and separates them into bunches of about $1.15 \cdot 10^{11}$ protons. The next-to-last step is the Super Proton Synchrotron (SPS), accelerating the proton bunches to 450 GeV, about 1/15th of the intended LHC beam energy. From the SPS, they are injected into the LHC, where they are accelerated to the chosen nominal energy. Maximally about 2800 bunches with 25 ns spacing can be circulated in the LHC. The full system is illustrated in Fig. 3.1.

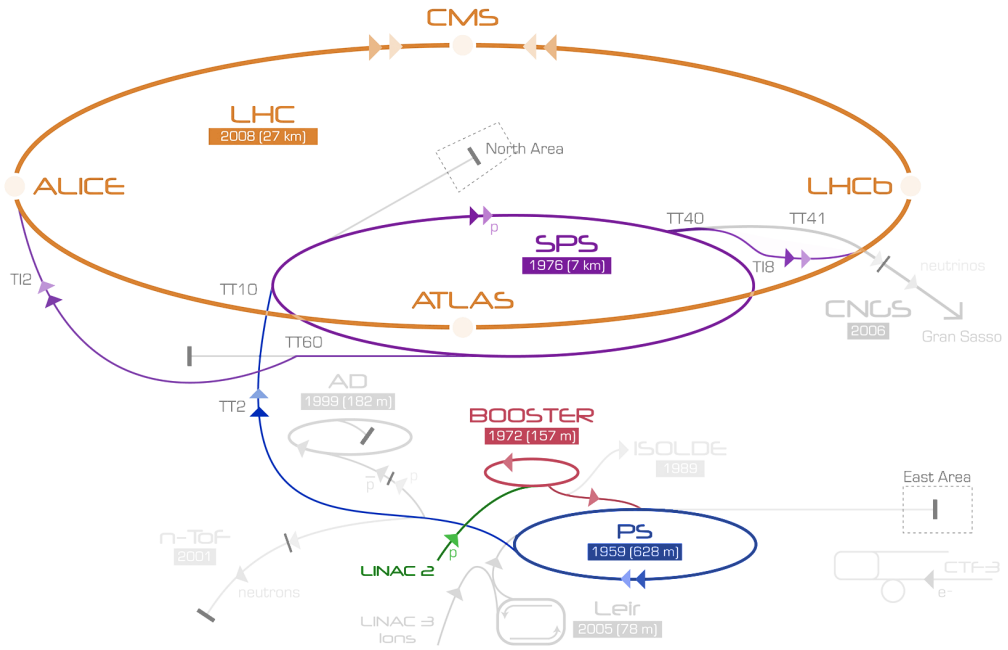


Figure 3.1: Schematic overview of the CERN LHC accelerator complex, showing the journey of the protons through Linac 2 (green), the Booster (red), the Proton Synchrotron (blue) and the Super Proton Synchrotron (purple), into the LHC (orange). On the LHC ring, the four interaction points are indicated: CMS, LHCb, ATLAS and ALICE, clockwise starting from the top. [44].

The flux of particles in an accelerator can be expressed in terms of (instantaneous) *luminosity*, which is proportional to several key accelerator parameters: the number of bunches in the beam, the number of protons in the bunches, the revolution frequency, and the focus of the beam. At the same time, interactions have a certain likelihood to occur, expressed by their *cross section*. The product of the luminosity and the cross section gives the collision rate: $L \cdot \sigma = dN/dt$. *Integrated luminosity* $\mathcal{L}$, the integral of the luminosity over time, can then be used to indicate the amount of produced or recorded events in a given period; it is commonly expressed in inverse femtobarns, fb^{-1}.

The LHC has a design luminosity of 10^{34} cm^{-2} s^{-1}. As the total inelastic proton-proton cross section is about 70 mb [45] at a centre-of-mass energy $\sqrt{s} = 8$ TeV, this results in an inelastic event rate of about 700 MHz. At 25 ns spacing, the maximum crossing rate in the 27 km ring is 40 MHz. With around 2800 proton bunches however, some larger gaps are allowed and the crossing rate comes to about 31.6 MHz. Comparing this to the 700 MHz inelastic event rate shows that multiple interactions will happen in each bunch crossing.

The primary, most energetic (hardest) interaction in a high luminosity bunch crossing (which is usually the interaction scientists try to select for study) will be clouded by additional, softer interactions, all of which can leave signals in the detectors. In addition, due to the small bunch spacing, also interactions from previous or subsequent bunch crossings may contribute to the event. The total of all extra activity in an event – on top of the primary interaction – is

called the pile-up. It can be split into In-time pile-up (itPU), originating from the same bunch crossing as the primary interaction, and out-of-time pile-up (ootPU), originating from previous or subsequent bunch crossings. During event reconstruction in a detector, it will be important to try and separate contributions of the primary and pile-up interactions, and to correct for the latter when reconstructing observables for the former.

As the instantaneous luminosity was increased over the course of three years of data taking, also the average number of pile-up interactions increased, from just a few in 2010 to roughly 20-25 in 2012. Fig. 3.2 shows the peak number of interactions versus time.

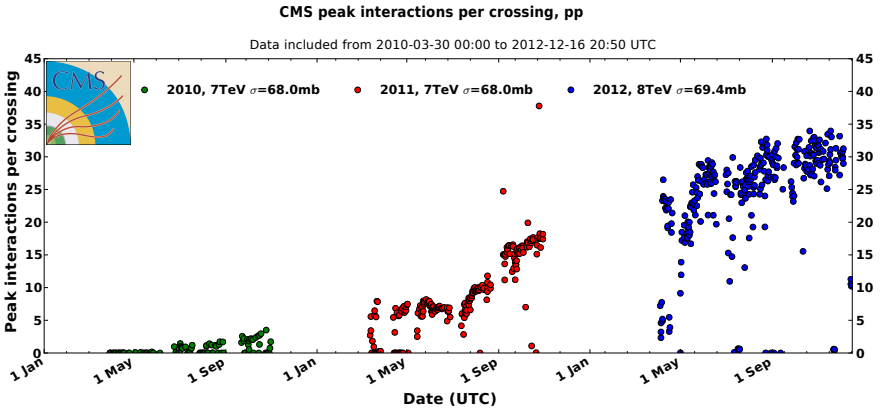


Figure 3.2: Peak number of interactions at the LHC versus time, from 2010 to 2012. [46].

3.1.2 Detector experiments

Protons can be brought to collision at four interaction points along the accelerator. Four large LHC detector experiments are housed in large caverns around these interaction points: ATLAS [47], CMS [48], LHCb [49] and ALICE [50][1]. Three smaller experiments, LHCf [51], TOTEM [52] and MoEDAL [53], share respectively the ATLAS, CMS and LHCb caverns. The locations of the four interaction points are indicated on Fig. 3.1.

The two large general purpose high luminosity detector experiments, ATLAS and CMS, were designed to achieve the same goals, independently, and via two different approaches. Their physics programme includes the search for the Higgs boson, precision testing of the SM, and probing for New Physics at increased collision energies. The other two large detector experiments are ALICE, devised primarily to study ion collisions and to learn about the strong force at large energy densities, including the formation of quark-gluon plasma; and LHCb, studying B-physics and the matter–anti-matter imbalance. The smaller TOTEM and LHCf experiments study forward physics, while MoEDAL looks for magnetic monopoles.

[1] Abbr. LHCb (Large Hadron Collider beauty), ALICE (A Large Ion Collider Experiment), LHCf (Large Hadron Collider forward), TOTEM (TOTal Elastic and diffractive cross section Measurement) and MoEDAL (Monopole and Exotics Detector At the LHC).

In order to tackle these research objectives efficiently, various indispensable core properties have been identified and implemented in the detector designs. As the collision rate at the LHC is very large, the detectors first of all need to have a very fast response, to identify and record the necessary data. Secondly, they need to be as hermetic as possible and must have a fine granularity; this to be able to measure all particles going out in all directions, such that for example energy balances can be made. Thirdly, they need to be able to withstand high levels of radiation over large periods of time, in order to avoid requiring frequent expensive maintenance.

Lastly, both ATLAS and CMS made it a priority goal to be able to measure charged particle momenta with high resolution; the resolution can be expressed as $(\Delta p)/p \approx 1/(BL^2)$, with p the momentum, B the magnetic field strength, and L the size. In the case of ATLAS, fraction $1/(BL^2)$ is made small through the size parameter: the ATLAS detector is very large. CMS on the other hand, chose to design their detector using a very large magnetic field B, and is more compact.

In the following section 3.2, the CMS detector will be discussed in more detail, as the analysis covered in this thesis makes use of CMS data.

3.1.3 Operation

Run I The LHC operation timeline allocates alternating periods of running and upgrades / maintenance. After obligatory maintenance in 2008 and commissioning in 2009, actual data taking finally took off in 2010-2011. This was the start of Run I. Late in 2010, the LHC was operated at $\sqrt{s} = 7$ TeV and a dataset of a little under 50 pb^{-1} was recorded by the experiments. This milestone was easily surpassed by over a factor of hundred in 2011, still running at the same collider energy. In 2012, the centre-of-mass energy was increased to $\sqrt{s} = 8$ TeV, and the LHC produced roughly 25 fb^{-1} of collision data. Run I was finally concluded early in 2013, with just under 30 fb^{-1} on tape. The progress of data recording at LHC is illustrated in Fig. 3.3.

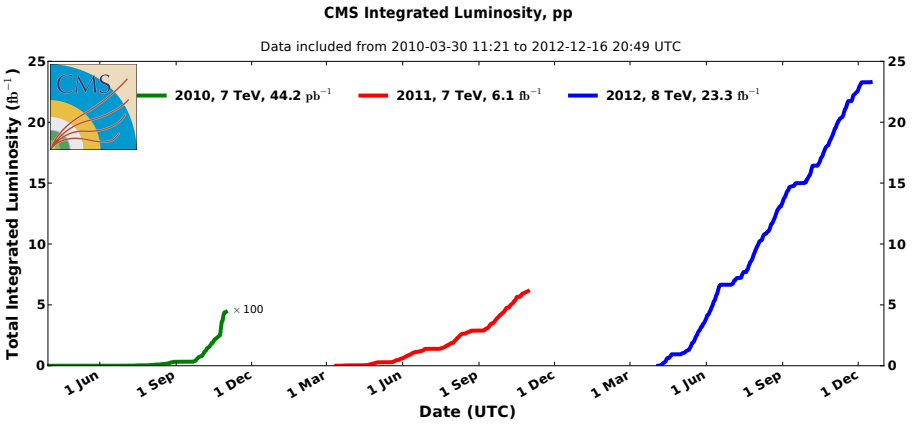


Figure 3.3: Total integrated luminosity over time for LHC Run I. [46].

Long Shutdown I Between early 2013 and mid 2015, the LHC had its first long shutdown period. During these two years, various parts of the accelerator were repaired, upgraded and optimised for use in Run II. The main priority was a serious refurbishment of the LHC magnets, but also the PS and the SPS pre-accelerators underwent upgrades. At the same time, also the experiments serviced their detectors and put new elements in place, in some cases looking as far forward as upgrades planned after Long Shutdown II. The CMS experiment for example replaced its beampipe (in anticipation of a future tracker upgrade), made changes which allow to operate the tracker at a colder temperature, performed repairs on the electromagnetic calorimeter, and made improvements to the muon detector. The work was completed in carefully planned stages, allowing the entire complex to be tested and restarted by May 2015.

Run II The second run of the LHC is planned to extend from mid 2015 to late 2018. The nominal beam energy will be increased first to 6.5 TeV, and finally to the design 7 TeV. The peak luminosity will exceed its design value by about a factor of two. Intending to run with about 2500 bunches at 25 ns spacing, the integrated luminosity should be almost 150 fb^{-1} over the course of the entire run period. Given this increased luminosity as well as the higher collision energy, also the average pile-up will be higher than in Run I, around 40.

The physics programme for Run II includes continued precision studies of the SM and the Higgs boson discovered in 2012, and further search for New Physics at higher energies.

3.2 The Compact Muon Solenoid

The CMS detector [48,54] is a compact and dense, cylindrical, modular, general purpose detector. It is composed of a barrel and two closing end-caps, and features a layered structure, designed to realise optimal detection and measurement of different types of particles. In addition, the detector is as hermetic as possible, covering the volume around the beam pipe with as little gaps as possible, and maximising the acceptance for particles created in the collisions and travelling outward from the interaction point. Figure 3.4 shows a schematic view of CMS detector.

A powerful superconducting solenoid and an iron return yoke act as the basis and frame of the experiment. Three subdetectors are located inside the solenoid: the tracker, the Electromagnetic CALorimeter (ECAL) and the Hadron CALorimeter (HCAL). The muon chambers, on the other hand, are integrated in the iron return yoke, outside the solenoid.

The subdetectors each serve specific purposes, and jointly provide an optimised understanding of collision events. Immediately around the beam pipe, the tracker measures the trajectories of charged particles. This happens in a non destructive manner, the particles travel onward through the rest of the detector. Next to the trajectories, also the locations where particles originate – the vertices – are derived. In the second and third layer, the ECAL and HCAL

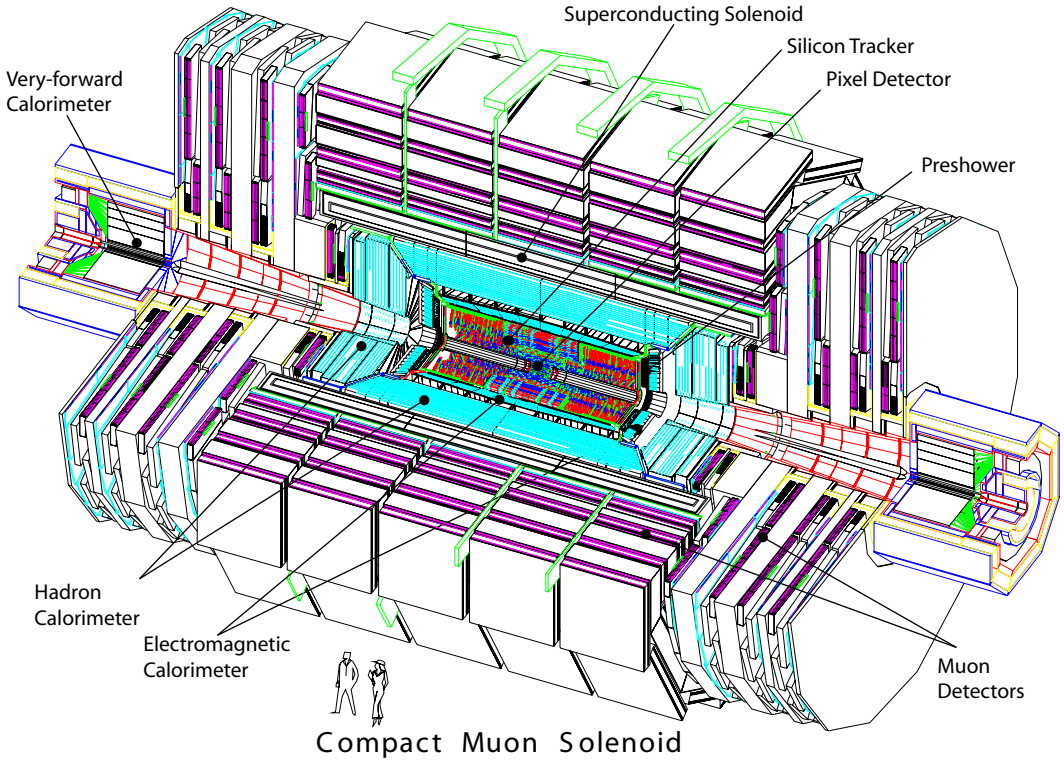


Figure 3.4: Schematic view of the CMS detector, showing the layered structure and the various submodules. [48].

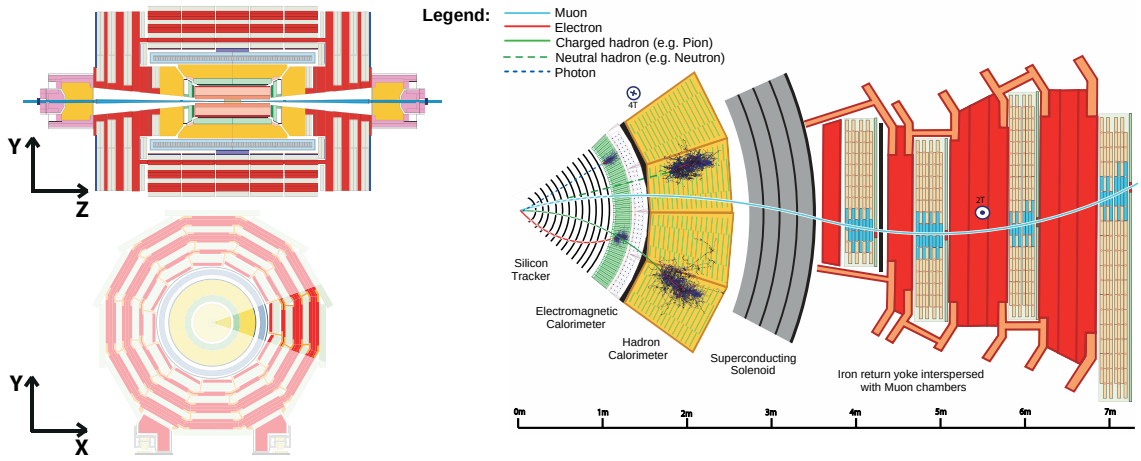


Figure 3.5: Illustration of the CMS coordinate system (left) and transverse slice of the detector (right), with example particle trajectories. Charged particles (*solid lines*) have curved trajectories which are measured in the silicon tracker. The energy of electromagnetically interacting particles (e.g. electrons (*red line*) or photons (*dashed blue line)*) is measured in the electromagnetic calorimeter. Hadronically interacting particles (e.g. pions (*green line*) or neutrons (*dashed green line*)) are not stopped there and have their energies measured in the hadron calorimeter. Finally the momentum of muons (*light blue line*) is measured in the muon chambers. [55].

calorimeters determine the energy of electromagnetically (e.g. electrons and photons) and hadronically (e.g. protons and pions) interacting particles. This involves interactions of the particles with detector material, steadily reducing their energy and effectively stopping most of them. The final fourth layer, outside the solenoid, consists of the muon detectors. These measure the momenta of the muons, minimum ionising particles which lose only little energy in the calorimeters and as such are not stopped there. The path of different types of particles through the detector is illustrated in Fig. 3.5.

There is also a category of particles not stopped or even seen by the detector: those that interact only through the weak force or through new physics processes, and hence don't leave signals in any of the modules. An example of a known undetected particle is the neutrino. It is however possible to infer the presence of neutrinos in a event, indirectly, through the use of conservation laws. As the CMS detector is not fully hermetic in the longitudinal direction (particles can escape through the beam pipe), one cannot use a 3-momentum balance. In the transverse plane though, this is not an issue, and a transverse energy balance can be made. The energy of all undetected particles is then covered in the missing transverse energy $\mathrm{E}_{\mathrm{T}}^{\mathrm{miss}}$.

Finally all collision events will be consistently described in the CMS coordinate system (cf. Fig. 3.5), a right-handed system with the origin at the nominal interaction point. The x-axis points to the centre of the LHC ring, the y-axis points vertically upward, and the z-axis points along the beam in counter-clockwise direction (as seen from the top). Azimuthal angle ϕ is defined from the x-axis in the x-y plane, with r as the radial component. Polar angle θ is defined similarly from the z-axis in the r-z plane. Instead of θ often pseudorapidity η is used, defined as $\eta = -\ln(\tan\frac{\theta}{2})$. Small values of η indicate elements transverse with respect to the beam

direction (central elements, $|\eta| < 2.5$), while large values of η indicate elements more parallel to the beam direction (forward elements, $|\eta| > 2.5$).

The following subsections give more information about the various subdetectors of CMS.

3.2.1 Superconducting solenoid

For CMS, one of the key elements in performing high precision measurements of charged particle trajectories, i.e. in achieving good particle momentum resolution, is the large bending power of the magnet, curving the trajectories of all charged particles. A superconducting solenoid was specifically designed to meet this task, and was at the time of constructing the largest magnet ever built. It measures 6.3 m in diameter and is 12.5 m long, and generates a 3.8 T magnetic field. This field is guided and closed by a 10 000 tonne iron yoke, which has the additional property of stopping almost all particles (except muons and neutrinos). The yoke extends to about 14.6 m in diameter and 21.6 m in length, and encloses all subdetectors.

3.2.2 Tracker

The CMS tracker [56, 57], used mainly to measure charged tracks (particle trajectories) and to reconstruct vertices (particle origins)[2] of particles created in LHC collisions, is setup centrally inside the solenoid. Track and vertex information is crucial for particle identification, especially when dealing with high pile-up level (many tracks and vertices) and long-lived particles (secondary vertices). A schematic view of the tracker is provided in Fig. 3.6. One of its main properties is that it is entirely silicon based. It consists of two parts: the silicon pixel tracker and the silicon strip tracker. In both cases, the design features a cylindrical barrel region and two end-cap regions closing the barrel, typical for the entire CMS detector. The tracker covers pseudorapidities up to $|\eta| < 2.5$.

In addition to being very precise, the CMS tracker is required to be very fast. Particle interactions occur at very high rates, and being able to measure all tracks in time is essential in the experiment. The tracker is designed to achieve efficiencies above 95% for tracks with transverse momentum $p_T > 1$ GeV. At the centre of the detector, where the number of hits per area is largest, pixel detectors are used. The pixel tracker extends outward from near the interaction point to a radius of 10.2 cm. Further out from there, the strip tracker is built with larger and more rectangular detector cells. The increased length limits the required number of readout channels.

[2] The vertex of the primary interaction is called the primary vertex (PV). Other types of vertices include secondary vertices (SV), vertices displaced from the PV, indicating long-lived particles; and pile-up vertices, vertices of additional pile-up interactions.

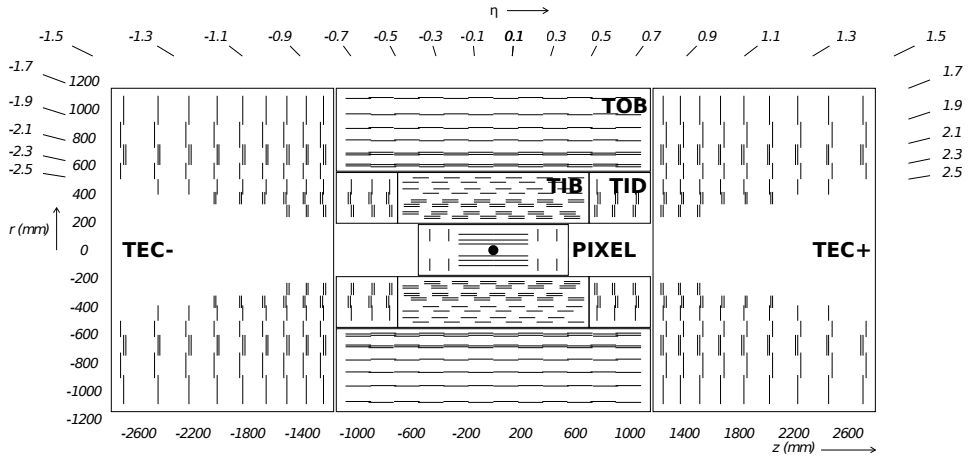


Figure 3.6: Schematic cross section through the CMS tracker. Each line represents a detector module. Double lines indicate back-to-back modules which deliver stereo hits. [48].

The pixel tracker has a triple barrel structure (layers at r = 4.4, 7.3 and 10.2 cm) and double layered end-caps. Each layer is meant to provide a measurement for each charged particle coming through. Around the pixel tracker sits the strip tracker, extending outward to approximately r = 1.15 m. The inner half of it features the tracker inner barrel (TIB, 4 layers) and two tracker inner disks (TID, 3 layers). Up to four further hits are expected to be registered as particles pass through these layers. The outer half and final tracker layer then has the tracker outer barrel (TOB, 6 layers) and two tracker end-caps (TEC, 9 layers). Finally, some of the layers are doubled, with back-to-back silicon cells, in order to provide a measurement of the second (r,z) coordinate as well. The tracker layout is summarised in Table 3.1.

In order to reconstruct tracks from the signals in tracker, as well as to derive particle momenta, various algorithms were developed. These algorithms, as well as the tracker performance, will be discussed in section 5.1.

Table 3.1: Summary of the tracker layout and dimensions.

subsystem	number of layers	pitch (μm)	location (cm)
pixel barrel	3 cylinders	100 x 150	$4.4 < r < 10.2$
TIB	4 cylinders	80-120	$20 < r < 55$
TOB	6 cylinders	122-183	$55 < r < 116$
pixel end-cap	2 disks	100 x 150	$34.5 < \|z\| < 46.5$
TID	3 disks	100-141	$58 < \|z\| < 124$
TEC	9 disks	97-184	$124 < \|z\| < 282$

3.2.3 Calorimeters

Around the tracker, two calorimeter layers take care of particle energy measurements and provide complementary track information. Calorimeters exist of mainly absorber material in which incidenting particles initiate secondary showers that can be measured. As the original particles are absorbed in the process, this is a destructive technique. Given the inherent differences between electromagnetic and hadronic interactions, two distinct subdetector designs are usually required to achieve optimal measurement of both electrons and photons, and strongly interacting particles.

Showers in electromagnetic calorimeters are formed either through pair-production initiated by a photon ($\gamma \to e^+e^-$), or through bremsstrahlung off of an electron ($e^\pm \to \gamma e^\pm$). A shower grows until the energy of the particles is sufficiently low for them to be captured and absorbed by the surrounding detector material. The amount of energy a particle looses when crossing through a certain amount of material can be expressed in terms of radiation length X_0, the average distance covered by an electron before its energy is reduced through bremsstrahlung by a factor $1/e$. One radiation length is also equivalent to 7/9th of the mean free path of the photon, the average distance covered before pair production occurs. In the transverse direction, the extent of a shower is characterised by Molière radius R_M; a cylinder of $r = R_M$ contains on average 90% of the shower energy. Both X_0 and R_M are naturally material dependent.

In a hadron calorimeter, hadrons shower through inelastic interactions, including multi-particle production (e.g. pion pair production) and nuclear decays. As hadron showers include neutral pions, which decay electromagnetically, they will include a component which can be registered in an electromagnetic calorimeter. The characteristic length used in relation to hadronic showers is the nuclear interaction length λ_I, the average distance crossed before undergoing an inelastic nuclear interaction. Also λ_I is material dependent, and generally larger than X_0.

Considering the larger characteristic length of hadronic showers as well as their inclusion of an electromagnetic component, it makes sense for electromagnetic calorimeters to precede hadronic calorimeters in the detector layout.

Technology-wise, two broad categories of calorimeters exist: homogeneous calorimeters, consisting entirely of active materials; and sampling calorimeters, alternating active and passive materials. Homogeneous calorimeters can provide excellent energy resolution, but are less suited for particle identification (lower position accuracy) or the measurement of hadron showers (large interaction length makes for large thickness requirements). Sampling calorimeters on the other hand, have lower energy resolution, but provide better spatial resolution (easier to segment) and thanks to absorber layers manage to contain hadron showers with smaller material thickness. The lower thickness requirement, which makes the detector cheaper, is often the main argument for the use of sampling calorimeters. The different CMS calorimeters are discussed in more detail in the following paragraphs, and illustrated in Fig. 3.7. More information can be found in [48, 58].

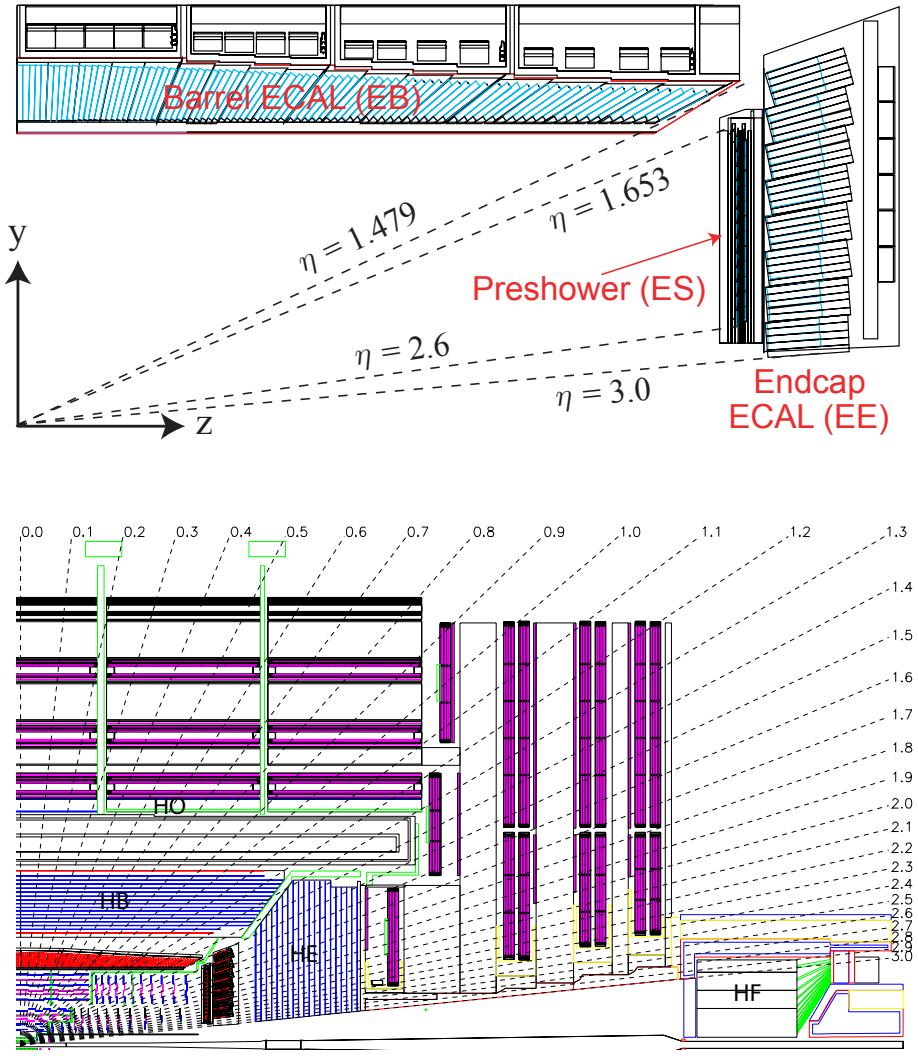


Figure 3.7: Overview in the y-z plan of (a) the CMS electromagnetic calorimeter and (b) the CMS hadron calorimeter. [48, 59].

Electromagnetic calorimeter

The CMS ECAL [59, 60] is a hermetic, homogeneous scintillating crystal calorimeter. The active material is lead tungstenate ($PbWO_4$). It features a barrel shape with two end-caps, and includes two preshower detectors in front of the end-caps. The barrel extends from $r = 1.3$ m to $r = 1.8$ m and the end-cap section extends from $z = 3$ m to $z = 3.8$ m. One of the main design goals of the ECAL was to be able to resolve the photons of e.g. a $H \to \gamma\gamma$ decay with high precision. A homogeneous crystal calorimeter allows this, and in addition meets the criteria of being fast, having fine granularity and being radiation resistant.

Particles entering the calorimeter instigate scintillation, which is registered using photodetectors. With 80% light emission withing 25 ns, the $PbWO_4$ scintillators are well suited to the LHC design bunch crossing time. The scintillation light is collected using silicon avalanche photodiodes and vacuum phototriodes. Shape, polishing and positioning of the crystals is optimised to maximise light collection; in both the barrel and the end-caps the crystals are oriented towards the interaction point, with a small tilt to optimise the coverage.

The ECAL barrel (EB) covers a pseudorapidity range of $|\eta| < 1.479$. It has a fine η-ϕ granularity, with a square cross section of 0.0174×0.0174 (1°) at the front of the crystal. The calorimeter is not segmented in the longitudinal direction, the crystals are 2.3 m long (25.8 X_0) and cover the entire thickness of the detector. The crystals are further grouped in submodules (10 crystals), modules (40-50 submodules) and supermodules (4 modules). Each supermodule then corresponds to 20° in ϕ.

The end-caps (EE) extend the pseudorapidity range from $|\eta| = 1.479$ to 3. Similar to the barrel, the end-cap is segmented in η-ϕ, and not in the longitudinal direction. The crystals have a front face of 28.6×28.6 mm^2 and are 22 cm long (24.7 X_0). They are grouped per 5×5 crystals in supermodules.

Photodiodes and -triodes take care of measuring the scintillation light. The diodes used in the barrel are more efficient than the triodes used in the end-caps, but the triodes have a larger active area, compensating for the lower efficiency. The main reason for the use of triodes, is the higher level of radiation in the end-cap regions; the use of diodes in such conditions would lead to too much electronics noise.

Finally preshower detectors (ES), consisting of two active silicon strip layers and passive lead absorber layers, are placed in front of the ECAL end-caps. They cover pseudorapidity range $1.653 < |\eta| < 2.6$. The first layer is about 2 X_0 thick, the second about 1 X_0, with the strips oriented orthogonal to the ones in the first layer.

The ECAL energy resolution was measured with test beams [48], resulting in:

$$\frac{\sigma(E)}{E} = \frac{2.8\%}{\sqrt{E}} \oplus \frac{12\%}{E} \oplus 0.3\%,$$

including stochastic (shower fluctuations), noise (electrical noise) and constant terms (instrumental effects). The resolution is illustrated in Fig. 3.8.

Hadron calorimeter

The CMS HCAL [61, 62] is a sampling calorimeter extending to $|\eta| = 3$, located between $r = 1.8$ m and the coil at $r = 3.0$ m, and between $z = 3.8$ m and $z = 5.6$ m. Measuring hadronic showers, the HCAL is particularly relevant in the study of jets and E_T^{miss}. It also has a barrel plus end-caps structure.

In general the structure alternates passive brass absorber plate layers with active plastic scintillator layers. The HCAL barrel (HB, $|\eta| < 1.3$) is segmented in wedges azimuthally; 18 wedges each cover 20°, in four sectors. Longitudinally the barrel is split in two (HB^+ and HB^- on each side of the IP), segmented further in 16 η sectors of 0.087 (5°).

The full barrel contains 16 passive layers (∼50 mm thick) and 17 active layers (∼5 mm thick). The amount of material corresponds to 6-11 λ_I, depending on θ, in addition to the 1.1 λ_I of the ECAL. The active layers consist of scintillator tiles, connected by wavelength shifting fibres to hybrid photodiodes, and grouped in ϕ sectors.

The HCAL end-caps (HE) cover the η region $1.3 < |\eta| < 3.0$. The final η sector of the barrel overlaps with the end-cap, making the HCAL as hermetic as possible. The HE layers are slightly thicker than the HB layers, at ∼80 mm (passive) and ∼9 mm (active). In total there is about 10 λ_I of material, including EE and ES. η-ϕ segmentation is 0.087×0.087 (5°) for $|\eta| < 1.6$ and 0.17×0.17 (10°) for $|\eta| \geq 1.6$. Each HE has 18 layers.

As the size of the HCAL is limited by its location inside the solenoid, and its 6-11 λ_I are not sufficient to contain all hadronic showers and avoid leakage into the muon layers, the inner HCAL is complemented by an outer HCAL layer, immediately around the solenoid, embedded in the iron return yoke. The solenoid itself adds $1.4/\sin(\theta)$ λ_I to the material budget.

The iron yoke is segmented in z, forming five wheels. In the middle one, closest to the IP, two additional layers of scintillator are added. In the other wheels one is sufficient. The total material budget then comes to 11.8 λ_I. The wheels have 12 ϕ sectors, separated by 75 mm thick beams that hold the iron yoke. The η-ϕ granularity is again 0.087×0.087.

Forward calorimeters

The HCAL system is further expanded in the forward direction, on both sides of the interaction point, by the forward calorimeter (HF, $11.1 < |z| < 14$ m, $3.0 < |\eta| < 5.2$). The largest challenge for such a detector is being able to cope with the very large particle flux observed in the forward direction.

HF is a sampling quartz-fibre Cerenkov calorimeter. The quartz fibres are sufficiently radiation hard and act as the active material. Charged particles passing through the detector with an energy above the Cerenkov threshold will generate Cerenkov light, of which a fraction is captured by photodetectors. The passive absorber material is provided by 5 mm thick steel plates. The full depth of the detector is about 10 λ_I (165 cm). A fiber layout with fibres of two lengths (one type runs over the full depth of the detector while the other is shorter and only starts after 22 cm), allows to distinguish electromagnetic (e/γ) and hadronic showers, as the former will deposit most of their energy in the first part, while for the latter energy will be equally distributed. The fibres are embedded in grooves in the absorber plates, in a square grid with 5 cm separation. The full structure is doubled and cylindrical. The modules

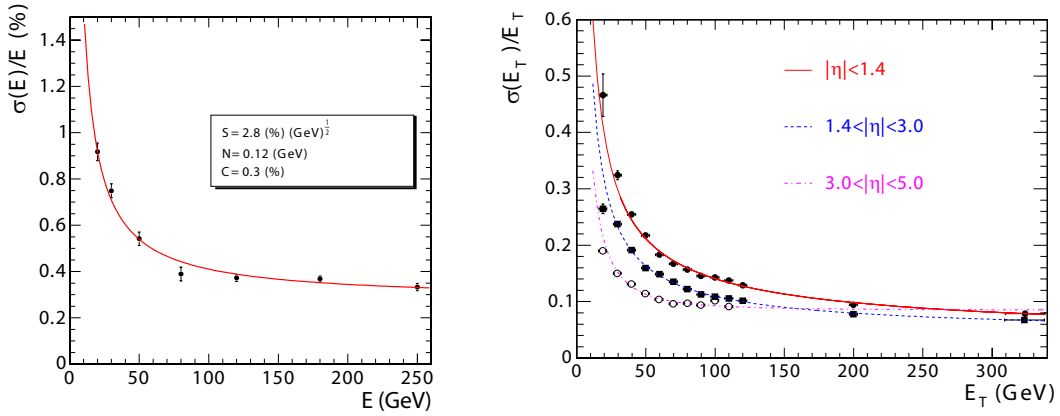


Figure 3.8: Energy resolution of the CMS electromagnetic (left) and hadron (right) calorimeters. [48].

are subdivided in 18 20° wedges. The η-ϕ segmentation is 0.175 × 0.175. In order to protect sensitive detector readout components from radiation, the entire subdetector is embedded in a 40/40/5 cm steel/concrete/polyethylene shield.

The energy resolution of the HCAL and HF calorimeters was measured using test beams [48]:

$$\frac{\sigma(E)}{E} = \frac{84.7\%}{\sqrt{E}} \oplus 7.4\% \qquad \text{(for HCAL)}$$

$$\frac{\sigma(E)}{E} = \frac{198\%}{\sqrt{E}} \oplus 9\% \qquad \text{(for HF).}$$

The resolution is illustrated in Fig. 3.8.

A final contribution to the calorimeter system is the CASTOR very forward quartz-tungsten Cerenkov sampling calorimeter, active in the $-6.6 < \eta < -5.2$ pseudorapidity region. It includes both an electromagnetic and a hadronic section. Build from alternating tungsten (passive) and quartz (active) plates, the EM section has a sampling depth of 20.1 X_0 (0.77 λ_I), while the hadronic section has a sampling depth of 9.24 λ_I; the total then comes to 10 λ_I.

3.2.4 Muon detection

The CMS muon system [63,64] serves a twofold purpose. On one hand it performs a measurement of the muon momentum and charge, independent of the inner tracker measurement. On the other hand, the muon system plays a large role in triggering single- and multi-muon events.

Composed of three types of gaseous detectors, interspersed in the open space inside the iron return yoke of the magnet, the muon spectrometer forms the outer layer of the CMS detector. It sits in the regions $3.8 < r < 7.3$ m and $5.6 < z < 10.6$ m. As mostly all electrons, protons and hadrons are stopped in the calorimeter layers, charged particles detected in the outer layer are highly likely to be muons, and muon identification efficiency is high.

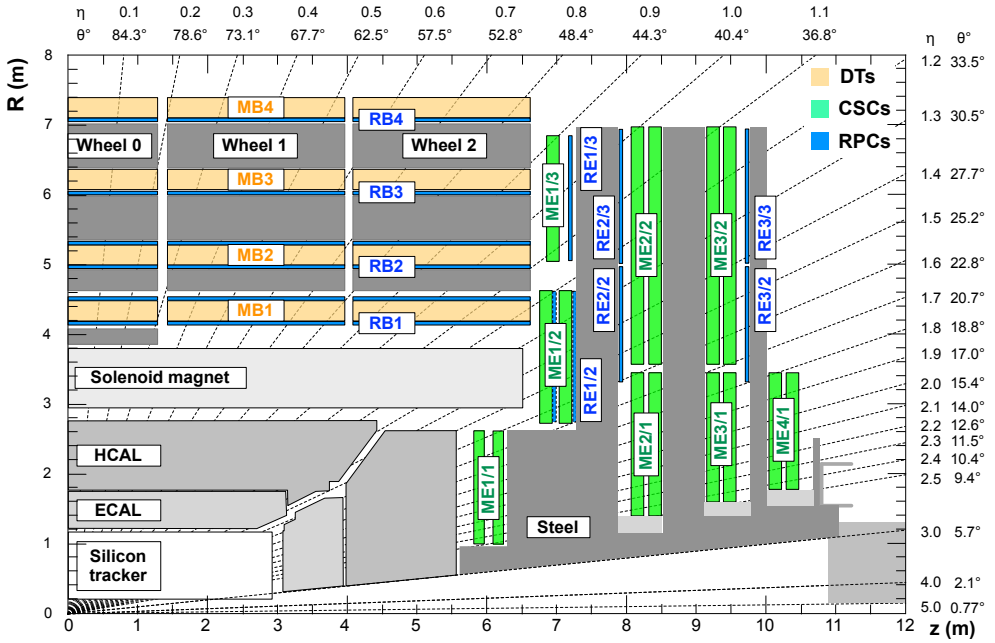


Figure 3.9: Overview of the CMS muon system in the y-z plane. [64]

Similar to the inner layers of the detector, also the muon system features a barrel plus end-caps design. The layout is illustrated in Fig. 3.9. In the barrel region ($|\eta| < 1.2$), where the muon flux, the residual magnetic field and the neutron background are small, drift tubes (DTs) are used. The barrel is subdivided in four stations in radial direction and five wheels in z direction. Each chamber provides two measurements for (r,ϕ) and one for (r,z).

In the end-cap region, where the muon flux, the residual magnetic field and the neutron background are instead high, cathode strip chambers (CSCs) are used. Covering pseudorapidity range $0.9 < |\eta| < 2.4$, the end-caps have four stations in z direction, each with two or three rings in radial direction. Each ring is in turn divided in 18 or 36 segments in ϕ. The chambers have six layers, and provide (r,ϕ) measurements.

The DTs and the CSCs have good spatial resolution and provide measurements of particle properties as well as trigger decisions. They are complemented in both the barrel and end-cap regions ($|\eta| < 1.6$) by resistive plate chambers (RPCs) which have poorer spatial resolution but excellent timing resolution. The RPCs allow for independent prompt trigger decisions and efficient association of tracks to a specific bunch crossing.

The track measurements made by the muon system are complementary to those made by the inner tracker. Making a combination of both leads to improved resolution, as is shown in Fig. 3.10. In the low p_T region (< 100 GeV) the gain achieved by including the muon track information is small, but in the region above $p_T = 100$ GeV the combination visibly improves the result.

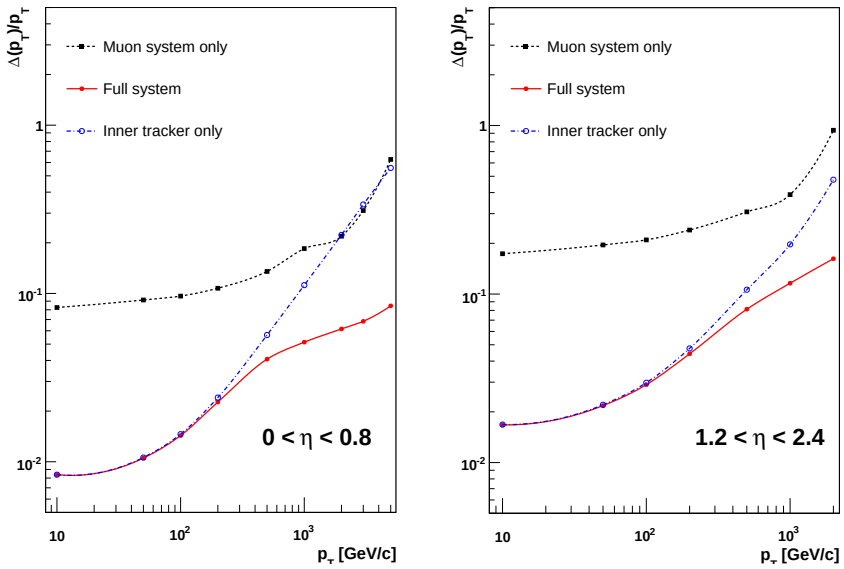


Figure 3.10: Muon transverse momentum resolution as a function of the transverse momentum for the muon system, the inner tracker, and the combination of both; (left) $|\eta| < 0.8$ and (right) $1.2 < |\eta| < 2.4$. [48].

3.2.5 Trigger and data acquisition

The LHC is designed to bring protons to collision with a bunch crossing rate of 40 MHz. At design luminosity, each crossing includes on average about 20 interactions. Even when taking into account occasional wider bunch spacing and counting a reduced rate of 32 MHz, the resulting amount of data is still far too large to be processed and stored for analysis in its entirety.

The average size of an event is around 1 MB. With a storage capacity of the order of 100 MB/s, this means the event rate needs to be reduced by a factor 10^7, from 1 GHz to 100 Hz approximately. To achieve the required reduction, experiments employ triggers: selection algorithms deciding on events' suitability for analysis. This is the first "online" step of the event selection; further "offline" selection will be applied when analyses are performed. The goal of the trigger system is to reject less interesting events, while accepting with as high an efficiency as possible the ones relevant for physics search and analyses. In most cases, decisions are made in several consecutive steps, using different kinds of hardware and accessing progressively more detailed event data. At CMS a two-step trigger system is implemented, with a level-1 (L1) trigger step (using electronics hardware), followed by a software high-level trigger (HLT) step.

Two important criteria for a trigger system are flexibility and efficiency. Flexibility is required to be able to cater to changes in the experiment, and the efficiency needs to be high as well as well understood, in order to guarantee a high yield and good purity for physics analyses.

The *L1 trigger system* uses custom hardware (e.g. application-specific integrated circuits (BASIC's), field-programmable gate array (FPGAs) and discrete logic) and provides real time decisions for all events. To do so, L1 triggers use information from calorimeter and muon systems. Due to the large size of tracker information and the way in which this is stored, it is not possible to use it at L1. The L1 triggers search for key signatures of interesting events: leptons, photons, jets and E_T^{miss}. All trigger objects are saved together with (η,ϕ) information, which means it becomes possible to trigger on objects near or opposite one another, as well as to vary selection thresholds based on object location. The L1 trigger achieves a reduction factor of about 10^4, sending around 100 kHz of data to be considered by the HLT.

The *HLT system* is implemented using commercial hardware; events are processed on a large computer farm. Given the already reduced data rate, it is possible for the HLT triggers to access more and higher resolution data than the L1 triggers. Ultimately the HLT system will use the full event data. Internally, three levels exist: L2, L2.5 and L3 decisions. At L2, still only calorimeter and muon data are used; L2.5 then concerns an intermediate step including partial tracks, and L3 uses full track reconstruction. The HLT reduces the rate with another factor 10^3, bringing the final amount of events processed and written to disk down to the order of 100 Hz.

An L1 decision can take up to 3.2 µs, during which time event data are stored in pipeline memories. Accepted events are stored in random-access memory and send for HLT processing, which takes around 1 s. Once a final decision has been made, all channels are read out and the event is reconstructed and stored on disk by the data acquisition (DAQ) system. The CMS trigger and DAQ system is illustrated in Fig. 3.11, and for more information the interested reader is referred to [65, 66].

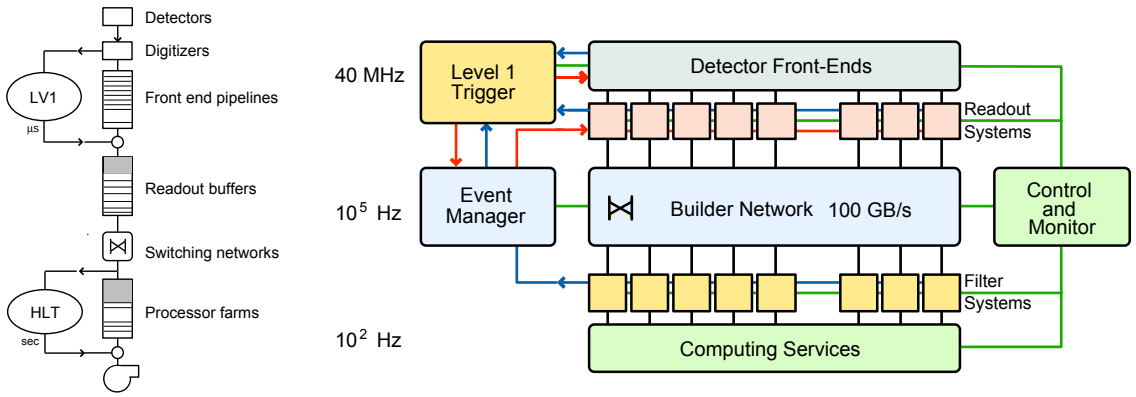


Figure 3.11: Sketches of the CMS trigger and DAQ system. (a) shows the different trigger levels [48], and (b) illustrates the layout of the DAQ system [66].

3.2.6 Detector control and monitoring

To ensure correct operation of the CMS experiment and the quality of the recorded data, it is important to monitor the state of the experiment continuously. This task is taken on by the run control and monitor system (RCMS) and the underlying detector control system (DCS). The RCMS is designed to at all times be able to monitor, configure and control the entire detector. As part of the monitoring infrastructure, the DCS provides information about power supplies, cooling, communication, etc. of all subdetector systems.

Another essential task is the monitoring of the beam conditions. As operation under adverse beam conditions or loss of control over them could lead to serious damaging of the detector, safety systems are in place that can trigger aborts. The LHC then deflects the beam at a specific location and the beam is dumped in a controlled manner.

Finally, the quality of recorded data is ensured by data quality monitoring (DQM) procedures. Both online during data taking (from the CMS control room) and offline at later dates (from remote locations), data is monitored using histograms and graphs specifically designed to quickly signal problems. Data sets can then be flagged in databases as good or bad, and potential issues can be communicated to detector experts. At the end of the DQM chain, a list of data sets certified for physics analysis is produced.

3.2.7 Software framework

The data of the CMS experiment – events accepted by the trigger system and read out through the DAQ system – need to be processed in real-time, moved to permanent storage, and analysed. The CMS software (CMSSW) framework bundles all code developed for these purposes. It imposes a flexible, modular software structure, implementing a large collection of C++ classes (related to various detector subsystems, processing steps and analysis elements), and allowing all of these to be configured and executed using python scripts. The framework structure also allows the release and tracking of different versions, which can be distributed for use in different environments on all kinds of computer systems. To take care of statistical analysis of data and graphical representation of results, CMSSW is interfaced with ROOT [67].

The computing power and storage capacity required by an experiment such as CMS are so large that they cannot feasibly all be made available in one location. Instead, the LHC experiments make use of grid computing: all tasks and data storage are distributed over many computing centres, linked through the worldwide LHC computing grid (WLCG) project. Three categories of computing centres (Tier-0, Tier-1 and Tier-2) deal with three types of CMS data (RAW, RECO and AOD).

The data coming directly from the experiment needs to be processed and stored permanently. Processing includes applying pattern recognition and reconstructing physics objects out of detector hits, applying event filtering, and taking care of data reduction and compression.

RAW format data contains the information as recorded by the detector, as well as the decisions taken by the trigger system; it will be classified automatically into distinct primary datasets based the applicable triggers. *RECO format* data is obtained after further object reconstruction and data reduction, and contains high-level physics objects as well as the collection of detector hits used to compute them. *AOD format* data stands for analysis object data and is the smallest format, meant to be used for physics analyses and containing only the high-level physics and some additional parameters.

The only Tier-0 centre, located at CERN, takes care of the very CPU intensive initial reconstruction of RAW detector data into RECO data, and manages the storage of the data in at least two Tier-1 locations (CERN + another one). The next level of the grid is made up of $\mathcal{O}(10)$ Tier-1 centres, located all over the world. They provide more storage and computing power, and take care of data reprocessing (about once a year, introducing important calibration or software updates). Each Tier-1 centre is also responsible for distributing data to a subset of Tier-2 centres. There are $\mathcal{O}(100)$ Tier-2 centres, responsible for most of the production of simulated data, as well as for making available computing and storage resources to CERN users from around the world.

This chapter provided an overview of the experimental setup. We discussed the LHC and the properties of the collisions produced by it, as well as the detectors recording these collisions, and the operational timeline of the accelerator. We then went into more detail about the CMS detector, since the VBF $H \to b\bar{b}$ search presented in this thesis uses CMS data. We covered the various CMS subdetectors, its trigger and data acquisition systems, and its software framework.

The following two chapters will focus on event simulation (chapter 4) and event reconstruction (chapter 5).

Chapter 4

Event simulation

Theoretical predictions are crucial to experimental research, and inversely experiments are crucial to theory development. For collider experiments, the required theoretical predictions exist in the form of simulated events, produced by event generators implementing the theory models. Experimental searches and measurements often require inputs from simulation, for example for the estimation of background components of signal-enriched data, or for the computation of efficiencies. In addition, simulations may assist in the validation of new experimental analysis techniques, or the development of future experimental setups. Taking the opposite point of view, simulations allow to test new theory models in an experimental context.

Event generators have the task of simulating complete events, taking into account all incoming, intermediate and outgoing particles, including all relevant subprocesses, and retaining various kinds of information about both real and virtual particles at different stages of the process. Fig. 4.1 illustrates the anatomy of an event, including the primary hard process, QCD radiation of coloured particles in both the initial and final states, secondary interactions, hadronisation of partons to colour neutral hadrons, and decays of particles unstable at the time-scale of a collider event; details will be given in the following section, and a review may be found in [68]. Full analytical treatment of these matters is not possible. The simulation cannot be approached as one single large computation, but needs to be broken up into manageable subprocesses, which require different treatments. Fortunately, QCD factorisation allows the simulation process to be split up into contributions involving different momentum scales. Some steps – the ones involving high momentum transfers – may be calculated perturbatively, while others – the ones at lower momentum scales – require non-perturbative approaches. The simulation of each one of these steps can be addressed using Monte Carlo (MC) methods. A key element of MC is the generation of pseudo-random numbers, exhibiting the same statistical fluctuations as true random numbers, but generated in a deterministic way. The pseudo-random numbers are then used to simulate physics processes according to probabilistic distributions, as well as for the computation of integrals. Following a review of the general simulation methodology in section 4.1, a few specific MC generators will be discussed in section 4.2.

When generating events, one is usually interested in a specific hard process. Often, the occurrence of that process is not very frequent, and simply generating random events and selecting the ones of interest would be very inefficient. The factorisation of the simulation process however, means that one is not obliged to go through the full simulation process in a time-ordered manner. Instead, MC generators approach event simulation starting from the hard process, and then work backwards from there to the incoming particles as well as forwards to the outgoing particles. Another important element in MC event generation, is the possibility to simulate

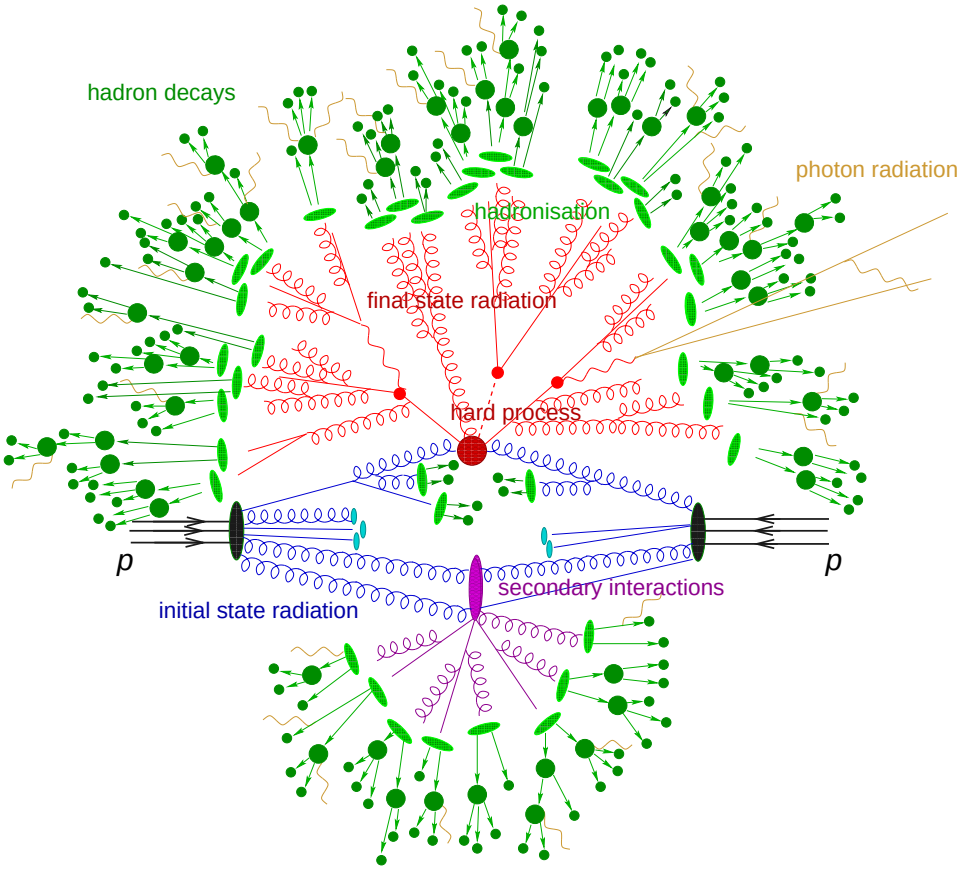


Figure 4.1: Anatomy of a simulated event, including incoming protons (black), hard process (dark red), secondary interactions (purple), final state radiation (red), initial state radiation (blue), hadronisation (light green), and hadron decays (green). Adapted from [69].

full events as they would be observed in a detector. This involves study of instrument response using detector simulation, which is discussed in section 4.3. Comparisons of simulated and experimental data may then happen at two levels: either at detector level, with the generated events passed through detector simulation, or at generator level, with the experimental data corrected for detector effects.

4.1 Methodology

As mentioned above, event generation may be split up in various stages, all of which can be simulated using MC techniques. This section discusses the key concepts of each stage, and how the stages are related.

4.1.1 Hard processes

The core part of event generation is the simulation of the hard process. As it takes place at large momentum scales (or short distances), it can be calculated perturbatively, using Feynman calculus. For hadron-hadron collisions, the inclusive scattering cross section can be written as:

$$\sigma_{ab\to n} = \sum_{a,b} \int dx_a dx_b \int d\phi_n \; f_a(x_a, \mu_F^2) f_b(x_b, \mu_F^2) \; \frac{1}{2\hat{s}} \left|\mathcal{M}_{ab\to n}\right|^2 \left(\phi_n; \mu_F^2, \mu_R^2\right), \tag{4.1}$$

describing the production of final state n, with initial partons a and b. Parton distribution functions (PDFs) $f_i(x_i, \mu_F)$ describe the probability of finding a parton i, with longitudinal momentum fraction x_i of the total proton momentum, at energy or resolution scale μ_F. As the description of partons in the proton is non-perturbative, these functions cannot be computed entirely from theory. Instead, the distributions at fixed energy scales are modelled and fitted to experimental data, and their scale evolution is described using evolution equations [16]. Factorisation and renormalisation scales μ_F and μ_R can be considered as resolution scales separating long and short distances, or non-perturbative and perturbative regions. Finally the differential cross section of the hard process is written as $|\mathcal{M}|^2 d\phi$; the matrix element (scattering amplitude) squared, integrated over phase space ϕ. Both the generation of the matrix element and the phase space integration are well suited to the use of MC methods.

Matrix elements are computed to fixed-order in QCD. First or leading order (LO) calculations are well-established and can be done for processes with a reasonably high number of outgoing particles (high multiplicity). Calculations at higher orders – next-to-leading order (NLO) and next-to-next-to-leading order (NNLO) – are feasible only for lower multiplicity processes. General purpose generators will commonly have a set of matrix element calculations and phase space parametrisations built-in for a set of standard processes at leading order (i.e. $2 \to 1$, $2 \to 2$ and to a certain extent $2 \to 3$). For computations of processes at higher order or including higher multiplicities, they defer to dedicated programs for matrix element generation and phase space integration. Such programs compute just the hard process, and are interfaced to a general purpose generator which will take care of the further steps in the event generation.

Parton distribution functions have been measured over decades, and the results have been made available in a dedicated toolkit, LHAPDF [70]. The user can indicate as part of the configuration which PDF to use. A subtlety to be taken into account is that the choice of the PDF will influence the optimal settings for parameters of non-perturbative (phenomenological) models used at various generator stages. A set of model parameters optimised to describe a data set well, is often called a tune. When configuring an event generator, it is therefore important to choose a PDF and tune in concordance with one another and the data set.

Also scales μ_F and μ_R need to be configured. There is never just a single correct value for them, but neither should they be varied entirely at random. Often the scales are set to $\mu_F = \mu_R = Q^2$. The energy scale, Q^2, is usually taken equal to the resonance mass squared of the process, or the transverse momentum squared in case of a massless particle ($Q^2 = M^2$ or ${p_T}^2$). While

fixed-order calculations depend on μ_F and μ_R, the cross section calculated to all orders does not. The dependence is largest at the lowest order. As such, variations of the scale parameters can be used as an indication of the accuracy of a calculation. If the observed variation of a result with μ is large, it is likely that higher-order corrections would still change the result significantly; if the variation is small, it is safe to disregard further corrections.

A property of LO QCD computations is that they manage to describe only the shape of distributions and not their normalisation. When comparing generated and experimental data, the normalisation of the generated distribution may be corrected using a multiplicative k-factor, the ratio between the NLO and LO total cross section for the process. Alternatively, one can attempt to move to higher-order calculations, in which case also the normalisations would follow from the computation. Note also that in some cases the NLO corrections are small, and an LO computation can be sufficient.

At NLO accuracy, a cross section can be described as the LO cross section ($\mathcal{B}$) plus two corrections: a virtual ($\mathcal{V}$) and a real one ($\mathcal{R}$).

$$d\sigma^{NLO} = d\phi_n \left(\mathcal{B}(\phi_n) + \alpha_s \mathcal{V}(\phi_n)\right) + d\phi_{n+1} \alpha_s \mathcal{R}(\phi_{n+1}). \tag{4.2}$$

The complication here lies in the treatment of some divergences, involving contributions from different phase spaces, which must cancel between the different terms. Details of higher order ME generation will not be discussed in this thesis; more information and further references can be found in [68].

4.1.2 Parton showers

For the description of a complete event, simulations should include the production of hadrons, as they are observed by detectors. Hadronisation of partons into hadrons as well as the decay of unstable hadrons to stable ones, take place at low momentum scales and have to be approached using non-perturbative models. These will be described in the next subsections. Prior to that, a discussion of the step connecting the hard process (at high momentum scales) to the hadronisation (at low momentum scales) is required [69, 71].

The partons involved in the hard scattering are colour charged. Much like Bremsstrahlung happens for electrically charged particles, the coloured partons will radiate gluons as they change momentum. The emitted gluons, unlike the photon which is electrically neutral, are again colour charged and may emit more gluons. The result is a cascade of parton branchings, also called a parton shower.

This shower is the continuation of the hard process discussed before: it incorporates the higher order corrections to the fixed-order hard process calculation. In a hypothetical fixed-order calculation of a shower, each subsequent branching taken into account would increase the order in perturbation theory. As the order to which fixed-order perturbative calculations can be

executed is severely limited, it appears that it is impossible to describe a parton shower in this way. An alternative way of viewing this would be to say that, if fixed-order calculations up to high orders were not an issue, the computation of a parton shower could simply be absorbed in the computation of the hard process, and a separate step would not be needed. They are however an issue, so a different approach is required.

There are ways to describe a parton shower in an approximate manner, considering it as a succession of parton branchings, step-by-step going through a scale evolution. The branchings are governed by scale dependent probability functions, and the evolution between scales is taken care of using evolution equations. Different equivalent variables can be used to keep an idea of 'scale' throughout the shower evolution. Said scale variable then allows to order the branchings, and to 'resum' or integrate the shower from an initial value (the hard process level) to a cutoff value (the hadronisation level). The result is an approximate, finite calculation of the shower, to all orders in perturbation theory. Examples of scale variables are the parton virtual mass q, the parton transverse momentum $\mathrm{p_T}$, or the angle between the partons θ.

Depending on the kinematic region, different types of QCD factorisation and matching evolution equations need to be adopted. The most common description uses collinear factorisation and DGLAP equations [16]. In this scenario, collinear splittings (higher energy, small angle) and soft gluon emissions (low energy, larger angle) are the dominant contributions in a parton shower. While collinear branchings can be considered independently, a description of soft gluon emissions needs to take into account interference[1] or colour coherence effects. Here the scale variable chosen matters. For some scale variables (angular ordering), the interference effects are automatically well described. In other cases (virtual mass or momentum ordering), manual fixes need to be implemented to ensure correct description.

Other than just using different scale variables, it is possible to describe a shower in another, more distinct manner. Instead of looking at branchings of individual particles, one can consider dipoles of colour connected particles and study gluon radiation off of these dipoles. This leads to a $\mathrm{p_T}$ ordered dipole shower picture, which is equivalent to the regular angular ordered shower picture, and also includes interference effects by construction. Note that event generators usually implement only one possible parton shower definition. Which one they adopt is and important part of what makes the different existing generators distinct.

Showering as discussed above, is immediately applicable to the parton shower between the hard process and the outgoing particles at hadronisation level, also called final state radiation (FSR). To connect the initial state to the hard process, any time ordered description of a shower would need to give the evolution from low scale incoming particles, to those participating in the hard process plus accompanying radiation. Since the hard processes of interest require very specific kinematic conditions, it would as mention before be inefficient to implement the simulation this way. Instead, event generators start with the hard process and simulate backwards from there to the incoming particles to generate what is called initial state radiation (ISR).

[1] The gluon emission probability contains contributions from different gluon emission sources and the interference between them.

Finally, one needs to note the delicacy of connecting hard process calculations (at higher energy scales), to shower calculations (at lower energy scales). This is usually done using techniques called matching [72] or merging [73]. The boundary between the end of the hard process and the beginning of the shower is not strictly defined. In connecting the two, some freedom is allowed, but there is a real danger of double-counting or under-counting contributions, or ending up with scale mismatches and non-smooth distributions. Striving to move to increasingly higher order calculations, this matter is highly relevant. For details we refer to the referenced documents.

4.1.3 Hadronisation

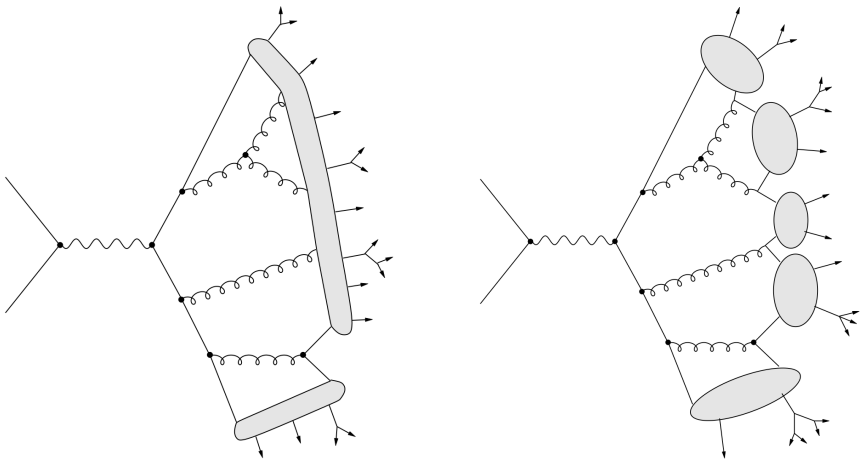
At the end of the final state parton shower, a low energy scale Q_0^2 is reached and we enter the non-perturbative regime. Confinement becomes relevant, and partons cluster into colour-neutral hadrons. As stated above, the hadronisation step needs to be described using a model. An exact non-perturbative calculation, e.g. in the framework of lattice QCD, would run into trouble with time-evolution [68]. Two options exist for hadronisation in event generators: the string model [74] and the cluster model [75]. Both have been extensively tested at low-energy hadron–hadron, hadron–lepton, and lepton–lepton colliders.

In the string model one uses the property that QCD confinement is a large distance effect. Consider a $q\bar{q}$ pair, moving apart, connected by a colour flux tube with potential $V(r) = \kappa r$. The connection can also be seen as a relativistic string with mass density κ. As the string is stretched, the potential energy grows and the kinetic energy drops. It may become energetically favourable for the string to break into two, leading to the creation of an additional $q\bar{q}$ pair. This process continues until the remaining energy is insufficient for more $q\bar{q}$ pairs to be created. The remaining partons can be identified with hadrons. Gluons complicate the picture: they are colour connected to two partons instead of one, and create "kinks" in the strings. Hadron production will be enhanced in the angular regions between quarks and gluons, as was shown experimentally in [76, 77]. The string model is illustrated in Fig. 4.2a.

In the alternative cluster model (Fig. 4.2b), partons are clustered towards the end of the parton shower, into colour-singlet groups or proto-hadrons. The formed clusters can then decay into hadrons following quasi two-body decay processes to less excited states.

4.1.4 Particle decays

Most hadron production occurs via short-lived resonances, unstable at the time-scales of a collider experiment. Their decays will determine the actual final state of the event. When generating these decays, just using particle properties doesn't always suffice. Choices concerning decay modes have to be made, especially in the case involving heavy quarks.



(a) String model **(b)** Cluster model

Figure 4.2: Schematic view of hadronisation according to the string model and the cluster model.

For a decay generation, the first task is to select which hadrons to use in decays. Subsequently choices have to be made concerning decay modes (one might want to study just one particular one) and how to simulate them, taking into account spin and parity effects. Some methods may be built-in in generators, while for others, generators will be interfaced with dedicated decay simulators.

Given the long list of possible decays and their correlations needing to be considered, the particle decay step is a laborious but essential part of the event generation. Also, it is not limited to the decay of hadrons after hadronisation: various heavy resonances may decay at various stages of the event generation (e.g. top quarks or tau leptons).

4.1.5 Secondary interactions

A bunch crossing at a proton-proton collider such as the LHC includes more activity than just the primary hard process, the particles leading into it, and the particles following from it. Secondary interactions need to be considered. To begin with, looking just at the main proton-proton collision in the bunch crossing, three types of additional activity can be listed:

- the beam remnant may include partons not involved in the hard process that are still colour-connected to the system and need to be dealt with accordingly
- there will be additional interactions between partons from the incoming hadrons, often called multiple parton interactions (MPI).
- ISR and FSR components not connected to the primary hard process will still add to the total activity

The whole of all activity in a proton-proton collision, minus the hard process and particles connected to it, is referred to as the underlying event (UE). The simulation of the underlying event activity typically involves phenomenological models, including various free parameters which need to be tuned to experimental data.

Also in the same bunch crossing will be pile-up interactions, additional softer proton-proton collisions of low interest, obscuring the main collision. In simulation, pile-up is introduced to a sample by superposing a chosen amount of minimum bias (MB) events. Such events are generated at random, with as only restriction that some small activity is present in them, thus effectively sampling the total proton-proton cross section, without biasing towards a specific type of events.

4.2 Event generators

To simulate a complete collision event, all stages described in section 4.1 need to be implemented. Some general purpose generators, e.g. Pythia, Herwig++ and Sherpa, can be used to simulate any or all of the stages, while other generators are dedicated to specific generator stages, e.g. MadGraph and Powheg (matrix elements), and Tauola (particle decays). Dedicated generators can be interfaced with a general purpose generator in order to obtain a complete event simulation. Below we briefly review three general purpose generators and three dedicated generators. They will be referred to in section 6.4, where the simulated data used in the VBF $H \to b\bar{b}$ analysis are discussed.

Pythia is a general purpose event generator implemented using FORTRAN (Pythia 6 [78]), and later C++ (Pythia 8 [79]). It includes LO $2 \to 2$ matrix element generation, virtual mass ordered parton showers and p_T ordered dipole showers in versions 6 and 8, respectively, and string model hadronisation. The underlying event model parameters are tuned to experimental data.

Herwig++ [80, 81] is also a general purpose C++ event generator with built-in LO $2 \to 2$ matrix element generation. The Herwig++ parton shower uses angular ordering, and hadronisation follows the cluster model. Underlying event model parameters are again tuned to experimental data.

Sherpa [82] is the newest of these three general purpose event generators, implemented using C++ and Python. It includes slightly more elaborate LO $2 \to n$ ($n < 6$) matrix element generation (like MadGraph), uses p_T ordered dipole showering, and has cluster model hadronisation. Underlying event model parameters are again tuned to experimental data. For connecting the matrix element to the parton shower, Sherpa relies by default on the CKKW method [73].

Table 4.1: Summary of generator properties concerning matrix element generation, parton showering and hadronisation.

generator	matrix element	parton shower ordering	hadronisation model
Pythia 6	LO $2 \to 2$	virtual mass	string
Pythia 8	LO $2 \to 2$	dipole p_T	string
Herwig++	LO $2 \to 2$	angular	cluster
Sherpa	LO $2 \to n$	dipole p_T	cluster
MadGraph	LO $2 \to n$	-	-
Powheg	NLO $2 \to 2$	-	-

MadGraph [83, 84] is a dedicated LO $2 \to n$ ($n < 8$) matrix element generator. For the computation of the parton shower, hadronisation and underlying event components, it needs to be interfaced with a general purpose generated; by default Pythia 6 is used. The connection between the matrix element and the parton shower is usually made using the MLM method [72]. An advantage of MadGraph is that it automates process generation and is not restricted to a list of specific processes implemented by the authors. The MadGraph family also includes aMC@NLO, which extends the automated process generation to NLO.

Powheg [85] is a dedicated NLO $2 \to 2$ matrix element generator, also requiring interfacing with a general purpose generator for parton shower, hadronisation and underlying event computations. The connection between the matrix element and the parton shower is usually made using the Powheg method [86].

Tauola [87] is a toolkit dedicated to the decay of tau leptons, often used with Pythia.

The properties of the above generators are summarised in Table 4.1.

4.3 Detector simulation

In a collision experiment, the particles as produced in the collision will be affected by possible interactions with detector material. Considering the simulation chain as described in previous sections, one additional step is required in order to be able to compare simulated collisions to observations: detector simulation.

For this purpose, the CMS collaboration uses Geant 4 [88], a toolkit capable of simulating the interaction of particles with matter. It allows to define a full geometry of active and passive

detector components, and to simulate the passage of particles and their interactions via their energy loss and creation of secondary radiation. To do so accurately, inputs from dedicated test beam experiments, conducted prior to the construction of the experiment, as well as inputs from dedicated calibration runs and physics control processes are used. After simulation of the electronics response, the simulated data has exactly the same structure as the experimental data, and it can be subjected to the same reconstruction algorithms required for analysis and interpretation of results. Data reconstruction will be discussed in chapter 5.

Detector simulation allows for experimental and simulated data to be compared at detector level. Alternatively, it is possible to correct experimental data for detector effects. Interpreting the data as being a fold (convolution) of the true particle four-momentum and the detector response, this technique is often referred to as unfolding (deconvolution). After unfolding, experimental data can be compared to simulated data at particle level, as produced by event generators directly, or to analytical predictions based on for example the SM. Note that both detector simulation and unfolding are computationally intensive tasks.

When comparing theoretical models with data from a single experiment, it is often easiest to use detector simulation and compare at detector level. This option is however not always available, as detector parameters and simulation programs are usually private to collaboration. As a result, it is often considered better for the experimentalists to provide unfolded results, making confrontation with a wider set of theory predictions or competing results possible at particle level. In practice, detector results are often obtained and published first, with particle level results following later.

In this chapter we discussed how theory predictions for collider experiments are obtained, introducing event generators and how they simulate the various components of a collision process. In addition to event generation, we considered detector simulation, required to be able to compare experimental data and theoretical predictions at the same level.

The experimental and simulated data as we considered it in this chapter, is still raw data. To allow physics analyses to use it efficiently, it needs to be reconstructed. Event reconstruction with the CMS detector is the topic of chapter 5.

Chapter 5

Event reconstruction

Before experimental or simulated data can be analysed and interpreted, it needs to be reconstructed. Signals as measured by the detector are transformed into physical "objects" that can be used for study: charged particle tracks, jets, electrons, etc. The following sections describe the algorithms used by the CMS collaboration for the reconstruction of different objects.

5.1 Track and primary vertex reconstruction

The first step in event reconstruction is often the reconstruction of charged particle trajectories and their origin, and the determination of particle momenta. The following sections discuss track reconstruction (section 5.1.1) and primary vertex reconstruction (section 5.1.3), as well as their performance (section 5.1.2 and 5.1.4).

5.1.1 Track reconstruction

The combinatorial track finder (CTF) [57, 89, 90], an adaptation of a combinatorial Kalman filter [91] that uses reconstructed pixel and strip hits as inputs, is used to reconstruct tracks. The full track reconstruction procedure can be split up into a few logical steps:

- hit reconstruction
- seed generation
- track finding / pattern recognition
- track fitting
- track selection

Track reconstruction is performed iteratively, running the CTF algorithm multiple times. After each iteration, The ***hit reconstruction*** step clusters signals recorded in adjacent tracker pixels or strips, and uses thresholds to filter noise. Such a cluster is called a hit. The collection of all hits, each with position and position uncertainty parameters, is passed on to the track finder.

Hits corresponding to tracks reconstructed in that iteration are removed, making it easier to reconstruct more tracks in the next iteration. The idea is to first find the easily reconstructable tracks (ones with large p_T, originating near the interaction point) and then to move on to more challenging ones (lower p_T, displaced from the interaction point).

In order to describe the expected helical path of a charged particle in the tracker, five parameters are required. At CMS d_0, z_0, ϕ, $\cot\theta$ and p_T are used. To determine these parameters, the input of ideally three hits, or alternatively two hits and an extra beam spot constraint, is required.

These inputs are the seeds to the track finding step. Possible seeds are identified in the ***seed generation*** step, by scanning through the hits and matching them into pairs or triplets. To improve the algorithm performance, some weak criteria on p_T and origin are set for the seeds. As the hit density is lowest in the pixel tracker, the choice was made to start seeding tracks from the inside out; an additional benefit is that the pixel detector provides 3D measurements. Finally, also low p_T tracks that are deflected away from the outer tracker, may be reconstructed this way.

In the ***track finding*** step, tracks are built starting from a given seed, extrapolating outwards and matching hits in subsequent detector layers. After each addition of a hit, the track candidates are updated. The procedure is continued until the outer layer is reached. In some cases with a minimal amount of matched hits for a track candidate, an additional inward search may be executed to find more hits. At the end, the track candidate collection is filtered to remove suspected duplicates.

Candidate trajectories are refitted using a Kalman filter and smoothed in the ***track fitting*** step. The filter again starts from the innermost hit. For each hit in the list, the position uncertainty is updated using the current track parameters. A second filter, using the output of the first and working backwards, acts as a smoothing step. Final track parameters can be obtained as a weighted average of both filters.

The above procedure produces some fake tracks. The ***track selection*** step reduces this number by setting thresholds on various parameters: the number of layers with (3D) hits, the number of layers lacking a hit where one was expected, the quality of the track fit, and the d and z uncertainty and resolution.

5.1.2 Performance of track reconstruction

The performance of the track reconstruction software was tested using simulated single muon data [89]. The resolution of the track parameters is studied in function of η and p_T; the results are shown in Fig. 5.1. At large $|\eta|$, the resolution deteriorates as the particles have traversed more material (and are subject to energy losses). Also at low p_T the resolution is worse, due to multiple scattering. The improvements in the d_z resolution at small $|\eta|$ can be explained by the fact that those particles create signals in multiple pixel cells, due to their travel direction, which leads to an improved measurement.

The performance of the tracker and the track reconstruction is especially important for the momentum measurement of (isolated) muons, leptons and photons up to very high energies, as well as for the tagging of b jets through the identification of secondary vertices. The p_T resolution is better than $\delta p_T/p_T = (20 \cdot p_T \oplus 0.5)\%$ for small $|\eta|$ values (with p_T in TeV), and better than $\delta p_T/p_T = (70 \cdot p_T \oplus 0.5)\%$ for more forward η. This shows that the design criteria are met and charged particle tracking with good p_T resolution is possible up to high energies. The spatial resolution in the transverse direction is better than $\delta(d_0) = 25$ µm, over the whole η range, for particles with $p_T > 10$ GeV, while the spatial resolution in the longitudinal direction is better than $\delta(z_0) = 75$ µm in the same range. This ensures good b-tagging efficiencies of at least 50% in the central η region, and at least 40% in the forward region, with mistag probabilities of $\mathcal{O}(1\%)$, for b jets with p_T = 50-200 GeV.

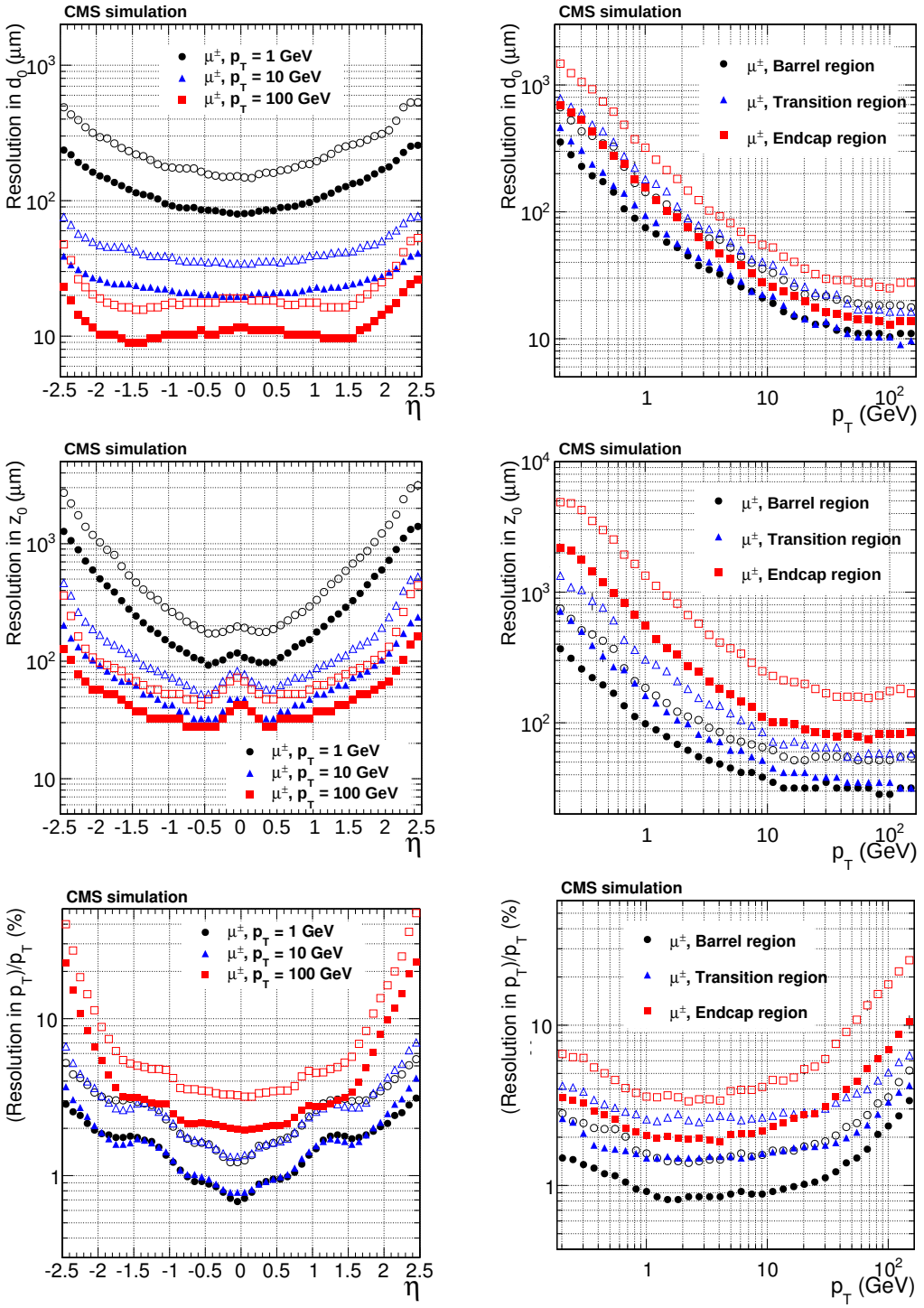


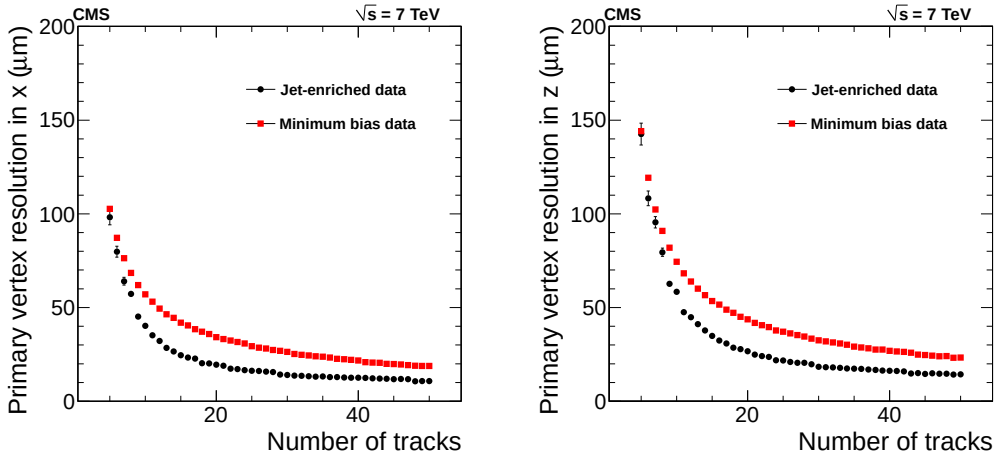
Figure 5.1: Spatial resolution and transverse momentum resolution of the tracker, as a function of pseudorapidity (left) and transverse momentum (right), measured using simulated data [57]. The solid (open) symbols correspond to the half-width of the 68% (90%) intervals, centred around the most probable value, for the residuals distribution (defined as reconstructed - simulated values).

5.1.3 Primary vertex reconstruction

Vertex reconstruction uses all tracks as input and attempts to reconstruct as many true vertices as possible, including the primary vertex (PV)[1], secondary vertices (SV), and pile-up vertices. Tracks are selected and clustered for originating from the same vertex. The vertex position is then fitted. For track clustering, a deterministic annealing algorithm [92] is used to find a global minimum, using z coordinates at closest approach to the beam line as a parameter. After clustering, an adaptive vertex fit (weighted tracks) [93] is performed to get the vertex parameters. The fit is iterative: repeat until convergence. From the fitted vertices, usually the one with the highest transverse momentum squared sum $\sum_{\text{tracks}} {p_T}^2$ is selected as the leading PV.

5.1.4 Performance of vertex fitting

The vertex fitting performance was measured using minimum bias (MB)[2] data as well as jet enriched data. Tracks are ordered by p_T, and alternatingly allocated to one of two sets (odd/even in order). Vertex reconstruction is then performed on each set separately. On average both sets should have the same properties. The resolution of the fitted parameters follows from the width of the obtained distributions. Results are derived in function of the number of tracks involved. As can be seen in Fig. 5.2, the resolution improves with the number of tracks. Additionally, the result is better for jet enriched data: jets on average have higher p_T, leading to better track parameter resolution and consequently better vertex resolution.



(a) PV resolution in x,y **(b)** PV resolution in z

Figure 5.2: Primary vertex resolution in x,y and z, in function of the number of tracks, for MB data (red squares) and jet enriched data (black circles). [57].

[1] Cf. section 3.2.2.

[2] A minimum bias sample is a sample of collisions recorded using a very general trigger, usually requiring just some activity in the event and nothing more restrictive. In this way, the total proton-proton cross section is sampled without biasing the selection towards a specific type of events. Similarly, a zero bias sample is recorded without any restrictions at all.

5.2 Jets

As briefly discussed in the introduction, due to QCD confinement and hadronisation, partons created in a collision will not be observed as individual particles. Instead, they present in the detector as collimated bundles of hadrons, called jets. In section 5.2.1, the algorithms used to obtain a consistent reconstruction of such jets are outlined. Following that, section 5.2.2 discusses the different possible input objects used for jet reconstruction with the CMS detector, and section 5.2.3 handles the corrections to the energy of the jets. Last, jet tagging methods are treated in sections 5.2.4 and 5.2.5.

5.2.1 Jet algorithms

The step from input objects to jets is handled by jet algorithms. Good jet algorithms need to fulfil a few requirements. First of all, they need to be applicable to both experimental data and simulations, and give consistent results in both cases; jet energy corrections needed to match theory and experiment should be small. Furthermore, additional collinear or soft emissions should not affect the outcome of the algorithm. If these requirements are fulfilled, the algorithm is labelled ultraviolet and infrared safe. Finally, the algorithm should be efficient; the computing time required to run it should be reasonable.

Many jet algorithms exist, not all of which adhere to all of the above requirements. Two broad groups can be defined: cone algorithms (iteratively looking for stable cones around a list of input directions) and sequential recombination algorithms (recombining objects based on distance parameters). Algorithms from the first group can be either seeded (starting from a narrowed down list of inputs) or seedless (using all possible inputs). Two clear problems arise: when seeded, usually ultraviolet and infrared safety are an issue, while when seedless, computing requirements become problematic. One seedless cone algorithm that manages to circumvent most problems is the SiSCone algorithm [94]. The second group of algorithms is ultraviolet and infrared safe by construction, as well as computationally efficient. During Run I, the CMS experiment made use of the anti-k$_t$ algorithm [95], an example of a sequential recombination algorithm.

The anti-k$_t$ algorithm uses distance parameters which are inversely proportional to the input momenta, effectively recombining objects in reversed order compared to most other algorithms. Two distance parameters are defined: between two objects (d_{ij}), and between an object and the beam (d_{iB}):

$$d_{ij} = \min\left(\frac{1}{p_{Ti}^2}, \frac{1}{p_{Tj}^2}\right)\frac{\Delta_{ij}^2}{R} \qquad \text{and} \qquad d_{iB} = \frac{1}{p_{Ti}^2}, \tag{5.1}$$

with $p_{T\,i}$ the input object transverse momenta, $\Delta_{ij} = \sqrt{(y_i - y_j)^2 + (\phi_i - \phi_j)^2}$ the distance in the (y,ϕ)-plane, and R a chosen cone radius ($R = 0.5$ in the case of CMS). The iterative procedure is straightforward:

- The set of all input objects, with transverse momenta p_{Ti}, serves as a list of jet candidates or protojets.
- The distances defined in Eq. 5.1 are calculated.
  - If there is a $d_{ij} < d_{iB}$, protojets i and j are merged using p_{Ti}-weighted sums for the properties. The new combination replaces i and j in the list of protojets.
  - If instead d_{iB} is the smallest, i is moved from the list of protojets to the list of jets.
- The process is repeated until all protojets are merged into jets.

Given the specific distance definition, soft protojets are combined with hard protojets and collinear protojets are merged, before a hard protojet is identified as a jet. This guarantees the ultraviolet and infrared safety of the algorithm.

5.2.2 Jet reconstruction

At CMS, four types of jets are commonly reconstructed [96, 97]: calorimeter (Calo) jets, jet-plus-tracks (JPT) jets, particle flow (PF) jets and track jets. The difference between them lies in the choice of input objects fed to the jet algorithm. In all cases the anti-k$_t$ algorithm is used.

Calojets are clustered starting from energy deposits in calorimeter towers, a calorimeter tower being a combination of HCAL and ECAL cells. The geometrical configuration of the cells in the combinations depends on the position in the detector (barrel, end-cap). Additionally, minimum energy thresholds are added to deal with readout noise and pile-up.

JPT jets include information from the tracker, and achieve improved transverse momentum response and resolution with respect to Calojets. Standard Calojets are reconstructed first, and then charged particle tracks are associated to each jet using a ΔR criterion. Based on the properties of the associated tracks, the energy and direction of the Calojet are then corrected.

Trackjets are independent from the other jet types, as they do not use calorimeter information, only tracker information [98]. They are efficient down to very low p_T, due to the higher energy over p_T ratio of the tracker, with respect to the calorimeters. A downside is that they only include charged particles, and are only useable up to $|\eta| < 2.5$.

PF jets are the most used jet type for Run I [99, 100]. The PF algorithm aims at reconstructing and identifying as many final state particles as possible, down to very low energies. To achieve this, information from all subdetectors of CMS is used. The following gives a summary of the PF algorithm.

In the ***first step*** tracker information is taken, allowing for a detailed determination of track p_T and track direction at the production vertex, for charged hadrons. An iterative tracking algorithm is used:

- A search for tracks is done using very tight criteria, obtaining moderate efficiency and low fake track rate.
- Tracker hits associated to tracks are removed and the criteria are relaxed.
- The previous steps are repeated multiple times to reach very high efficiency, finding tracks with as little as three hits and as far down in momentum as $p_T = 150$ MeV.

The ***second step*** adds calorimeter data and thus includes the clustering of neutral particles, making the identification of photons possible. An added benefit is that low quality tracks or high-p_T tracks from the previous step can be improved. The algorithm again has three steps:

- Calorimeter energy maxima (above a certain threshold) serve as "cluster seeds".
- "Topological clusters" are then grown by aggregating the neighbouring cells to existing clusters, as long as the cells have sufficient energy.
- "PF clusters" are created from topological clusters; one PF cluster per cluster seed.

A particle travelling through the detector is expected to create some or all of the following: a *charged particle track*, some *calorimeter clusters*, and a *muon track*. In the ***third step*** – the linking algorithm – all these elements are linked to avoid double counting, with link quality depending on distance. After linking, blocks of on average just 1-3 elements become inputs to the final **PF reconstruction and identification algorithm**, which reconstructs and identifies the sets of particles in each block, providing the final description of the event.

For each block, the algorithm proceeds iteratively, working its way through the various particle types and identifying the particles of each type. As the iterative process continues, the tracks of each reconstructed object are removed from the input list.

1. *PF muons* are reconstructed by matching tracks to information from the muon chambers; the momentum of the tracks is required to be compatible with the measurements from the muon chambers.
2. Electrons are pre-identified in the tracker, and traced to the ECAL, where the final identification is done. A *PF electron* is then reconstructed.
3. Tighter criteria are applied to the remaining tracks. Some are rejected, most of which are fakes. The small amount of energy that is dropped due to the rejection of non-fakes, is recovered from the calorimeter information. From the remaining elements, charged hadrons, photons, neutral hadrons or extra muons may be reconstructed.

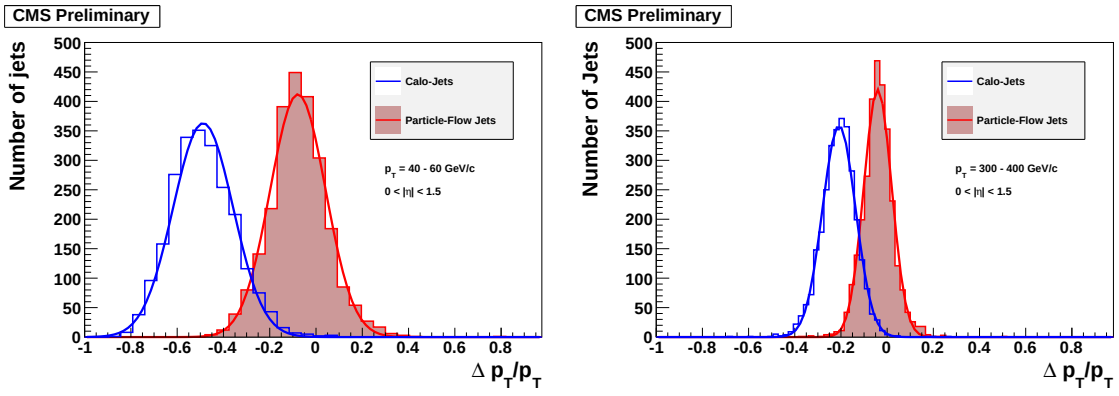
4. The remaining charged particle tracks immediately give rise to *PF charged hadrons*, with properties taken from the track and possibly improved with calorimeter data.

5. If subsequently calorimeter clusters are found with energies exceeding those of associated tracks or not associated to any tracks, this leads to the reconstruction of *PF photons* (ECAL+HCAL excess) and *PF neutral hadrons* (HCAL excess only).

6. Using the full list of obtained particles, finally the *PF missing transverse energy* can be reconstructed by the algorithm.

The performance of the PF algorithm was tested using simulated multijet events [99]. First of all, jets are clustered at generator level from all stable particles except neutrinos ("genjets"). Secondly both PF and Calojets are reconstructed ("recojets"), and matched to the genjets using $\Delta R = 0.1$. Fig. 5.3 shows the distribution of $(p_T^{reco} - p_T^{gen})/p_T^{gen}$, the difference between reconstructed and generated p_T, weighted with the generated p_T. The result is shown for two p_T^{gen} slices.

The jet response and resolution are derived from a Gaussian distribution fitted to the results in Fig. 5.3. It is clear that the PF algorithm, using information from all subdetectors, manages to reconstruct more of the particles and has a better response. Also the resolution is improved with respect to the Calojet result, mainly due to the inclusion of tracker data.



(a) $40 < p_T^{gen} < 60$ GeV **(b)** $300 < p_T^{gen} < 400$ GeV

Figure 5.3: Distribution of $(p_T^{reco} - p_T^{gen})/p_T^{gen}$ for two slices of p_T^{gen} ((a),(b)), for Calojets *(blue)* and PF jets *(red)*. Results are shown for the barrel region of the detector ($|\eta| < 1.5$). [99].

5.2.3 Jet energy calibration

The need for jet energy calibration (JEC) stems from the presence of pile-up, and non-uniform and non-linear detector response. The goal is to calibrate such that, on average, the energy measured matches the energy of the corresponding true particle. This is done using a factorised calibration [96, 101]:

1. an offset correction to deal with the excess energy from pile-up
2. a response correction derived from simulation to deal with detector response
3. a residual response correction derived from data

In the following, the calibration is discussed in function of PF jets. Calibration for the other jet types follows the same methods. A graphical illustration of the calibration steps is given in Fig. 5.4. Figure 5.5 shows the jet response before calibration, after pile-up correction, and fully calibrated.

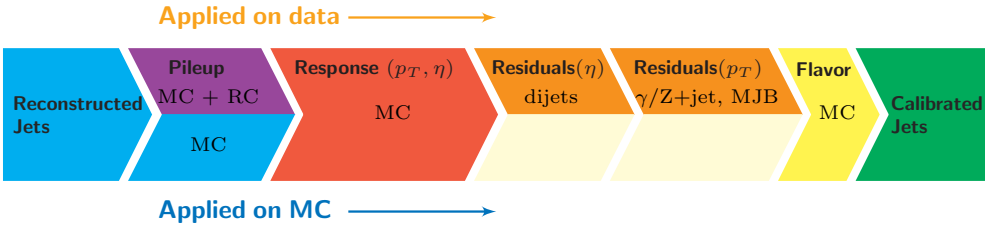


Figure 5.4: Graphical illustration of the calibration steps, applicable for simulated samples and data. [101].

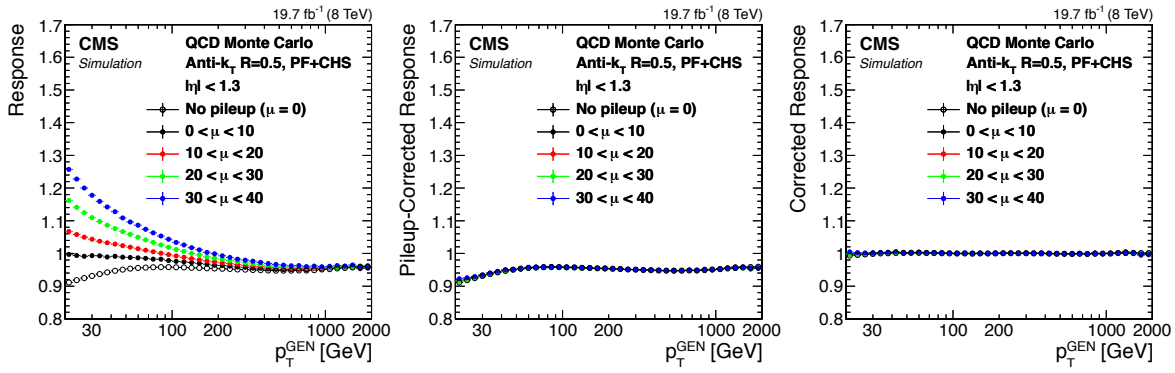


Figure 5.5: Jet response (p_T^{reco}/p_T^{gen}) before calibration (left), after pile-up correction (centre) and after full calibration (right). The response is given for PF jets with CHS applied, for the central region of the detector ($|\eta| < 1.3$), as a function of the genjet transverse momentum, and for different amounts of average pile-up. [101].

Pile-up offset correction Three variables are commonly used to describe pile-up: the number of primary vertices N_{PV}, the average number of pile-up events per bunch crossing μ, and the diffuse offset energy density ρ [102]. As mentioned in chapter 3, pile-up can be divided into out-of-time pile-up and in-time pile-up.

Out-of-time pile-up is affected by the time integration window for the signal, as well as by the separation between bunches. Careful signal handling based on detector timing information allows reasonable control of out-of-time pile-up. In-time pile-up from charged particles in turn can be reduced through charged hadron subtraction (CHS). This optional step is inserted into the PF clustering algorithm. The method considers all PF object candidates and traces them back to their primary vertex. If they originate from a pile-up vertex, they are dropped from the event; if they originate from the leading primary vertex, or can't be associated with a primary vertex, they are kept. After the CHS removal of PF candidates, the PF clustering continues normally. For the new Run II data taking period, various techniques have been developed or optimised in order to improve even further the reduction of in-time pile-up [103–105].

The remaining effect of pile-up, including residual out-of-time pile-up and pile-up from neutral particles, is corrected using the hybrid jet area method [101]. The energy offset is determined as the product of the effective jet area and the median energy density, in function of the jet η and the jet p_T. The offset is first calculated for simulated events, by comparing samples with and without pile-up. Subsequently, the offset calculation is amended with a data/simulation scale factor, also as a function of η, derived from zero bias data.

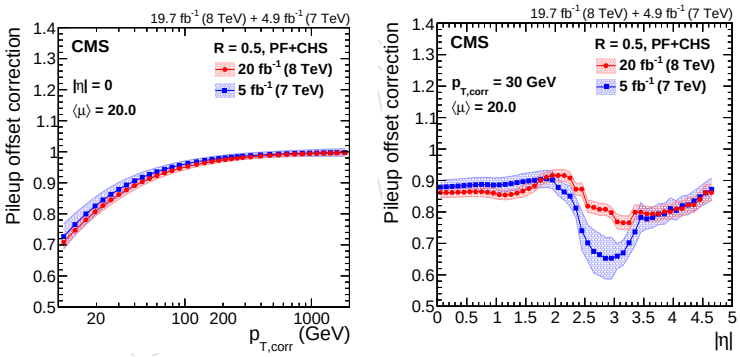
Note that pile-up can also be responsible for the appearance of fake hard jets, due to overlapping soft jets. These fake jets are generally still softer than real jets, and can be distinguished from them using multivariate techniques. One implementation is the pile-up jet identification method (PUjetId) [106]. While fake hard jets are important in analysis interpretations, they only play a minor role in jet calibration.

The total pile-up offset correction factor is shown in Fig. 5.6a as a function of p_T and η. Pile-up adds energy of the order of 10 GeV to jets; the effect thereof is larger for low p_T jets than for higher p_T jets. With respect to η the effect is more uniform. The fluctuations for $2.0 < |\eta| < 3.5$ can be attributed to varying out-of-time pile-up conditions during 2011 data-taking.

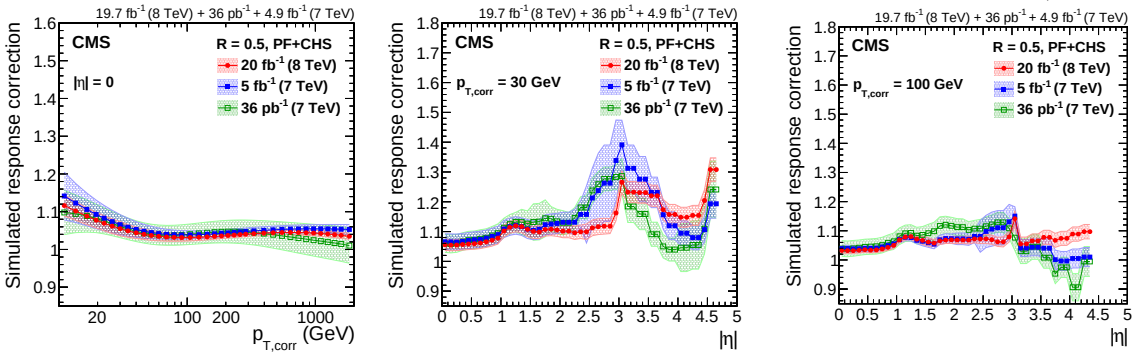
Response correction from simulation The second component of the JEC is a response correction taken from simulation. The study uses a high-statistics QCD dijet sample, including pile-up simulation and corrections for it. The response is defined as:

$$R\left(\langle p_T^{reco}\rangle, \eta\right) = \frac{\langle p_T^{reco}\rangle}{\langle p_T^{gen}\rangle}\left(p_T^{gen}, \eta\right), \tag{5.2}$$

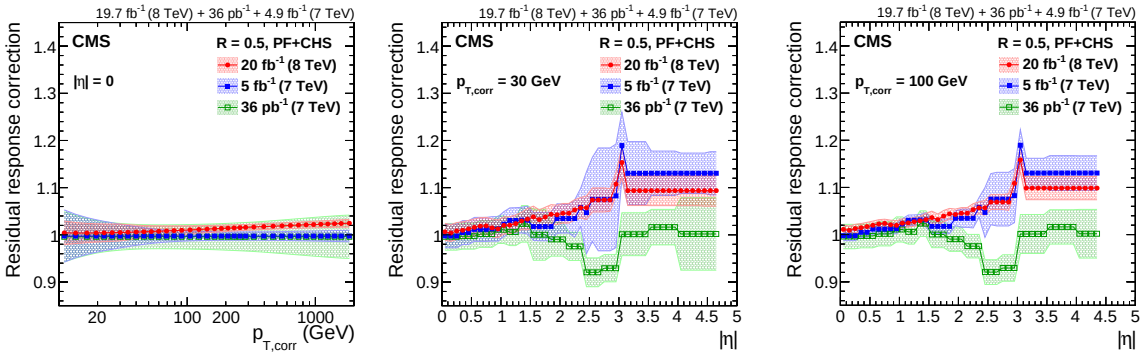
with p_T^{gen} the generator level p_T and p_T^{reco} the reconstructed p_T. Correction factors are derived in bins of $\langle p_T^{reco}\rangle$ and η; the result is shown in Fig. 5.6b. The correction required at low p_T is due to small energy deposits not recorded in the calorimeters because they fall below certain



(a) Pile-up offset correction



(b) Response correction from simulation



(c) Residual response correction from data

Figure 5.6: Summary of contributions to the jet energy calibration, as a function of jet p_T (for central η) and jet η (for different p_T values). [101].

thresholds. With respect to η, larger corrections are required in the subdetector transition regions at $|\eta| = 1.3$ and $|\eta| = 3.0$, where the particles need to cross more detector material and are subject to more interactions.

Residual response correction from data The response correction is further refined with factors derived from data. First a relative correction dependent on η is calculated using a tag and probe technique in a dijet sample. One jet is required to lie in the central region ($|\eta| < 1.3$), while the other is left free. The correction is derived relative to the central region:

$$R_{rel}^{p_T} = \frac{1 - \langle \Delta \mathrm{p_T} \rangle}{1 + \langle \Delta \mathrm{p_T} \rangle} \quad \text{with} \quad \Delta \mathrm{p_T} = \frac{\mathrm{p_T^{probe}} - \mathrm{p_T^{tag}}}{2\,\mathrm{p_T^{ave}}}. \tag{5.3}$$

A small dependence of the correction factor on $\mathrm{p_T}$ is observed, so it is again binned in $\mathrm{p_T}$.

Secondly an absolute correction dependent on $\mathrm{p_T}$ is added. To derive this component, Z+jet and γ+jet samples are used. A momentum balance is made between the resonance (Z or γ) and the recoiling jet:

$$R_{abs}^{p_T} = \frac{\mathrm{p_T^{jet}}}{\mathrm{p_T^{\gamma,Z}}}. \tag{5.4}$$

The correction factor, shown in Fig. 5.6c, is again function of $\mathrm{p_T}$ and η.

The residual correction as a function of $\mathrm{p_T}$ is small. The correction as a function of η is small in the central region and increases slightly towards higher η values.

Jet energy resolution Finally the energy resolution for jets (JER) is determined from simulation (for varying amounts of pile-up), and data vs. simulation scale factors are derived using dijet and γ+jet events [101]. Both are given in Fig. 5.7.

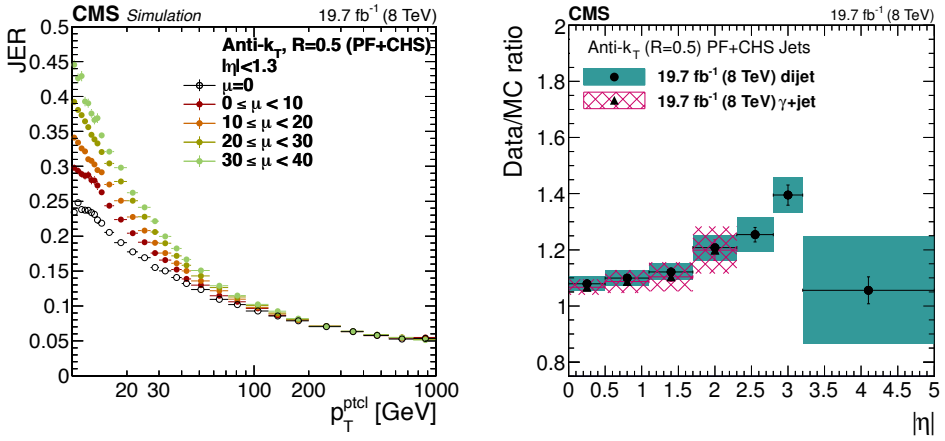


Figure 5.7: Jet energy resolution in the central region ($|\eta| < 1.3$), measured from simulation, for varying amounts of average pile-up. [101].

5.2.4 Jet tagging: identification of b jets

In the study of decays which have one or more b jets in their final states, the identification of these jets is crucial in reducing various backgrounds with jets of different origin (light quarks, gluons or c quarks). For this purpose, detectors implement b-tagging algorithms. Several types exists, making use of different b-quark properties. One can compare the algorithms based on tagging efficiency (real b jets) and mistag rate (light flavour or gluon jets).

All b-tagging algorithms considered here [107] provide a single discriminant value for each jet. Working points are globally defined on the mistag probability vs. discriminant value curve, at probabilities of 10% (loose), 1% (medium) and 0.1% (tight). (These probabilities are valid for average $p_T \sim 80$ GeV). Most algorithms are intended for use with PF jets, but some can be used with any reconstructed jet.

Primary vertex candidate are identified by applying a vertex fit, and the one with the highest $\sum {p_T}^2$ is selected. For finding secondary vertices, a reduced set of tracks is used (limited to a cone $\Delta R = 0.3$ around a jet, and of high purity). Again a vertex fit is applied. All tracks get a weight between 0 and 1, and tracks with weight > 0.5 are assigned to the vertex and removed from the collection. This is repeated until no more vertices can be identified.

Two distinct types of tagging algorithms are discussed. The first type uses impact parameter information (IP between track and primary vertex). Added to this value is the impact parameter significance $S_{IP} = IP/\Delta_{IP}$, with Δ_{IP} the impact parameter uncertainty.

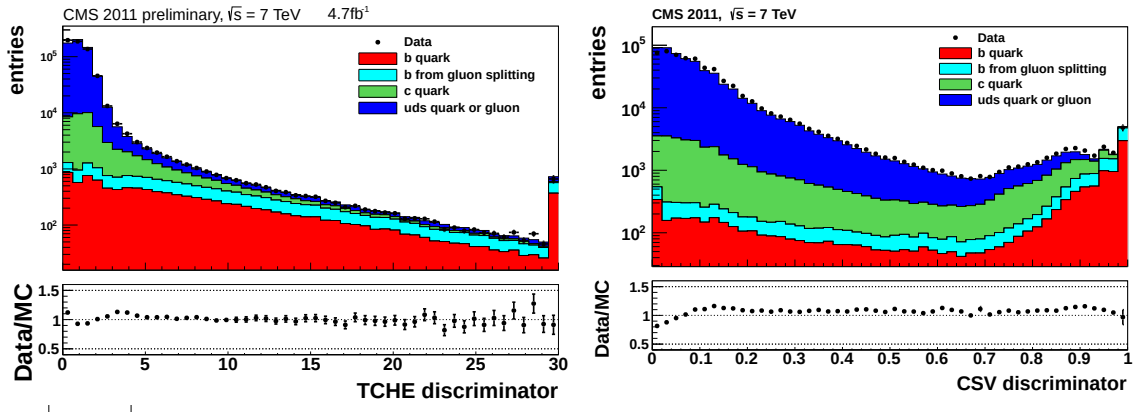
The track counting (TC) algorithm has all tracks sorted by S_{IP} in decreasing order. The first track is at risk for a positive bias, so the final high efficiency (TCHE) and high purity (TCHP) algorithms use the S_{IP} of the 2nd and 3rd track respectively. Multitrack extensions of the TC algorithm are the jet probability (JP) and jet b-quark probability (JBP) algorithms, which calculate the probability that all tracks associated to a jet originate from the primary vertex, without and with S_{IP} weight respectively.

The second type uses secondary vertex (SV) information and variables related to this SV. Before use, the SV candidates are filtered to enhance purity. In the simplest form, the simple secondary vertex (SSV) algorithm, flight distance significance S_{FD} is the discriminating variable. Again there are high efficiency (HE) and high purity (HP) alternatives of the algorithm (depending on the number of vertices associated with a track).

Up from SSV, there is the combined secondary vertex (CSV) algorithm, which adds track-based lifetime information. An extra quality of the CSV algorithm is that also in the event of no SV match, a discriminant value can be obtained. The final CSV discriminant is computed as the combination of a b vs. light quark and a b vs c quark likelihood, weighted with a factor 0.75 and 0.25 respectively.

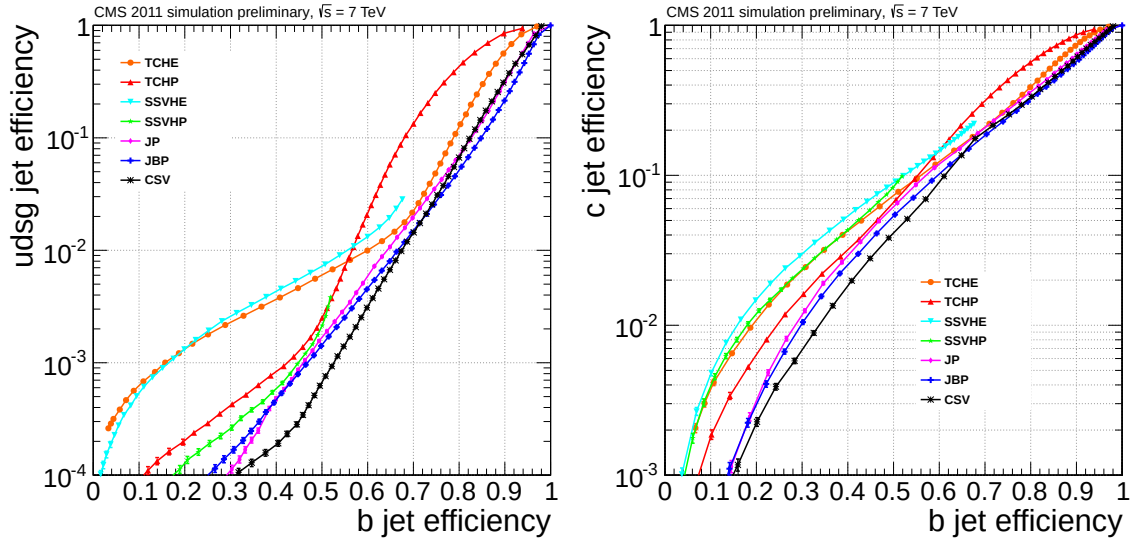
Fig. 5.8 shows the discriminant value of the TCHE and CSV algorithms for a multijet sample. The efficiency of the different methods is given in Fig. 5.9.

The agreement between data and simulation in Fig. 5.8 is good. The distributions for light quarks and gluons peak at low values, while the distribution for b-quarks peaks at high values. In Fig. 5.9 one can see that the CSV algorithm has a higher b-jet efficiency than most other algorithms, for the same light quark and gluon rejection efficiency. The same is true for the b-jet efficiency versus c-quark rejection efficiency.



(a) TCHE **(b)** CSV

Figure 5.8: Discriminant distributions for the THCE and CSV discriminants, derived from data. Expected contributions of different components are taken from QCD multijet simulation and plotted on top. [107].



(a) w.r.t. light quark jets **(b)** w.r.t. c jets

Figure 5.9: Discriminant performance: light quark jet rejection efficiency (left) and c quark jet rejection efficiency (right) versus b-jet efficiency, for various b-tagging algorithms. [107].

5.2.5 Jet tagging: discrimination of quark and gluon jets

In addition to b-jet identification, also identification of the jet origin (quark or gluon) may help reduce backgrounds in analyses with light quark jets in the final state. Three obvious properties can be exploited to build a quark-gluon likelihood (QGL) that allows to filter events:

- Jets originating from gluons tend to have more constituents[3]
- Constituents of jets originating from gluons tend to carry a smaller fraction of the jet transverse momentum
- Jets originating from gluons tend to be less collimated

In the CMS experiment, a quark-gluon discriminant is implemented [109, 110] that does allows such filtering. It makes use of PF jets and aims to be complementary to b-tagging and pile-up identification. As such, b jets and pile-up jets were filtered out from the samples using loose rejection criteria, before construction of the likelihood. Furthermore, to deal with dependencies, the likelihood is binned in η, $\mathrm{p_T}$ and ρ.

Either three or five variables are used (depending on the likelihood version):

- **(i)** The total multiplicity (larger for gluons)
- **(ii)** Shape variables: major and minor axes describing the jet cone (larger for gluons)
- **(iii)** $\mathrm{p_T^D}$ and R (larger for gluons)
- **(iv)** pull: asymmetry (smaller for gluons)

The final likelihood is a product of the variables, with underlying PDFs computed in bins of η, $\mathrm{p_T}$ and ρ. The output can be considered as a quark probability, as shown in Fig. 5.10. The likelihood peaks at one for quark originating jets, and at zero for gluon originating jets. Alternatively, a gluon probability could be defined, with the peak positions swapped. The performance of the likelihood is best at central rapidities, in the higher $\mathrm{p_T}$ region, decreasing slightly for the lower $\mathrm{p_T}$ region and the more forward rapidities.

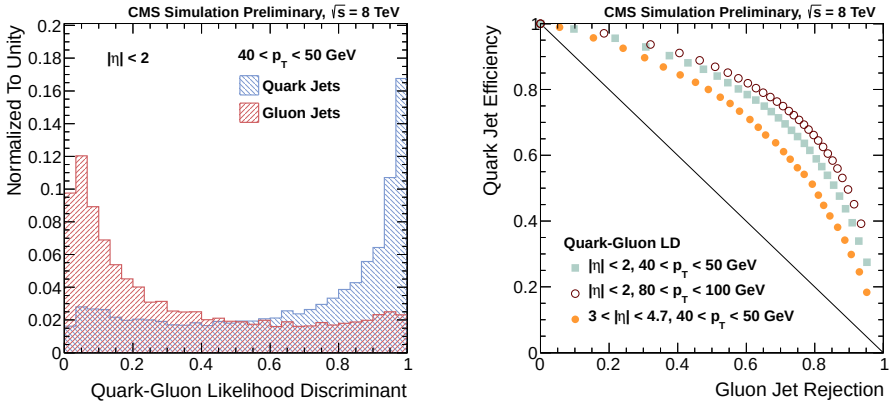
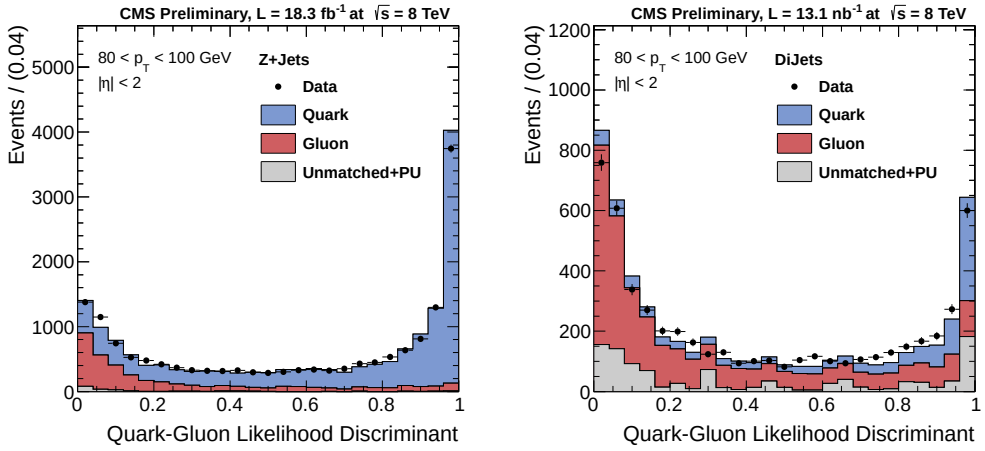


Figure 5.10: QGL discriminant distribution (a) for simulated events, in the $40 < \mathrm{p_T} < 50$ GeV bin, and for $|\eta| < 2$, and discriminant performance (b).

[3]The probability of gluon emission by a quark ($q \to qg$) or a gluon ($g \to gg$) is proportional to colour factors $C_F = 4/3$ and $C_A = 3$, respectively. As a result, the ratio of the multiplicity of a quark and gluon jet is proportional to the ratio of the colour factors: $\frac{<N>_q}{<N>_g} = \frac{C_F}{C_A} = \frac{4}{9}$ [108].

The performance of the quark-gluon likelihood discriminant was also evaluated in data, using Z+jets (containing mostly quark originating jets) as well as dijet (containing more gluon originating jets) samples [110, 111]. Figure 5.11 shows the QGL discriminant distribution for data and simulation, for both samples.



(a) Z+jets **(b)** Dijet

Figure 5.11: QGL discriminant distribution for data and simulation, for jets with $80 < p_T < 100$ GeV and central pseudorapidity, in Z+jets (a) and dijet events (b). [110].

5.3 Other objects

Next to jets, also electrons, muons and missing transverse energy are reconstructed with the CMS detector. The VBF $H \to b\bar{b}$ analysis discussed in this thesis however, studies a hadronic final state, and does not require these objects explicitly. Hence, only a brief description of their reconstruction is given here.

5.3.1 Electron reconstruction

Electrons are reconstructed [112] combining information from the tracker and the electromagnetic calorimeter. Energy losses in the tracker are modelled using a specific filter, before moving on to clustering and electron energy determination in the calorimeter. This method achieves a better resolution than the PF reconstruction algorithm. Additional quality tags can be attributed to reconstructed electrons. Electrons are expected to be somewhat isolated from other activity in the collision. In cases where other objects are mistakenly reconstructed as electrons (e.g. jets), the opposite is true, meaning that isolation criteria can be used to improve the purity of the reconstructed electron collection. The quality of electron candidates can be further refined using multivariate identification methods, matching the candidate properties with the expectations for electrons. Several electron isolation and identification criteria have been implemented by CMS [113].

5.3.2 Muon reconstruction

Two approaches are used for muon reconstruction [114]: tracker muon reconstruction and global muon reconstruction. In the first method, tracks from the centre of the detector are extrapolated outwards and matched with muon system signals. The latter method makes a global track fit of track segments in the muon system and track segments in the tracker. For (very) low momentum muons, only few hits will be observed in the muon system, and tracker muon reconstruction will be more efficient. At higher momenta, several tracks segment will be identifiable in the muon system, and global muon reconstruction will perform best. Entirely similar to the electron case, several muon isolation and identification criteria are available, allowing for reconstructed muon collection quality improvement.

5.3.3 Missing transverse energy reconstruction

Missing transverse energy E_T^{miss} [115], originating from particles not leaving signals in the detector, can be traced using a momentum balance. It is defined as the negative of the vector sum of the transverse momenta of all reconstructed particles. This can be done for different reconstruction algorithms, e.g. using Calo or PF reconstruction. With the PF algorithm being able to reconstruct more particles, also PF E_T^{miss} is usually more accurate than Calo E_T^{miss}.

This concludes the review of object reconstruction for the CMS detector. Having studied in previous chapters the relevant physics theory aspects, the experimental setup, and the simulation and reconstruction software, we now have all building blocks available to discuss the search for the SM Higgs boson produced by vector boson fusion and decaying to bottom quarks. This analysis will be covered in part two of this thesis.

Part Two

Chapter 6

Introduction

This chapter introduces the search for the SM Higgs boson produced by vector boson fusion (VBF) and decaying to a bottom quark pair ($b\bar{b}$) [116], the main topic of this thesis.

At the LHC, the dominant Higgs boson production mechanism is gluon-gluon fusion (GF, cf. section 2.1), followed by vector boson fusion, associated production with a vector boson (VH) and associated production with a top quark pair ($t\bar{t}H$). For low to moderate Higgs boson masses ($m_H \lesssim 135$ GeV), the most prominent decay mode is to two bottom quarks; at $m_H = 125$ GeV, the branching fraction for $H \to b\bar{b}$ is approximately 58%. Even though the GF $H \to b\bar{b}$ channel as such has the largest cross section times branching fraction, the very large irreducible QCD multijet background to this signal means that it is not a viable search channel.

The first observations of the SM Higgs boson by CMS were made in the ZZ, $\gamma\gamma$ and WW decay channels [117–119]. While these channels have lower branching fractions than $H \to b\bar{b}$, they present much cleaner, more identifiable and fully reconstructable leptonic final states, resulting in a higher search sensitivity. Since the dominant GF Higgs boson production happens through a top quark loop, these measurements mostly shed light on the coupling of the Higgs boson to up-type quarks.

The study of the $b\bar{b}$ decay channels is hence important, as it offers an opportunity to probe also the coupling to down-type quarks. At CMS, a search was first performed in the VH production mode (with the vector boson decaying leptonically), the most sensitive $b\bar{b}$ channel [120]. In addition searches were executed in the $t\bar{t}H$ production mode [121, 122]; its cross section is much smaller, but again the semileptonic final states allow for a reasonable sensitivity. Combined with a search in the $\tau\tau$ decay channel [123], these results provide evidence for the direct decay to fermions [124].

In the following we will discuss in more detail the most recent $H \to b\bar{b}$ search: the one in the VBF production mode. While this channel has a larger cross section times branching fraction than the previously studied $b\bar{b}$ channels, it has a lower sensitivity: the fully hadronic final state provides less of a handle for triggering than (semi)leptonic final states, and the large QCD multijet background can only be partially suppressed. The result of this search is very important in the continued testing of the properties of the Higgs boson, allowing to probe the strength of the VBF mechanism, as well as to help establish (in combination with the other $b\bar{b}$ results) the coupling of the Higgs boson to down-type quarks.

6.1 Signal properties

The VBF $H \to b\bar{b}$ process starts with two quarks (one from each of the two colliding protons) radiating a vector boson. The quarks thereby exchange energy and momentum with the vector boson, and are slightly scattered away from the beam direction. The radiated vector bosons subsequently fuse to form a Higgs boson, which finally decays to a pair of bottom quarks. As the longitudinal momentum of the produced Higgs boson is expected to be low, the bottom quarks will be produced centrally. This all results in a hadronic four-jet final state with the two energetic light quark jets (from the VBF process initiating quarks) travelling in forward and backward direction close to the beam pipe, and two central b jets from the Higgs boson decay.

Since the VBF process is purely electroweak, specific expectations hold for the average QCD colour evolution in the events, and further signal versus background discrimination is possible. Most commonly, the light-quark jets will colour-connect to the proton remnants along the beam direction, while the b-quark jets originating from the colour neutral Higgs boson will reconnect among themselves. Hence, there should be only very little hadronic activity in the pseudorapidity region between the two light-quark jets, with the exception of the b-quark jets.

In Fig. 6.1 the transverse momentum and pseudorapidity distributions of the four final state partons are shown, taken from a POWHEG+PYTHIA 6 simulation of the VBF $H \to b\bar{b}$ signal, for $m_H = 125$ GeV. All distributions are normalised to unity. It can be noted that the transverse momentum of the b quarks is slightly harder, but the overall difference is small. At the same time, it is clear that the b quarks are more central, mostly contained within the $|\eta| < 2.5$ tracker acceptance. The light quarks, with an average $|\eta| = 2.1$ and the tail of the distribution extending to $|\eta| \sim 5$, fall mostly within the $|\eta| < 5$ calorimeter acceptance.

Fig. 6.2 again shows the transverse momentum and pseudorapidity distributions of the four final state partons, this time ordered by parton transverse momentum and parton pseudorapidity respectively. The average transverse momentum of the quarks lies between $\langle {p_T}^{(0)} \rangle \approx 104$ GeV for the hardest one, and $\langle {p_T}^{(3)} \rangle \approx 37$ GeV for the softest one; there is no strong correlation between the transverse momentum ordering and parton ID. In the pseudorapidity distributions one can observe the natural symmetry around zero between the first and the fourth, and the second and the third quark, when ordered in pseudorapidity. Furthermore, while not perfect, there is a strong correlation between the pseudorapidity ordering and the parton ID.

For the analysis of experimental data, our understanding of the signal has to extend to reconstructed jet level. In simulated data, reconstructed jets can be matched to partons using a matching algorithm. For each parton/jet combination, the distance $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ is computed, and jets are matched to the parton with the smallest ΔR; if no parton with a ΔR below a certain cutoff (often 0.2-0.3) is found for a jet, no match is made. Fig. 6.3 shows the same distributions as Fig. 6.2, for jets instead of partons. Overlaid on the VBF $H \to b\bar{b}$ signal distributions are the same distributions for QCD multijet background, taken from a PYTHIA 6 simulated sample with the requirement of having four jets. The QCD multijet background is the

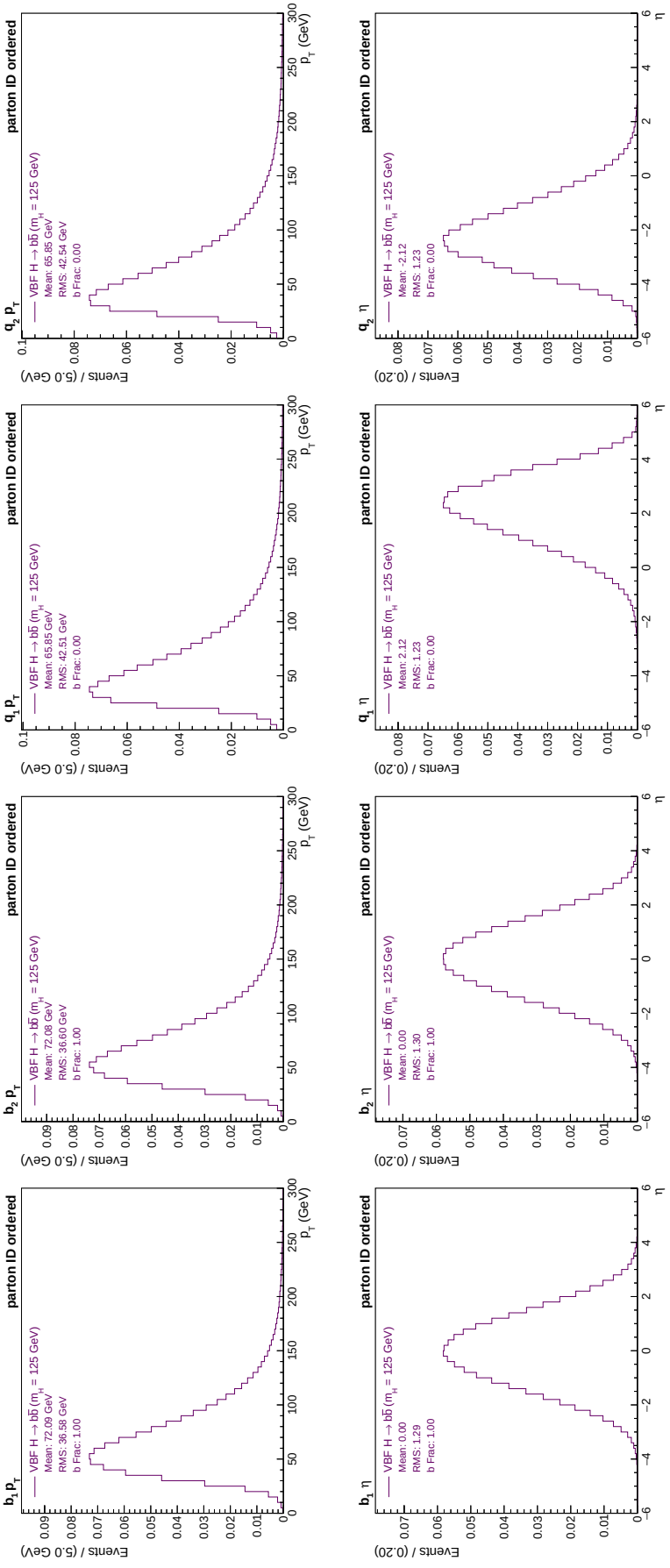


Figure 6.1: Transverse momentum (top) and pseudorapidity (bottom) of the four final state partons, ordered by particle ID. From left to right the two b quarks b_1 and b_2, and the two light quarks q_1 and q_2. All distributions are taken from POWHEG+PYTHIA 6 VBF $H \to b\bar{b}$ signal simulation and normalised to unity. The fraction of b quarks versus other quarks (b Frac) is indicated for each distribution.

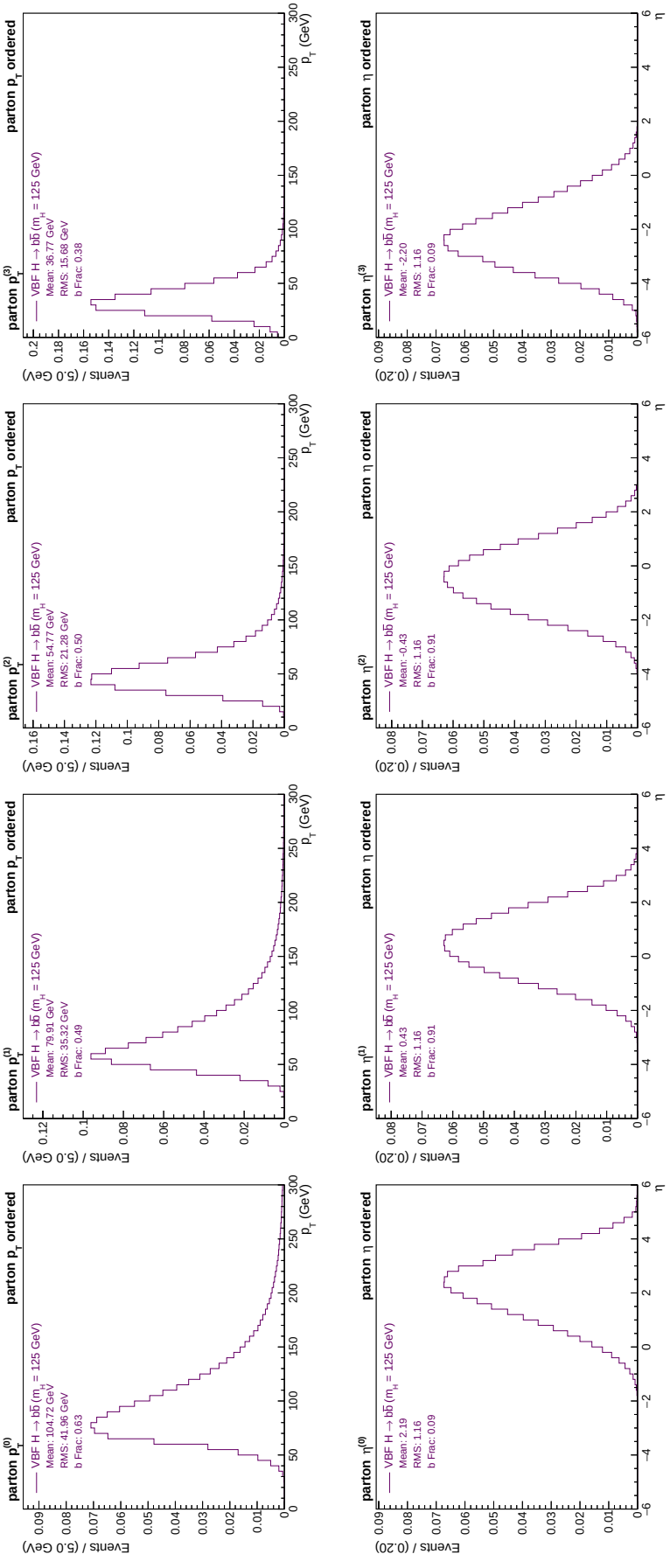


Figure 6.2: Transverse momentum of the four final state partons, ordered by parton transverse momentum (top), and pseudorapidity of the four final state partons, ordered by parton pseudorapidity (bottom). All distributions are taken from POWHEG+PYTHIA 6 VBF H → bb̄ signal simulation and normalised to unity. The fraction of b quarks versus other quarks (b Frac) is indicated for each distribution.

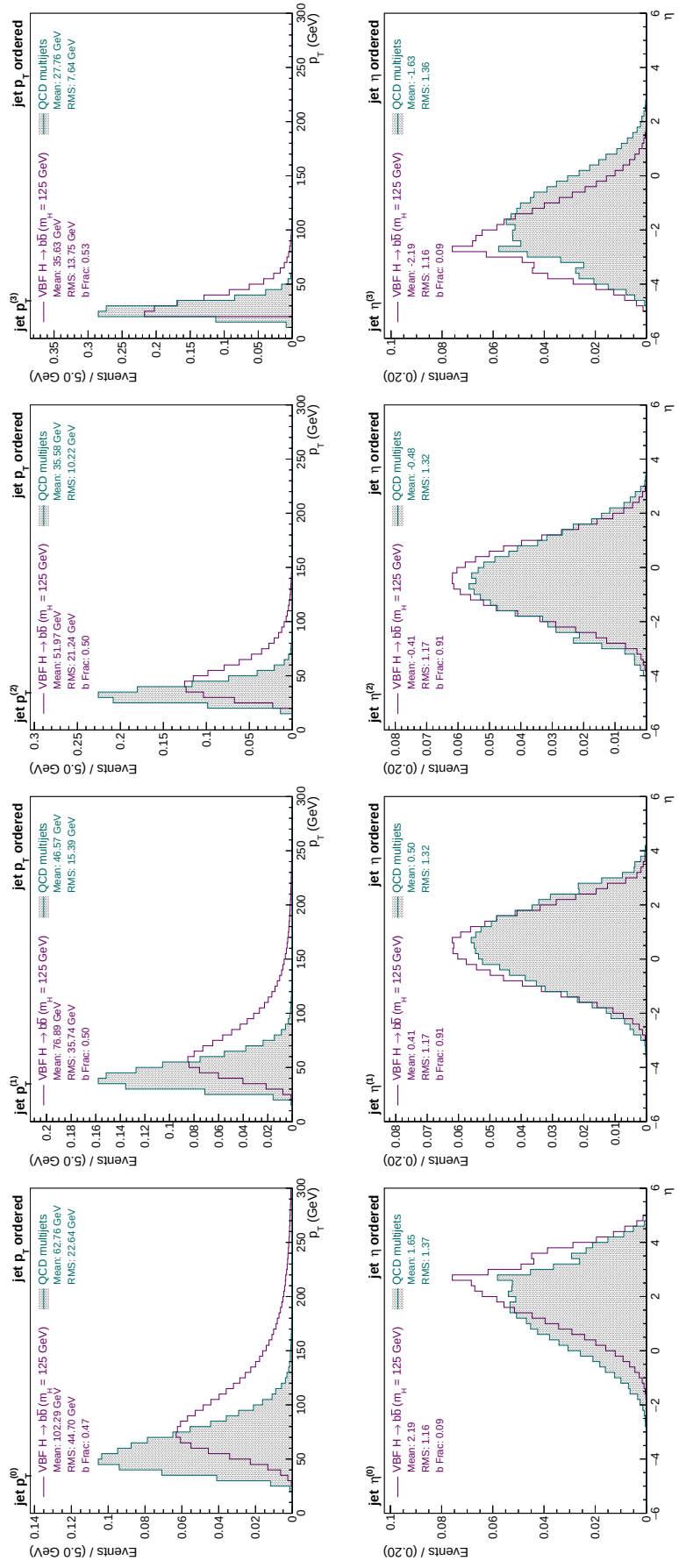


Figure 6.3: Transverse momentum of the four final state jets, ordered by jet transverse momentum (top), and pseudorapidity of the four final state jets, ordered by jet pseudorapidity (bottom). Distributions are taken from POWHEG+PYTHIA 6 VBF $H \to b\bar{b}$ signal simulation (purple) and PYTHIA 6 QCD multijet background simulation (green filled), and normalised to unity. The fraction of b quarks versus other quarks (b Frac) is indicated for each distribution.

largest one by far, or the order of $\mathcal{O}(10^4)$ larger than the signal. All distributions are normalised to unity. The same trends remain visible. The p_T is slightly softer for QCD multijets than for signal. Fluctuations around $|\eta| = 3$ can be attributed to the transition between the barrel and endcap of the detector. Note also that the most forward and most backward quark ($\eta^{(0)}$ and $\eta^{(3)}$) show a slightly larger $|\eta|$ for signal than for background.

Given the fact that no definitive identification of the partons is possible for experimental data, it is necessary to choose an ordering according to which to identify the final state particles. Especially when considering composite variables such as the invariant mass of the light-quark pair, the pseudorapidity separation between the two light quarks, the radial separation between the two light quarks, etc., this can be important. Fig. 6.4 diagrams the particle identification efficiency for different orderings (CSV b-tag value, pseudorapidity and transverse momentum). The top row shows the fraction for which a specific quark is attributed to the b-quark pair, the second row shows the fraction for which the quark is attributed to the light-quark pair. The columns show the specific quarks, ordered either by parton ID (b_1, b_2, q_1 & q_2), CSV b-tag value (decreasing), pseudorapidity (decreasing) or transverse momentum (decreasing). In the b-tag ordered case, the two most b-tagged quarks (the two quark with the highest b-tag value) are expected to form the b-quark pair. For the most b-tagged quark, the identification efficiency is good, for the second most b-tagged one, there is a larger amount of misidentification. In the pseudorapidity ordered case, the quarks with the lowest and the highest value are expected to form the light-quark pair, the middle ones are the b-quark pair. The identification efficiency is symmetric by construction, and is reasonably good. In the transverse momentum ordered case, the ordering is less clear. In Fig. 6.1 the light-quark pair was shown to be fractionally softer, but the difference is too small for the transverse momentum ordering to be useful for identification purposes.

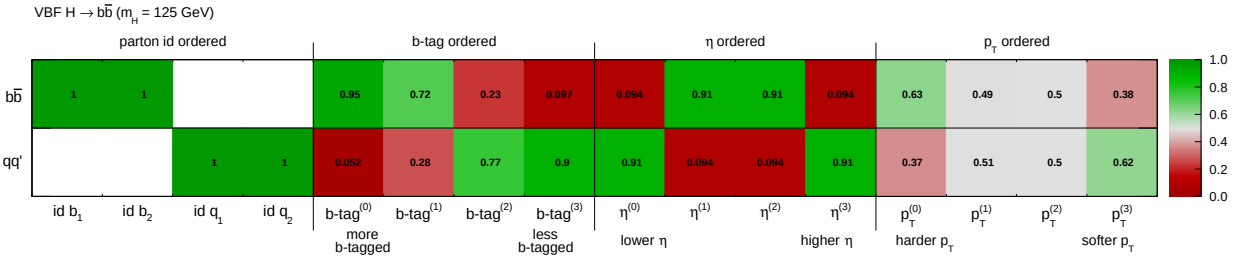
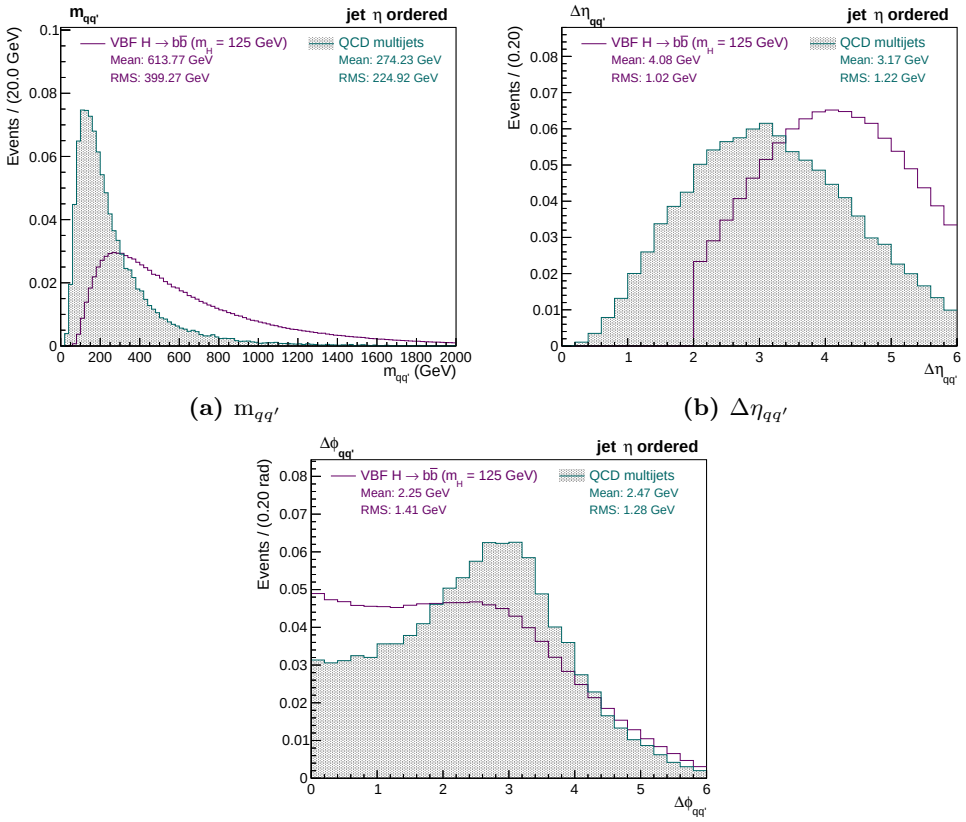


Figure 6.4: Particle identification efficiency for the CSV b-tag value, pseudorapidity and transverse momentum ordering or the partons. Values are taken from POWHEG+PYTHIA 6 VBF $H \to b\bar{b}$ signal simulation.

Fig. 6.5 shows the light-quark jet pair invariant mass, pseudorapidity separation and radial separation, for both the signal and the QCD background. The invariant mass spectrum is steeply falling for QCD multijet background, while it has a large tail towards higher values for signal. The pseudorapidity separation[1] again is larger for signal than for QCD multijet background. The radial separation peaks at π for QCD multijet background (indicating back-to-back jets), while for signal it is mostly uniform between 0 and π.

[1]The hard cutoff at $\eta = 2$ is a manual truncation implemented to limit file size and not a physics feature.



(a) $m_{qq'}$ **(b)** $\Delta\eta_{qq'}$

(c) $\Delta\phi_{qq'}$

Figure 6.5: Distributions of the light-quark jet pair (a) invariant mass, (b) pseudorapidity separation and (c) radial separation, using pseudorapidity ordered parton identification. Distributions are taken from POWHEG+PYTHIA 6 VBF $H \to b\bar{b}$ signal simulation (purple) and PYTHIA 6 QCD multijet background simulation (green filled), and normalised to unity.

6.2 Background contributions

As briefly mentioned above, the largest background to the VBF $H \to b\bar{b}$ signal arises from mostly irreducible QCD multijet production, with either true or misidentified b jets. Most QCD jet production is induced by gluons. As a result, a quark-gluon discriminant as discussed in chapter 5 can be used to increase the signal versus background discrimination. As can be noted in Figs. 6.3 and 6.5, most QCD multijet spectra are steeply falling, a property which can also be used in event selection.

Further backgrounds arise from hadronic decays of vector bosons produced in association with additional jets (Z+jets/W+jets), hadronic decays of top-quark pairs in association with additional jets (tt+jets), and hadronic decays of singly produced top quarks (single top).

Data versus simulation comparison plots including all background contributions will be shown after trigger and offline selection in section 8.3.

6.3 Analysis strategy

With QCD multijet production being the largest background to the VBF $H \to b\bar{b}$ signal, the trigger and selection criteria will be aimed at exploiting the differences between both types of events. To select a subsample of events and improve the signal over background ratio, selection criteria are imposed on the data samples: i.e. lower or upper boundaries are defined for several variables, which the selected events need to satisfy.

The first step is to select four jets to match the expected final state. Jets are considered in decreasing order of transverse momentum; the leading four are selected, if they additionally pass four progressive transverse momentum thresholds. Next, one or two requirements can be added with respect to the b-tagging value of the jets. Finally, the light-quark jet pair is identified according to a chosen ordering (b-tag value or η), and selection criteria are imposed on the light-quark jet pair invariant mass and pseudorapidity separation.

These requirements are implemented in both the trigger and the offline selection for this analysis (cf. chapter 7, sections 7.2 and 7.5). The thresholds for the offline selection will generally lie slightly above the thresholds at trigger level, and are optimised to obtain good trigger efficiency while minimising signal rejection. In addition, for the offline selection, events will be reinterpreted using a multivariate discriminant combining both b-tagging and pseudorapidity information, in order to optimise the light-quark pair and b-quark pair identification efficiency (cf. section 7.4).

After preselection, events are again filtered using a multivariate discriminant to further improve the signal versus background separation. On top of the kinematic properties and b-tagging information used for the trigger and selection algorithms, the discriminant also takes into account variables representing the hadronic activity, and quark/gluon-tagging values (cf. section 8.1). The discriminant is however constructed without using b-quark jet pair kinematic properties. This ensures that the shape of the b-quark jet pair invariant mass distribution remains as unbiased as possible, allowing it to be fit in the last step of the analysis.

One more improvement is made using a multivariate regression targeting the b-quark jet transverse momentum. Jet-by-jet correction factors lead to an improved jet transverse momentum as well as an improved b-quark pair invariant mass resolution. Finally a search for a resonance in the $m_{b\bar{b}}$ distribution is performed on top of the smooth QCD multijet background. Upper limits at 95% confidence level are derived for the production cross section times branching fraction, and compared to SM expectations, in the mass range $m_H = 115$-135 GeV.

6.4 Data and simulation samples

6.4.1 Data samples

The data used in this analysis corresponds to the full 2012 dataset, recorded at a centre-of-mass energy $\sqrt{s} = 8$ TeV. CMS recorded data is organised into primary datasets based on the High Level Trigger it was selected by. Since the VBF $H \to b\bar{b}$ signal contains b quarks, the VBF $H \to b\bar{b}$ triggered data falls under the BJetPlusX primary dataset, except for a small part from the beginning of 2012, which is organised in the MultiJet primary dataset. In addition to the nominal data recorded with dedicated VBF $H \to b\bar{b}$ triggers, this analysis also uses data from a parked dataset, i.e. data which was recorded during 2012, but only processed and made available for analysis later. For this second dataset, only a general purpose VBF trigger was available, and no dedicated VBF $H \to b\bar{b}$ one. Note that in the following chapters, the nominal dataset will be referred to as Set A, while the parked dataset will be referred to as Set B. Lastly, data from the Jet(Mon) data streams was used for trigger studies.

The reconstruction of raw data for these datasets, as well as the data quality certification is handled centrally by CMS. In total, the integrated luminosity of the data used in the VBF $H \to b\bar{b}$ analysis corresponds to 19.8 fb^{-1} for Set A, and 18.3 fb^{-1} for Set B. A summary of the data samples is given in Table 6.1.

Table 6.1: Summary of data samples.

Tags	Samples	Integrated lumi (fb^{-1})
Set A	/MultiJet/Run2012A-22Jan2013-v1/AOD /BJetPlusX/Run2012B-22Jan2013-v1/AOD /BJetPlusX/Run2012C-22Jan2013-v1/AOD /BJetPlusX/Run2012D-22Jan2013-v1/AOD	19.784
Set B	/VBF1Parked/Run2012B-22Jan2013-v1/AOD /VBF1Parked/Run2012C-22Jan2013-v1/AOD /VBF1Parked/Run2012D-22Jan2013-v1/AOD	18.281
Trigger control	/Jet/Run2012A-22Jan2013-v1/AOD /JetMon/Run2012B-22Jan2013-v1/AOD /JetMon/Run2012C-22Jan2013-v1/AOD /JetMon/Run2012D-22Jan2013-v1/AOD	-

6.4.2 Simulation samples

Next to the experimental data, the analysis makes use of simulated signal and background samples for comparisons and event yield estimations. The samples are produced using various MC event generators: MadGraph (v5.1.3.2) [84], Pythia 6 (v6.4.26) [78], Powheg (v1.0) [86] and Tauola (v2.7) [87]. All simulations were performed within the CMSSW software framework (v5_3_X). For all simulated samples, pile-up interactions were added to the main interactions, in order to match the vertex multiplicity distributions in data.

The largest background, QCD multijets, was simulated in four H_T ($\sum p_T$) slices (100-250, 250-500, 500-1000 and 1000-Inf GeV). As the occurrence of events with higher H_T decreases exponentially for QCD multijets, it would very inefficient to generate one sample with the full H_T spectrum in one go; the total number of generated events required to still get reasonable statistics in the tail would then be extremely high. Instead the simulation was performed in slices, which can be combined with appropriate scaling. The simulation used MadGraph for matrix element generation and Pythia 6 for the parton shower and underlying event components. The generator tune was set to Z2*[2] , which uses the CTEQ6L PDFset [126] and was the CMS default at the time.

Also the Z+jets, W+jets and $t\bar{t}$ background samples were simulated using MadGraph+Pythia 6. For the single top samples, Powheg was used, interfaced with Pythia 6 and Tauola. The PDFset used in the case of MadGraph was CTEC6L1 [126], while for the Powheg simulations CT10 [127] was used. The cross sections of these samples were rescaled to NNLO using the FEWZ [128–132].

VBF $H \to b\bar{b}$ (and GF $H \to b\bar{b}$) signal samples were simulated for five mass hypotheses (m_H = 115, 120, 125, 130 and 135 GeV), using Powheg interfaced with Pythia 6 and Tauola. A few variations of these signal samples were produced in order to study the systematic uncertainties due to energy scale (μ_F, μ_R variations using Powheg) and the underlying event and parton showers (Pythia 8 instead of Pythia 6).

6.5 Event reconstruction

The reconstruction of physics objects from raw detector signals was discussed in chapter 5. The VBF $H \to b\bar{b}$ analysis studies a fully hadronic final state, and as such uses mainly jets; other objects however (e.g. leptons and missing transverse energy) are used for selection vetoes. The same reconstruction is applied for data and simulated events. For simulated events, an additional event weight is derived to deal with the different pile-up conditions present in data and simulation (cf. section 6.5.1).

[2] Pythia 6 tune Z2* is an adaption of tune Z1 [125].

Table 6.2: Summary of simulated samples; backgrounds samples (green) and signal samples (purple).

Tags / Samples	Generator	m_H (GeV)	σ (pb)	N	Eff. lumi (fb^{-1})
QCD	MadGraph+Pythia 6				
/QCD_HT-100To250_TuneZ2Star_8TeV-madgraph-pythia6			1 036 000	50 117 340	4.84e-3
/QCD_HT-250To500_TuneZ2Star_8TeV-madgraph-pythia6			276 000	27 062 076	9.81e-2
/QCD_HT-500To1000_TuneZ2Star_8TeV-madgraph-pythia6			8 426	30 599 286	3.63
/QCD_HT-1000ToInf_TuneZ2Star_8TeV-madgraph-pythia6			204	13 843 863	67.9
Z+jets	MadGraph+Pythia 6				
/ZJetsFullyHadronic_Ht100_Pt50_Pt30_deta22_Mqq200_8TeV-madgraph			650	8 141 421	12.5
W+jets	MadGraph+Pythia 6				
/WJetsFullyHadronic_Ht100_Pt50_Pt30_deta22_Mqq200_8TeV-madgraph			1446	4 951 861	3.42
TT+jets	MadGraph+Pythia 6+Tauola				
/TTJets_MassiveBinDECAY_TuneZ2star_8TeV-madgraph-tauola			245.8	6 923 652	28.2
Single top	Powheg+Pythia 6+Tauola				
/T_t-channel TuneZ2star_8TeV-powheg-tauola			56.4	3 753 227	66.5
/T_tW-channel TuneZ2star_8TeV-powheg-tauola			11.1	497 658	44.8
/T_s-channel TuneZ2star_8TeV-powheg-tauola			3.79	259 961	68.6
/Tbar_t-channel TuneZ2star_8TeV-powheg-tauola			30.7	1 935 072	63.0
/Tbar_tW-channel TuneZ2star_8TeV-powheg-tauola			11.1	493 460	44.5
/Tbar_s-channel TuneZ2star_8TeV-powheg-tauola			1.76	139 974	79.5
VBF $H \to b\bar{b}$	Powheg+Pythia 6+Tauola				
/VBF_HToBB_M-115_8TeV-powheg-pythia6_ext		115	1.215	4 733 000	3.90e+3
/VBF_HToBB_M-120_8TeV-powheg-pythia6_ext		110	1.069	4 790 650	4.48e+3
/VBF_HToBB_M-125_8TeV-powheg-pythia6_ext		125	0.911	4 794 398	5.26e+3
/VBF_HToBB_M-130_8TeV-powheg-pythia6_ext		130	0.746	998 890	1.34e+3
/VBF_HToBB_M-135_8TeV-powheg-pythia6_ext		135	0.585	2 486 405	4.25e+3
GF $H \to b\bar{b}$	Powheg+Pythia 6+Tauola				
/GluGluToHToBB_M-115_8TeV-powheg-pythia6		115	15.93	4 987 881	3.13e+2
/GluGluToHToBB_M-120_8TeV-powheg-pythia6		120	13.52	4 999 559	3.70e+2
/GluGluToHToBB_M-125_8TeV-powheg-pythia6		125	11.12	4 496 923	4.04e+2
/GluGluToHToBB_M-130_8TeV-powheg-pythia6		130	8.82	4 799 551	5.44e+2
/GluGluToHToBB_M-135_8TeV-powheg-pythia6		135	6.69	4 188 257	6.26e+2
VBF $H \to b\bar{b}$	Powheg+Pythia 8				
/VBF_HToBB_M-125_8TeV-powheg-pythia8		125	0.911	1 000 000	1.10e+3
GF $H \to b\bar{b}$	Powheg+Pythia 8				
/GluGluToHToBB_M-125_8TeV-powheg-pythia8		125	11.12	250 000	2.25e+1

The analysis uses:

- **Jets**: The main analysis is built on PF jets. Loose jet identification criteria are set to reject noise jets, and at least a loose pile-up ID tag is requested to suppress pile-up jets. To calibrate the jets the official CMS prescriptions [101] are implemented. Finally, jets are only considered for the analysis if they satisfy $p_T > 30$ and $|\eta| < 4.7$. In the trigger, also calojets are used, and a few reconstructed variables are based on trackjets.
- **b-Tagging**: Jets are tagged as b jets using the CSV b-tagging algorithm. Working points are set at 0.244 (loose, 10% mistag probability), 0.679 (medium, 1% mistag probability) and 0.898 (tight, 0.1% mistag probability). Different thresholds are enforced depending on the application in the analysis. In the trigger, also the TCHE b-tagging algorithm is used (cf. section 5.2.4).
- **Leptons**: Muons are reconstructed combining tracker and muon system information, and electron are reconstructed combining tracker and ECAL information. All leptons are required to pass tight identification criteria, and have a relative isolation below 0.15. Leptons are only considered if they satisfy $p_T > 20$ and $|\eta| < 2.5$.
- **Missing transverse energy**: The E_T^{miss} is reconstructed using the PF algorithm; no further corrections are considered.

6.5.1 Pile-up reweighting

When simulation samples are generated, preset amounts of pile-up are superimposed in order to emulate the conditions in experimental data. As the simulations are often made before the final data conditions are known, there will be differences between the pile-up distributions in data and simulation. To optimise the match between both, the pile-up distribution in simulation is reweighted to that found in data. As mentioned before, several variables can be used to describe pile-up: the number of reconstructed primary vertices N_{PV}, the diffuse offset energy ρ, and the average amount of pile-up $\langle\mu\rangle$. In simulation the exact number of generated pile-up interactions is given by variable N_{PU}. The pile-up reweighting scale factors (weights) are extracted from the ratio of the pile-up distribution taken from data and the N_{PU} distribution from simulation; different weights can be obtained for different data samples. The simulation samples are then reweighted on an event-by-event basis. Fig. 6.6 shows the effect of the reweighting procedure for the N_{PV} distribution, for two data selections Set A and Set B, which are defined in section 7.2. The slope visible in the ratio curves is corrected reasonably well by the pile-up reweighting.

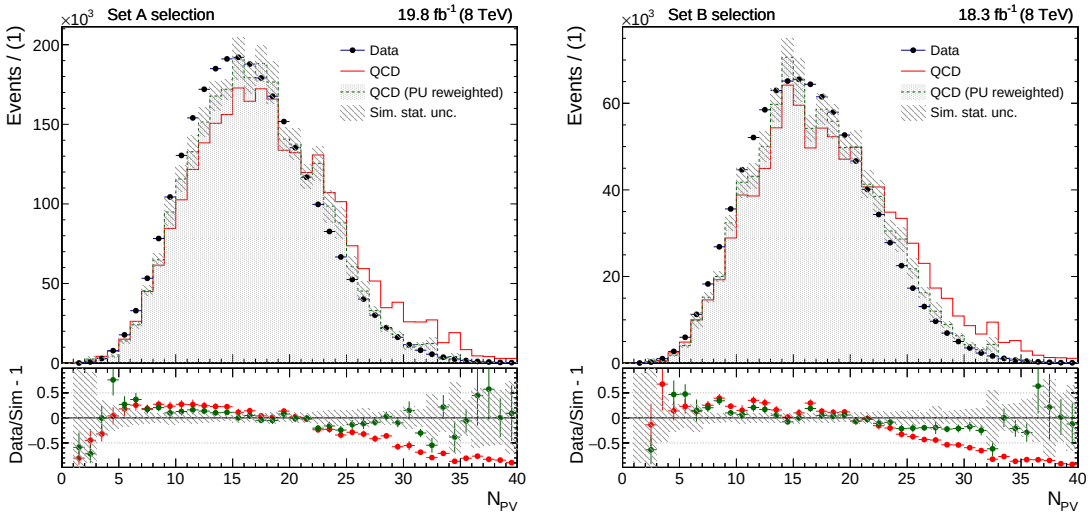


Figure 6.6: N_{PV} distributions for QCD simulation, before (red) and after (green) pile-up reweighting, compared to data (black). The result is shown for both the Set A and Set B selections.

After this summary of the key properties of the VBF $H \to b\bar{b}$ channel, the introduction of the analysis strategy, and the overview of the data samples used in the analysis, we can move on a more detailed review of the analysis. The first step, the data selection for the analysis, is covered in Chapter 7. Chapter 8 then focuses on the second step, the multivariate signal versus background discrimination. Chapter 9 discusses the last step of the analysis, the fit of the b-quark pair invariant mass spectrum. The final results are presented in chapter 10.

Chapter 7

Trigger and event selection

This chapter deals with the selection of the data used in the VBF $H \to b\bar{b}$ search, the first step of the analysis. Sections 7.1 to 7.3 cover the online selection, at trigger level. The development of a dedicated VBF $H \to b\bar{b}$ is outlined, and the trigger configuration as well as performance of all triggers used in the analysis is presented. Section 7.5 covers the offline selection, applied after event reconstruction and interpretation (chapter 5 and section 7.4 respectively). The CMS trigger hardware was discussed in section 3.2.

7.1 Trigger development

When a VBF $H \to b\bar{b}$ analysis was being considered in 2011, just before the start of Run I data taking in 2012, several existing multijet L1 and HLT triggers were considered for use. Due to poor signal efficiency[1] for the VBF $H \to b\bar{b}$ signal (in the order of 0.5% at best) and the fact that the triggers would fire at quite high rates[2] and require prescaling[3] at higher luminosity data taking in the future, none of the options were quite viable. Instead it was decided to develop dedicated VBF $H \to b\bar{b}$ triggers, based on the signal vs. background discriminating properties discussed in chapter 6.

A *triplejet L1 trigger*, setting three progressive transverse momentum thresholds as well as pseudorapidity requirements, was used in combination with *quadjet HLT triggers*, setting four progressive transverse momentum thresholds, b-tag requirements and requirements on the kinematics of the light-quark jet pair. A schematic view of the trigger chain is given in Fig. 7.1. To determine the optimal configuration of the triggers, the trigger rate and signal efficiency were computed for large sets of variations (of thresholds as well as additional variables) using trigger simulation as well as early recorded data. The aim when developing a trigger is to maximise the signal efficiency, while keeping the rate at an acceptable level with respect to the experiment's bandwidth; also the trigger's processing time needs to be sufficiently low. For the VBF $H \to b\bar{b}$ analysis, the allocated bandwidth was 2.5-5 kHz at L1, and 5-10 Hz at HLT level. In order to be able to predict trigger performance under the various luminosity conditions expected for the Run I data taking period, all triggers were tested for at least two luminosity configurations.

As mentioned in chapter 6, two distinct sets of triggers were used to record data for the VBF $H \to b\bar{b}$ analysis. The data selected with the dedicated VBF $H \to b\bar{b}$ triggers forms the nominal dataset, Set A. The additional parked dataset, Set B, contains data selected with general purpose

[1] The signal efficiency indicates the amount of signal events passing the trigger criteria.
[2] The frequency at which a trigger selects events for recording is called the trigger rate.
[3] When trigger rates exceed handleable values, it can be decided to randomly drop a certain percentage of triggered events, thus reducing the amount of data effectively recorded; this is called prescaling.

VBF triggers, not optimised for the VBF $H \to b\bar{b}$ signal. In the following subsections, the outcome of the optimisation of the dedicated Set A triggers is briefly discussed. Subsequently, section 7.2 covers the configuration of the final set of triggers (both Set A and Set B), and section 7.3 details their performance.

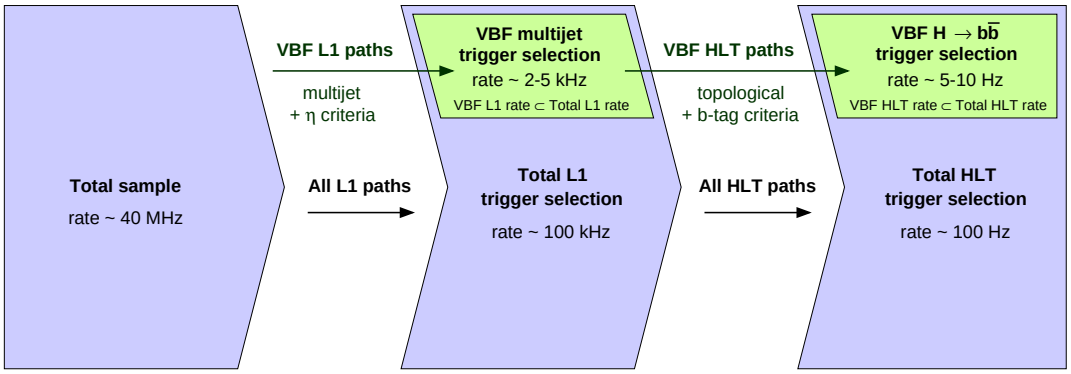


Figure 7.1: Schematic overview of the trigger chain, starting with a collision rate of 40 MHz and showing progressive rate reductions at L1 and HLT level. The VBF triggers are part of a much larger set of triggers, used by many different analyses, and as such can only use part of the experiment's bandwidth.

7.1.1 Dedicated level-1 triggers

The L1 trigger infrastructure allows to look for jets and to set transverse energy thresholds, as well as to select specific event topologies (thanks to the stored (η, ϕ) information). This means that in addition to focusing on jets with a standard multijet trigger path (n jets with transverse energy above certain thresholds), it is possible to further target the VBF signal properties and to set requirements on the pseudorapidity of the jets. To begin with, standard multijet paths were found to result in very poor signal efficiency and were dropped. For the multijet plus topological requirement paths, both dijet and triplejet paths were studied further. While showing reasonable rate and efficiency outcomes, the dijet paths were identified as suboptimal when used together with the quadjet paths planned for the HLT trigger. With a dijet L1 seed, the thresholds set at HLT level are required to be higher to sufficiently reduce the rate, resulting in a significantly reduced efficiency. This lead to the choice of developing an optimised triplejet plus topological requirement path of the form L1_TripleJet_X_Y_Z_VBF. First, the trigger asks for three jets with transverse energies above thresholds X, Y and Z GeV. For technical reasons, the thresholds implemented in the trigger system always change with 4 GeV increments; this needs to be taken into account when optimising the values for X, Y and Z. Secondly, the pseudorapidity of the jets passing the thresholds is considered, and the trigger asks for either three central jets ($|\eta| < 2.6$), or two central and one forward jet ($2.6 < |\eta| < 5.2$); in the latter case the more forward jet is required to pass at least the second highest transverse energy threshold Y. Finally, a $|\Delta\eta_{qq'}| > 1.5$ cut was investigated, which significantly reduces the trigger rate but also strongly affects the signal efficiency. This last cut is too tight for general data taking (low signal efficiency), but would allow the trigger to still be used during high luminosity runs of the experiment.

7.1.2 Dedicated high-level triggers

At HLT trigger level, selections on many more reconstructed variables are possible. The VBF $H \to b\bar{b}$ has a four jet event topology, suggesting the use of a quadjet trigger path (four transverse momentum thresholds), complemented with b-tagging and kinematic restrictions matching the signal properties ($\Delta\eta_{qq'}$ and $m_{qq'}$). In order to reconstruct these kinematic variables, the light-quark jet pair needs to be identified. As mentioned earlier, this is done using either the b-tag value (qq′ pair = two jets with lowest b-tag values), or the η value of the jets (qq′ pair = two jets with largest η separation). For the HLT paths, it is especially important to study the processing time requirements of the various elements of the trigger. Naturally more thorough reconstructions allow higher efficiency, but the time window in which a trigger decision needs to be provided is limited (about 1s). A common workaround is to apply matching cuts at different levels of reconstruction: setting looser thresholds for rough reconstructed variables (Level 2.5) to reduce the amount of events requiring to be processed, and then setting tighter thresholds for the remaining events, after more detailed reconstruction (Level 3). This is applicable to b-tagging, but also to the transverse momentum requirements. Initially the trigger paths were developed using calojets. It is however possible to improve the efficiency of the trigger while keeping the same rate, by setting slightly lower thresholds for calojets, and then enforcing the original higher cuts on particle-flow jets, which are more accurate but require more processing. In addition to using two types of jets, also two types of b-tagging algorithms are used: TCHE b-tagging in the case of calojets, and CSV b-tagging in the case of PF jets.

Two types of paths were fully developed, optimised and commissioned for use: HLT_QuadJet-_W_X_Y_Z_BTagIP_VBF and HLT_QuadPFJet_W_X_Y_Z_BTagCSV_VBF, using respectively calojets, and calojets as well as PF jets. Figure 7.2 (left) shows the evolution of the rate and efficiency with the application of each additional cut, for an example of a calojet quadjet path. The rate is reduced from about 2.5 kHz, to about 3 Hz, and the total signal efficiency is 8%. In the right-hand figure the rate and efficiency change is shown each time with

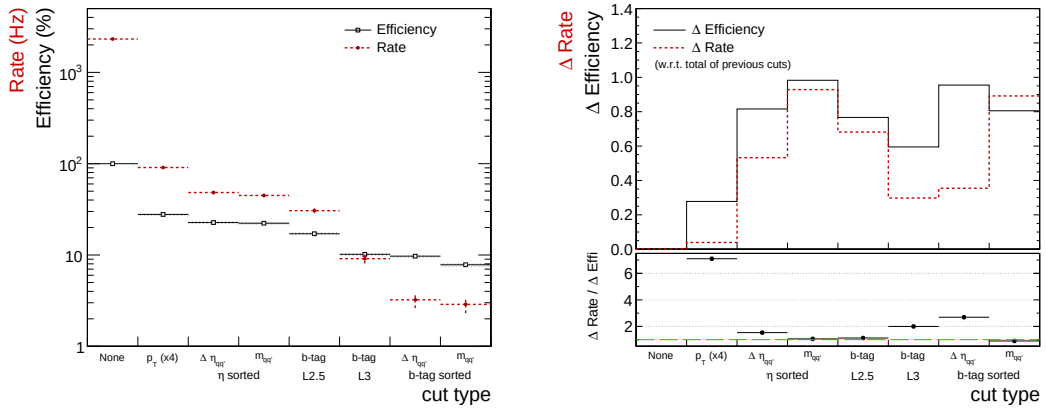


Figure 7.2: Illustration of rate and efficiency reduction with progressive cuts, for a quadjet path using calorimeter jets with cuts $p_T > 80$, 58, 45 and 20 GeV, η ordered $\Delta\eta_{qq'} > 2.5$ and $m_{qq'} > 200$ GeV, TCHE b-tag value > 2.5 at L2.5 and > 7.5 at L3, and b-tag ordered $\Delta\eta_{qq'} > 2.5$ and $m_{qq'} > 200$ GeV. (left) Reduction with each progressive cut. (right) relative change with respect to the previous step in the sequence.

respect to the previous step in the sequence. The PF jet quadjet paths follow the same idea: first the selection is applied on calojets with reduced thresholds, subsequently on PF jets with the original thresholds.

Next to the transverse momentum thresholds, b-tagging and $\Delta\eta_{qq'}$ and $m_{qq'}$ requirements mentioned before, a few other variables were tested during the development of the trigger: e.g. $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, $\mathrm{sign}(\eta_{q_1} \cdot \eta_{q_2})$, or $\sum \mathrm{p_T}$, but none of the options shows a real improvement in rate reduction over the standard set of variables. As a result, using them would needlessly complicate the trigger (which is relevant in the study of the trigger performance, cf. section 7.3), and they are not used. Finally, variables related to the kinematics of the b-quark jet pair, e.g. $m_{b\bar{b}}$, $\Delta\phi_{b\bar{b}}$, are avoided on purpose, to avoid introducing a possible bias in the search for the $\mathrm{H} \to \mathrm{b\bar{b}}$ signal.

7.2 Trigger configuration

7.2.1 Level-1 triggers

Set A Three versions of the dedicated L1_TripleJet_X_Y_Z_VBF trigger path were used for the VBF $\mathrm{H} \to \mathrm{b\bar{b}}$ analysis during Run I data taking (L1 A1, A2 and A3), to seed the dedicated HLT trigger paths recording data for the nominal dataset, Set A. Instantaneous luminosity conditions increased throughout run periods A, B, C and D, requiring the thresholds in the trigger path to be tightened to maintain manageable trigger rates. Table 7.1 lists the used threshold values and the corresponding efficiencies and trigger rates.

Most data was gathered using the second path, L1 A2, which was active and largely unprescaled throughout the entire data taking period. L1 A1 was important at lower luminosity conditions, during run period A, but was prescaled afterwards. L1 A3 acted as a backup path with a reduced rate, which was required for peak luminosity data taking in run period D, when also L1 A2 had to be prescaled.

Set B From about the middle of run period B onward, additional trigger paths were activated to record data which would be stored and processed at a later date. While no dedicated VBF $\mathrm{H} \to \mathrm{b\bar{b}}$ trigger paths were available in this parked data campaign, the analysis could make use of general purpose VBF trigger paths. The logical OR of four L1 paths of the form L1_ETMX and L1_HTTY was used as a seed (L1 B1, B2, B3 and B4). Paths of the first type require missing transverse energy $\mathrm{E_T^{miss}}$ above threshold X, while paths of the second type require a transverse energy sum $\mathrm{H_T}$ above threshold Y. Table 7.2 lists the used threshold values and the corresponding efficiencies and trigger rates.

During run period B, with instantaneous luminosity conditions of the order of $5 \cdot 10^{33}\ \mathrm{cm^{-2}\,s^{-1}}$, combination (B1 or B2) was prevalent. At higher luminosity, during run periods B-D, combination (B1 or B3) took over; this setup provided the bulk of the Set B dataset. Finally backup combination (B1 or B4) could be used at peak luminosity during run period D.

Table 7.1: Configuration, efficiency and rates for the dedicated VBF $H \to b\bar{b}$ L1 trigger paths used for the nominal dataset, Set A. (Renormalised to $5 \cdot 10^{33}$ $cm^{-2}\,s^{-1}$).

Path no.	p_T thresholds (GeV)			Efficiency	Rate	Luminosity
	X	Y	Z	(%)	(kHz)	($cm^{-2}\,s^{-1}$)
L1 A1	64	44	24	62	5.0	$5 \cdot 10^{33}$
L1 A2	64	48	28	56	3.5	$5 \cdot 10^{33}$
L1 A3	68	48	32	50	2.5	$5 \cdot 10^{33}$

Table 7.2: Configuration, efficiency and rates for the general purpose L1 trigger paths used for the parked dataset, Set B. (Renormalised to 5 and $7 \cdot 10^{33}$ $cm^{-2}\,s^{-1}$).

Path no.	E_T^{miss}	H_T	Efficiency	Rate	Luminosity
	X (GeV)	Y (GeV)	(%)	(kHz)	($cm^{-2}\,s^{-1}$)
L1 B1	40	-	20.6	3.0 - 5.5	5 - 7 $\cdot$ 10^{33}
L1 B2	-	150	27.9	1.2 - 2.2	5 - 7 $\cdot$ 10^{33}
L1 B3	-	175	-	0.8 - 1.2	5 - 7 $\cdot$ 10^{33}
L1 B4	-	200	-	0.5 - 0.8	5 - 7 $\cdot$ 10^{33}
L1 B1 or B2	40	150	38.5	-	-

7.2.2 High-level triggers

Set A The dedicated L1 triplejet paths seed dedicated HLT quadjet paths. As mentioned there are two types of paths: HLT_QuadJet_W_X_Y_Z_BTagIP_VBF (HLT A1 and A2) and HLT_QuadPFJet_W_X_Y_Z_BTagCSV_VBF (HLT A3, A4, A5 and A6), using respectively calojets, and calojets as well as PF jets. Again, due to changing luminosity conditions, the various thresholds were updated throughout the different run periods. The thresholds used are shown in Table 7.3, together with the resulting efficiencies and trigger rates.

The paths are always deployed in pairs: HLT A1+A2, A3+A4 and A5+A6. Since the second path in each pair uses slightly tighter cuts, events triggered by the second path will also be triggered by the first path; as the luminosity increases however, the first path could be prescaled, while the second path would stay active up to higher values without prescaling. The bulk of the Set A dataset was recorded with paths HLT A4 and A6.

Set B The parked dataset is recorded using three general purpose dijet VBF trigger paths of the form HLT_DiJetX_MJJY_AllJets_DEtaZ_VBF (HLT B1, B2 and B3). Transverse momentum threshold $p_T > X$ is enforced for the two leading jets, m_{jj} is required to be above Y, and $\Delta\eta_{jj}$ has to be greater than Z. The jj quark jet pair here is identified as the one with the largest invariant mass. The used variations of the thresholds are given in Table 7.4, together with the resulting efficiencies and rates.

The bulk of the Set B dataset was recorded with HLT B2, but the overlap between the three path is very large. Note also that the high m_{jj} and $\Delta\eta_{jj}$ thresholds, as well as the lack of b-tagging, are suboptimal when using these triggers in the search for the VBF $H \to b\bar{b}$ signal.

Table 7.3: Configuration, efficiency and rates for the dedicated VBF $H \to b\bar{b}$ HLT trigger paths used for the nominal dataset, Set A. (Renormalised to $5 \cdot 10^{33}$ cm^{-2} s^{-1}).

Path No.	p_T thresholds (GeV)				η ordered		b-tagging[4]		b-tag ordered		Effi.	Rate	Lumi.
	W	X	Y	Z	$\Delta\eta_{qq'}$	$m_{qq'}$	L2.5	L3	$\Delta\eta_{qq'}$	$m_{qq'}$	(%)	(Hz)	(cm^{-2} s^{-1})
HLT A1 Calo	75	55	35	20	2.5	200	2.5	6.8	2.5	200	9.2	8.0	$5 \cdot 10^{33}$
HLT A2 Calo	75	55	38	20	2.5	200	2.5	7.9	2.5	200	8.0	6.0	$5 \cdot 10^{33}$
HLT A3 Calo	64	44	24	18	2.5	180	-	-	2.5	200	11.0	12.0	$5 \cdot 10^{33}$
+ PF	75	55	35	20	2.5	200	0.6	0.8	2.5	-			
HLT A4 Calo	64	44	24	18	2.5	180	-	-	2.5	200	9.6	9.0	$5 \cdot 10^{33}$
+ PF	78	55	38	20	2.5	200	0.6	0.9	2.5	-			
HLT A5 Calo	64	44	24	18	2.5	180	-	-	2.5	200	8.3	3.5	$5 \cdot 10^{33}$
+ PF	78	61	44	31	2.5	220	0.6	0.8	2.5	-			
HLT A6 Calo	66	50	30	22	2.5	200	-	-	2.5	200	6.5	3.5	$5 \cdot 10^{33}$
+ PF	82	65	48	35	2.5	240	0.6	0.8	2.5	-			

Table 7.4: Configuration, efficiency and rates for the general purpose HLT trigger paths used for the parked dataset, Set B. (Renormalised to 5 and $7 \cdot 10^{33}$ cm^{-2} s^{-1}).

Path no.	p_T thresholds (GeV)		m_{jj}	$\Delta\eta_{jj}$	Effi.	Rate	Lumi.
	X1	X2	Y (GeV)	Z	(%)	(Hz)	(cm^{-2} s^{-1})
HLT B1 Calo	35	35	650	3.5	12.4	150	$5 \cdot 10^{33}$
HLT B2 Calo	35	35	700	3.5	10.6	120 - 160	$5 - 7 \cdot 10^{33}$
HLT B3 Calo	35	35	750	3.5	9.8	100 - 130	$5 - 7 \cdot 10^{33}$

7.2.3 Reference triggers

For the evaluation of trigger efficiencies, data vs. simulation scale factors and systematic uncertainties, an unbiased reference trigger path is required. The reference trigger should be 100% efficient for the selected events; when comparing the trigger efficiency in simulation with and without the reference trigger applied, there should not be a significant difference. Both for Set A and Set B, HLT_DiPFJetAve80 is a viable choice. This trigger path requires two PF jets with an average transverse momentum above 80 GeV. It is available in all datasets, but was heavily prescaled throughout the entire data taking period; this will affect the statistical uncertainty of trigger performance measurements.

[4]Note that for the calojet paths, the TCHE b-tagging algorithm is used, while for the Calo + PF jet paths, the CSV b-tagging algorithm is used.

7.3 Trigger efficiency

The main element in studying the performance of the triggers is to study their efficiency. This is done using data from the JetMon data streams, for each set of trigger separately (Set A and Set B), within the phase space defined by the preselection criteria matching the dataset (cf. section 7.5). These preselection criteria are defined to constrict the pool of selected events to the regions in phase space where the triggers perform in a well-defined and controlled manner. When studying the efficiency of a trigger as a function of a parameter, one will find a typical shape, starting at zero, picking up quickly from a certain value onward, and then reaching a plateau; this is often referred to as a 1D turn-on curve. The preselection criteria are then optimised to reproduce the trigger requirements (e.g. four p_T thresholds, $m_{qq'}$ and $\Delta\eta_{qq'}$ requirements, and b-tagging for Set A), and to fall sufficiently in or close to the trigger turn-on plateau, without reducing the signal acceptance. In order to avoid run dependency, the preselection criteria are kept fixed for the entire 2012 data taking period.

The trigger efficiency can be estimated in two ways: either by *using a trigger simulation* in combination with a large simulated sample of preselected multijet QCD events, or by *using data* recorded with a reference trigger that is 100% efficient for our preselected events. The reference trigger used is listed in section section 7.2.3. The minor fraction of background processes other than QCD, present in the data after preselection, is small enough to allow the comparison of efficiencies for just data and QCD to be sufficient. The trigger efficiency is computed as a function of various parameters relevant to the trigger and preselection criteria, in the form of one dimensional turn-on curves. If the variable with respect to which the efficiency is being computed is part of the preselection criteria, the cut on this variable is left out in order to make the turn-on effect visible ((N-1) cuts are applied); for other variables all N preselection cuts are applied. It is demonstrated in this section that the simulated trigger efficiency agrees within 10% with the one that is directly extracted from data. In order to avoid having to apply many different data driven trigger efficiencies for various luminosity conditions and run periods, the analysis relies on trigger simulation, but correct the mismodelling (between data and simulation) by applying scale factors.

The scale factors are computed separately for each dataset, as the ratio of the data driven trigger efficiency and the simulated efficiency obtained from the most common HLT trigger paths. For Set A this means using HLT A2 and A6, which together cover about 93% of all data recorded with dedicated triggers; for Set B HLT B2 is used. The mismodelling observed between data and simulation is not linear, and is also correlated between many different variables. For the derivation of a correction method, a two dimensional approach was found to produce better results than a one dimensional one; the trigger efficiency scale factors are hence computed in function of two parameters. For Set A, in order to cope with the different heavy-quark content in data and in the VBF signal events, the scale factors were determined in several bins of the CSV b-tag discriminant of the most b-tagged jet in the event. Additionally, the scale factors were binned in $m_{qq'}$, the invariant mass of the light quark pair. Both these parameters are part of the original trigger requirements, so it makes sense to have a correction method dependent on

them. The result is a two dimensional map of trigger efficiency scale factors that will be applied as a weight to each event that is triggered by the trigger simulation. For Set B, the correction was computed using bins in $m_{qq'}$ and $\Delta\eta_{qq'}$, both parameters featuring in the original trigger logic.

7.3.1 Trigger turn-on curves and efficiency scale factor maps

We demonstrate that the preselection requirements are generally within or very close to the trigger turn-on plateaus, and that there is a reasonable agreement between the efficiencies obtained from trigger simulation applied to QCD multijet simulation and those directly extracted from data by means of a reference trigger. It was checked that the reference triggers applied here do not introduce any bias, by comparing the efficiency obtained from simulation with and without the reference trigger applied.

The trigger efficiencies are defined as the ratio of events triggered by the trigger and the reference trigger as well as passing the preselection criteria, over the events triggered by the reference trigger and passing the preselection criteria:

$$\varepsilon = \frac{\text{(TRG)+(REFTRG)+(PRESEL)}}{\text{(REFTRG)+(PRESEL)}}. \tag{7.1}$$

The results of the reference trigger validation are shown for Set A and Set B in Fig. 7.3.

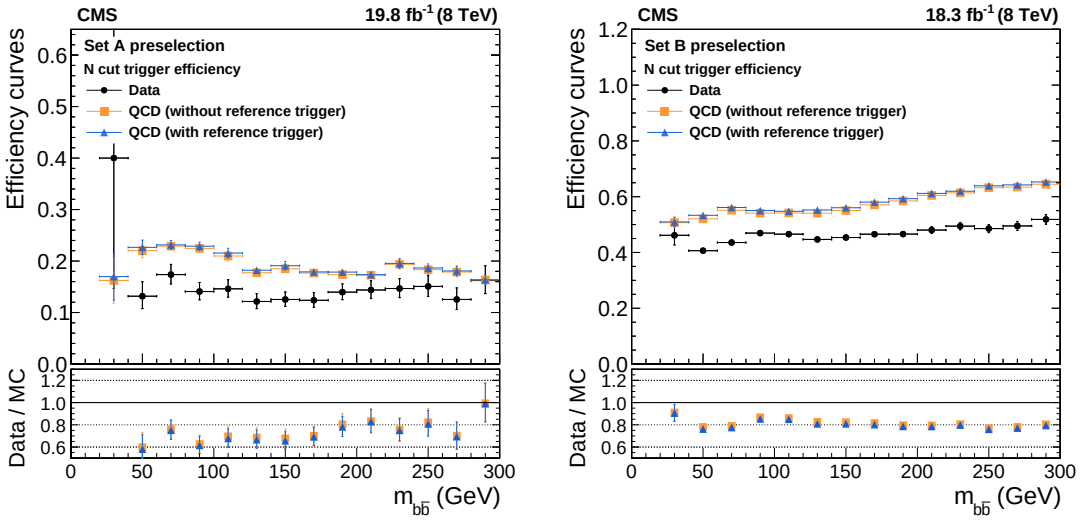


Figure 7.3: Validation of HLT_DiPFJetAve80 reference trigger for both the Set A and Set B selection. Trigger efficiency curves are shown for the QCD multijet sample excluding (orange) and including (blue) the reference trigger; the efficiency in data is shown in black. The efficiency is calculated with respect to the b-quark pair invariant mass $m_{b\bar{b}}$, with the full preselection applied.

A complete set of trigger efficiency curves with and without correction is shown for Set A in Figs. 7.6a and 7.6b, and for Set B in Figs. 7.7a and 7.7b. The efficiencies are shown as a function of the preselection variables.

Set A the four jet transverse momenta (jet $p_T^{(0-3)}$), the pseudorapidity difference between the light-quark jets (following the b-likelihood interpretation, cf. section 7.4) ($\Delta\eta_{qq'}$) and their invariant mass ($m_{qq'}$), and the CSV b-tag value of the two most b-tagged jets (b-jet$^{(0,1)}$ CSV). In addition to these eight variables from the preselection, three other important variables are tested: the b-quark jet pair invariant mass (following the b-likelihood interpretation) ($m_{b\bar{b}}$), and it's ϕ separation ($\Delta\phi_{b\bar{b}}$), and the multivariant discriminant for Set A (cf. section 8.2).

Set B the average transverse momentum ($p_T^{\mathrm{ave}(0,1)}$) of the two leading jets, the transverse momentum of the fourth jet, the pseudorapidity difference between the light-quark jets (following the b-likelihood interpretation) ($\Delta\eta_{qq'}$) and their invariant mass ($m_{qq'}$), the pseudorapidity difference between the jj jet pair (following the trigger logic) ($\Delta\eta_{jj}$) and it's invariant mass (m_{jj}), and the CSV b-tag of the two most b-tagged jets (b-jet$^{(0,1)}$ CSV). Also for Set B the three extra variables are tested.

Several observations can be made from these figures:

- the trigger efficiencies in Set A are small, $\mathcal{O}(15\%)$, because the heavy-quark content of the preselected sample is low; in Set B the efficiencies are closer to $\mathcal{O}(50\%)$
- the preselection criteria lie within or very close to the efficiency plateaus (marked with a red dashed line)
- the efficiencies extracted from data and obtained from simulation before correction agree within 50% of their values
- the ratio between the efficiencies obtained from data and simulation are often largely independent of the considered variables, in particular for the $m_{b\bar{b}}$ invariant mass spectrum

Choices for the binning of the correction factors are made based on the flatness of the trigger efficiency ratio for each variable. For the Set A triggers the largest effects are visible in the $m_{qq'}$ and b-jet$^{(0)}$ CSV variables. For Set B, $m_{qq'}$ and $\Delta\eta_{qq'}$ are the most obvious choices. The dependence of trigger efficiency on the chosen variables can be seen for data and simulation in Fig. 7.6b (top left) and Fig. 7.6a (bottom right) for Set A and in Fig. 7.7a (middle row) for Set B.

The trigger efficiency scale factor maps derived as the ratio of the data and simulation trigger efficiency maps are shown in Fig. 7.4 (Set A) and Fig. 7.5 (Set B). To appreciate the effect of the calculated correction factors, one can compare the uncorrected (blue triangles) and corrected (green diamonds) efficiency curves in Figs. 7.6 and 7.7.

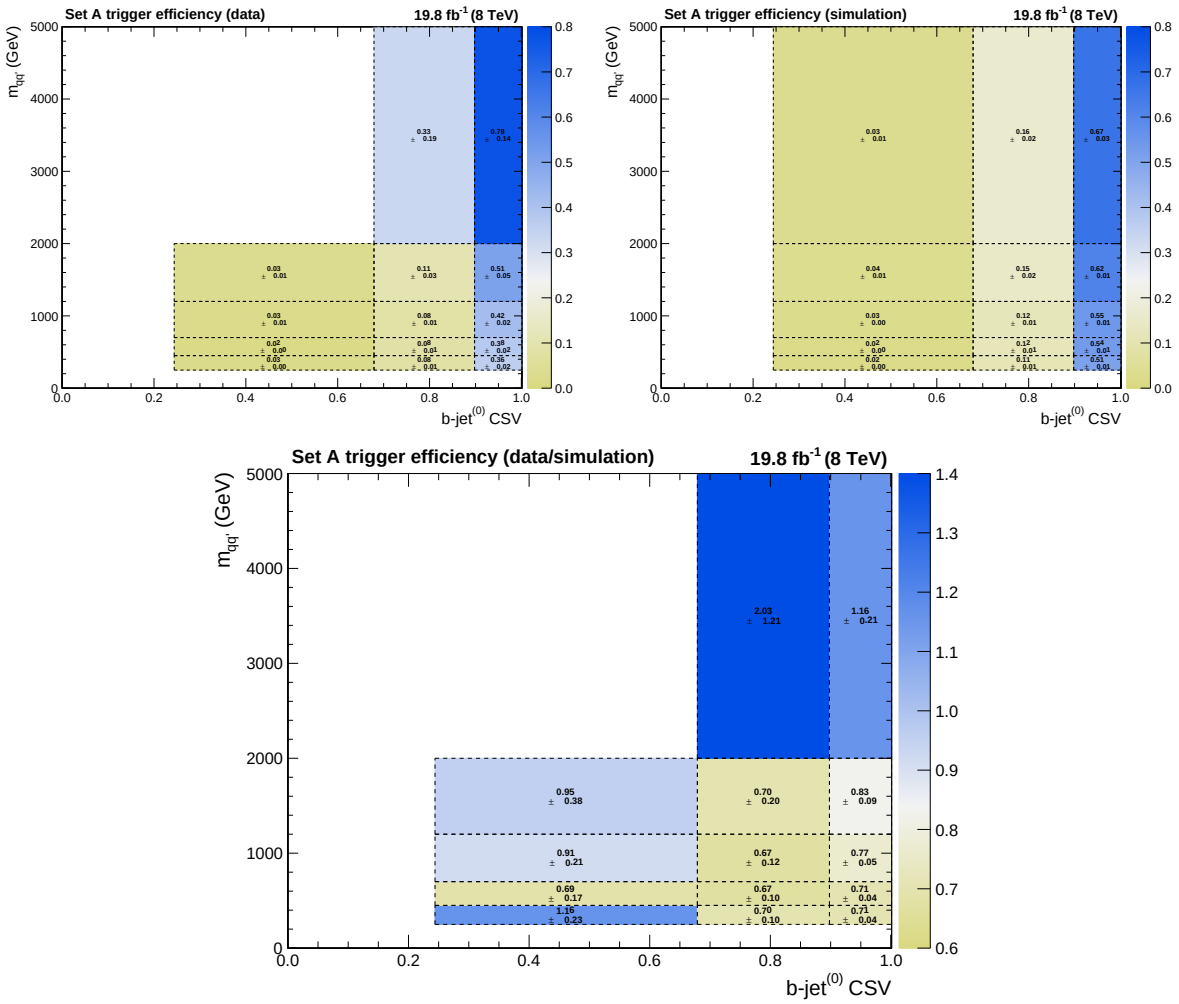


Figure 7.4: Two dimensional trigger efficiency scale factor map for the Set A selection (bottom), derived as the ratio of the data (top left) and simulation (top right) trigger efficiency maps. The scale factors are calculated with respect to the leading b-jet CSV b-tag value and the light-quark jet pair invariant mass.

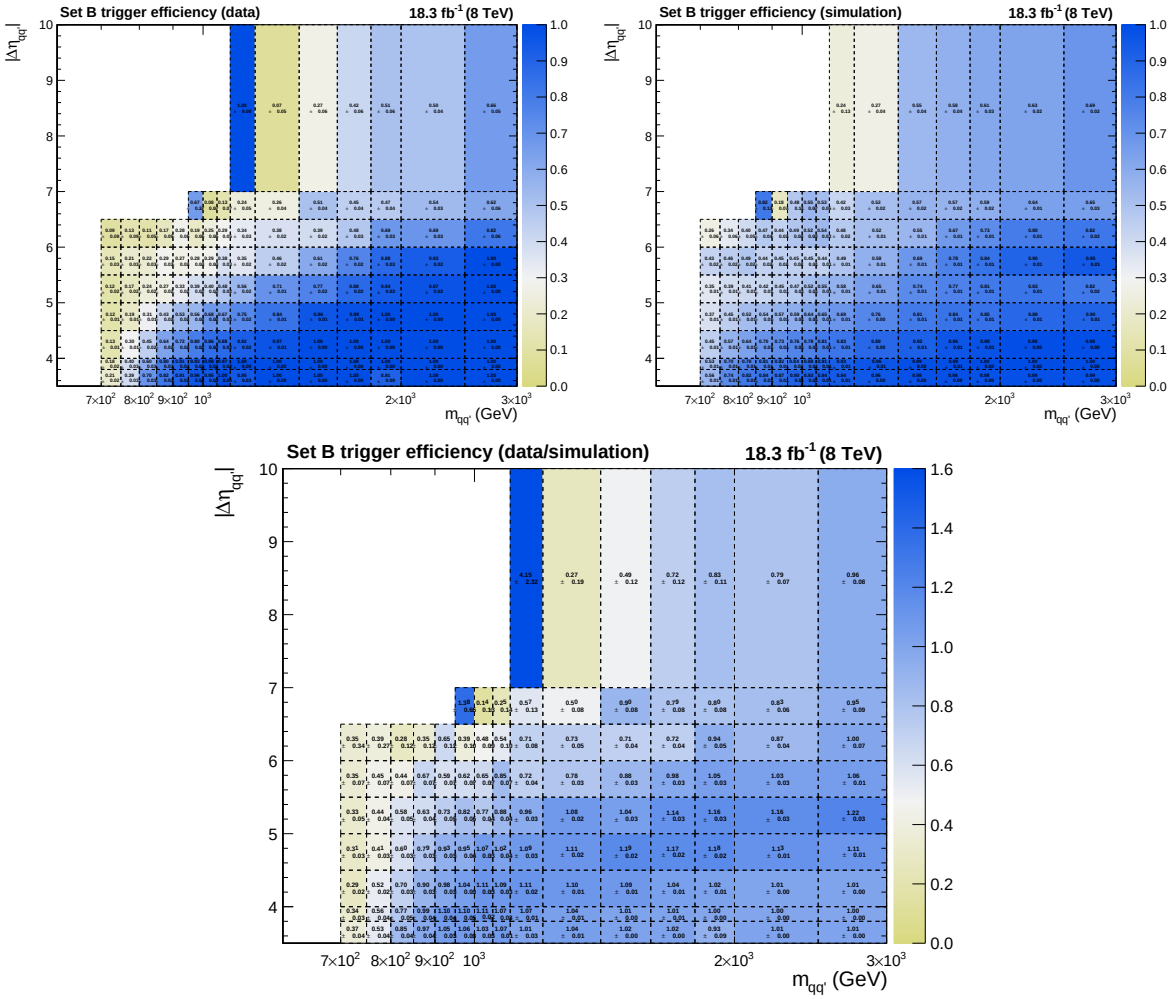


Figure 7.5: Two dimensional trigger efficiency scale factor map for the Set B selection (bottom), derived as the ratio of the data (top left) and simulation (top right) trigger efficiency maps. The scale factors are calculated with respect to the light-quark jet pair invariant mass and η separation.

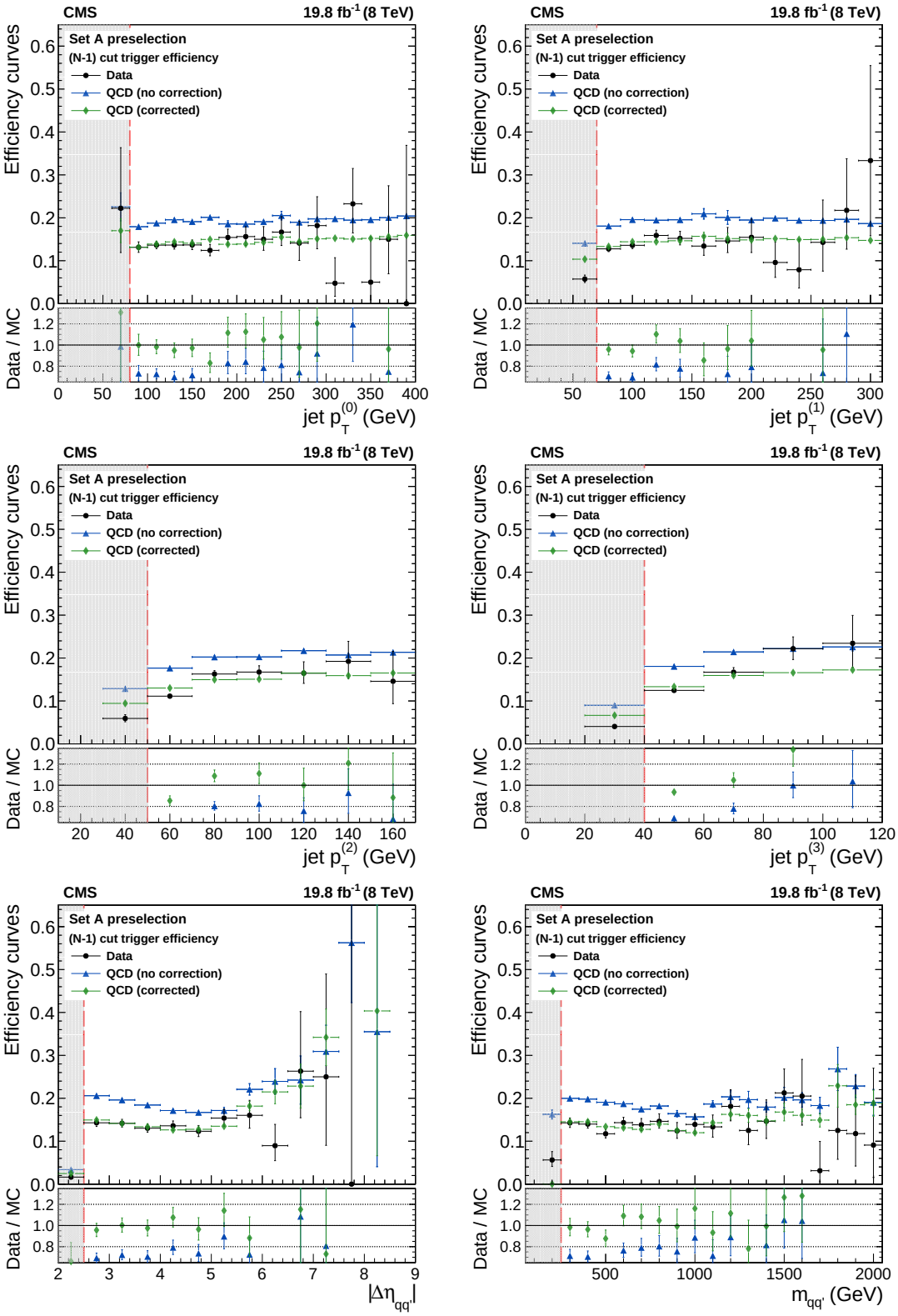


Figure 7.6a: Validation of trigger efficiency scale factors for the Set A selection. Trigger efficiency curves are shown for the QCD multijet sample without (blue) and with (green) the correction; the efficiency in data is shown in black. The efficiencies are calculated with respect to eleven different variables relevant to the analysis, with the full preselection applied. If the variable relative to which the efficiency is calculated is part of the preselection, the cut on this variable is removed ((N-1) cuts are applied); the removed cut is indicated with a grey band. (1/2)

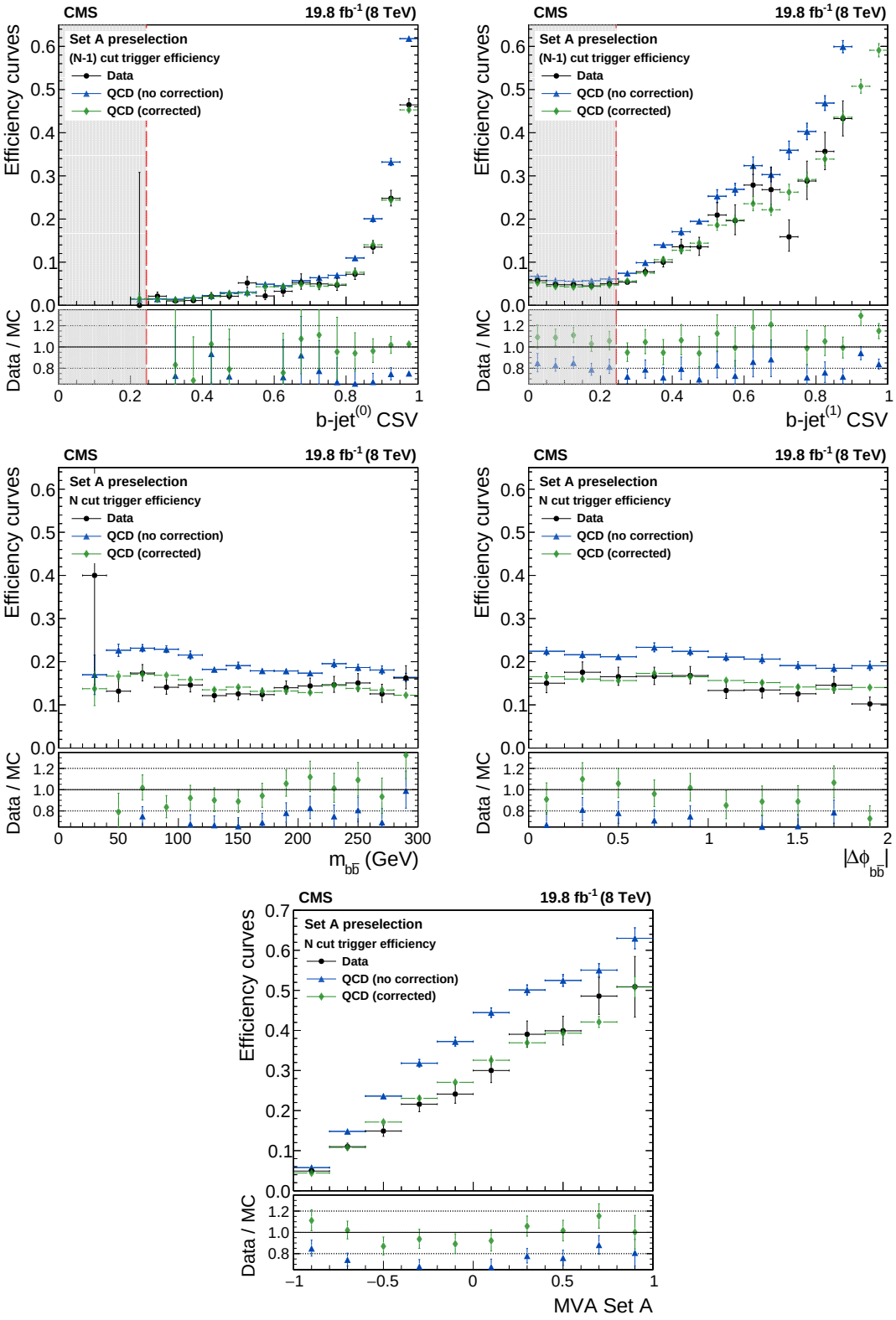


Figure 7.6b: Validation of trigger efficiency scale factors for the Set A selection. Trigger efficiency curves are shown for the QCD multijet sample without (blue) and with (green) the correction; the efficiency in data is shown in black. The efficiencies are calculated with respect to eleven different variables relevant to the analysis, with the full preselection applied. If the variable relative to which the efficiency is calculated is part of the preselection, the cut on this variable is removed ((N-1) cuts are applied); the removed cut is indicated with a grey band. (2/2)

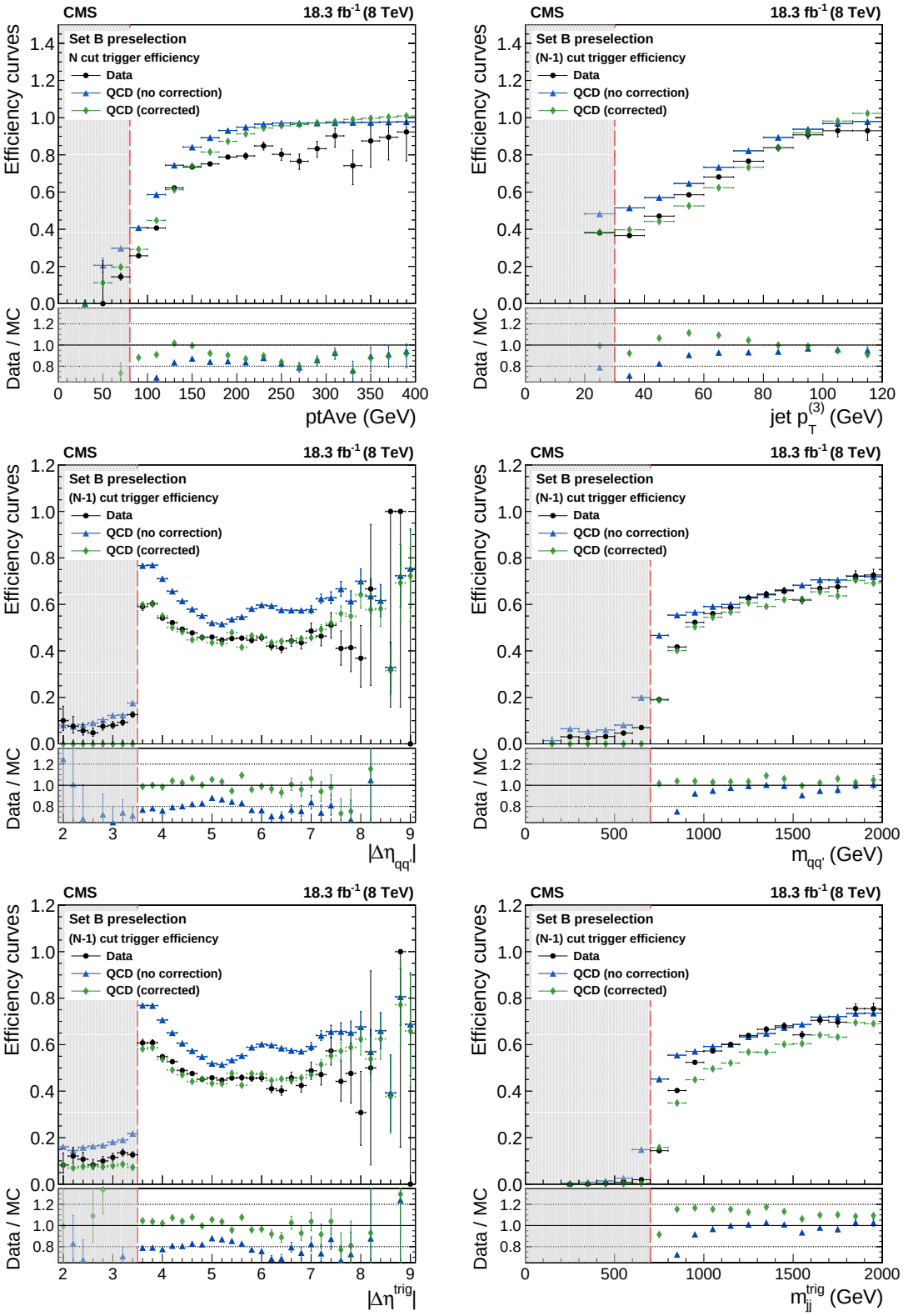


Figure 7.7a: Validation of trigger efficiency scale factors for the Set B selection. Trigger efficiency curves are shown for the QCD multijet sample without (blue) and with (green) the correction; the efficiency in data is shown in black. The efficiencies are calculated with respect to eleven different variables relevant to the analysis, with the full preselection applied. If the variable relative to which the efficiency is calculated is part of the preselection, the cut on this variable is removed ((N-1) cuts are applied); the removed cut is indicated with a grey band. (1/2)

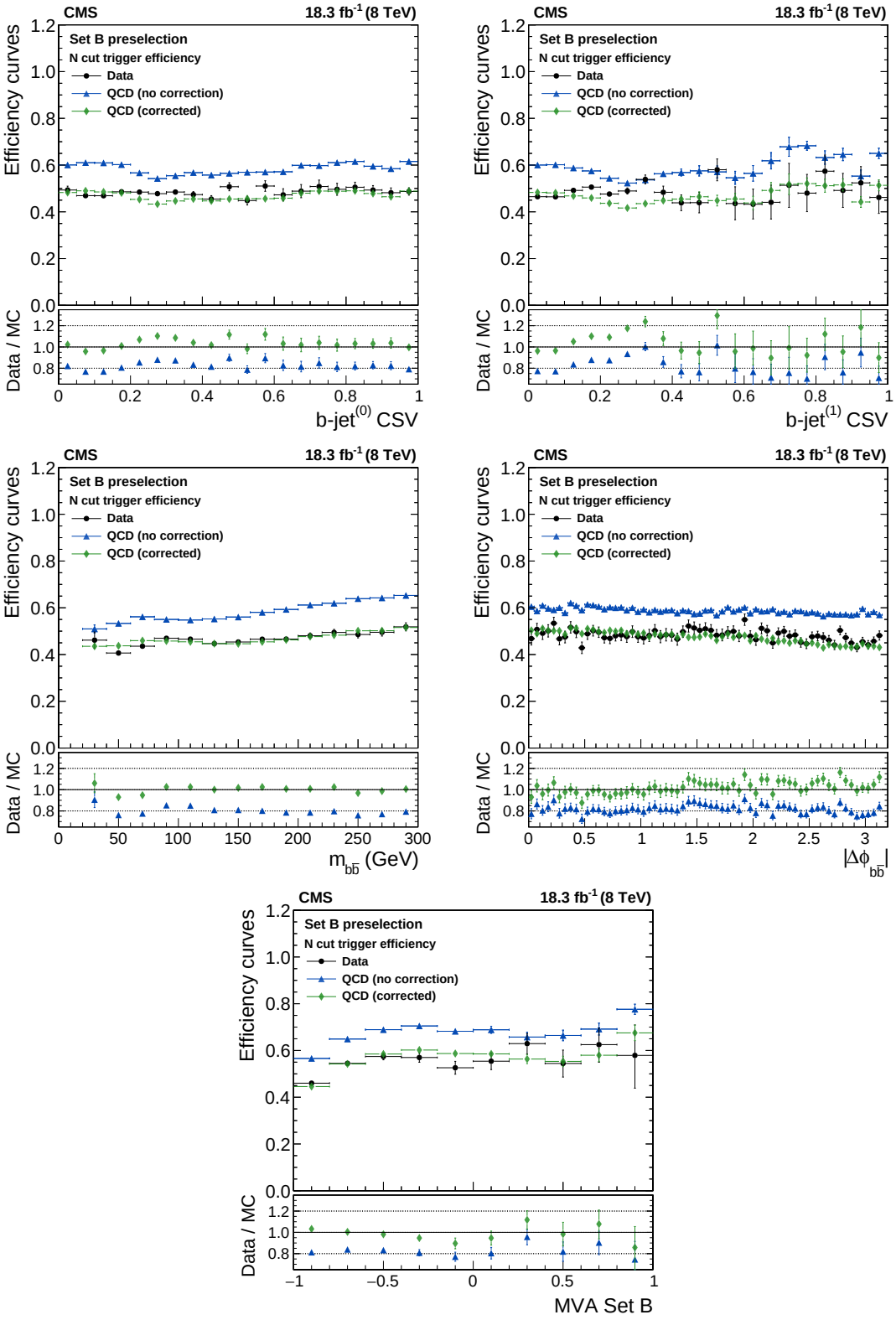


Figure 7.7b: Validation of trigger efficiency scale factors for the Set B selection. Trigger efficiency curves are shown for the QCD multijet sample without (blue) and with (green) the correction; the efficiency in data is shown in black. The efficiencies are calculated with respect to eleven different variables relevant to the analysis, with the full preselection applied. If the variable relative to which the efficiency is calculated is part of the preselection, the cut on this variable is removed ((N-1) cuts are applied); the removed cut is indicated with a grey band. (2/2)

7.4 Event interpretation

The VBF $H \to b\bar{b}$ analysis studies four jet final states. In addition to selecting the four jets with the highest transverse momentum, the two light-quark jets and the two b-quark jets need to be identified. The result of the analysis will depend on the quality of the interpretation. At trigger level, events are interpreted following the most naive and straightforward way: according to their CSV b-tag ordering. The identification efficiency for the most b-tagged jet is reasonably good, but especially for the second most b-tagged jet, the misidentification probability is not negligible. Hence offline, where one has the opportunity to perform further processing, events are reinterpreted using a multivariate b-likelihood discriminant, aiming to maximise the interpretation quality. The multivariate discriminant uses four input parameters:

- CSV b-tag value
- pseudorapidity value
- CSV b-tag rank
- pseudorapidity rank.

Together, these four parameters cover the properties of the VBF $H \to b\bar{b}$ signal well: the b-quark jets should have high b-tag values, the should have central pseudorapidity values, and they should be situated inbetween the light-quark jets in pseudorapidity. To fully optimise the performance, the discriminant is trained separately for the Set A and Set B datasets.

The multivariate discriminant is trained with the ROOT TMVA toolkit [133], using the VBF $H \to b\bar{b}$ simulated signal sample with $m_H = 125$ GeV. All reconstructed jets in the sample are matched to partons using a matching criterion $\Delta R < 0.25$. Those jets matching b-partons (parton id $= \pm 5$) form the signal sample for the discriminant training, while the others form the background sample. The setup uses a boosted decision tree (600 trees, boost type=gradient, shrinkage=0.2) [133] and the training is done using an equal amount of signal and background entries (100k).

The input variables to the discriminant are shown for Set A and Set B in Figs. 7.8 and 7.9. The b-quark matching efficiency is suboptimal especially for the second most b-tagged jet. The linear correlation between the variables is given in Figs. 7.10 and 7.11.

Figures 7.12-7.15 illustrate the performance of the discriminant. Fig. 7.12 shows the background rejection versus signal efficiency. The result is best for the b-likelihood interpretation. In Figs. 7.13 and 7.14 the b-likelihood distributions are given, showing improved signal versus background separation. Finally Fig. 7.15 shows the effect of the corresponding improvement on the b-quark pair invariant mass $m_{b\bar{b}}$. The tail into higher $m_{b\bar{b}}$ values is much reduced in the b-likelihood interpretations, compared to the b-tag ordered interpretation. Overall, the improved event interpretation quality, reducing the amount of wrong quark pair combinations, leads to an increase in signal acceptance of the order of 10%.

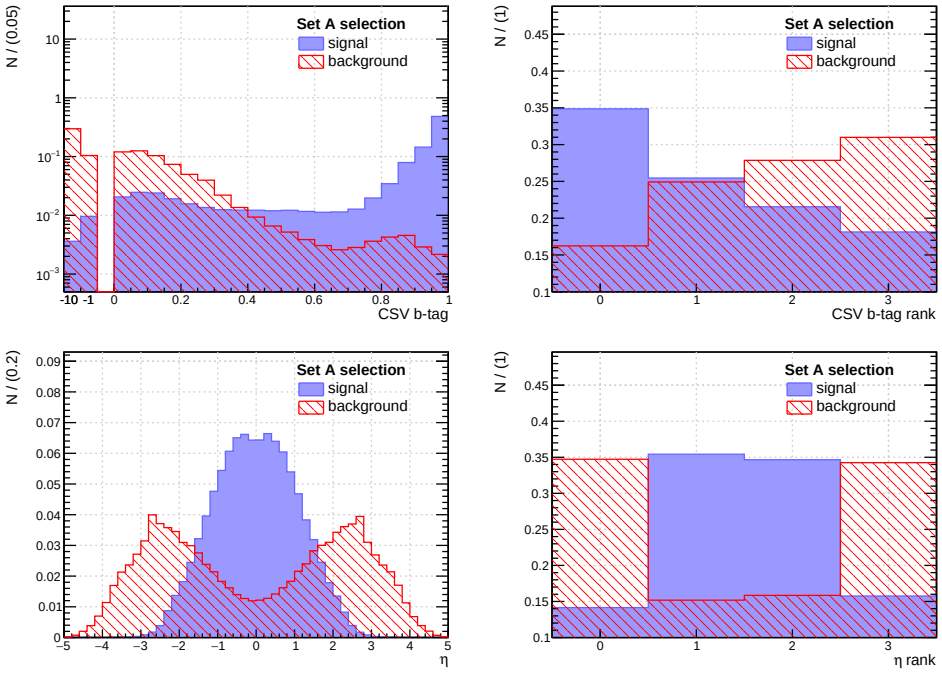


Figure 7.8: Input variables for the event interpretation b-likelihood discrimininant, for Set A (after trigger selection). Signal consists of jets matched to b-partons, background contains all other jets; using the VBF $H \to b\bar{b}$ signal sample for $m_H = 125$ GeV.

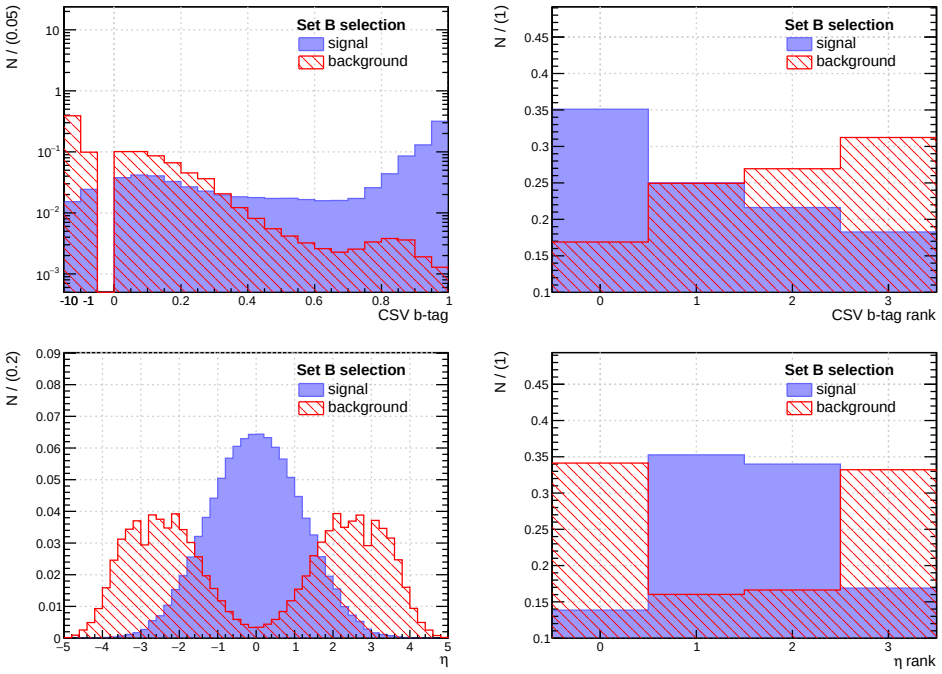


Figure 7.9: Input variables for the event interpretation b-likelihood discrimininant, for Set B (after trigger selection). Signal consists of jets matched to b-partons, background contains all other jets; using the VBF $H \to b\bar{b}$ signal sample for $m_H = 125$ GeV.

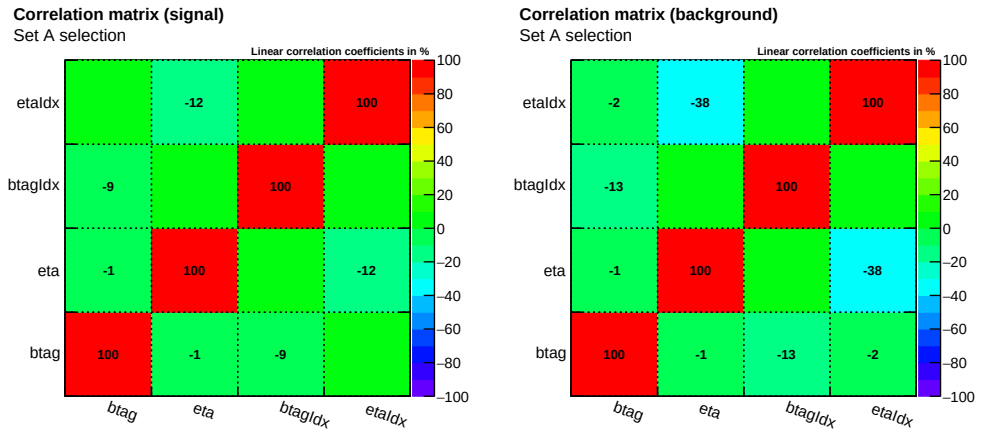


Figure 7.10: Linear correlation coefficients for the input variables to the event interpretation b-likelihood discriminant, for Set A (after trigger selection).

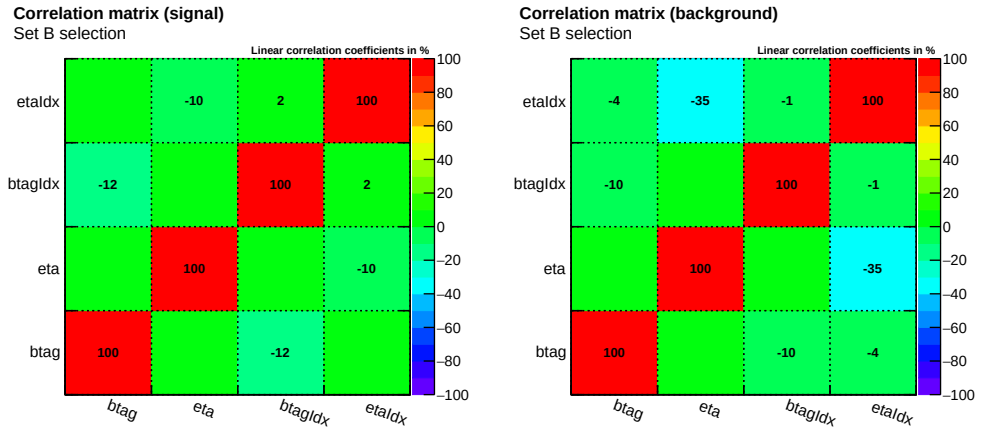


Figure 7.11: Linear correlation coefficients for the input variables to the event interpretation b-likelihood discriminant, for Set B (after trigger selection).

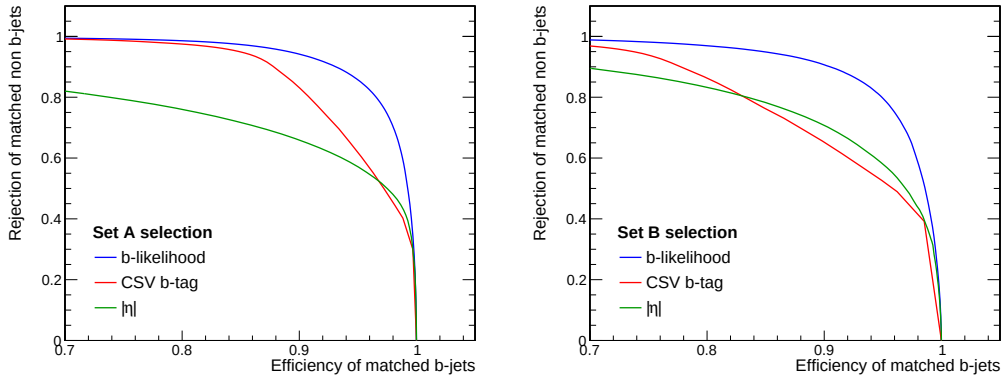


Figure 7.12: Background rejection vs. signal efficiency for the event interpretation b-likelihood discriminant, for Set A and Set B (after trigger selection).

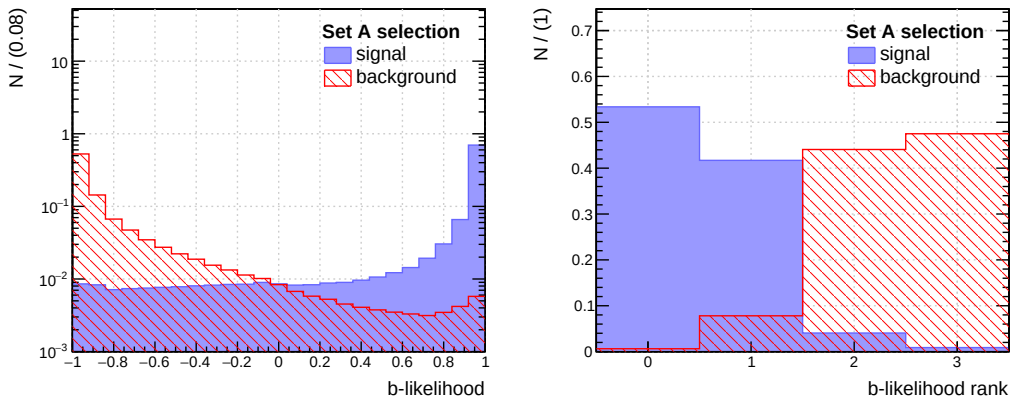


Figure 7.13: Output of the event interpretation b-likelihood discriminant, for Set A (after trigger selection).

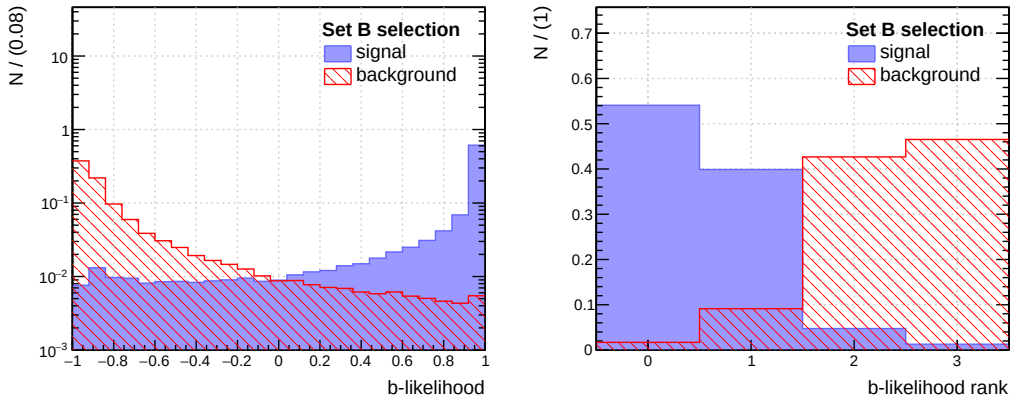


Figure 7.14: Output of the event interpretation b-likelihood discriminant, for Set B (after trigger selection).

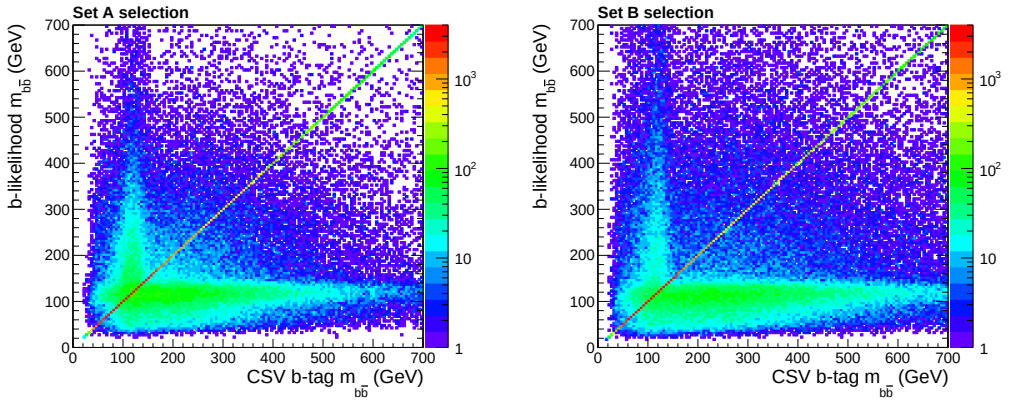


Figure 7.15: Comparison of b-quark pair invariant mass $m_{b\bar{b}}$ in the CSV b-tag interpretation and the b-likelihood interpretation, for Set A and Set B (after trigger selection).

7.5 Event selection

The event selection criteria are optimised to reproduce the trigger requirements and to fall close to or in the trigger plateau regions, to select a phase space where the trigger efficiency is well-defined, while reducing the signal acceptance as little as possible. Two selections are defined: the Set A selection matching the Set A triggers, and the Set B selection matching the Set B triggers.

For Set A, following the structure of the dedicated VBF $H \to b\bar{b}$ trigger, the four leading jets must have transverse momenta $p_T^{(0,1,2,3)} > 80, 70, 50$ and 40 GeV. At least two jets must be loosely b-tagged. Subsequently, the event is interpreted according to the Set A b-likelihood discriminant values of the four leading jets. The kinematic properties of the VBF $H \to b\bar{b}$ signal are enforced by requiring pseudorapidity separation $|\Delta\eta_{qq'}| > 2.5$ and invariant mass $m_{qq'} > 250$ GeV, for the light-quark jet pair. An additional angular separation requirement $\Delta\phi_{b\bar{b}} < 2.0$ is set for the b-quark jet pair, in order to suppress back-to-back b-jet QCD background. Finally, events with identified and isolated leptons are explicitly rejected. This selection forms the Set A data sample.

For Set B, events selected from the parked data streams using the general-purpose VBF trigger are considered. Offline emulation of the online selection here requires the use of additional variables, following the trigger logic. Jet pairs are constructed according to all possible combinations of the four leading jets (in transverse momentum), with the additional requirement that $p_T > 35$ GeV for both jets in a pair. The pair with the largest invariant mass m_{jj} is selected, and for the same pair the pseudorapidity separation $|\Delta\eta_{jj}|$ is derived. In parallel, the invariant mass $m_{qq'}$ and pseudorapidity separation $|\Delta\eta_{qq'}|$ for the light-quark pair are computed following the Set B b-likelihood interpretation. The kinematic requirements are set twice: once on the variables following the trigger logic and once on the variables following the b-likelihood interpretation. The offline selection requires $m_{qq'}, m_{jj} > 700$ GeV and $|\Delta\eta_{qq'}|, |\Delta\eta_{jj}| > 3.5$. Furthermore, the average transverse momentum of the two leading jets has to satisfy $p_T^{ave} > 80$ GeV. Given the general-purpose online VBF selection, no b-jet requirement have yet been made. The offline selection ensures b-jet enrichment by requiring at least one medium and one loosely identified b-jet. Again, an angular separation requirement $\Delta\phi_{b\bar{b}} < 2.0$ is set for the b-quark jet pair, in order to suppress back-to-back b-jet QCD background, and events with identified and isolated leptons are explicitly rejected. This selection forms the Set B data sample.

The Set A and Set B selections are not orthogonal, there is a significant overlap. Fig. 7.16 illustrates the relative size and overlap of the two selections, based on the VBF $H \to b\bar{b}$ simulated signal sample for $m_H = 125$ GeV, after trigger and offline selection requirements. Since the Set A data sample is built on the dedicated VBF $H \to b\bar{b}$ trigger, it is considered the primary

data sample and is used inclusively (including the overlap). In order not to double-count events, the Set B data sample is defined exclusively (excluding the overlap), namely events satisfying the Set A trigger and selection criteria are vetoed. The exclusive Set B dataset contributes approximately 35% of additional data.

Table 7.5 summarises the selection requirements for the two samples.

Table 7.5: Summary of selection requirements for Set A and Set B.

	Set A	Set B
trigger	OR of Set A triggers	Set B trigger
jet p_T	$p_T^{(0,1,2,3)} > 80, 70, 50, 40$ GeV	$p_T^{(0,1,2,3)} > 30$ GeV and $p_T^{ave(0,1)} > 80$ GeV
b-tagging	2x CSVL (> 0.244)	1x CSVM (> 0.679) and 1x CSVL (> 0.244)
interpretation	Set A b-likelihood	Set B b-likelihood and trigger logic
VBF topology	$m_{qq'} > 250$, $\lvert\Delta\eta_{qq'}\rvert > 2.5$	$m_{qq'}, m_{jj} > 700$, $\lvert\Delta\eta_{qq'}\rvert, \lvert\Delta\eta_{jj}\rvert > 3.5$
QCD suppression	$\Delta\phi_{b\bar{b}} < 2.0$	$\Delta\phi_{b\bar{b}} < 2.0$
vetoes	lepton veto	lepton veto and Set A veto

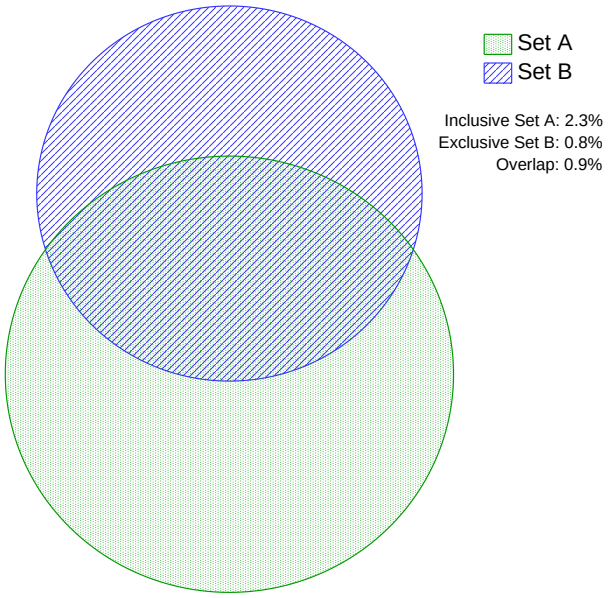


Figure 7.16: Selection efficiency for the Set A and Set B selections, based on the VBF $H \to b\bar{b}$ simulated signal sample for $m_H = 125$ GeV. The relative size and overlap are illustrated.

Chapter 8

Signal vs. background discrimination

The second major step in the VBF $H \to b\bar{b}$ analysis, after trigger and offline selection, is the development and use of a multivariate discriminant, optimising the signal vs. background discrimination. Section 8.1 highlights three techniques that exploit final state properties and bring significant improvement to the analysis sensitivity: jet transverse momentum regression, quark-gluon discrimination, and measuring additional hadronic activity. These affect the sensitivity either directly, by improving the b-quark jet pair invariant mass resolution, or indirectly, by providing variables that contribute to the performance of the multivariate discriminant. Section 8.2 then discusses the details and performance of the multivariate discriminant. Finally section 8.3 concerns data validation, in the form of a series of data vs. simulation comparisons, for both the Set A and Set B selections.

8.1 Signal properties

8.1.1 Jet transverse momentum regression

The standard CMS jet calibration [101] is performed as a function of jet transverse momentum and pseudorapidity, and is an average calibration by construction. When working with PF jets, many more jet properties are accessible than are used in the standard calibration, and an improved result can be obtained by using this information in a multidimensional recalibration. A jet regression technique was developed in collaboration with the VH $b\bar{b}$ analysis group, using the additional jet composition properties and targeting the jet transverse momentum; it is performed on b-jets only, after the standard calibration. One of the main problems addressed by the recalibration is that of semileptonic b decays: due to energy losses via undetected neutrinos, the response for this type of events is lower, and there is a significant mismeasurement of the jet transverse momentum.

The regression is done at particle level (jets clustered from all stable particles), using the VBF $H \to b\bar{b}$ simulated signal sample for $m_H = 125$ GeV. It is implemented using the TMVA toolkit [133], as a regression boosted decision tree (BDT) with thirteen inputs:

- **(i-iii)** *jet properties:* the jet transverse momentum, pseudorapidity and mass
- **(iv-v)** *energy fractions:* the neutral hadron and photon energy fractions
- **(vi-vii)** *secondary vertex properties (when present):* the mass and uncertainty in the decay length of the secondary vertex

- **(viii-ix)** *missing transverse energy:* the missing transverse energy and its azimuthal direction relative to the jet
- **(x)** *multiplicity:* the total number of jet constituents
- **(xi-xii)** *soft-lepton candidates (when present):* the transverse momentum of the candidate and its component perpendicular to the jet axis
- **(xiii)** *pile-up:* the average transverse momentum density ρ.

The training of the BDT is done separately for the Set A and Set B selections. The recalibration of the b-jets leads to an improvement in the jet transverse momentum, which in turn leads to an improvement in the b-quark pair invariant mass resolution. Figure 8.1 shows the b-quark pair invariant mass resolution before and after jet transverse momentum regression, for both the Set A and Set B selections. The resolution improvement is around 17% for the Set A selection, and around 18% for the Set B selection. In Fig. 8.2, the b-quark pair invariant mass distribution is shown for data and simulation, for both selections.

The jet transverse momentum regression was validated in data, measuring the response in Z + 1 b-jet and Z + 2 b-jets data, with the Z boson decaying to two leptons. The Z boson is reconstructed from the leptons, and a balance is made between the Z boson transverse momentum and that of the b-jet or b-jet system. The balance distribution is observed to be narrower and centred nearer to zero after jet transverse momentum regression correction. The scale and resolution uncertainties derived in this study are used in the final fit of the invariant mass spectrum.

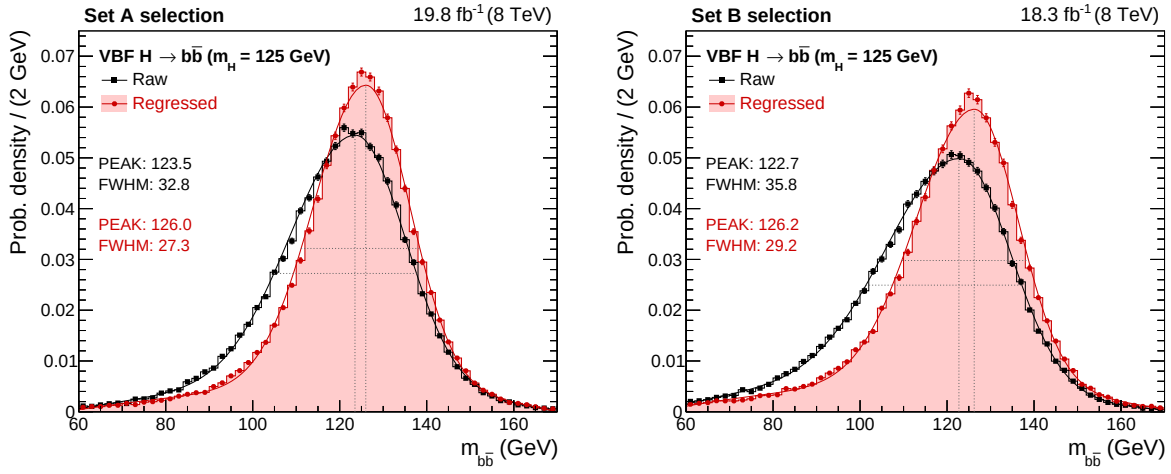


Figure 8.1: Comparison of the b-quark pair invariant mass distribution before and after jet p_T regression, as measured from VBF $H \to b\bar{b}$ signal simulation, for the Set A (left) and Set B (right) selections. The peak value and the width of the distribution at half of its maximum value are indicated in the figures.

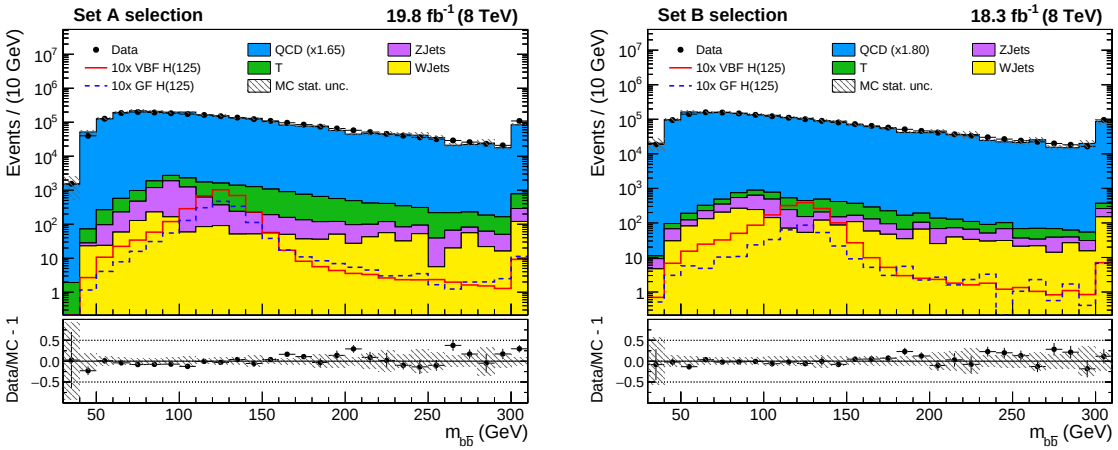


Figure 8.2: Distribution of the b-quark jet pair invariant mass, after jet p_T regression, for the Set A (Set B) selection. Data points are shown with black markers. Background contributions are stacked, weighted with their respective cross sections, and scaled to the luminosity in data. The LO QCD cross section is multiplied by a k-factor of 1.65 (1.80) such that the total number of background events matches the number of events in data. The VBF and GF signal contributions are overlaid, their cross sections scaled by a factor ten to improve visibility. All simulated events are corrected for pile-up and trigger efficiency effects. The last bin is an overflow bin. The bottom plot shows the fractional difference between data and background simulation, with the shaded band representing the statistical uncertainty in the simulated samples.[1]

8.1.2 Quark-gluon likelihood discriminant

For the QCD multijet background, typically about 70% of the jets originate from gluons, about 25% originate from light quarks, and the remaining ones originate from heavy quarks. The VBF $H \to b\bar{b}$ signal however, is a purely electroweak process, meaning all jets originate from quarks only. Consequently, applying quark-gluon tagging (as discussed in section 5.2.5) allows to improve the signal vs. background discrimination significantly. The discriminant uses jet composition properties to distinguish between quark and gluon originating jets: gluon originating jets tend to have higher constituent multiplicity, softer constituent p_T and wider jet cones. For this analysis, the gluon likelihood version of the discriminant is used, peaking at zero for quark originating jets, and at one for gluon originating jets.

Figures 8.3 and 8.4 show the QGL discriminant distribution for each of the final state jets, in data and simulation, for both the Set A and the Set B selections. Especially for the two light-quark jets (lowest b-likelihood, b-jet$_2$ and b-jet$_3$), the signal distribution clearly peaks towards lower values. For the b-quark jets (highest b-likelihood, b-jet$_0$ and b-jet$_1$), the signal distribution is more uniform, with equal quark/gluon probability. The background distribution shows increased gluon content especially in the b-quark jet cases.

The QGL distributions of all four jets are entered as inputs to the multivariate signal vs. background discriminant (cf. section 8.2).

[1] All subsequent data vs. simulation comparison plots follow the conventions detailed here; the are only printed once. A broader discussion of the comparisons follows in section 8.3.

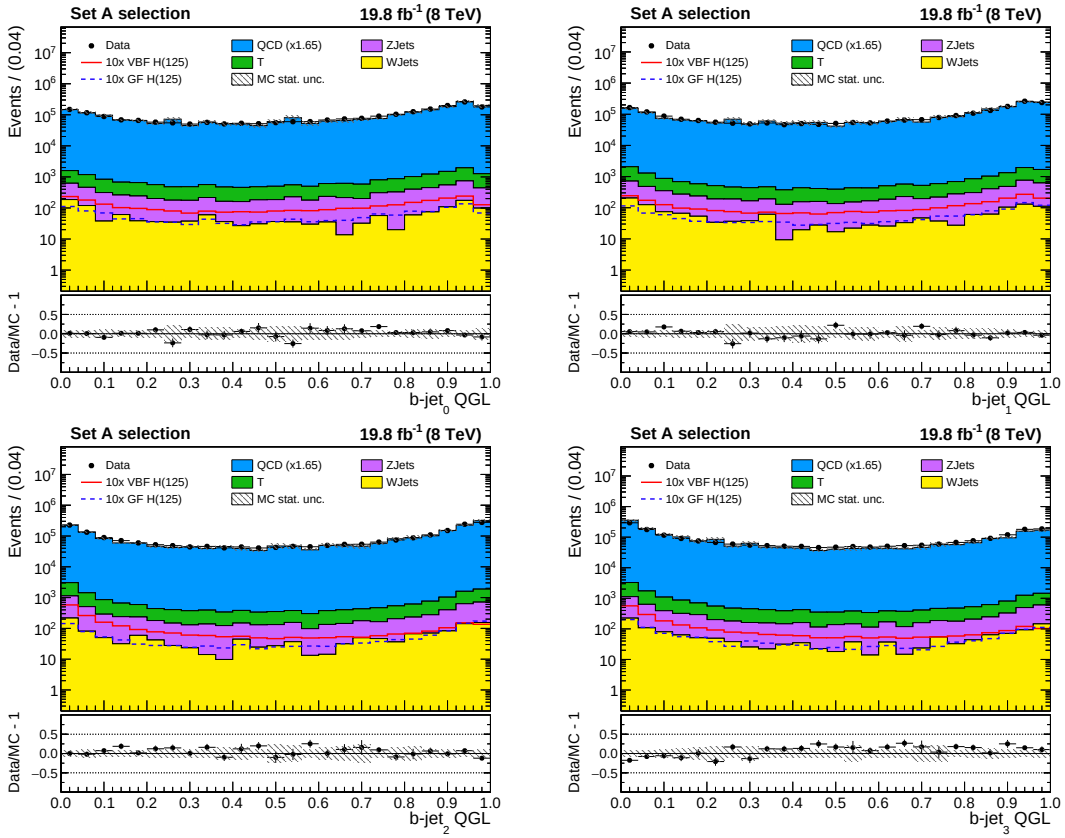


Figure 8.3: Distribution of the quark-gluon likelihood discriminant value of the jets, ordered by b-likelihood value, for the Set A selection.

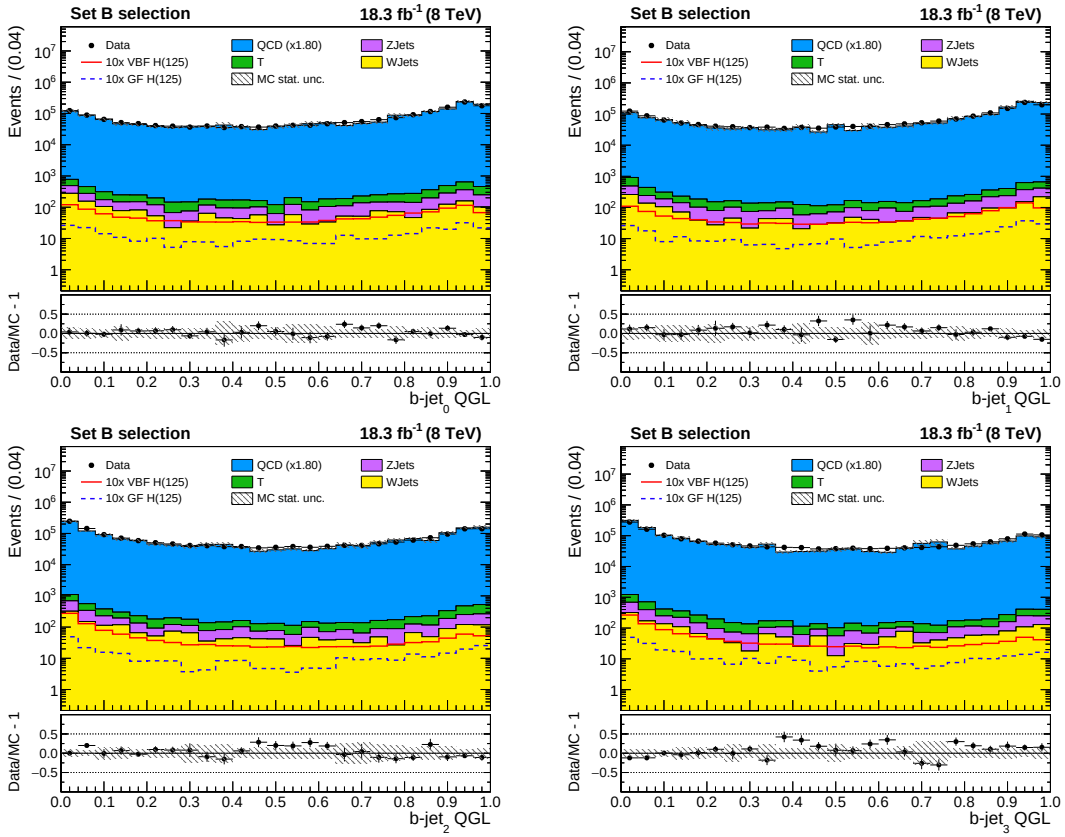


Figure 8.4: Distribution of the quark-gluon likelihood discriminant value of the jets, ordered by b-likelihood value, for the Set B selection.

8.1.3 Additional hadronic activity

As mentioned in the introduction (cf. section 6.1), the QCD colour structure of the events is quite specific for the VBF $H \to b\bar{b}$ signal. Since the signal is a purely electroweak process, no colour exchange is expected other than the light-quark VBF-tagging jets colour reconnecting with the beam remnant (along the z-axis), and the b-quark jets colour reconnecting between themselves. Hence, in the $\eta - \phi$ region between the light-quark jets and the more central b-quark jets, no hadronic activity is expected. The QCD multijet background however, concerns QCD processes, and will include additional hadronic activity in this region.

To distinguish between signal and background, variables are constructed that measure the additional hadronic activity in the region between the light-quark jets and the b-quark jets. For this purpose, an additional set of tracks is assembled, after pile-up subtraction, using only high purity tracks [90] that:

- **(i)** satisfy $p_T > 300$ MeV,
- **(ii)** are not associated to any of the four leading jets,
- **(iii)** fall outside the elliptical $\eta - \phi$ cone formed by the b-quark jets,
- **(iv)** have a minimal longitudinal impact parameter $|d_z(PV)|$ with respect to the main primary vertex,
- **(v)** satisfy $|d_z(PV)| < 2\,\text{mm}$ and $|d_z(PV)| < \sigma_{d_z}$ (the uncertainty on $|d_z(PV)|$).

The $\eta - \phi$ cone formed by the b-quark jets is defined with major axis length $\Delta R + 1$ ($\Delta R = \sqrt{(\Delta\eta_{b\bar{b}})^2 + (\Delta\phi_{b\bar{b}})^2}$) and minor axis length 1. Not counting the tracks in this cone makes a significant difference in the activity level measured for the signal, while having less of an effect on the QCD result.

The extra tracks are finally clustered into soft trackjets (cf. section 5.2.2) using the anti-k_T clustering algorithm with distance parameter 0.5. Trackjets are chosen for their good performance down to very low energies. In this case all jets satisfying $p_T > 1$ GeV are taken into account. From this set of newly reconstructed trackjets, the scalar transverse momentum sum (H_T^{soft}) and the soft multiplicity of all tracks with $p_T > 2$ GeV (N_2^{soft}) are constructed.

In Figs. 8.5 and 8.6, the H_T^{soft} and N_2^{soft} distributions are shown in data and simulation, for both the Set A and Set B selections. While the background distribution has a sizeable tail in H_T^{soft}, the signal distribution peaks at very small values. The same is true for the soft multiplicity, with the signal distribution decreasing quickly, and the background distribution extending towards higher values.

Both H_T^{soft} and N_2^{soft} are entered as inputs to the multivariate signal vs. background discriminant (cf. section 8.2).

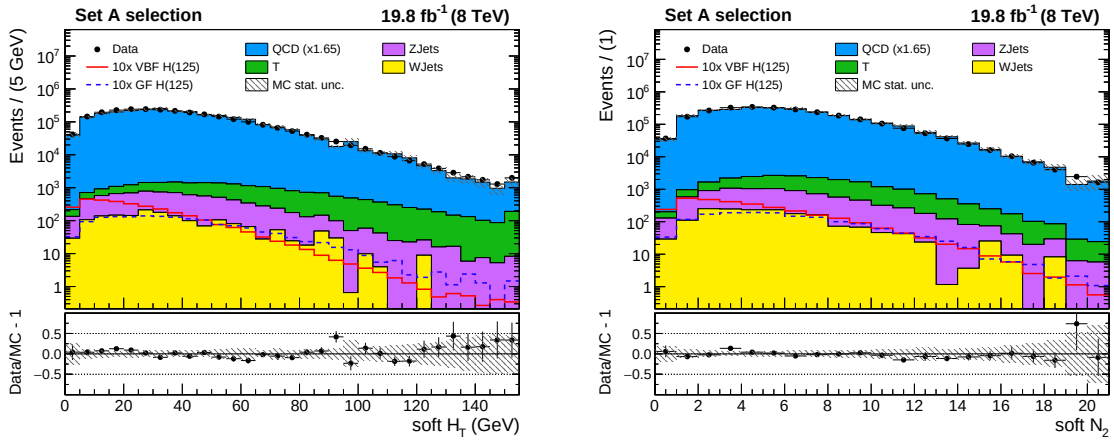


Figure 8.5: Soft track activity momentum sum and multiplicity distributions, for the Set A selection.

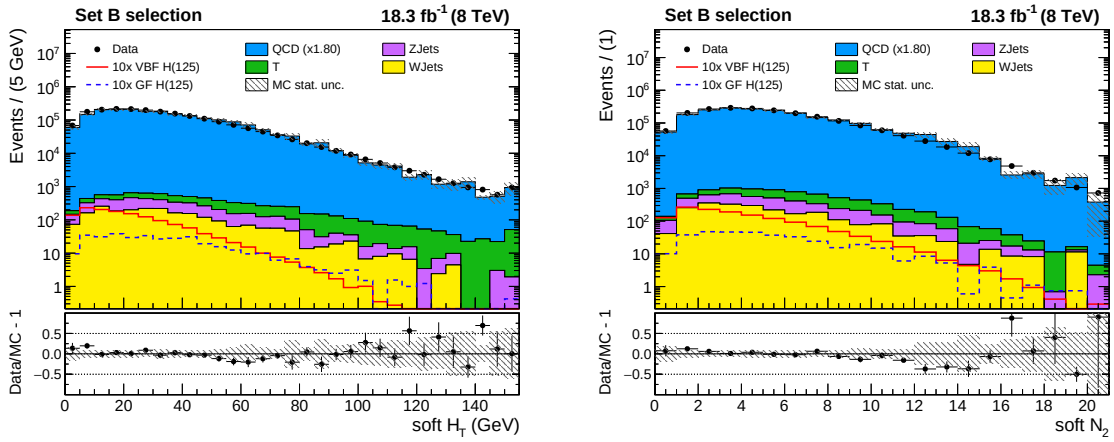


Figure 8.6: Soft track activity momentum sum and multiplicity distributions, for the Set B selection.

8.2 Signal vs. background discrimination

In order to maximise the separation between the VBF $H \to b\bar{b}$ signal and the (mostly QCD) background, it is important to use and combine all discriminating features in an optimal way. A multivariate discriminant is a good solution for this case. The inputs to the discriminant are chosen with two types of merit in mind: the discriminating power of the individual variables should be good, but they should also be as minimally dependent on the b-quark pair invariant mass as possible. In the third part of the analysis, the events will be classified into categories based on the discriminant value, and in each category the mass distribution will be fitted separately (cf. chapter 9). For this to work well, it is important that the distributions in each of said categories are as unaffected as possible by the split into categories. If complete independence between the discriminant and the mass could effectively be maintained, the shape of the distribution would be identical in each category. As it stands, some interdependence cannot be avoided, and a small residual correlation exists; this will be taken into account when performing the fit.

8.2.1 Input variables

The multivariate signal vs. background discriminant uses twelve input variables, all of which have good discriminating power, are minimally correlated, and are minimally dependent on the b-quark pair invariant mass. The variables are:

1. $\mathbf{m_{qq'}}$ – invariant mass of the light-quark pair,
2. $\mathbf{|\Delta\eta_{qq'}|}$ – pseudorapidity separation of the light-quark pair,
3. $\mathbf{|\Delta\phi_{qq'}|}$ – azimuthal angle separation of the light-quark pair,
4. **b-jet$_0$ CSV** – CSV b-tag value of the most b-tagged jet,
5. **b-jet$_1$ CSV** – CSV b-tag value of the second most b-tagged jet,
6. **b-jet$_0$ QGL** – QGL discriminant value of the jet with the highest b-likelihood score,
7. **b-jet$_1$ QGL** – QGL discriminant value of the jet with the 2nd highest b-likelihood score,
8. **b-jet$_2$ QGL** – QGL discriminant value of the jet with the 3rd highest b-likelihood score,
9. **b-jet$_3$ QGL** – QGL discriminant value of the jet with the lowest b-likelihood score,
10. $\mathbf{H_T^{soft}}$ – transverse momentum sum of the additional soft trackjet collection,
11. $\mathbf{N_2^{soft}}$ – multiplicity of the additional soft trackjet collection,
12. $\mathbf{|\cos\theta_{(b\bar{b},qq')}|}$ – angle between the (b_1,b_2) and (q_1,q_2) planes in the centre-of-mass frame of the four leading jets.

The above variables can be divided into five groups: (1-3) cover the typical VBF properties of the light-quark pair, (4-5) have information about b-tagging, (6-9) contain quark-gluon tagging information, (10-11) are a measure of the additional hadronic activity, and (12) is connected to the dynamics of the event production.

As a test during the development of the discriminant, several versions were trained using the same configuration, first using just the first variable group, then progressively adding the other groups, in order to get an idea of how much discriminating power each group adds to the discriminant. A similar test was done for each variable separately, training the discriminant with all variables but the one under study.

8.2.2 Training

The multivariate discriminant is configured with the TMVA toolkit as a BDT (600 trees, boost type=gradient, shrinkage=0.2) [133], and is trained for the Set A and Set B phase spaces separately. Other configurations, such as an artificial neural network, were found to be slower and don't improve the separation. For the signal, a combination of three VBF $\mathrm{H} \to \mathrm{b\bar{b}}$ signal samples with $\mathrm{m_H} = 115$, 125 and 135 GeV is used. For the background, the available amount of simulated QCD events is too small to provide good statistics, so instead a small data sample from the 2012 data taking period is used. The signal contamination in this small data sample is

negligible, and the sample is not reused in the final signal search. The size of the signal samples is 220k for Set A and 64k for Set B. The same amount of events is used for the background samples. Both the signal and the background samples are then split into two equal parts, one for training and one for testing the discriminant.

Figures 8.7 and 8.8 show the linear correlation coefficients between the twelve input variables to the discriminant, for both selections. None of the variables are strongly correlated, except for the $(m_{qq'}, |\Delta\eta_{qq'}|)$ and (b-jet$_0$ CSV, b-jet$_1$ CSV) pairs.

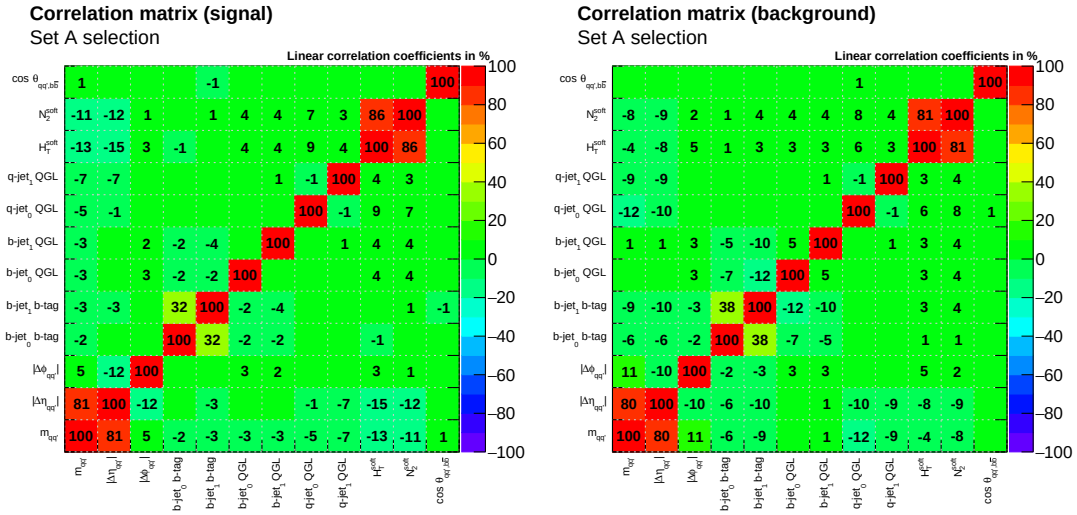


Figure 8.7: Linear correlation coefficients for the input variables to the multivariate discriminant, for Set A.

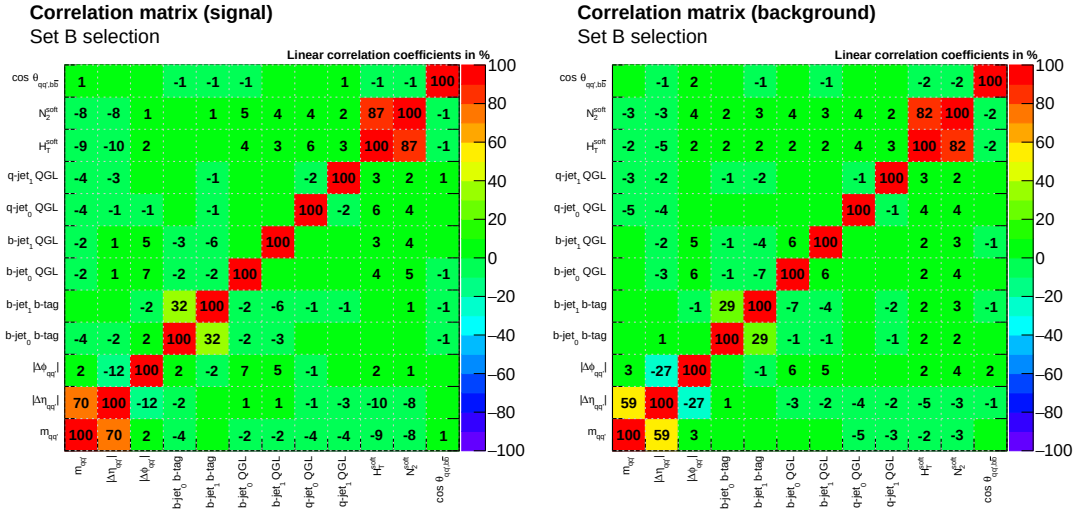


Figure 8.8: Linear correlation coefficients for the input variables to the multivariate discriminant, for Set B.

8.2.3 Performance

The output distribution of the discriminant is shown for signal and background, for both selections, in Fig. 8.9. The agreement between the training and test distributions is good, and there are no indications of overtraining. Figure 8.10 then illustrates the performance of the discriminant, in the form of the background rejection vs. signal efficiency curve[2]. The different curves indicate the different versions of the discriminant, each time trained including one additional group of variables. The final discriminant conforms to the red curve, including all five variable groups. The discriminating power can be quantified by the area under the curves, which is also indicated in Fig. 8.10. For the final discriminant the ROC area is approximately 0.85.

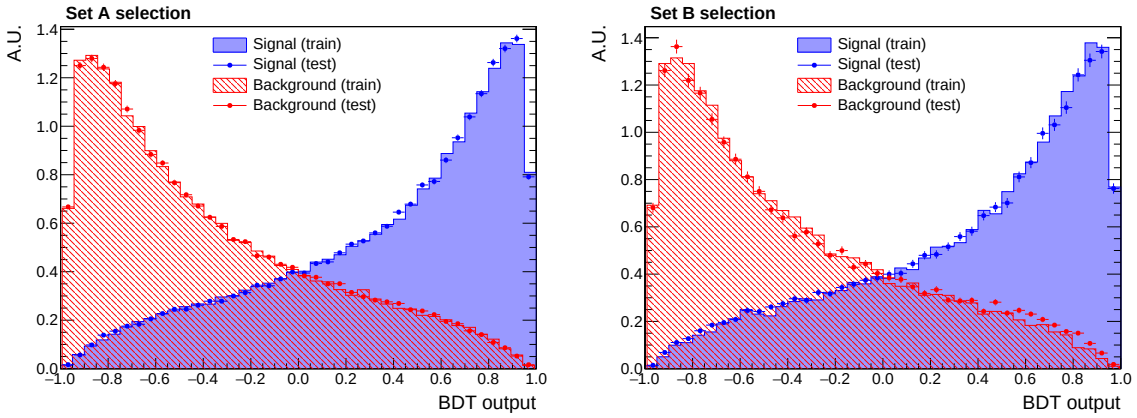


Figure 8.9: Multivariate discriminant output distributions for signal and background, for Set A and Set B.

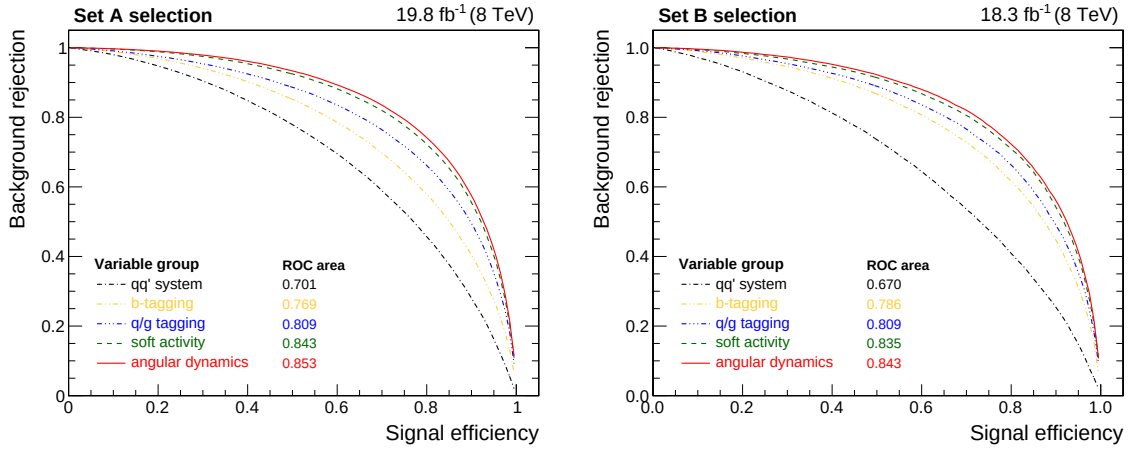


Figure 8.10: Background rejection vs. signal efficiency for the multivariate discriminant, for Set A and Set B. The five curves show the result of the progressive inclusion of the different variable groups. For each step the area under the curve is given as a measure of the discriminant performance.

[2] A common name for this type of curve is receiver operating characteristic or ROC curve.

8.3 Data vs. simulation comparisons

In the last section of this chapter, we show data vs. simulation comparisons for a selection of variables relevant to the VBF $H \to b\bar{b}$ analysis. On one hand this allows us to validate the data, i.e. to confirm that there are no pathologies present in the data caused by for example noise or pile-up. On the other hand, it allows us to check whether the simulated events manage to adequately describe the properties measured in data.

First, the comparisons are made for the standard selections Set A and Set B, validating the data at the level of the events selected for the VBF $H \to b\bar{b}$ search. In addition, comparisons are made for a top control sample (in which QCD events are heavily suppressed), providing a more detailed validation of the simulated events that will effectively be used to perform the analysis.

The QCD component is the largest background component by far. Since the QCD simulation is only done at leading order, one can expect some discrepancies to be present in the data vs. simulation comparisons for the standard selections, in cases where the QCD simulation does not manage to describe the event dynamics well. The QCD simulation however, is not used in the final fit of the b-quark pair invariant mass spectra (cf. chapter 9); the QCD shapes are extracted from data. The validation of the other simulated samples, which are used to provide inputs to the final fit, can be done in a much clearer manner in the top control phase space.

The standard set of comparisons is shown in sections 8.3.1 and 8.3.2, while the additional set can be found in appendix B.

In the comparisons, data points are shown with black markers. Background contributions are stacked, weighted with their respective cross sections, and scaled to the luminosity in data. The LO QCD cross section is multiplied by a k-factor of 1.65 (Set A) or 1.80 (Set B), such that the total number of background events matches the number of events in data. The VBF and GF signal contributions for $m_H = 125$ GeV are overlaid (not stacked), also weighted with their cross sections, and scaled to the luminosity in data. They are scaled with an additional factor ten to improve the visibility. All simulated events have been corrected with pile-up and trigger efficiency scale factors, in order to match the pile-up distribution and trigger efficiency in data.

8.3.1 Set A selection

Figures 8.11-8.19 show the comparisons for the Set A selection. The following groups of variables are included:

- jet transverse momenta (Fig. 8.11)
- jet pseudorapidities (Figs. 8.12 and 8.13)
- qq' system properties (Fig. 8.14)
- $b\bar{b}$ system properties (Fig. 8.15)
- CSV b-tag values (Fig. 8.16)
- b-likelihood scores (Fig. 8.17)
- multivariate discriminants (Fig. 8.18)
- event properties (Fig. 8.19)

For most comparisons, the agreement between data and background simulation is good. Only the transverse momentum distributions in Fig. 8.11 show slightly larger mismodelling. For all four jets, a residual slope can be observed in the ratio plots, in the turn-on of the distributions. This is caused by the difference in trigger efficiency in data and simulation. Trigger efficiency scale factors were applied to correct for this effect, but they were derived as a function of the CSV b-tag value of the most b-tagged jet, and the light-quark pair invariant mass (cf. section 7.3)[3]. This explains why the correction of the transverse momenta distributions is of lower quality.

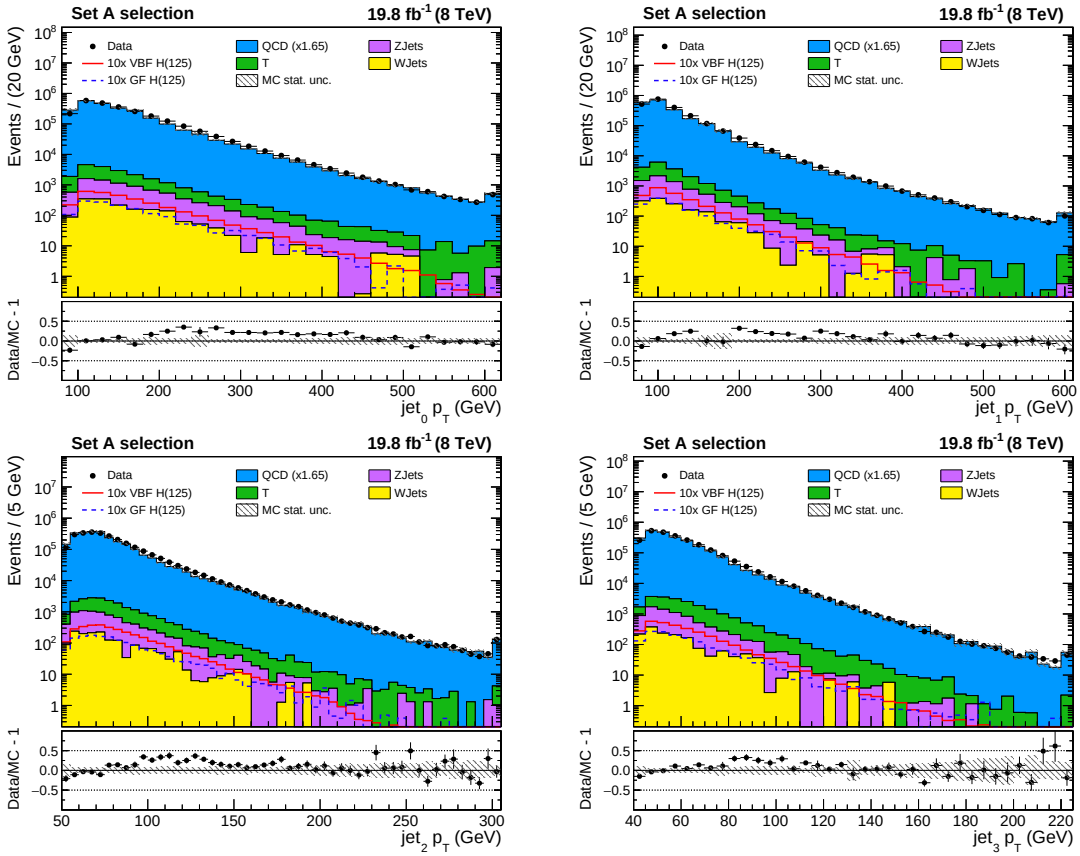


Figure 8.11: Transverse momentum distributions of the jets, ordered by decreasing p_T, for the Set A selection.

[3]The jet transverse momenta are not used as inputs to the multivariate signal vs. background discriminant, while the CSV b-tag value of the most b-tagged jet and the light-quark pair invariant mass are. Therefore, the latter variables are considered more important and they were prioritised when deriving the correction factors.

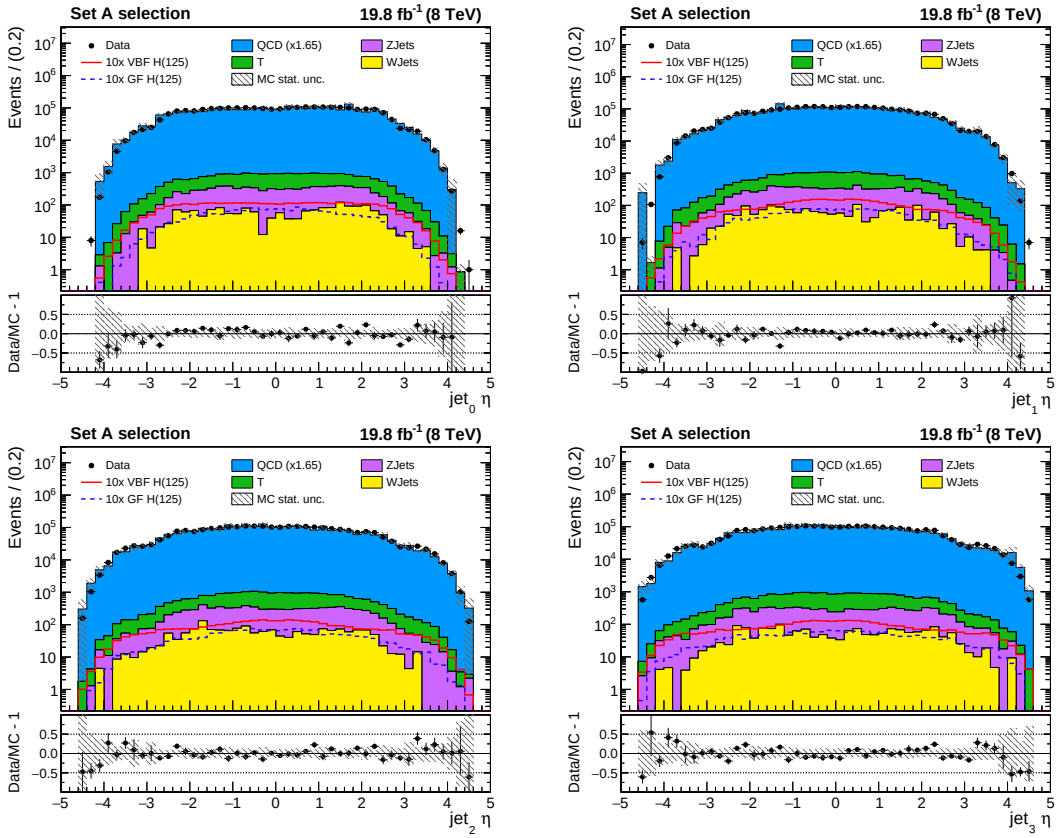


Figure 8.12: Pseudorapidity distributions of the jets, ordered by decreasing p_T, for the Set A selection.

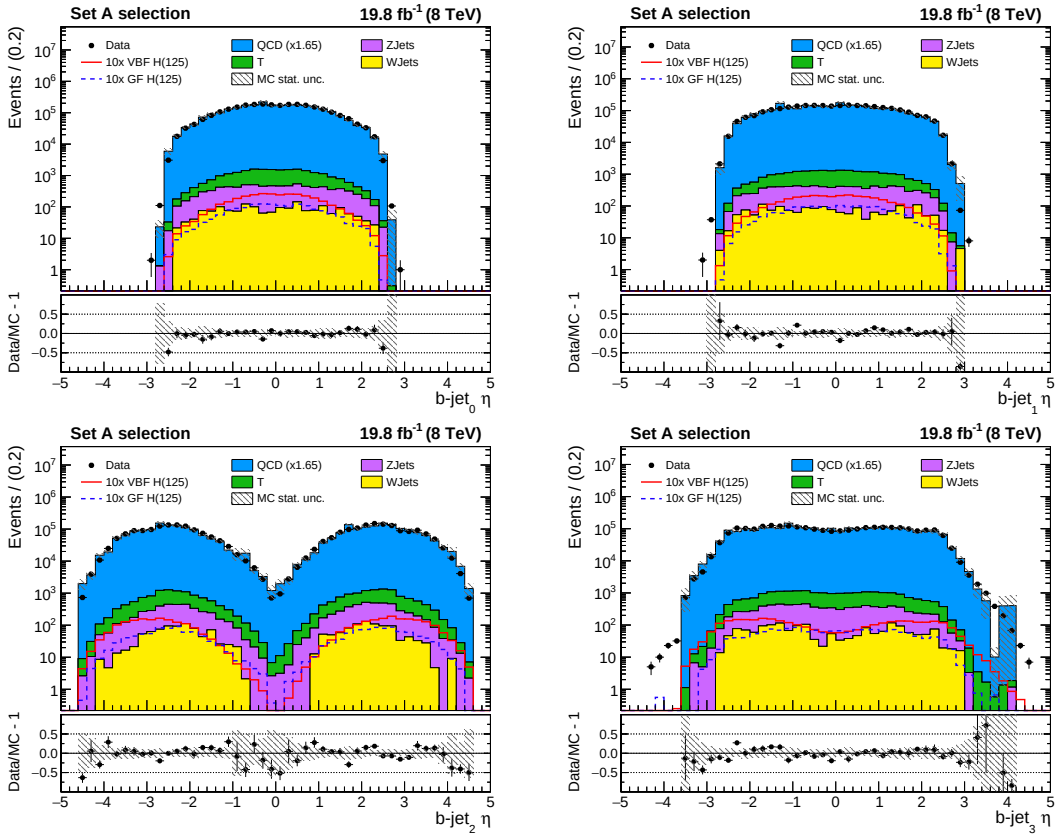


Figure 8.13: Pseudorapidity distributions of the jets, ordered by b-likelihood value, for the Set A selection.

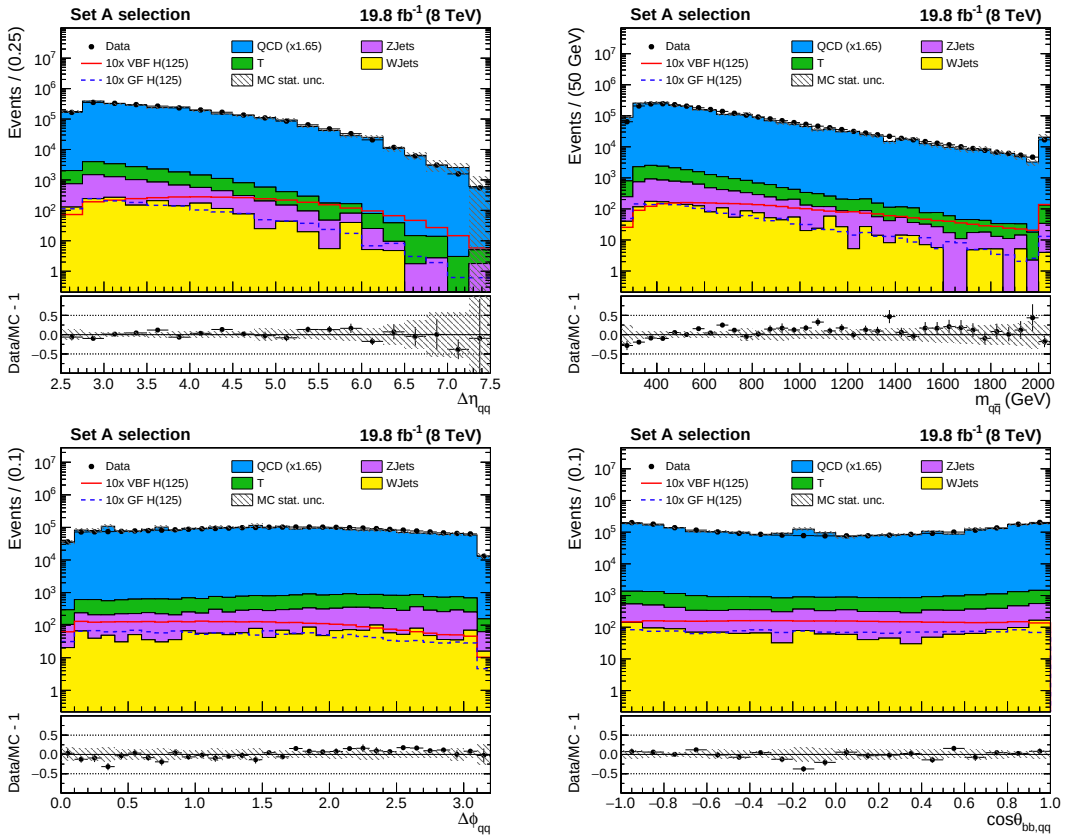


Figure 8.14: Kinematic properties of the qq′ system (pseudorapidity separation, invariant mass, angular separation, and angle with the $b\bar{b}$ system, for the Set A selection.

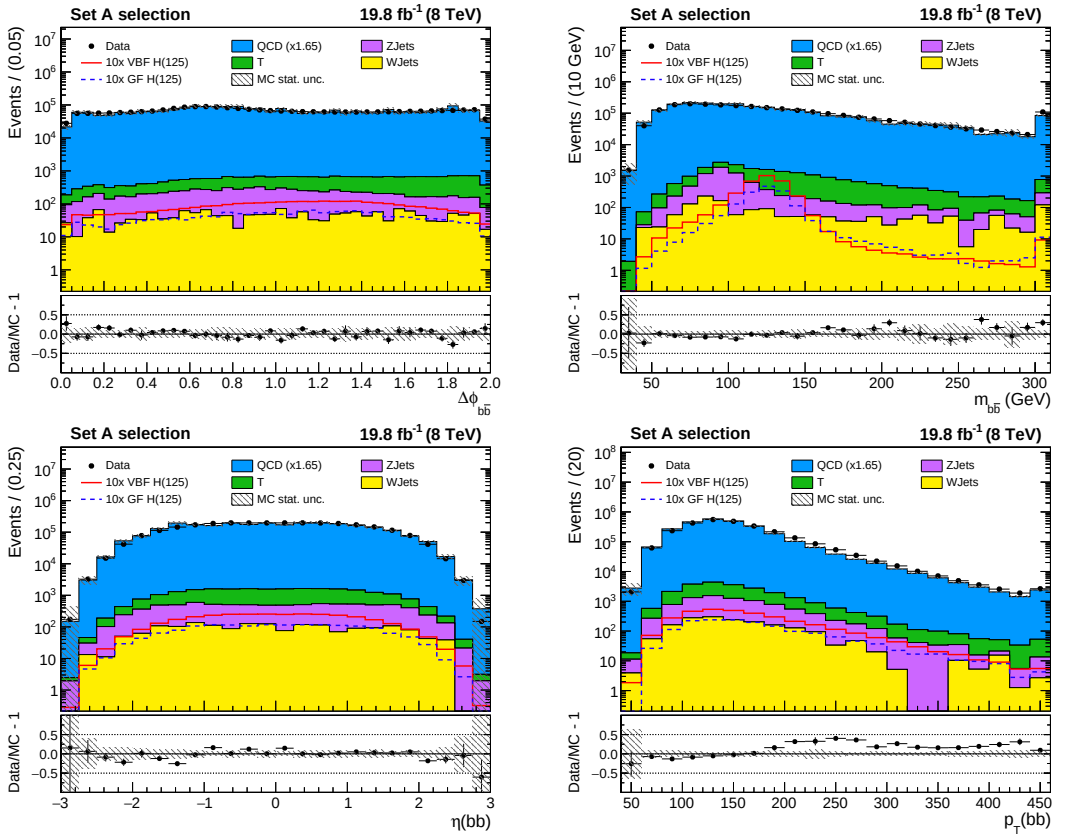


Figure 8.15: Kinematic properties of the $b\bar{b}$ system (angular separation, invariant mass, pseudorapidity, and transverse momentum), for the Set A selection.

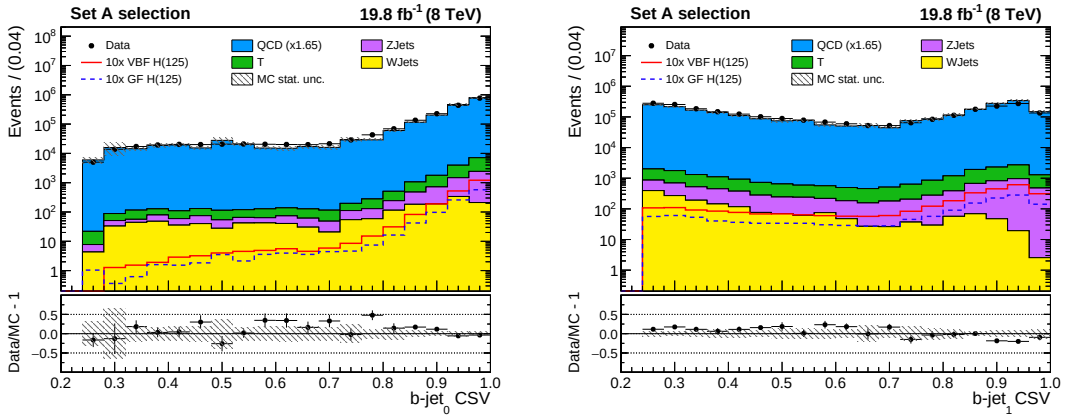


Figure 8.16: CSV b-tag value distributions of the two leading b-jets, ordered by CSV b-tag value, for the Set A selection.

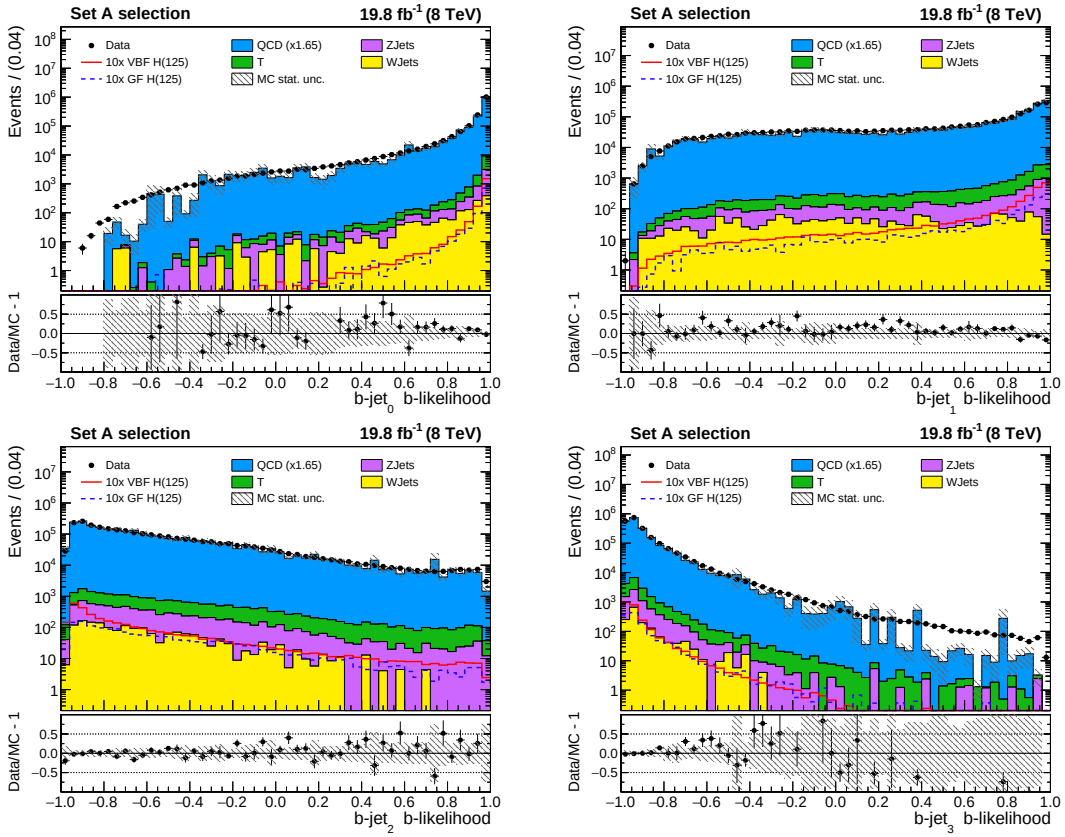


Figure 8.17: Distributions of the b-likelihood discriminant values of the jets, ordered by b-likelihood discriminant value, for the Set A selection.

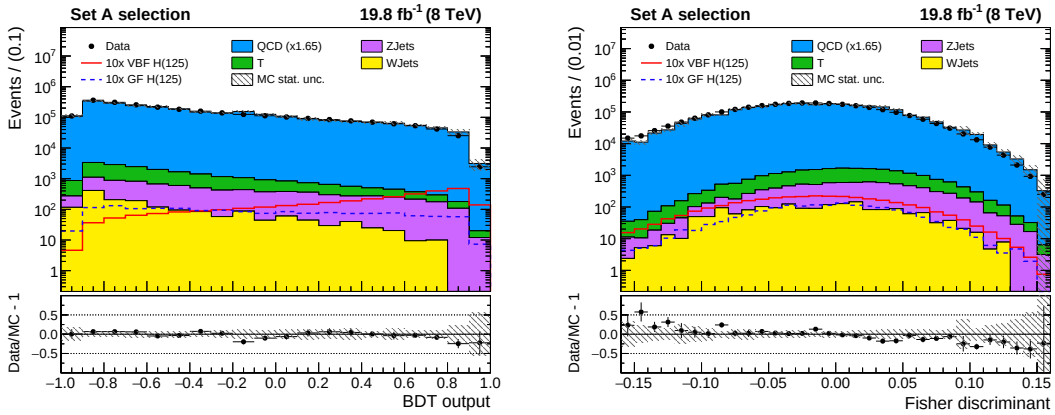


Figure 8.18: Multivariate discriminant output values (BDT and Fisher discriminant), for the Set A selection.

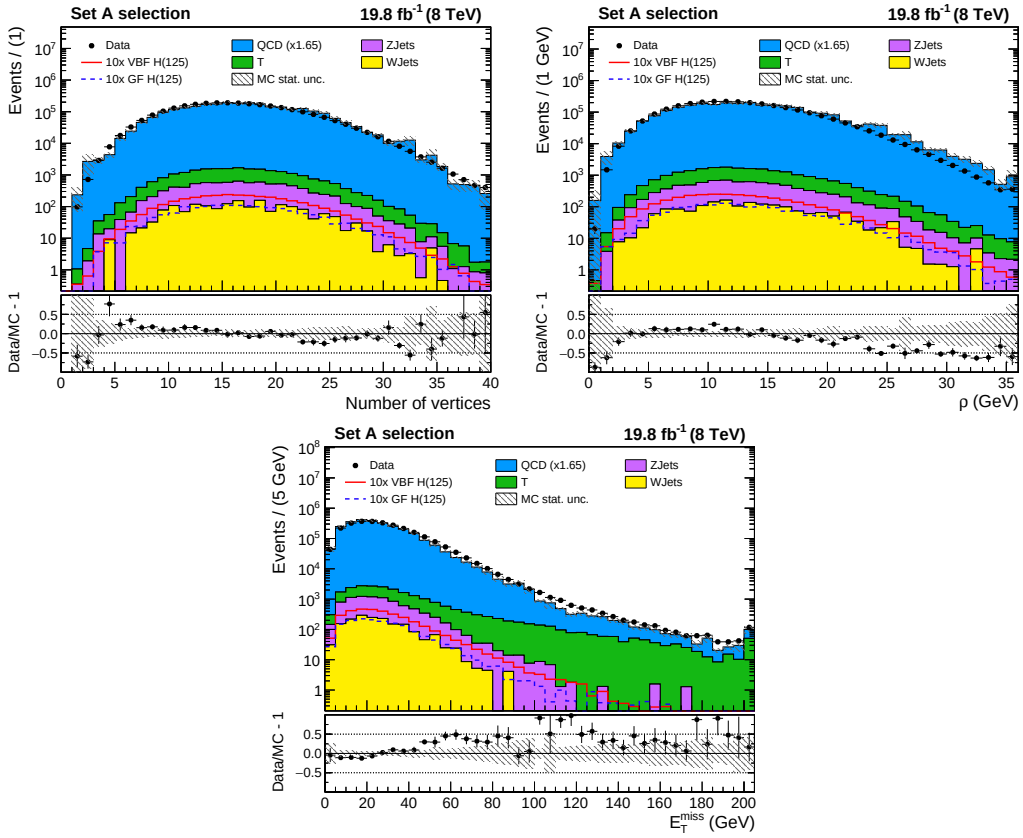


Figure 8.19: Number of reconstructed vertices, event p_T density (ρ) and PF missing transverse energy distributions, for the Set A selection.

8.3.2 Set B selection

Figures 8.20-8.28 show the comparisons for the Set B selection. The following groups of variables are included:

- jet transverse momenta (Fig. 8.20)
- jet pseudorapidities (Figs. 8.21 and 8.22)
- qq′ system properties (Fig. 8.23)
- $b\bar{b}$ system properties (Fig. 8.24)
- CSV b-tag values (Fig. 8.25)
- b-likelihood scores (Fig. 8.26)
- multivariate discriminants (Fig. 8.27)
- event properties (Fig. 8.28)

The same conclusions are true for the Set B selection comparisons as for the Set A selection ones. The overall agreement between data vs. background simulation is good, except in the jet transverse momenta distributions.

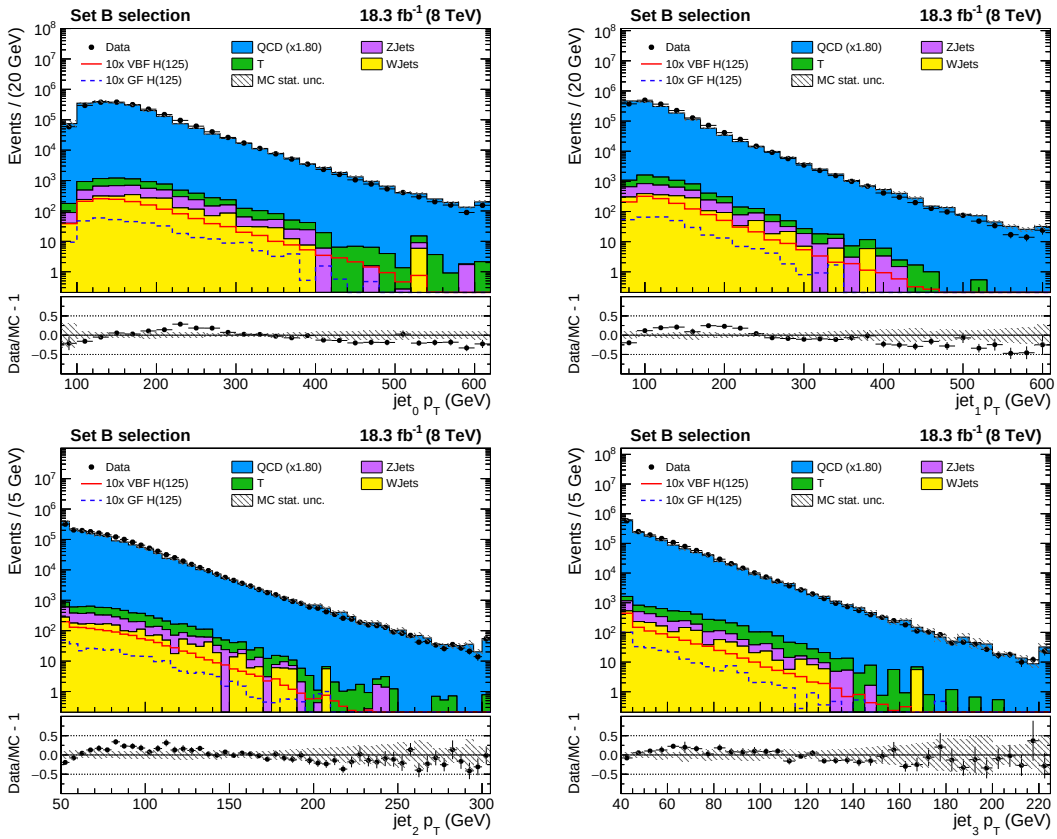


Figure 8.20: Transverse momentum distributions of the jets, ordered by decreasing p_T, for the Set B selection.

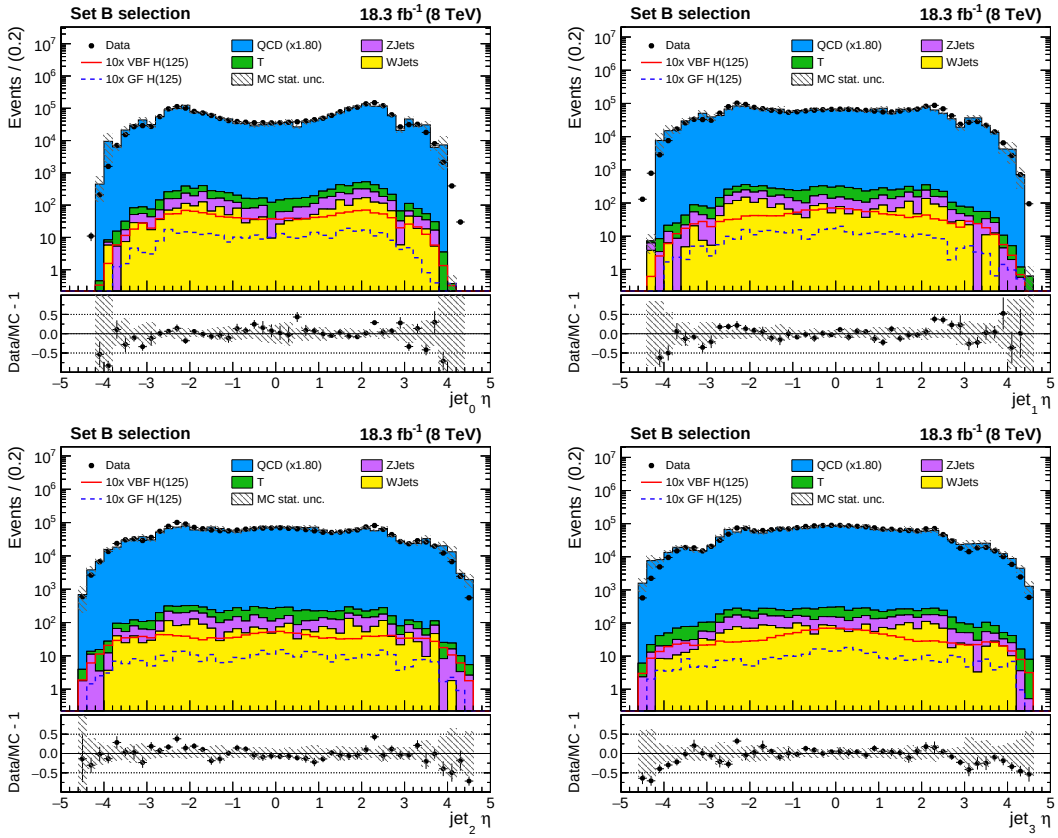


Figure 8.21: Pseudorapidity distributions of the jets, ordered by decreasing p_T, for the Set B selection.

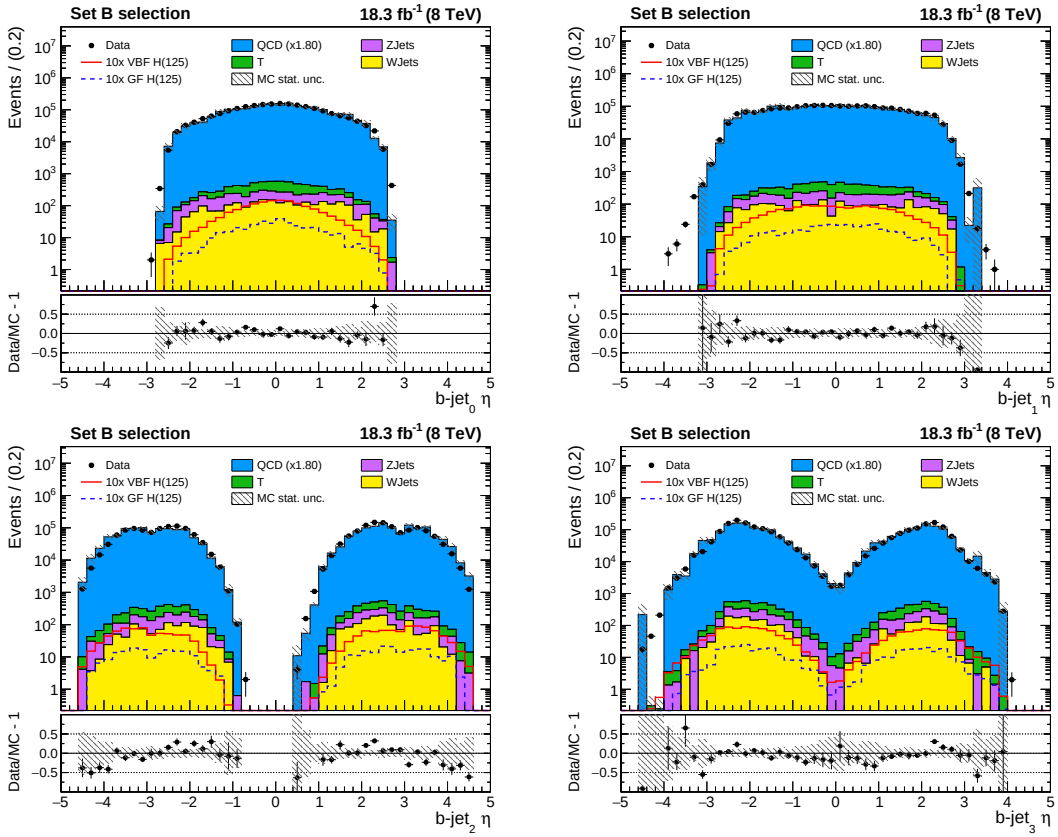


Figure 8.22: Pseudorapidity distributions of the jets, ordered by b-likelihood value, for the Set B selection.

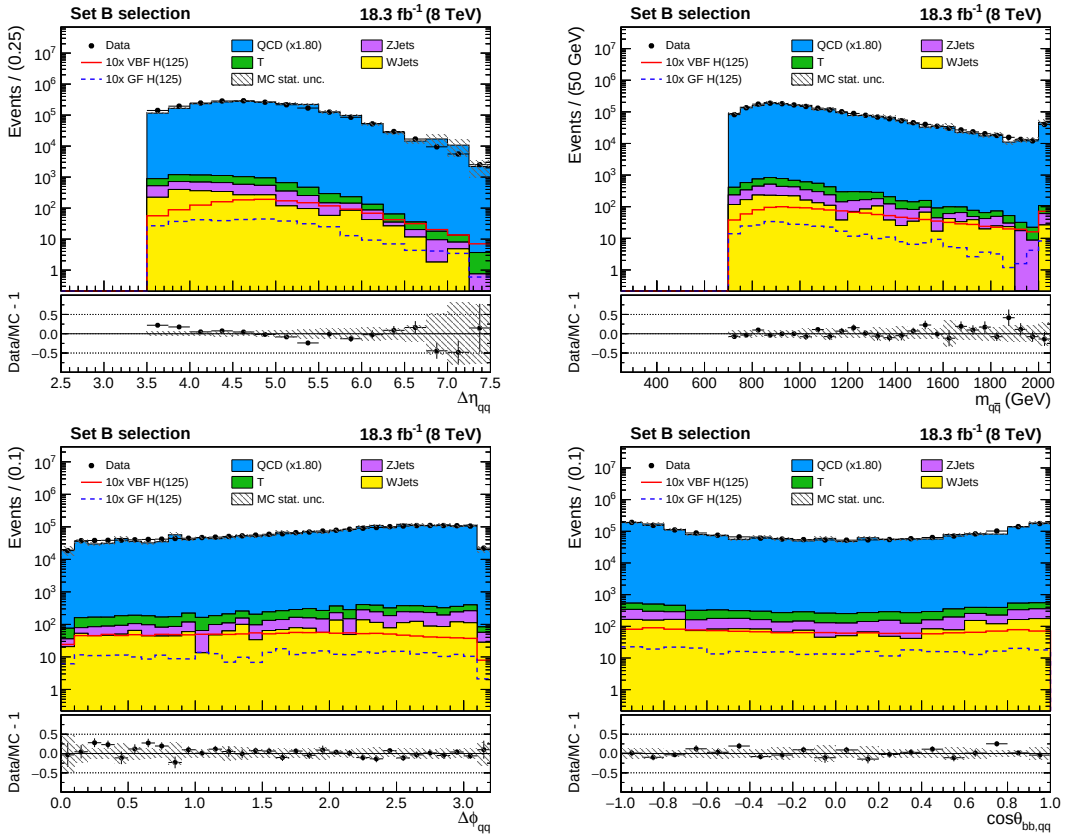


Figure 8.23: Kinematic properties of the qq′ system (pseudorapidity separation, invariant mass, angular separation, and angle with the $b\bar{b}$ system, for the Set B selection.

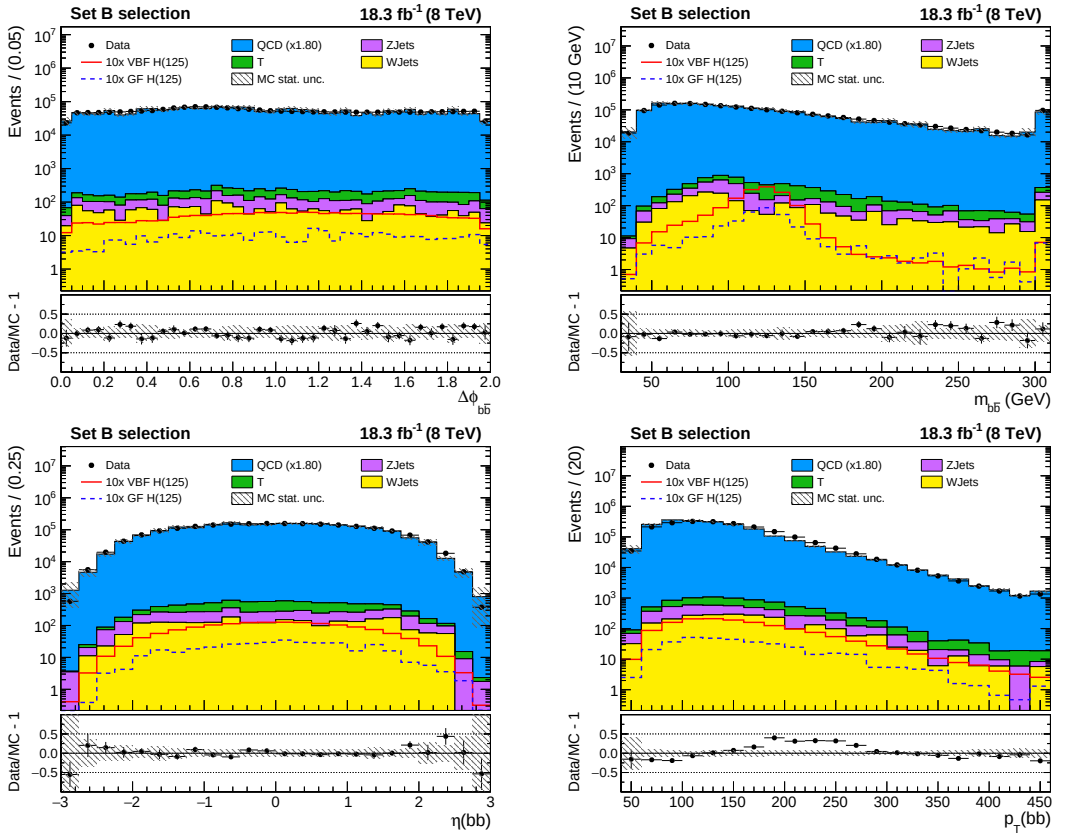


Figure 8.24: Kinematic properties of the $b\bar{b}$ system (angular separation, invariant mass, pseudorapidity, and transverse momentum), for the Set B selection.

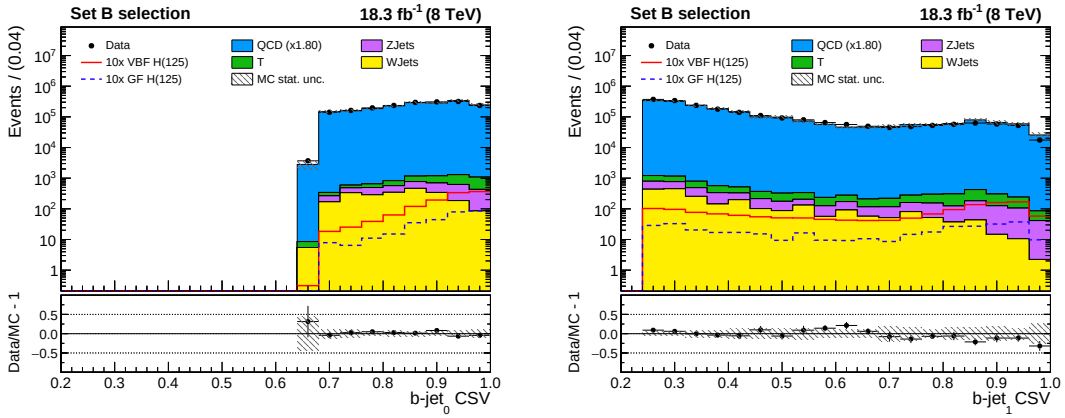


Figure 8.25: CSV b-tag value distributions of the two leading b-jets, ordered by CSV b-tag value, for the Set B selection.

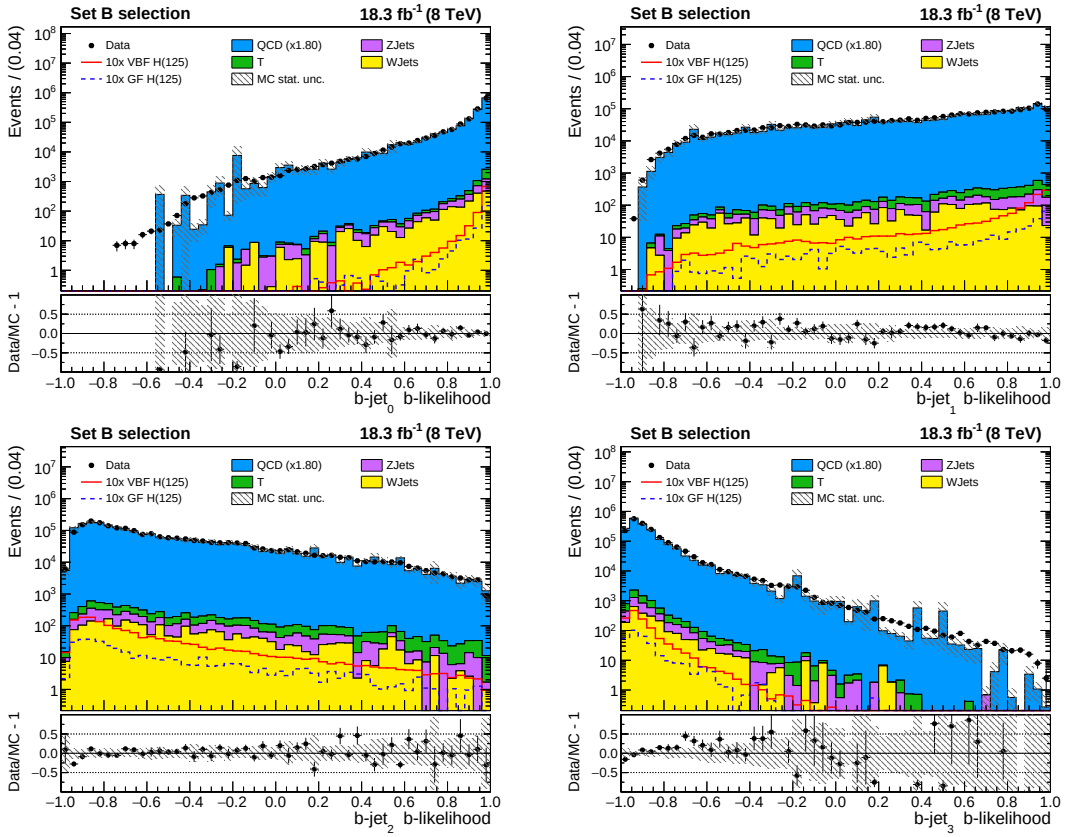


Figure 8.26: Distributions of the b-likelihood discriminant values of the jets, ordered by b-likelihood discriminant value, for the Set B selection.

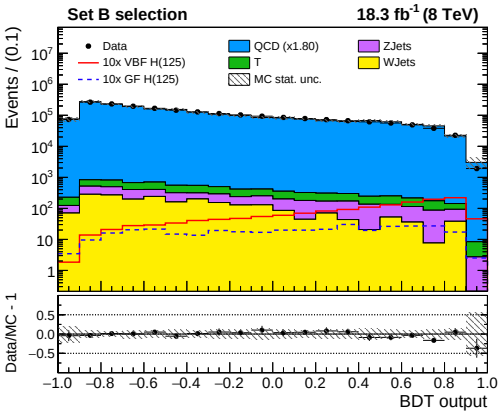


Figure 8.27: Multivariate discriminant output value (BDT), for the Set B selection.

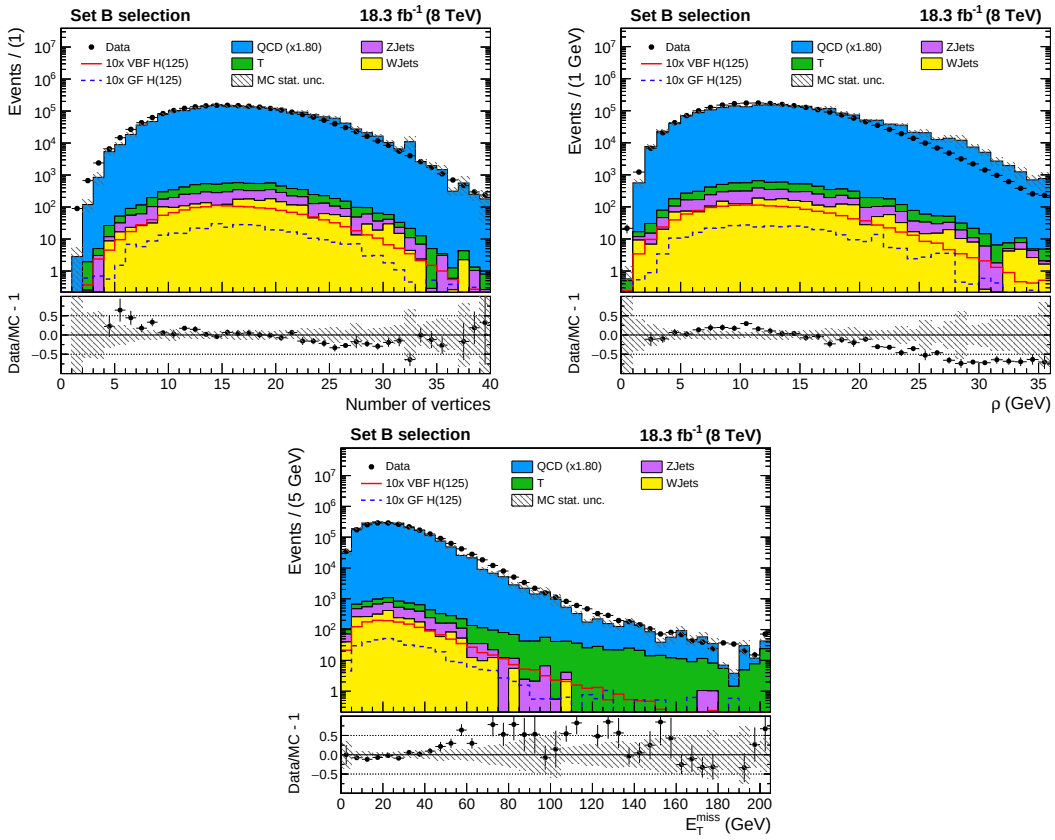


Figure 8.28: Number of reconstructed vertices, event p_T density (ρ) and PF missing transverse energy distributions, for the Set B selection.

Chapter 9

Fit of the b-quark pair invariant mass distribution

The third and final step in the VBF $H \to b\bar{b}$ analysis consists of the fit of the b-quark pair invariant mass distribution for the selected datasets, which I worked on extensively during the second half of my PhD. Given the magnitude of the QCD multijet background, fitting a small dijet resonance mass peak on top of it is highly nontrivial. The limited precision of the QCD multijet background simulation, as well as the moderate invariant mass resolution ($\mathcal{O}(10\%)$), further complicate the situation. A specific fit strategy is devised to tackle this challenge; it is sketched in section 9.1. The statistical procedure followed in the fit is detailed in section 9.2. Before applying this fit method to the selected datasets for the VBF $H \to b\bar{b}$ analysis and searching for the Higgs boson resonance, it is validated by performing a fit of a known resonance, namely that of the Z boson. This validation is discussed in section 9.3. The discussion of the actual fit of the Higgs boson resonance then follows in section 9.4, and section 9.5 covers the systematic uncertainties taken into account in this fit. The final results are presented in chapter 10.

9.1 Fit strategy

The strategy used to fit the Z and Higgs boson resonance in the b-quark pair invariant mass distribution can be split up into several steps:

Definition of event categories Based on the output of the multivariant signal vs. background discriminant, the events are divided into categories. The lower categories will be background dominated, while the higher categories will be signal dominated. The boundaries of the categories are defined such that the expected significance of the fit result is maximised, while at the same time taking care that the amount of statistics in each category remains sufficient. Templates for different signal and background components will be derived per category.

Template for the QCD multijet background The first challenge is to derive a template for the large QCD multijet background. As the precision of the QCD simulation is too low, the template cannot be taken from simulation. The solution is to derive QCD template from data. Extracting different templates for each event category was found to be suboptimal: the amount of statistics drops in the higher event categories, and the quality of the QCD templates would deteriorate. Since the multivariate signal vs. background discriminant was constructed

to be minimally correlated to the b-quark pair invariant mass, the variation in the shape of the invariant mass distribution between the different event categories, is however expected to be small. This allows us to use just one QCD template, taken from the lowest, most background-like category, the reference category. The function used to describe the QCD shape is a Bernstein polynomial [134][1]. The residual shape difference between the categories is then taken into account using linear or quadratic transfer functions between the reference category and any given other category. A similar technique was used in the ATLAS $Z \to b\bar{b}$ measurement [135].

Signal and background templates Templates are also derived, per category, for the top background component, the Z/W+jets background component, and the $H \to b\bar{b}$ signal components. The top component consists of the sum of all $t\bar{t}$ and single-top background components, and is described using a broad Gaussian. The Z/W+jets component is the combination of the Z+jets and W+jets background components, and is described using a Crystal Ball function (a Gaussian core plus a power law tail) [136] for the resonance peak, and a Bernstein polynomial for the combinatorial background contribution (misidentification of the b-quark pair). For the signal, the main component is the VBF $H \to b\bar{b}$ signal. It was found however that there is a secondary contribution to the $H \to b\bar{b}$ signal: the GF $H \to b\bar{b}$ signal (cf. section 9.4.4). The total $H \to b\bar{b}$ signal is described using the same Crystal Ball plus polynomial shape as used for the Z/W+jets component.

Simultaneous binned maximum likelihood fit of all event categories The final global fit is made using the CMS combine toolkit [137], based on the ROOT RooFit and RooStats packages [138]. It is set up as a binned maximum likelihood fit, fitting all event categories simultaneously. The statistical procedures used to interpret the data, implemented in the previously mentioned toolkit, are discussed in the next section.

[1] Bernstein polynomials are a class of polynomials based on the Bernstein basis polynomials. The basis polynomials are non-negative and don't oscillate. The linear combinations are smooth and very adaptable, and as such a good option to describe the smooth QCD background distribution.

9.2 Statistical procedure

The statistical procedures used to interpret data in the context of Higgs boson searches and combinations, are governed by the LHC Higgs combination group. They follow the methods as described in [139–142], which have been implemented in the ROOT RooFit and RooStats packages [138]. In this section we briefly cover the concepts used to obtain and present the results of this analysis.

To interpret a dataset with measurements x, one considers models $f(x|\boldsymbol{\alpha})$, which define the probability of measuring x under a hypothesis $\boldsymbol{\alpha}$. Hypothesis $\boldsymbol{\alpha} = (\mu, \boldsymbol{\theta})$ includes theoretical parameters μ, which one might like to measure, as well as nuisance parameters $\boldsymbol{\theta}$, which cover the model parameters and the systematic uncertainties on the predictions for signal and background. In the case of a search for new physics, such as a Higgs boson search, parameter of interest μ is often taken as the ratio of the measured cross section over the predicted standard model cross section: $\mu = \sigma/\sigma_{\mathrm{SM}}$. The nominal signal hypothesis then is $\mu = 1$, while the null-hypothesis is $\mu = 0$. The nuisance parameters are usually taken to follow either a log-normal distribution (for multiplicative uncertainties on parameters which are always positive, e.g. normalisations), a log-uniform distribution (for parameters allowed to float between two boundaries), or a Gaussian distribution. Examples from the VBF $\mathrm{H} \to \mathrm{b\bar{b}}$ analysis are the QCD background component normalisation (log-uniform), the Z/W+jets and top normalisations (log-normal), and the template parameters (Gaussian). The nuisance parameters for systematic uncertainties are inserted as multiplicative factors (log-normal).

A likelihood for parameters μ and $\boldsymbol{\theta}$ of hypothesis $\boldsymbol{\alpha}$ can be constructed using the probability density functions $f(x|\boldsymbol{\alpha})$. As analytical shapes for f are often not available, they are estimated from simulation in the form of histograms, and a binned likelihood is used. Counts in histograms, i.e. sums of contributions of all signal and background processes, are distributed according to Poissonian distributions. For the likelihood one can therefore write:

$$\mathcal{L}(x|\mu, \boldsymbol{\theta}) = \mathcal{L}(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N_{bins}} \text{Poissonian}\left(n_i | E_i(\mu, \boldsymbol{\theta})\right), \tag{9.1}$$

where n_i are the measurements in each bin, E_i the expectations given hypothesis $\boldsymbol{\alpha}$, and the product is taken over all histogram bins.

Estimates of parameters μ and $\boldsymbol{\theta}$ are obtained by maximising the likelihood function, hence the name *binned maximum likelihood fit*. In practice, this is usually done by minimising the negative logarithm of the likelihood: $-\ln \mathcal{L}(\mu, \boldsymbol{\theta})$. The parameter values that maximise the likelihood are called maximum likelihood estimators, and are denoted as $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$. Similarly, conditional maximum likelihood estimators $\hat{\hat{\boldsymbol{\theta}}}(\mu)$ are the parameter values that maximise the likelihood given a fixed value of μ. Maximum likelihood estimators have the advantage of being asymptotically unbiased, i.e. they converge to their true value. In addition, the parameter variance can be obtained from the covariance matrix of the fit, which includes the effects of all nuisance parameters in a correlated manner.

Quantitative inferences about hypothesis $\boldsymbol{\alpha}$ can subsequently be made using hypothesis testing. First, the profile likelihood ratio $\lambda(\mu)$ is defined:

$$\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}})}, \tag{9.2}$$

which is bounded between 0 and 1. Using the profile likelihood ratio, a test statistic t_μ is constructed:

$$t_\mu = -2 \ln \lambda(\mu). \tag{9.3}$$

A value of $\lambda = 1$ corresponds to good agreement between the data and hypothesis $\boldsymbol{\alpha}$. A lower value of λ (or a higher value of t_μ), indicates an increased level of disagreement.

The level of disagreement between the data and hypothesis $\boldsymbol{\alpha}$ is usually quantified by the p-value:

$$p_{\mu,obs} = \int_{t_{\mu,obs}}^{\infty} f(t_\mu|\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu_{obs}))\, dt_\mu, \tag{9.4}$$

where f is the sampling distribution of t_μ [2]. The p-value indicates the probability of making an observation, under the same conditions, that is less likely than the observation made in data. For a good hypothesis (small disagreement between the data and the hypothesis) the p-value will be high, for a bad hypothesis (large disagreement) it will be low.

A matching median expected p-value can be computed, quantifying the median expected probability to reject the hypothesis $\boldsymbol{\alpha}$ if an alternative hypothesis $\boldsymbol{\alpha}'$ is true:

$$p_{\mu,exp} = \int_{t_{\mu',exp}}^{\infty} f(t_\mu|\mu', \hat{\hat{\boldsymbol{\theta}}}(\mu'_{obs}))\, dt_\mu. \tag{9.5}$$

Any p-value can also be expressed as an equivalent significance, defined such that a Gaussian distributed value found Z standard deviations above its mean value, has an upper tail probability equal to the p-value:

$$Z = \phi^{-1}(1-p) = \sqrt{2}\,\mathrm{erf}^{-1}(1-2p), \tag{9.6}$$

with ϕ^{-1} the quantile or inverse cumulative distribution of the Gaussian.

For a discovery in a signal search, the general rule in particle physics is to require a rejection of the background hypothesis with a significance of at least five standard deviations, which corresponds to a p-value of $p \leq 2.87 \cdot 10^{-7}$, or at least 99.99994% confidence level. This is

[2] If this distribution needed to be computed using Monte Carlo simulation for many different scenarios, it could become rather CPU intensive; luckily, it can be approximated [140].

equivalent to saying that the probability for the background to fluctuate up to the observed excess is smaller than $2.87 \cdot 10^{-7}$. Conversely, to rule out a signal hypothesis, 95% confidence level is sufficient, which corresponds to a p-value of $p \leq 0.05$, or a significance $Z \geq 1.64$. In this case there is a 5% probability that one would falsely exclude a real signal.

To cover different use cases, alternative test statistics can be derived by making small amendments to the definition of λ, such that unphysical estimates of μ cannot contribute evidence for the rejection of the hypothesis. For the study of a μ parameter which is bounded at zero, an alternative $\tilde{t}_\mu$ can be defined, by setting $\hat{\mu} = 0$ for $\hat{\mu} < 0$. To set upper limits on the value of μ (indicating the largest value of μ which is not rejected), an alternative test statistic q_μ can be used, which sets $\hat{\mu} = \mu$ for $\hat{\mu} > \mu$. An additional bound at zero can be included to define $\tilde{q}_\mu$. The p-values and significances for the different test statistics are derived in the same manner as they were for t_μ.

In this Higgs boson signal search, upper limits on the signal strength are computed using the modified Frequentist $\mathrm{CL_S}$ method [142]. The limits are derived at 95% confidence level, $\mathrm{CL} = 1 - p_\mu = 0.95$. They are based on the $\tilde{q}_\mu$ test statistic, and are defined as:

$$\mathrm{CL_S} = \frac{p_\mu}{1 - p_b} \leq 0.05, \tag{9.7}$$

where p_μ and p_b are derived from $\tilde{q}_\mu$, as the p-values following the nominal and the null hypothesis respectively (in a signal search, the null hypothesis is equal to the background-only hypothesis). The ratio comes down to a reweighting of the signal plus background CL with the background CL, and is more conservative and less susceptible to fluctuations. One and two standard deviation uncertainty bands can be derived around the limit, following the distribution of the test statistic, which for $\tilde{q}_\mu$ is a χ^2-distribution.

In section 9.3.9 and chapter 10, upper limits on the production cross section, p-values and significances, as well as best-fit values will be presented for the Z and Higgs boson global fits.

9.3 Fit of the Z boson resonance

9.3.1 Introduction

The Z+jets background component ($Z \to b\bar{b}$) can, in the same phase space as selected for the Higgs boson search, be studied as though it were signal. A fit can be performed for the known Z boson dijet resonance. The gain from performing this fit is twofold: the fit procedure is validated, and a measurement is made of the b-quark invariant mass distribution scale and resolution (peak position and width).

The event selection used for the Z boson fit is kept as close as possible to the Set A selection. This is suboptimal for the measurement of the Z boson resonance, but required for the validation of the fit procedure in the Higgs boson search phase space. The final selection criteria are identical to the Set A selection, except for one tighter b-jet requirement: one CSV medium and one CSV loose as opposed to two CSV loose in the Set A selection. The Z selection is summarised in Table 9.1.

Table 9.1: Summary of Z selection requirements.

	Z selection
trigger	OR of Set A triggers
jet p_T	$p_T^{(0,1,2,3)} > 80, 70, 50, 40$ GeV
b-tagging	1x CSVM (> 0.679) and 1x CSVL (> 0.244)
interpretation	Set A b-likelihood
VBF topology	$m_{qq'} > 250$, $\lvert\Delta\eta_{qq'}\rvert > 2.5$
QCD suppression	$\Delta\phi_{b\bar{b}} < 2.0$
vetoes	lepton veto

9.3.2 Multivariate discriminant

A separate multivariate signal vs. background is trained for the Z selection. Again making sure to minimise the correlation with the b-quark pair invariant mass, the following variables are used:

- **(i)** the light quark pair pseudorapidity separation, $|\Delta\eta_{q\bar{q}}|$
- **(ii)** the b-quark pair pseudorapidity, $|\eta_{b\bar{b}}|$
- **(iii)** the CSV b-tag value of the most b-tagged jet, CSV_0
- **(iv-vi)** the quark-gluon likelihood of all four jets.

The discriminant is trained using the Z+jets sample for the signal, and a small data sample for the background. The size of the Z+jets sample is insufficient to allow training a non-linear multivariate discriminant. Instead a linear multivariate Fisher discriminant is trained [133]. Since however the correlation between the input variables is found to be small, and never larger than $\mathcal{O}(10\%)$, no significant gain would be obtained from the use of a non-linear discriminant.

The Fisher discriminant output distribution is shown for the Z selection background and signal components in Fig. 9.1.

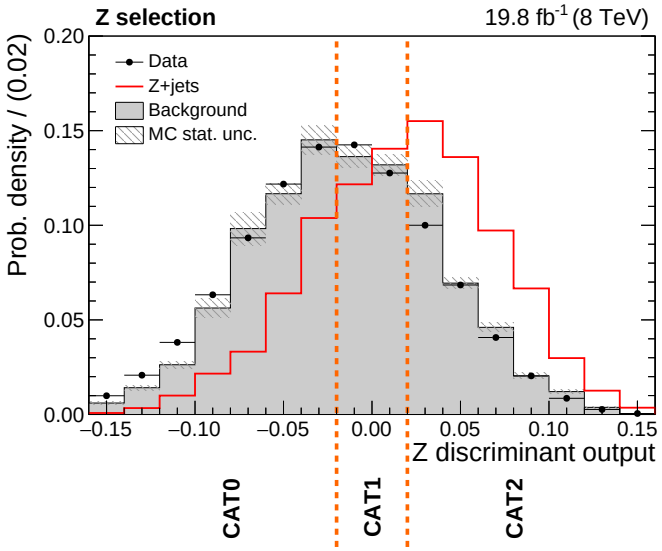


Figure 9.1: Fisher discriminant output distribution for the Z selection, with definition of event categories.

9.3.3 Categories

The selected events are divided into categories based on the multivariate discriminant output distribution. Three categories are defined for the Z selection. The category boundaries were optimised to maximise the expected significance of the fit result, while allowing for enough statistics in each individual category to derive templates for the fit. The boundaries are given in Table 9.2, and shown overlaid on the discriminant distribution in Fig. 9.1. Category 0 is the reference category, which is background-dominated.

Table 9.2: Summary of Z selection category boundaries.

Category	Discriminant range
0	$\text{FD} \leq -0.02$
1	$-0.02 < \text{FD} \leq 0.02$
2	$0.02 < \text{FD}$

9.3.4 QCD template and transfer functions

In the development of the fit strategy, the presumption was that the correlation between the b-quark pair invariant mass and the multivariate discriminant output would be small, such that the QCD template could be derived once in the background-dominated reference category, for use in all categories. Figure 9.2 shows the average b-quark pair invariant mass, in the Z fit mass window [60,170] GeV, as a function of the discriminant output, for data and QCD multijet background. The observed dependence is indeed small.

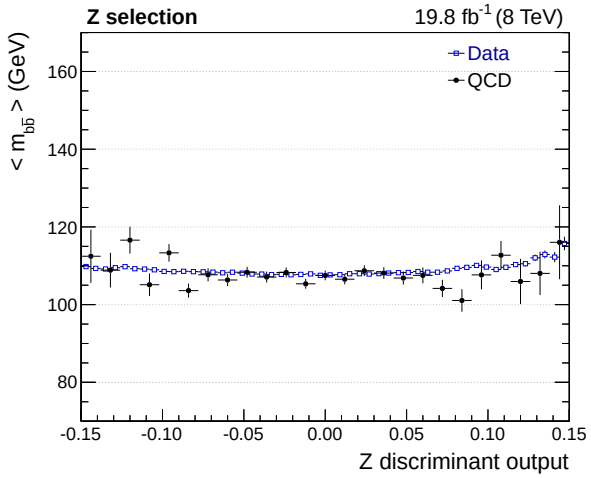


Figure 9.2: Profile histogram of the b-quark pair invariant mass as a function of the BDT output value, for the Z selection.

The QCD template is derived from data using an 8th order Bernstein polynomial[3] in CAT0 and used for all categories. The residual correlation, or the shape change from CAT0 to the other categories, is accounted for by using linear, multiplicative transfer functions. Figure 9.3 shows the shape of the b-quark pair invariant mass distribution for data, per event category, blinded in the peak region. Transfer functions are derived as the ratio of the distribution for a given category CATi to the distribution for reference category CAT0. The transfer function for CAT0 is 1 by construction. For each transfer function systematic uncertainties are assigned such that all local variations are covered. The transfer functions are shown in Fig. 9.4.

9.3.5 Background templates

Next to the QCD templates, also templates for the top background are required. They are derived from simulation, per event category, as broad Gaussians. The obtained templates shown in Fig. 9.5 The event yields of the different samples are given per category in Table 9.3.

[3]The order of the Bernstein polynomial is determined by a bias study. This study considers various alternative QCD models and evaluates which functions allow to fit the spectrum without introducing a bias in a possible signal measurement.

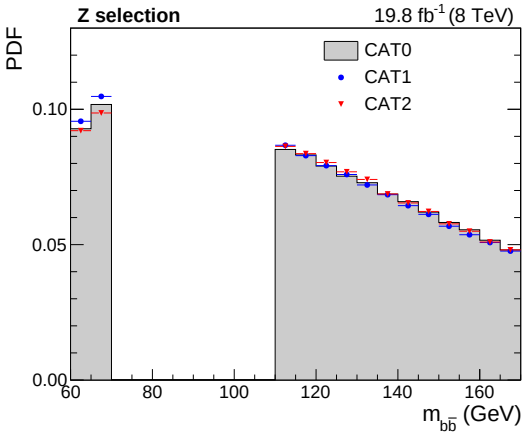


Figure 9.3: Shape comparison of the b-quark pair invariant mass distribution for the different categories of the Z selection.

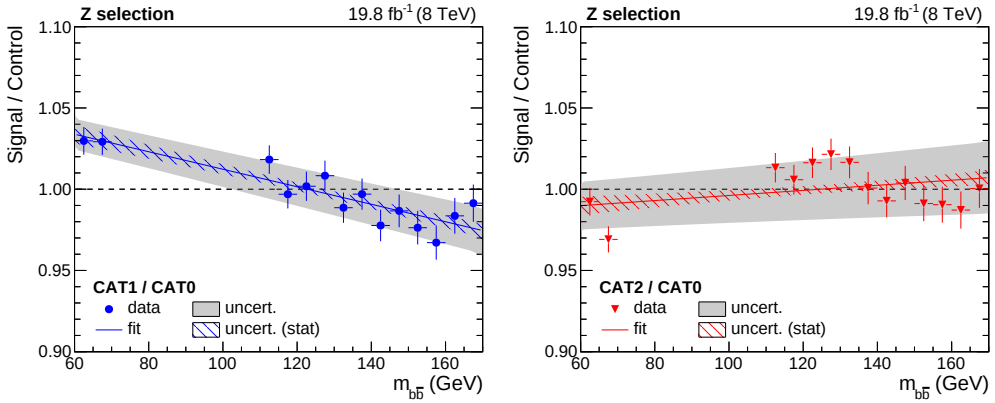


Figure 9.4: Linear transfer functions for the b-quark pair invariant mass distribution between the reference category (CAT0) and the other categories (CAT1-2) of the Z selection, derived from data.

Table 9.3: Sample yields for the various event categories of the Z selection, in the [60,170] GeV mass window.

	CAT0 FD $\leq$ -0.02	CAT1 -0.02 $<$ FD $\leq$ 0.02	CAT2 0.02 $<$ FD
Data	659873	374797	342931
Z+jets	1374	1467	2783
$t\bar{t}$	2124	1821	2327
Single top	657	569	812

9.3.6 Signal templates

The templates for the signal are also derived from simulation, using a combination of a Crystal Ball function (Gaussian core plus power law tail) for the resonance peak and a Bernstein polynomial for the combinatorial background. The obtained templates are shown in Fig. 9.6. The event yields for the Z+jets signal sample are included in Table 9.3.

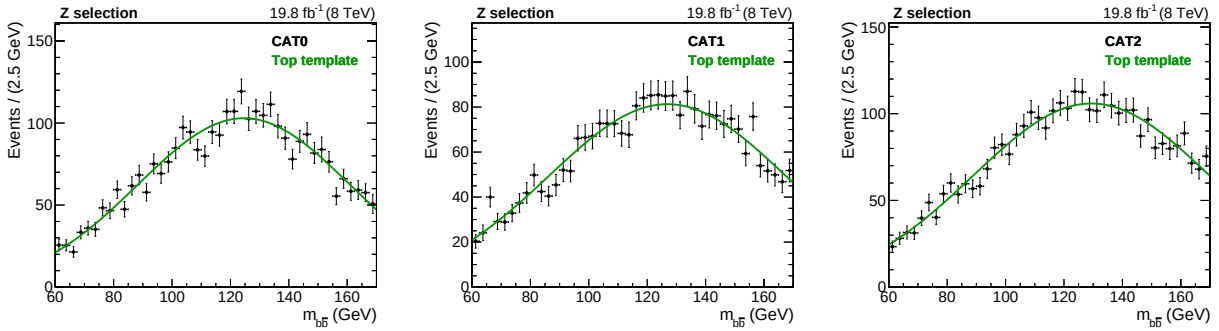


Figure 9.5: Templates for the b-quark pair invariant mass distribution for the Z selection, for the top ($t\bar{t}$ + single-top) background component.

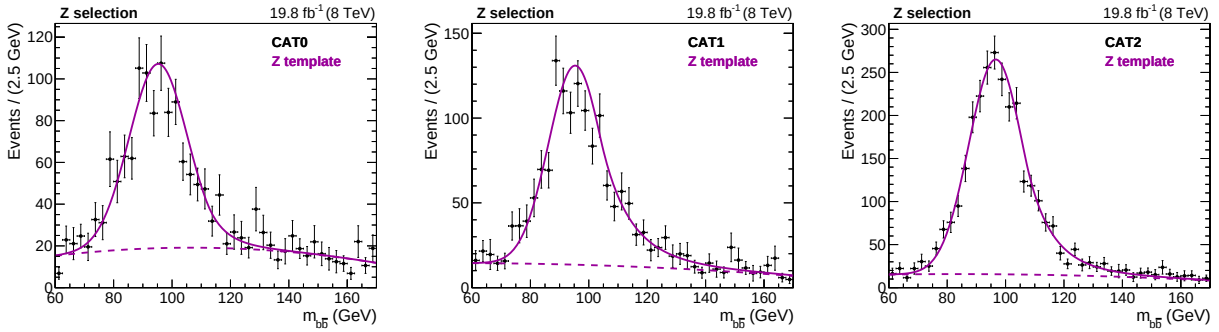


Figure 9.6: Templates for the b-quark pair invariant mass distribution for the Z selection, for the $Z \to bb$ signal. The solid line shows the total signal shape, while the dashed line indicates the combinatorial background component.

9.3.7 Fit of the b-quark pair invariant mass spectrum

With a QCD template for the reference category, the transfer functions for all categories, top background templates and Z+jets signal templates, all components are available to form the full fit model. The model is constructed as follows:

$$f_i(m_{b\bar{b}}) = N_i^{\text{qcd}} \cdot R_i(m_{b\bar{b}}) \cdot Q(m_{b\bar{b}}; \vec{p}) + N_i^{\text{top}} \cdot T_i(m_{b\bar{b}}) + \mu_Z \cdot N_i^{\text{Z}} \cdot Z_i(m_{b\bar{b}}; k_{\text{JES}}, k_{\text{JER}}), \qquad (9.8)$$

with normalisations N_i^{qcd}, N_i^{top} and N_i^{Z}, transfer functions $R_i(m_{b\bar{b}})$ and shapes $Q(m_{b\bar{b}}; \vec{p})$, $T_i(m_{b\bar{b}})$ and $Z_i(m_{b\bar{b}}; k_{\text{JES}}, k_{\text{JER}})$. For the fit, the QCD normalisation is allowed to float freely, while the top normalisation is allowed to float within 20% of its expectation; the Z normalisation is fixed to its expectation and modified by signal strength parameter $\mu_Z = \sigma/\sigma_{\text{SM}}$. The QCD shape is an 8th order Bernstein polynomial (as decided by a bias study, cf. section 9.3.8), derived from data, common for all categories, modified per category with linear transfer functions $R_i(m_{b\bar{b}})$. The parameters of the transfer functions are left free, and the parameters of QCD shape, $\vec{p}$, are allowed to float according to their uncertainties. The Gaussian top shape is kept fixed as derived from simulation. The Z+jets shape is also derived from simulation, but includes two nuisance parameters k_{JES} and k_{JER} that modify the signal scale and resolution, or in other words

the peak position and width. The scale and resolution are allowed to vary within 4 and 10% of their expected value, respectively, the uncertainties as derived from the Z+jets transverse momentum regression validation study (cf. section 8.1.1).

A binned maximum likelihood fit with bin size 0.5 GeV is performed to data, in the [60,170] GeV mass window, simultaneously in all event categories, taking into account the uncertainties as listed above. The results are presented in section 9.3.9.

9.3.8 Bias study for the Bernstein polynomial order

In the context of validating the fit procedure, a bias study is performed, studying whether the procedure allows to measure signal in an unbiased way.

The QCD multijet component accounts for a large fraction of the selected events, while the signal is just a small resonance. Given the large size of the QCD component, changes in its parametrisation can lead to unwanted differences in the measured amount of signal. Unfortunately, the shape of the QCD multijet background cannot be definitely determined using physics arguments; instead it is described using functions that cover the observed shapes well. A good fit function would be one that manages to describe the QCD multijet shape adequately, no matter which underlying model is responsible for its production, and when included in a full fit model, allows to measure the correct amount of signal.

Polynomials are known to be good choices for QCD fit functions. In the VBF analysis, a Bernstein polynomial is used. The order of the polynomial is decided according to the results of this bias study. In the study, various parametrisations are considered for the QCD background included in the full fit model as described in section 9.3.7. The QCD shape is taken from a fit to data in the [65-165] GeV mass window and its normalisation is taken from simulation, while for other components of the fit model both the shape and normalisation are taken from simulation as described above. The pseudo-experiments are repeated for signal injections from zero to four times the standard model expectation. For each parametrisation, 1000 pseudo-datasets are generated. After generating the pseudo-datasets, each dataset is fitted using a simultaneous binned maximum likelihood fit in all event categories, as would be performed in the actual analysis. At this stage, the fit functions used for the QCD component in the full fit model are Bernstein polynomials of varying orders.

For each fitted pseudo-dataset the bias on the signal strength is determined, defined as the difference between the measured amount of signal and the injected amount of signal. This bias is compared to the RMS of the measured signal distribution, the statistical uncertainty on the measured signal for one fit. If the bias is smaller than 20% of the RMS value, the fit is considered acceptable, if it is larger, the fit is considered biased. A full fit model using a specific QCD parametrisation is considered acceptable if it manages to fit the pseudo-datasets for all different generating parametrisations and for all considered amounts of injected signal, without biases above 20% of the RMS value.

For the fit of the Z boson resonance, it is found that the minimal order of the Bernstein polynomial required to reduce all biases below 20% of the RMS value, is 8th order.

Table 9.4 lists the alternate QCD parametrisations used for generating the pseudo-experiments, as well as the ones used to fit the pseudo-datasets. Figure 9.7 shows the signal bias as a function of the injected signal, for fit models with a 7th or 8th order Bernstein polynomial QCD component, and for various alternate generating models.

Table 9.4: List of alternate QCD parametrisations, included in the full fit model, used for pseudo-data generation and fitting, for the Z boson fit Bernstein polynomial order bias study.

Generating functions	expPow(x)	$(x-30)^a \cdot \exp(-bx^c)$	
	modG(x)	$\exp(-ax) \cdot \mathrm{erfc}(b-cx)$	
	tanhPol2(x)	$\tanh[a \cdot (x-b)] \cdot (cx^2+dx+e)$	
	erfPol2(x)	$\mathrm{erf}[(x-a)/b] \cdot (cx^2+dx+e)$	
Fitting functions	B^7(x)	$\sum_{\nu=0}^{7} \beta_\nu b_{\nu,7}(x)$	with $b_{\nu,n}(x) = \binom{n}{\nu} x^\nu (1-x)^{n-\nu}$
	B^8(x)	$\sum_{\nu=0}^{8} \beta_\nu b_{\nu,8}(x)$	

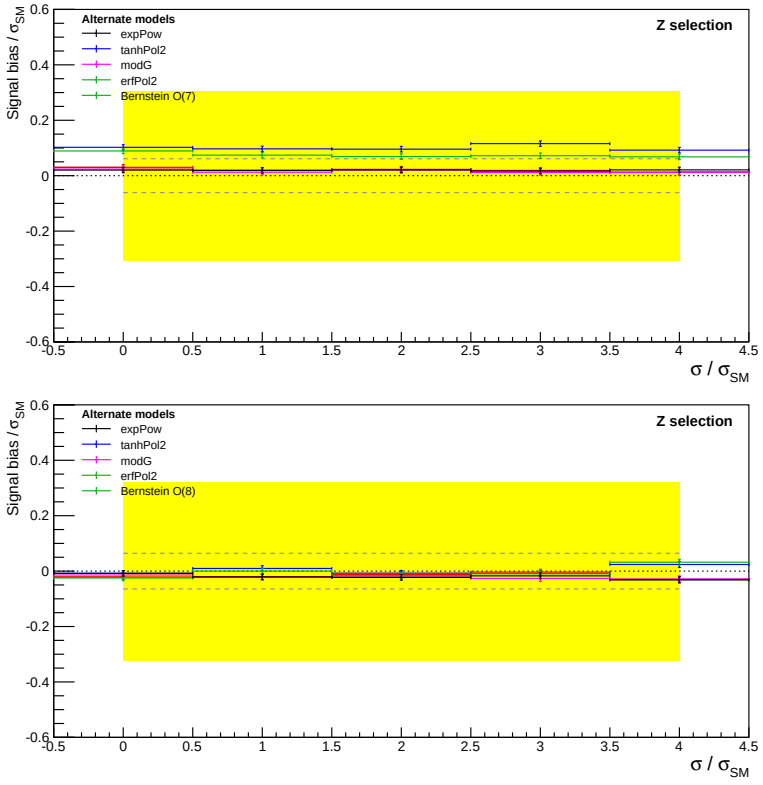


Figure 9.7: Signal bias as a function of the injected signal with a 7th (top) and 8th (bottom) order Bernstein polynomial as QCD component in the fit model for the Z boson fit. Different alternate QCD models have been used for generating the pseudo-data. The yellow band indicates the RMS of the fitted signal distribution and the dashed line corresponds to 20% of this value.

9.3.9 Results of the Z boson fit

The global fit of the b-quark pair invariant mass spectrum, performed simultaneously in all event categories of the Z selection, yields a signal strength $\mu_Z = 1.10^{+0.44}_{-0.33}$. This corresponds to an observed significance of 3.6 standard deviations, for an expected significance of 3.3 standard deviations. The results of the fit, including nuisance parameters k_{JES} and k_{JER}, are summarised in Table 9.5. The fitted distributions are shown in Fig. 9.8, together with the background-subtracted results. The measured (expected) mean and width of the Z shape are 97.7 (96.6) GeV and 9.3 (9.1) GeV respectively, in the most signal-like category CAT2. The observed and expected significance are illustrated in Fig. 9.9 (left), while the righthand panel of the same figure shows the individual results of the different categories included in the simultaneous fit.

The shift of the Z peak with respect to the theoretical Z boson mass (91.2 GeV) is due to the fact that the Z selection (chosen to allow validation of the fit method in a phase space similar to the one in which the Higgs boson fit will be performed), is subideal for a Z boson measurement. The chosen Z selection criteria lie very close or even still in the turn-on region of the trigger, and the lower efficiency at the low end of the mass spectrum causes the measured peak position to be pushed to higher mass values.

Table 9.5: Results of the Z boson fit.

Nuisance	Fitted value
μ_Z	$1.10^{+0.44}_{-0.33}$
k_{JES}	1.01 ± 0.02
k_{JER}	1.02 ± 0.10

This fit of the Z boson signal validates the fit model constructed for use in the VBF $\mathrm{H} \to \mathrm{b\bar{b}}$ signal search. At the same time it provides a data-driven measurement of the b-quark pair invariant mass distribution scale and resolution, which will be used to constrain the matching uncertainties in the VBF $\mathrm{H} \to \mathrm{b\bar{b}}$ signal search. In the following sections the inputs to the Higgs boson fit and the uncertainties considered in it are covered. The results are presented in the next chapter.

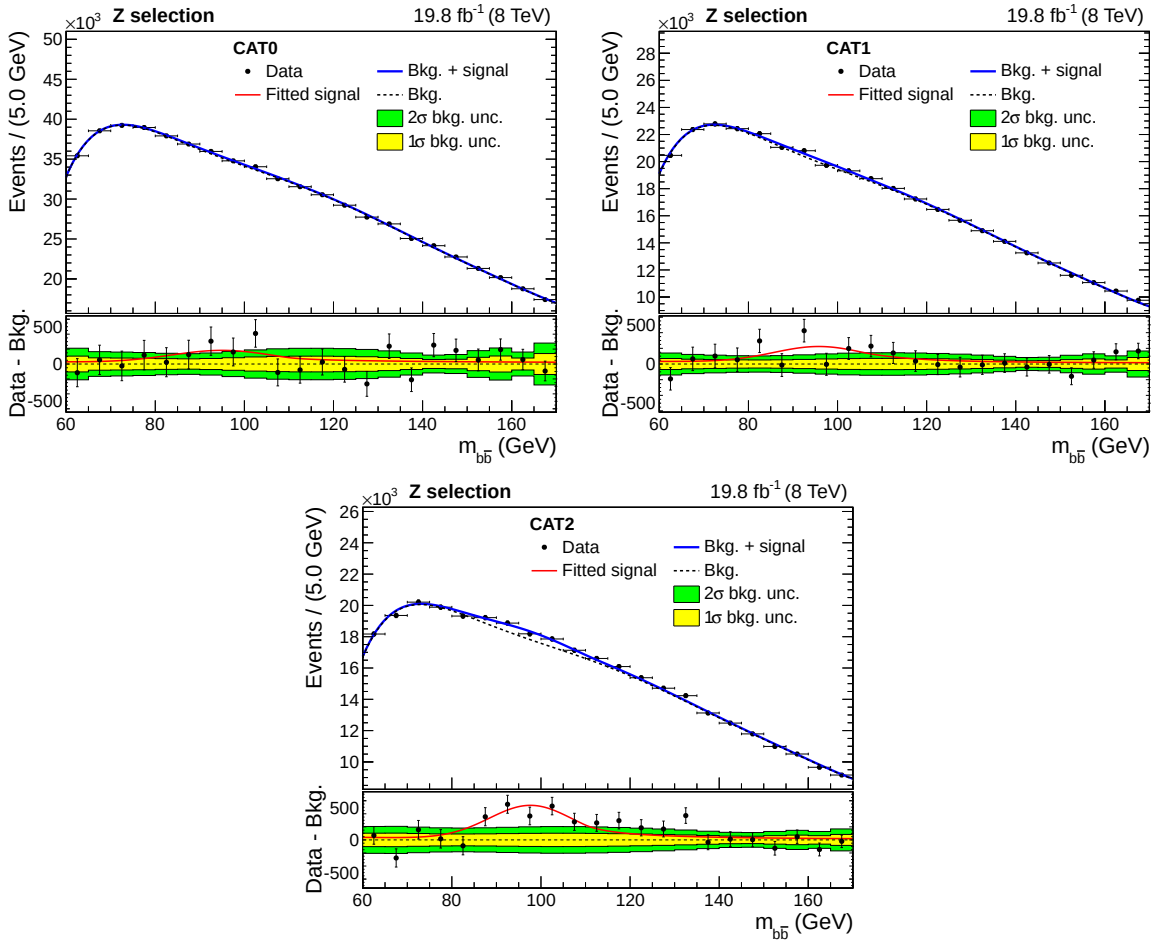


Figure 9.8: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Z selection. The top panels show the data (black markers), as well as the fitted curves: total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

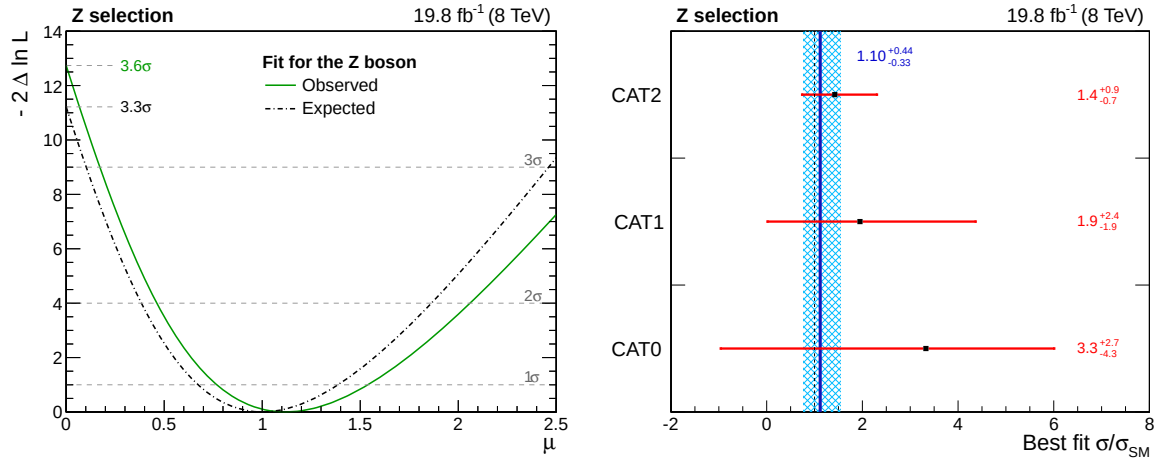


Figure 9.9: Expected and observed likelihood profile of signal strength $\mu = \sigma/\sigma_{\mathrm{SM}}$ (left) and channel compatibility check of the fitted signal in the three categories (right), for the Z boson fit.

9.4 Fit of the Higgs boson resonance

The VBF $H \to b\bar{b}$ signal search uses the same fit procedure as described for the Z boson fit in the previous section. In the following subsections the inputs to the Higgs boson fit are discussed, including the QCD background template, the transfer functions, the top and Z+jets background templates, and the VBF and GF $H \to b\bar{b}$ signal templates. The multivariate discriminant for the Set A ad Set B selections was already covered in section 8.2.

9.4.1 Categories

The Higgs boson signal search is executed using seven event categories, four for the Set A selection and three for the Set B selection. The categories are again based on the multivariate discriminant output values, with boundaries optimised to maximise the expected significance of the signal search, while making sure sufficient statistics are available in each category to derive fit templates. The boundaries of the categories are given in Table 9.6 and overlaid on the BDT output distributions for Set A and Set B in Fig. 9.10. Figure 9.11 shows the average b-quark pair invariant mass as a function of the BDT output, in the Higgs boson fit mass window [80,200] GeV, for both the Set A and Set B selections. The figure shows that the dependence of the invariant mass distribution on the BDT output is somewhat larger at the lowest BDT output values. As the fit strategy relies on having only a small dependence, CAT-1 and CAT-2 (the lowest category of each selection), are not used in the signal search. Since the signal contribution is negligible at the lowest BDT output values, leaving these categories out has no negative effect the significance of the result. A QCD template will be derived separately for the Set A and Set B selections, in background-dominated reference categories CAT0 and CAT4.

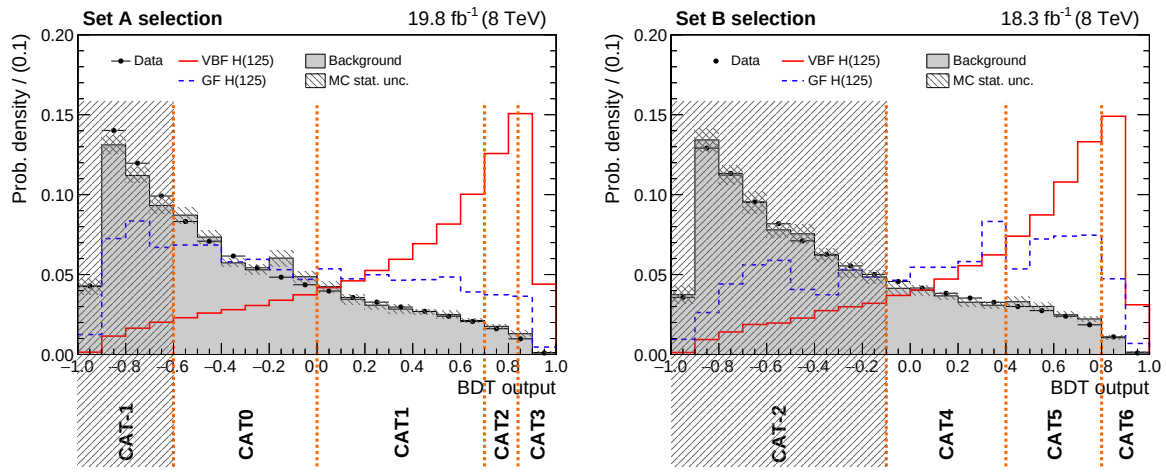


Figure 9.10: Multivariate discriminant output distributions for the Set A and Set B selections, with definition of event categories.

Table 9.6: Summary of the Set A and Set B selection category boundaries.

Set A		Set B	
Category	Discriminant range	Category	Discriminant range
-1	BDT ≤ -0.6	-2	BDT ≤ -0.1
0	$-0.6 <$ BDT ≤ 0.0	4	$-0.1 <$ BDT ≤ 0.4
1	$0.0 <$ BDT ≤ 0.7	5	$0.4 <$ BDT ≤ 0.8
2	$0.7 <$ BDT ≤ 0.84	6	$0.8 <$ BDT ≤ 1.0
3	$0.84 <$ BDT ≤ 1.0		

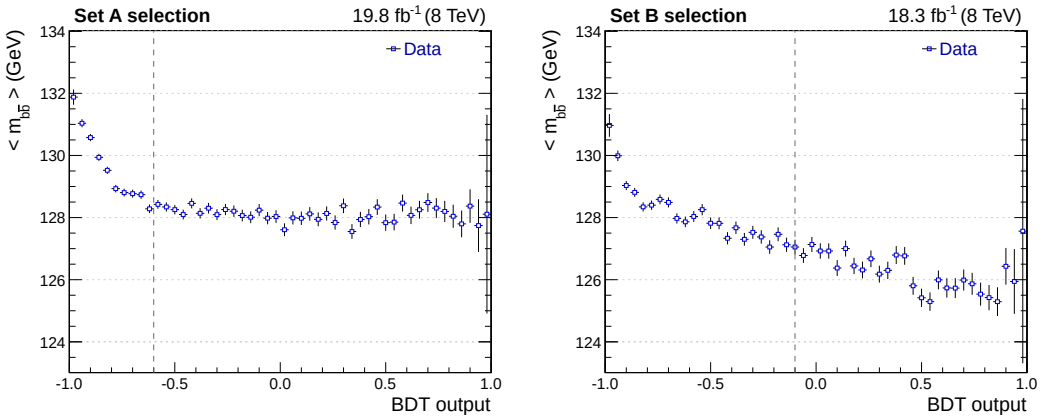


Figure 9.11: Profile histogram of the b-quark pair invariant mass as a function of the BDT output value, for the Set A and Set B selections. The dashed line corresponds to the boundary between the unused and the used data categories.

9.4.2 QCD templates and transfer functions

The QCD templates are derived from data, separately for the Set A and Set B selections, each time in the background-dominated reference category, CAT0 for Set A and CAT4 for Set B. The chosen parametrisation for the QCD component is a Bernstein polynomial, 5th order for Set A and 4th order for Set B. The residual shape differences between the categories, shown in Fig. 9.12 (blinded in the peak region), are accounted for by multiplicative transfer functions, derived as the ratio of the distribution for a given category CATi to the distribution for the matching reference category, CAT0 or CAT4.

The transfer functions for Set A are derived with respect to CAT0, as linear functions, while those for Set B are derived with respect to CAT4, as quadratic functions. For each transfer function systematic uncertainties are assigned such that all local variations are covered. The result is shown in Figs. 9.13 and 9.14. The order of both the Bernstein polynomial for the QCD shape and the order of the transfer function polynomials, is decided according to the results of bias studies as discussed in section 9.4.6.

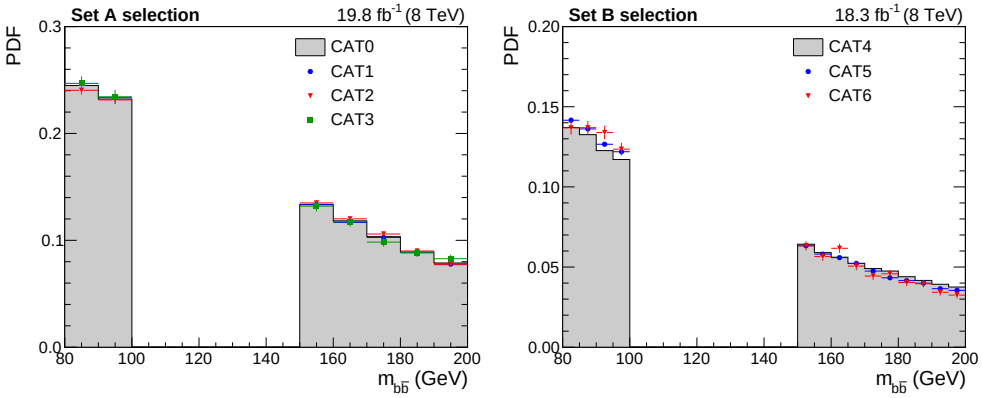


Figure 9.12: Shape comparison of the b-quark pair invariant mass distribution for the different categories of the Set A (left) and Set B (right) selections.

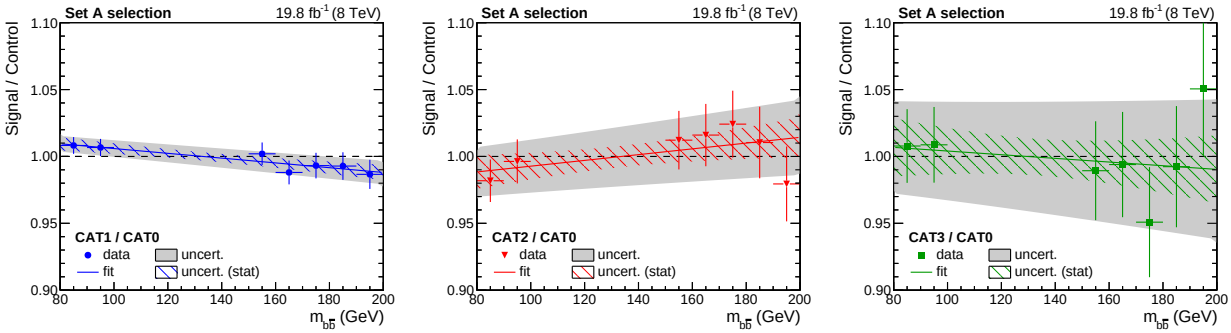


Figure 9.13: Linear transfer functions for the b-quark pair invariant mass distribution between the reference category (CAT0) and the other categories (CAT1-3) of the Set A selection, derived from data.

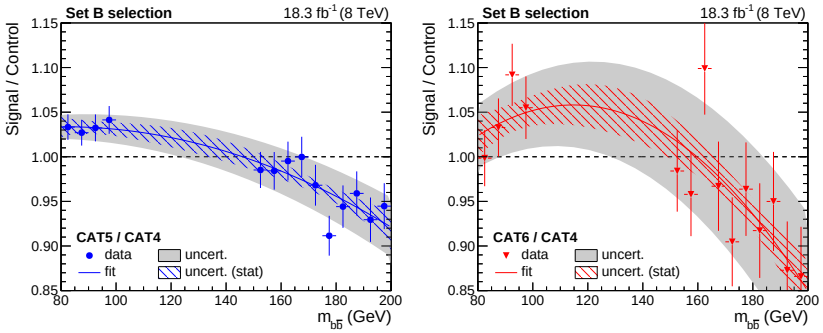


Figure 9.14: Quadratic transfer functions for the b-quark pair invariant mass distribution between the reference category (CAT4) and the other categories (CAT5-6) of the Set B selection, derived from data.

As a cross check, the transfer functions derived from data for the Higgs boson signal search, were compared to transfer functions derived from QCD simulation using the same selection criteria. Due to low statistics in the QCD samples, only one coarse signal category could be used for the comparison, CAT1-3 and CAT5-6 were merged and compared to CAT0 and CAT4 respectively. The result is shown in Fig. 9.15. Especially for Set B the statistical uncertainties in the QCD result are large, but the overall agreement is acceptable.

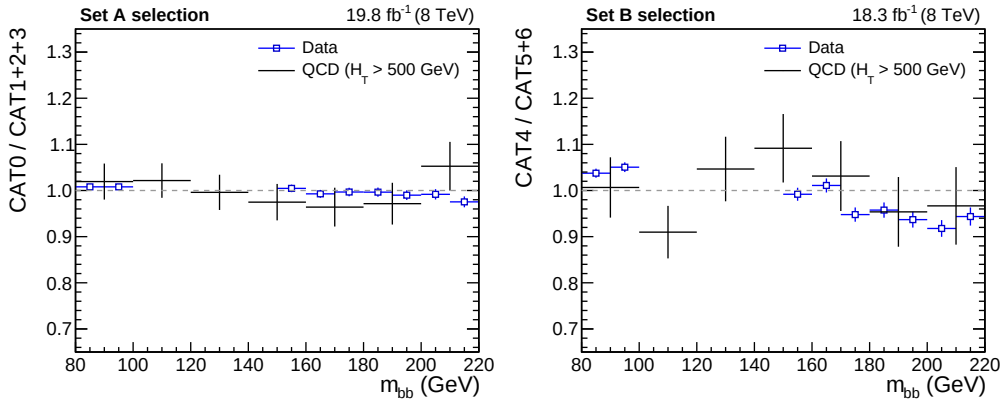


Figure 9.15: Comparison of transfer functions derived from data and QCD multijets, for the Set A and Set B selections. As QCD multijet sample is small, only one coarse signal category is used to reduce the statistical fluctuations.

9.4.3 Background templates

For the Higgs boson signal search, templates are required for the top and Z/W+jets background components. They are derived from simulation, for each category separately, using a broad Gaussian shape and a Crystal Ball plus polynomial shape, respectively. The obtained templates are shown in Figs. 9.16 and 9.17 for the top component and in Figs. 9.18 and 9.19 for the Z/W+jets component. The event yield per category for each of the samples is listed in Table 9.7.

Table 9.7: Sample yields for the various event categories of the Set A and Set B selections, in the [80,200] GeV mass window. Category CAT-1 and CAT-2 are not used in the signal search.

	Set A					Set B			
	CAT-1	CAT0	CAT1	CAT2	CAT3	CAT-2	CAT4	CAT5	CAT6
	< -0.6	$[-0.6, 0.0]$	$[0.0, 0.7]$	$[0.7, 0.84]$	$[0.84, 1.0]$	< -0.1	$[-0.1, 0.4]$	$[0.4, 0.8]$	$[0.8, 1.0]$
Data	651209	546121	321039	32740	10874	756231	203865	108279	15151
Z+jets	1661	2038	1584	198	71	918	435	280	45
W+jets	546	282	135	4	<1	966	225	92	17
$t\bar{t}$	4407	2818	839	45	14	1302	342	169	21
Single top	683	960	633	64	25	361	194	159	30
VBF 125	14	53	140	58	57	21	33	57	31
GF 125	36	53	51	8	5	10	9	10	2
VBF 115	17	65	171	71	69	27	41	73	39
GF 115	42	68	63	10	4	12	10	11	3
VBF 120	16	60	158	66	64	25	37	66	35
GF 120	38	57	54	9	6	11	9	10	3
VBF 130	12	46	121	50	48	18	27	47	26
GF 130	29	48	40	7	4	9	8	8	3
VBF 135	9	36	100	41	38	15	21	39	20
GF 135	24	38	34	6	3	8	6	7	2

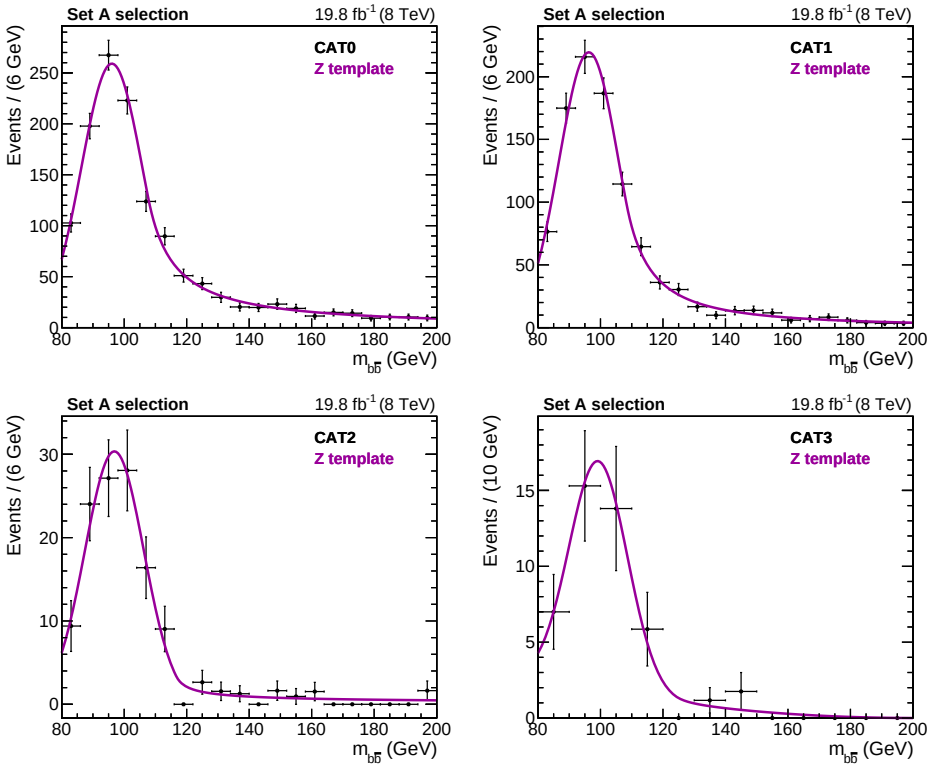


Figure 9.16: Templates for the b-quark pair invariant mass distribution for the Set A selection, for the Z/W+jets background component.

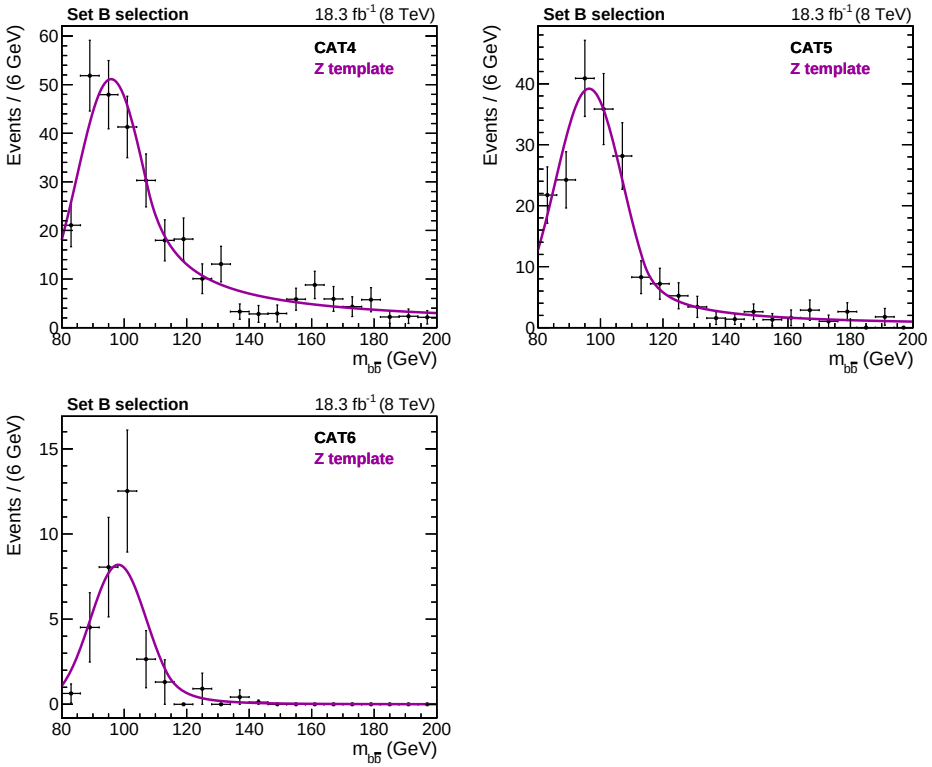


Figure 9.17: Templates for the b-quark pair invariant mass distribution for the Set B selection, for the Z/W+jets background component.

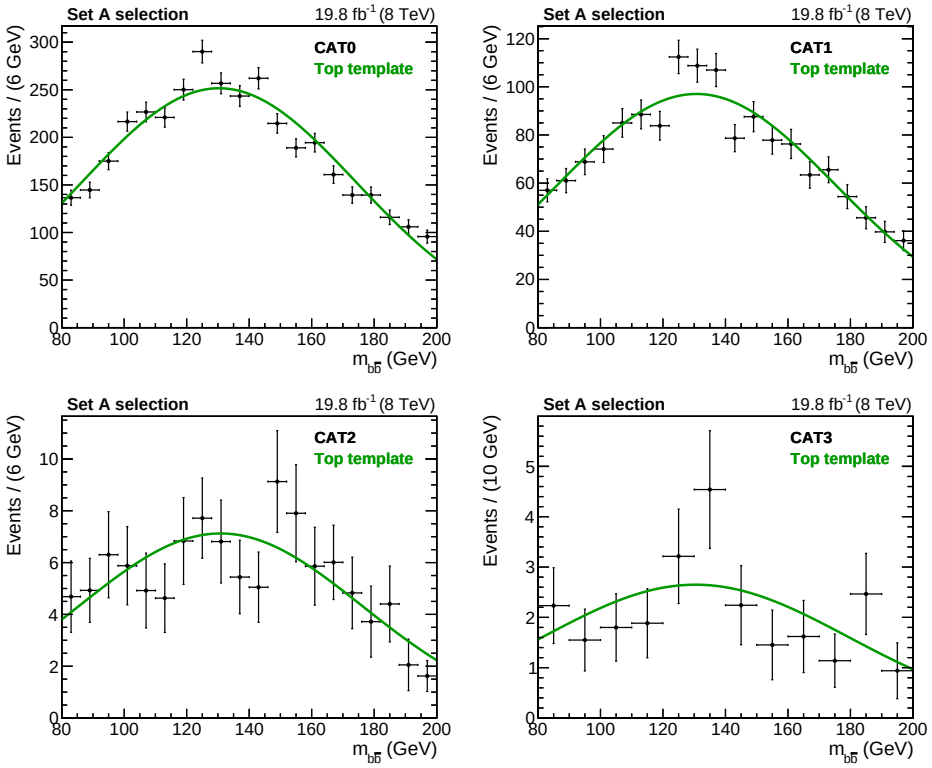


Figure 9.18: Templates for the b-quark pair invariant mass distribution for the Set A selection, for the top ($t\bar{t}$ + single-top) background component.

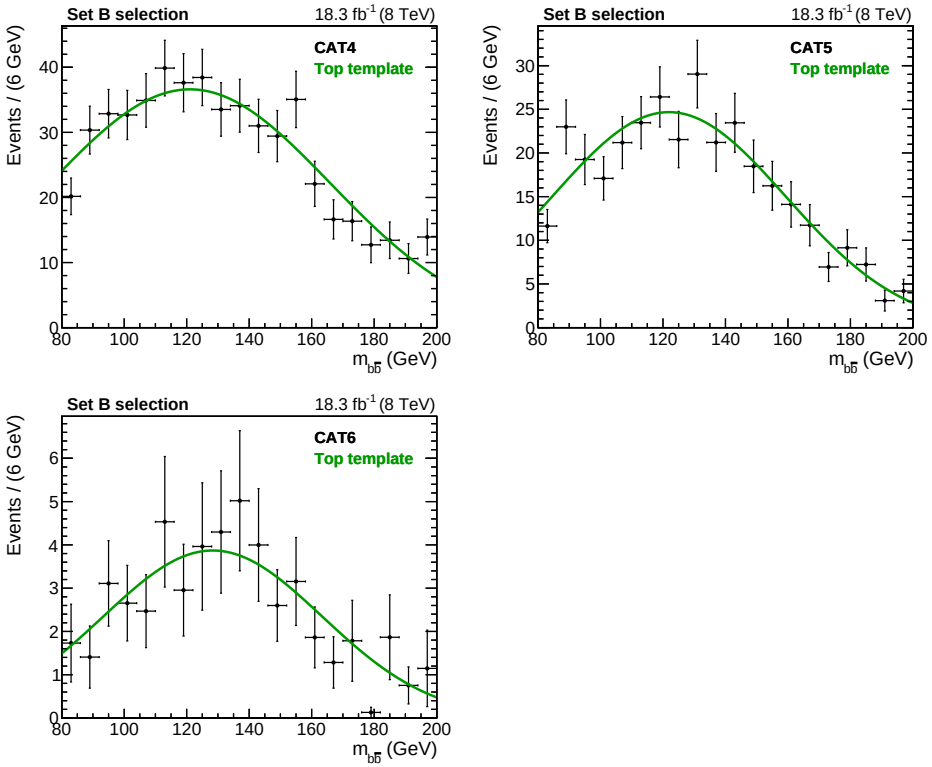


Figure 9.19: Templates for the b-quark pair invariant mass distribution for the Set B selection, for the top ($t\bar{t}$ + single-top) background component.

9.4.4 Signal templates

The main signal component in this Higgs boson signal search is the VBF $H \to b\bar{b}$ signal. There is however a non negligible, irreducible contribution in the same selection from the GF $H \to b\bar{b}$ signal as well. Figure 9.20 illustrates the fractional GF $H \to b\bar{b}$ contribution to the total signal yield, per event category and per Higgs boson mass hypothesis (115-135 GeV). It is highest in the more background-like categories, and smaller than 10% in the most signal-like categories. The signal templates are derived using the total signal distributions, after adding together the VBF and GF $H \to b\bar{b}$ components[4]. As the fitting function, a Crystal Ball plus polynomial shape is used. Templates are derived per event category, and per Higgs boson mass hypothesis. The templates for the main mass hypothesis ($m_H = 125$ GeV) are given in Figs. 9.22 and 9.23; for the other mass points we refer to appendix C. The properties of the fitted shapes are indicated in the figures, and summarised in Fig. 9.21. The expected signal yields per category are included in Table 9.7, and the expected signal vs. background ratios per category are listed in Table 9.8.

Table 9.8: Signal over background value per category, for the Set A and Set B selections.

Set A			Set B		
Category	S / B	S / $\sqrt{B}$	Category	S / B	S / $\sqrt{B}$
0	0.0005026	0.193	4	0.0004532	0.116
1	0.0016245	0.469	5	0.0014861	0.264
2	0.0057116	0.523	6	0.0057749	0.375
3	0.0170886	0.876			

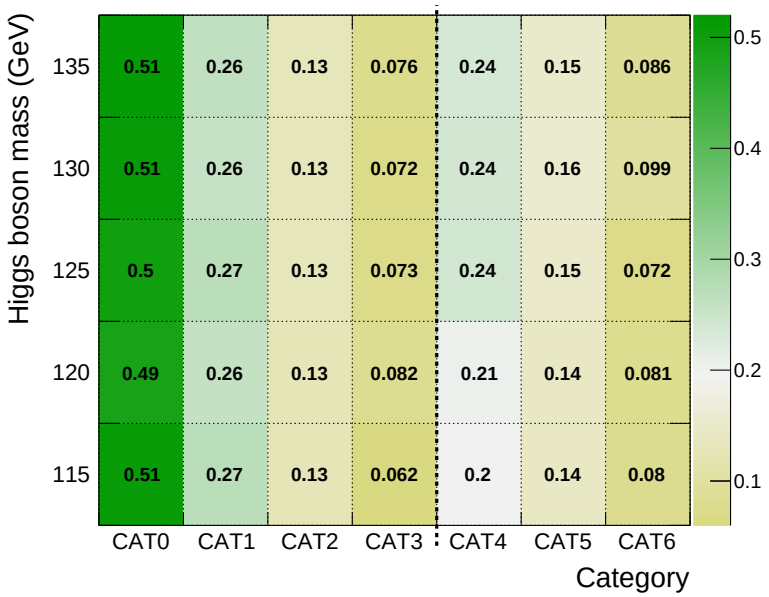


Figure 9.20: GF $H \to b\bar{b}$ contribution to the total $H \to b\bar{b}$ signal, per Higgs boson mass and per event category, given as fraction (GF / (GF + VBF)).

[4]Both components are fit together, since the individual contributions in some event categories are too small to obtain good quality individual templates.

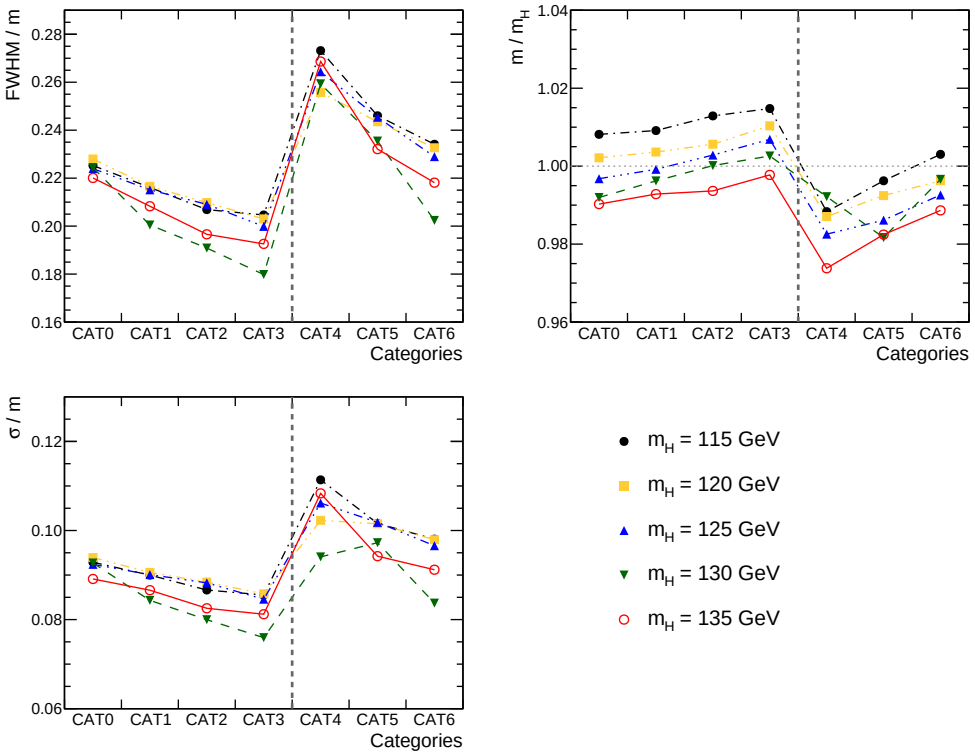


Figure 9.21: Summary of the Gaussian core parameters of the $H \to b\bar{b}$ signal templates, per event category and per mass point.

9.4.5 Fit of the b-quark pair invariant mass spectrum

The full fit model for the Higgs boson fit is constructed in the same manner as the one for the Z boson fit (a QCD template for the reference category, transfer functions for all categories, top background templates and Z/W+jets templates), plus additional VBF + GF $H \to b\bar{b}$ signal templates. The complete functional form is:

$$\begin{aligned} f_i(m_{b\bar{b}}) = {} & N_i^{\text{qcd}} \cdot R_i(m_{b\bar{b}}) \cdot Q(m_{b\bar{b}}; \vec{p}) + N_i^{\text{top}} \cdot T_i(m_{b\bar{b}}; k_{\text{JES}}, k_{\text{JER}}) \\ & + \cdot N_i^{\text{Z}} \cdot Z_i(m_{b\bar{b}}; k_{\text{JES}}, k_{\text{JER}}) + \mu_H \cdot N_i^{\text{H}} \cdot H_i(m_{b\bar{b}}; k_{\text{JES}}, k_{\text{JER}}), \end{aligned} \tag{9.9}$$

with normalisations N_i^{qcd}, N_i^{top}, N_i^{Z} and N_i^{H}, transfer functions $R_i(m_{b\bar{b}})$ and shapes $Q(m_{b\bar{b}}; \vec{p})$, $T_i(m_{b\bar{b}}; k_{\text{JES}}, k_{\text{JER}})$, $Z_i(m_{b\bar{b}}; k_{\text{JES}}, k_{\text{JER}})$ and $H_i(m_{b\bar{b}}; k_{\text{JES}}, k_{\text{JER}})$. For the fit, the QCD normalisation is allowed to float freely, while the top and Z/W+jets normalisation is allowed to float within 30% of its expectation; the Higgs boson signal normalisation is fixed to its expectation and modified by signal strength parameter $\mu_H = \sigma/\sigma_{\text{SM}}$. The QCD shape is a 5th order Bernstein polynomial for the Set A categories and a 4th order one for the Set B categories (as decided by a bias study, cf. section 9.4.6), derived from data, common for all categories of a specific set, modified per category with transfer functions $R_i(m_{b\bar{b}})$ (linear for Set A and quadratic for Set B, cf. section 9.4.6). The parameters of the QCD shape, $\vec{p}$, are allowed to float according to

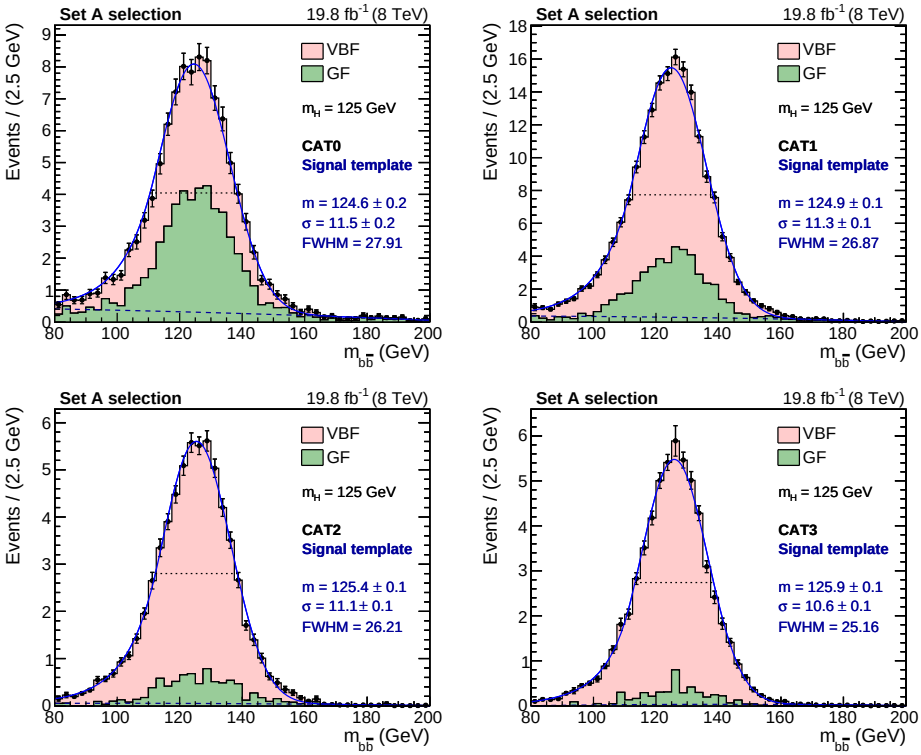


Figure 9.22: Templates for the b-quark pair invariant mass distribution for the Set A selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 125$ GeV. The VBF and GF contributions are weighted with their respective cross sections and stacked. The total histogram is scaled to the expected number of events.

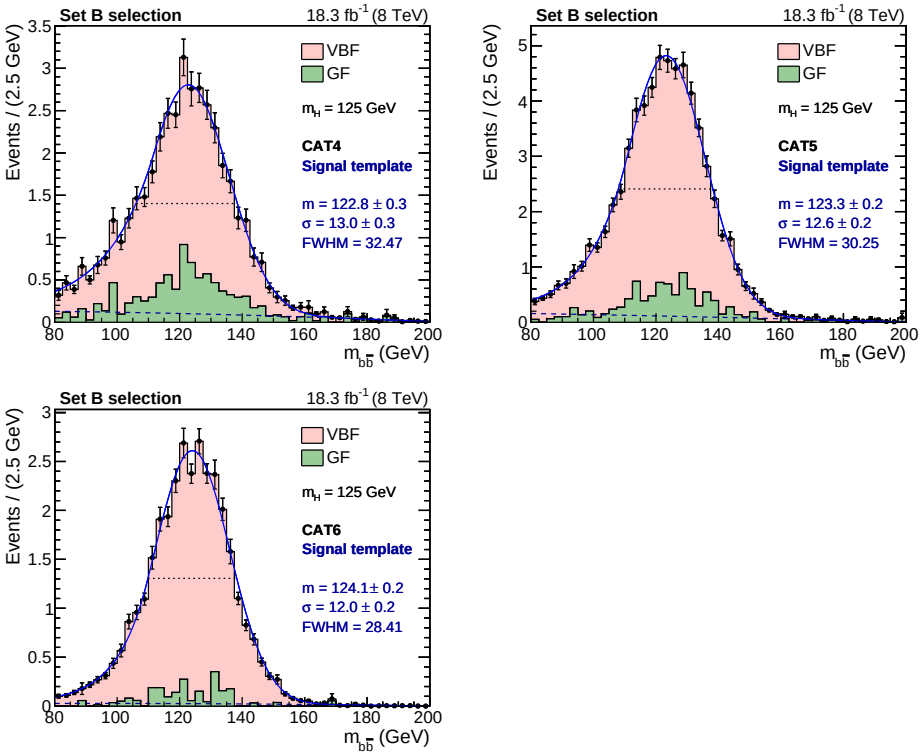


Figure 9.23: Templates for the b-quark pair invariant mass distribution for the Set B selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 125$ GeV. The VBF and GF contributions are weighted with their respective cross sections and stacked. The total histogram is scaled to the expected number of events.

their uncertainties. The Gaussian top shape, the Z/W+jets shape and the signal shape (both a Crystal ball + polynomial shape) are derived from simulation, and all include two nuisance parameters k_{JES} and k_{JER} that modify the signal scale and resolution. The scale and resolution are allowed to vary within 2 and 10% of their expected value, respectively, the uncertainties resulting from the Z boson fit.

A binned maximum likelihood fit with bin size 0.1 GeV is performed to data, in the [80,200] GeV mass window, simultaneously in all event categories of Set A and Set B. The uncertainties taken into account in the global fit will be discussed further in section 9.5.

9.4.6 Bias studies

A Bias study for the Bernstein polynomial order

The bias study as performed to determine the order of the Bernstein polynomial for the Z boson fit is repeated for the Higgs boson fit. The study is done separately for the Set A and Set B selections, i.e. one simultaneous fit is performed for all categories of the Set A selection, and another simultaneous fit is performed for all categories of the Set B selection. For each fit, the full fit model as described in section 9.4.5 is taken into account, including a QCD component (which is varied), the top and Z/W+jets components, and the signal component.

As in the Z boson bias study, pseudo-experiments are generated using various QCD parametrisations, and are then refitted with a binned maximum likelihood fit, using Bernstein polynomials of varying orders for the QCD component; each time this is done once for Set A and once for Set B. The entire procedure is repeated for signal injections from zero to four times the standard model expectation. The list of QCD parametrisations used in the generating and fitting stages is given in Table 9.9.

For each fitted pseudo-dataset the bias on the signal strength is determined, defined as the difference between the measured amount of signal and the injected amount of signal. It is found that for the Higgs boson fit, for Set A, a 5th order Bernstein polynomial is required to constrain the bias, while for Set B, a 4th order one is sufficient. Figure 9.24 shows the signal bias as of function of the injected signal and as a function of the Higgs boson mass, for a fit model with a 5th (Set A) or 4th (Set B) order Bernstein polynomial QCD component, and for various alternate generating models.

During the bias study for the Bernstein polynomial order also the fit range was optimised to [80,200] GeV, to guarantee minimal bias and good sensitivity.

Table 9.9: List of alternate QCD parametrisations, included in the full fit model, used for pseudo-data generation and fitting, for the Higgs boson fit Bernstein polynomial order bias study.

Generating functions	expPow(x)	$(x-30)^a \cdot \exp(-bx^c)$	
	modG(x)	$\exp(-ax) \cdot \mathrm{erfc}(b-cx)$	
	tanh(x)	$a-\ \tanh(b \cdot x - c)$	
	sine(x)	$1 + a \cdot\ \sin(b \cdot x - c)$	
Fitting functions	B^4(x)	$\sum_{\nu=0}^{4} \beta_\nu b_{\nu,4}(x)$	with $b_{\nu,n}(x) = \binom{n}{\nu} x^\nu (1-x)^{n-\nu}$
	B^5(x)	$\sum_{\nu=0}^{5} \beta_\nu b_{\nu,5}(x)$	
	B^6(x)	$\sum_{\nu=0}^{6} \beta_\nu b_{\nu,6}(x)$	

B Bias study for the transfer functions

Another study of the robustness of the fitting procedure is performed, verifying that the use of a transfer function (between a reference category and other categories) of a specific order does not introduce a significant bias in the fitted signal strength. The study uses the full fit model as described in section 9.4.5.

In the same manner that for the QCD distribution the underlying model cannot be determined from physics arguments, also the definitive shape of the residual invariant mass distribution shape differences between the event categories, described by the transfer functions, is not known. It is necessary to make sure the final fit model uses a fit function that can describe the shapes and fit signal without bias no matter what the true underlying model is. To do so, a bias study is executed similar to the bias studies determining the required order of the Bernstein polynomials used for the QCD component (cf. sections 9.3.8 and 9.4.6A).

Again pseudo-datasets are generated using alternate models, and are then refit using various fit models. The study is performed separately for the Set A and Set B selections, since in the data distributions for both selections indicate that there are differences in slope and possible curvature of the transfer function. The Set B dataset shows larger variations than the Set A one, so the Set B results will most likely be more affected by the choice of transfer function.

Three alternate models are used to generate pseudo-datasets. They are copies of the full fit model, with changes in the way the transfer function is set up. The three options are: a linear or quadratic polynomial with constrained parameters (FixPOL1 and FixPOL2), or an exponential function (Expo). The parameters of the functions are allowed to vary within the error taken from an independent fit of the transfer functions, performed before the global fit.

Four different models are used to fit the generated pseudo-experiments. On one hand the linear and quadratic polynomials with constrained parameters (FixPOL1 and FixPOL2) are again tested as transfer functions in the full fit model, while on the other hand linear and quadratic polynomials with free parameters (POL1 and POL2) are tested. In the latter case, the transfer function parameters are left completely free, and fit at the same time as the full model, in the global fit.

For each fitted pseudo-dataset the bias on the signal strength is determined, defined as the difference between the measured amount of signal and the injected amount of signal. The use of the different models and the outcome of the different combinations is illustrated schematically in Table 9.10. Figures 9.25 and 9.26 show the fitted signal strength for each alternate generating function, fitted using the different fit functions, for Set A and Set B.

Comparing the different outcomes, a few remarks can be made:

- The additional degree-of-freedom for quadratic polynomials worsens the RMS of the measurements compared to the matching linear polynomial (compare left/right frames of Fig. 9.25 or 9.26).
- For Set A, the choice of transfer function only matters very little: the bias falls well below 20% of the RMS value in all cases; it is however visibly smaller using free transfer functions (cf. Fig. 9.25).
- For Set B, fit models with constrained parameters show bias between linear and quadratic alternates (linear can't fit quadratic and quadratic can't fit linear). Also, linear transfer functions turn out to be inadequate, no matter whether they are constrained or free. In Fig. 9.26 only the bottom right option proves to be good.

Adding up these considerations, the best choice for the transfer functions is one with free parameters, of the lowest order that doesn't introduce any bias. For Set A this is a linear polynomial with free parameters, while for Set B a quadratic polynomial is required.

Figure 9.27 shows a cross check fit, simultaneous in all categories (Set A plus Set B), using the fit model with transfer functions as chosen from the study: linear for Set A and quadratic for Set B. Also for the full fit, the bias is constrained with 20% of the RMS value.

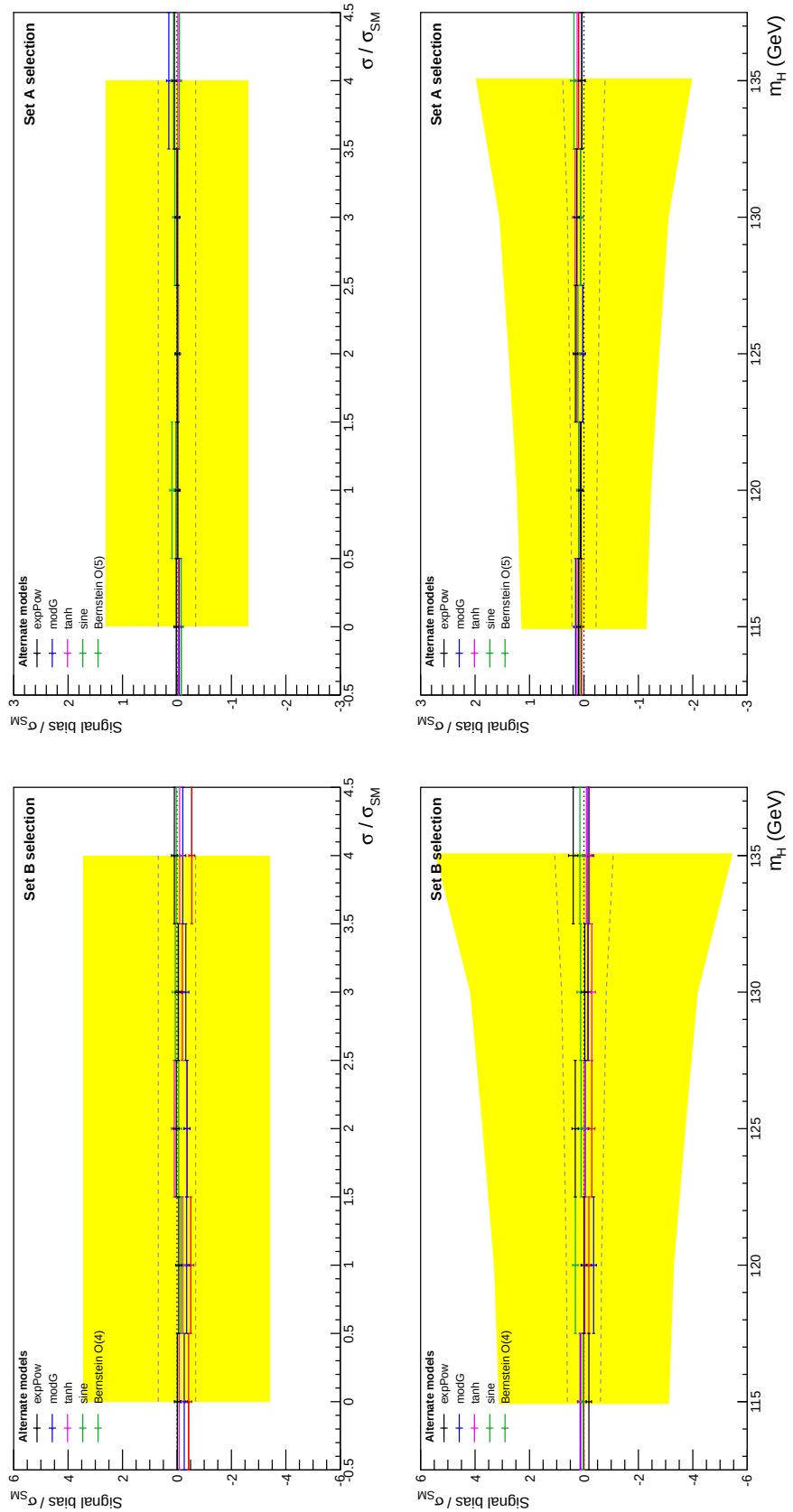


Figure 9.24: Signal bias as a function of the injected signal (top row) and the Higgs boson mass (bottom row), with a 5th (left column, Set A) and 4th (right column, Set B) order Bernstein polynomial as QCD component in the fit model for the Higgs boson fit. Different alternate QCD models have been used for generating the pseudodata. The yellow band indicates the RMS of the fitted signal distribution and the dashed line corresponds to 20% of this value.

Table 9.10: List of functions, included in the full fit model, used for pseudo-data generation and fitting, for the Higgs boson fit transfer function bias study.

			Functions used for pseudo-data generation					
			FixPOL1 $p_0 + p_1 x$		FixPOL2 $p_0 + p_1 x + p_2 x^2$		Expo $\exp(p_0 + p_1 x)$	
			Set A	Set B	Set A	Set B	Set A	Set B
Functions used for fitting	FixPOL1	$p_0 + p_1 x$	✓	✓	✓	×	✓	✓
	FixPOL2	$p_0 + p_1 x + p_2 x^2$	✓	×	✓	✓	✓	×
	POL1[4]	$1 + p'_1 x$	✓✓	✓	✓✓	×	✓✓	✓
	POL2[4]	$1 + p'_1 x + p'_2 x^2$	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓

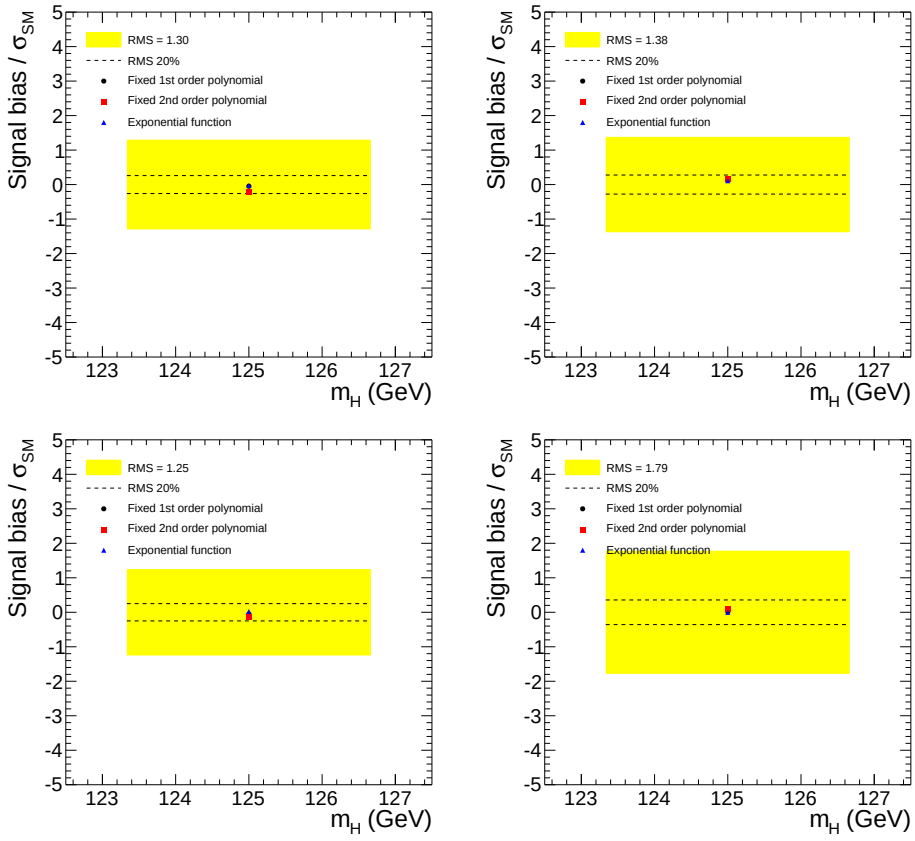


Figure 9.25: Fitted signal strength for the alternate pseudo-datasets, fitted with the different fit models, for the Set A selection. Each figure shows the result for each alternate generating function. The fits were done using a constrained linear transfer function (top left), a constrained quadratic one (top right), a free linear one (bottom left), and a free quadratic one (bottom right). The yellow bands show the RMS, with the dashed lines indicating 20% RMS.

[4]For technical reasons, when the function parameters are left free, functions have to be implemented in the form $(1 + p'_1 x + \ldots)$, where the normalisation gets absorbed in the global normalisation. If not done like that, the fits will fail because the normalisation cannot be determined.

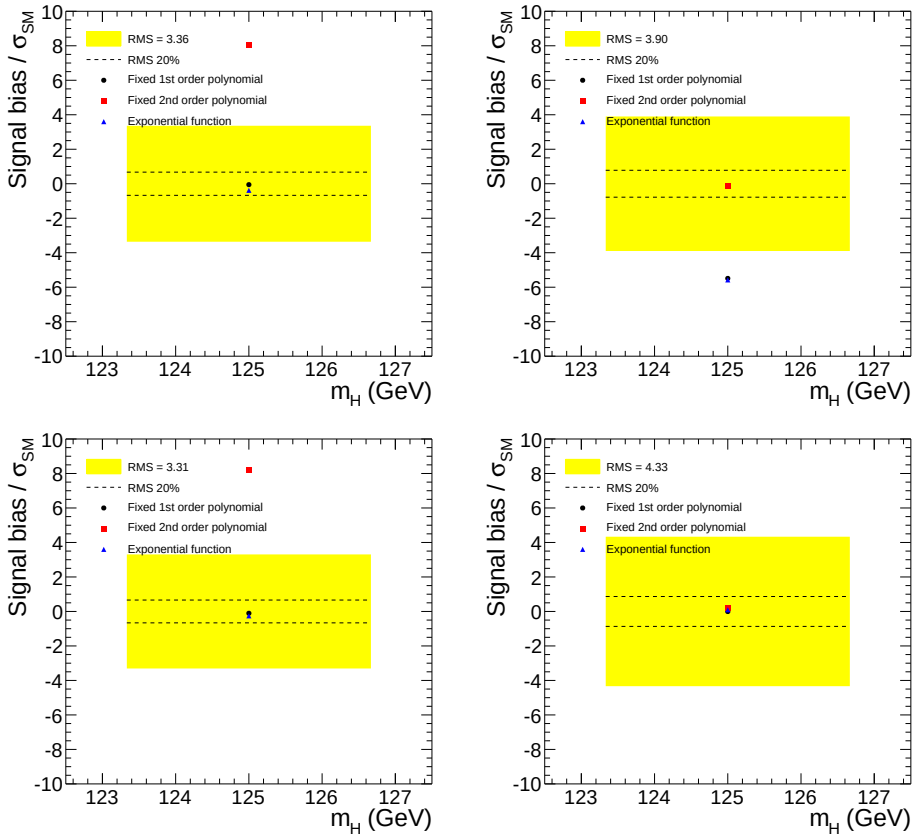


Figure 9.26: Fitted signal strength for the alternate pseudo-datasets, fitted with the different fit models, for the Set B selection. Each figure shows the result for each alternate generating function. The fits were done using a constrained linear transfer function (top left), a constrained quadratic one (top right), a free linear one (bottom left), and a free quadratic one (bottom right). The yellow bands show the RMS, with the dashed lines indicating 20% RMS.

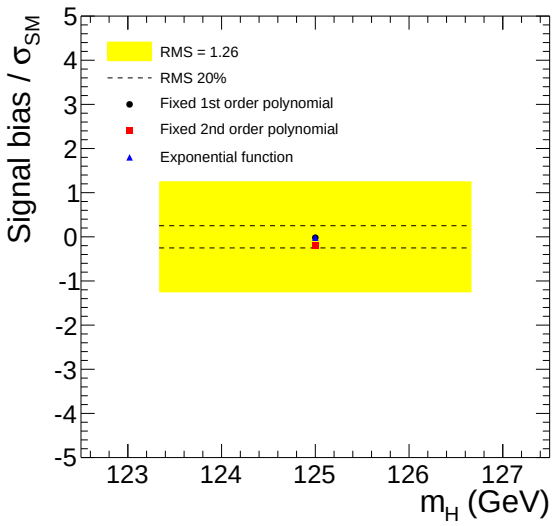


Figure 9.27: Fitted signal strength for the alternate pseudo-datasets for all categories together, using a linear free polynomial for the Set A categories and a quadratic free polynomial for the Set B categories. The yellow bands show the RMS, with the dashed lines indicating 20% RMS.

9.5 Systematic uncertainties

The systematic uncertainties identified for the VBF $H \to b\bar{b}$ analysis can be divided into three groups: uncertainties affecting the background, uncertainties affecting the signal, and theory uncertainties. For each uncertainty we briefly discuss what its impact is on the signal or background processes, and where relevant how it was evaluated. The final values of each uncertainty, as used in the global fit, are listed in Table 9.11.

9.5.1 Uncertainties affecting the background

The largest uncertainty by far follows from the QCD background description: both the shape and normalisation parameters are free and left to be determined in the global fit; the uncertainties follow from the post-fit covariance matrix. Additional uncertainties can be identified for the top and Z/W+jets background components: their normalisation is left to float within 30% of the expected values.

9.5.2 Uncertainties affecting the signal

The uncertainties affecting the signal are evaluated separately for the VBF $H \to b\bar{b}$ signal and for the GF $H \to b\bar{b}$ signal, using the simulated samples with $m_H = 125$ GeV. They can affect both the signal acceptance and the shape of the multivariate discriminant output, which may lead to (anti)correlation between the event categories.

- **Jet energy scale and resolution:** To study this uncertainty the JES and JER are varied according to their measured values [96], after which the events are reinterpreted. The effect on the b-quark pair invariant mass distribution normalisation and shape are evaluated separately. Figures 9.28 and 9.29 show the acceptance uncertainty due to the jet energy scale and resolution variation respectively, per category, for both the Set A and Set B selections, for the VBF $H \to b\bar{b}$ signal; the shape uncertainty due to the jet energy scale variation is shown in Fig. 9.30. Figures 9.31, 9.32 and 9.33 show the same results for the GF $H \to b\bar{b}$ signal. The 10% for the shape uncertainty due to the jet energy resolution variation was evaluated from Z+jets data in the jet transverse momentum regression validation study (cf. section 8.1.1).
- **Jet flavour tagging:** The uncertainty due to b-tagging and quark/gluon tagging is studied by performing a reshaping of the tagging discriminants (CSV and QGL discriminants, cf. sections 5.2.4 and 5.2.5), covering any observed differences between data and simulation. The result is shown in Figs. 9.34 to 9.37. The effect on the signal acceptance is largest for the CSV discriminant reshaping, since b-tagging is used in both the event selection and as

an input to the main analysis multivariant discriminant, while QGL tagging is only used for the discriminant.

- **Trigger:** The trigger uncertainty affects the signal acceptance and is derived from the uncertainty on the data vs. simulation trigger efficiency scale factors (cf. section 7.3). The two dimensional scale factor distributions are convoluted with matching two dimensional signal distribution maps, and the uncertainties are evaluated per event category. The result is shown for both the Set A and Set B selections, in Fig. 9.38 for the VBF $H \to b\bar{b}$ signal, and in Fig. 9.39 for the GF $H \to b\bar{b}$ signal.
- **Luminosity:** the integrated luminosity uncertainty is taken equal to the default CMS value of 2.6%.
- **Branching fraction ($H \to b\bar{b}$):** the uncertainty on the $H \to b\bar{b}$ branching fraction is taken from the handbook of LHC Higgs cross sections [143].

Table 9.11: Sources of systematic uncertainty and their impact on the shape and normalisation of the background and signal processes. All values are given in %.

Background uncertainties		
QCD shape parameters	determined by the fit	
QCD bkg. normalisation	determined by the fit	
Top quark bkg. normalisation	30	
Z/W+jets bkg. normalisation	30	
Uncertainties affecting the signal	**VBF signal**	**GF signal**
JES (signal shape)	2	
JER (signal shape)	10	
Integrated luminosity	2.6	
Branching fraction ($H \to b\bar{b}$)	2.4–4.3	
JES (acceptance)	6–10	4–12
JER (acceptance)	1–4	1–9
b-jet tagging	3–9	3–10
Quark/gluon-jet tagging	1–3	1–3
Trigger	1–6	5–20
Theory uncertainties	**VBF signal**	**GF signal**
UE & PS	2–7	10–45
PDF (global)	2.8	7.5
PDF (categories)	1.5–3	3.5–5
Scale variation (global)	0.2	7.7–8.1
Scale variation (categories)	1–5	1–5

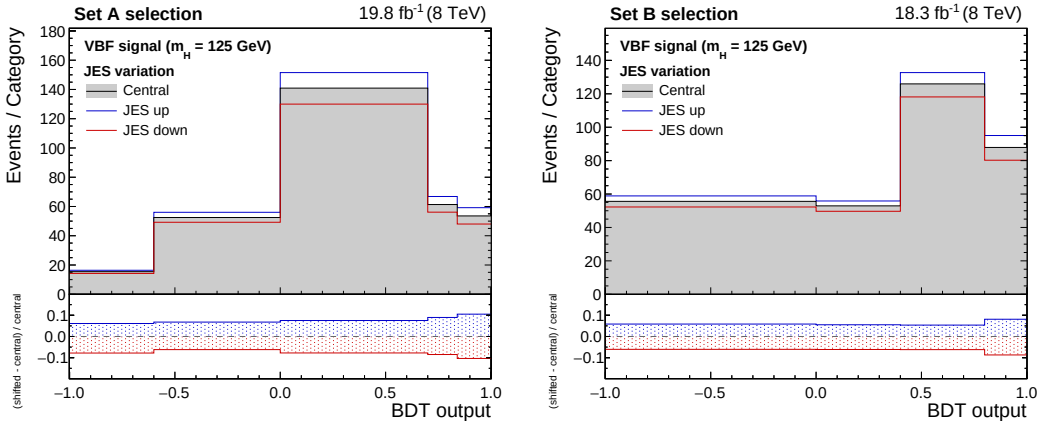


Figure 9.28: Jet energy scale uncertainty per category for the Set A (left) and Set B (right) selections, for the VBF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

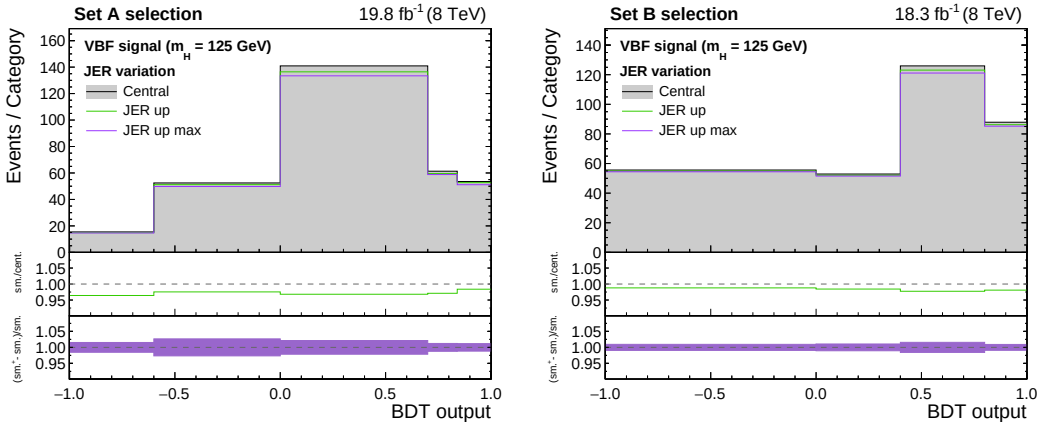


Figure 9.29: Jet energy resolution uncertainty per category for the Set A (left) and Set B (right) selections, for the VBF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

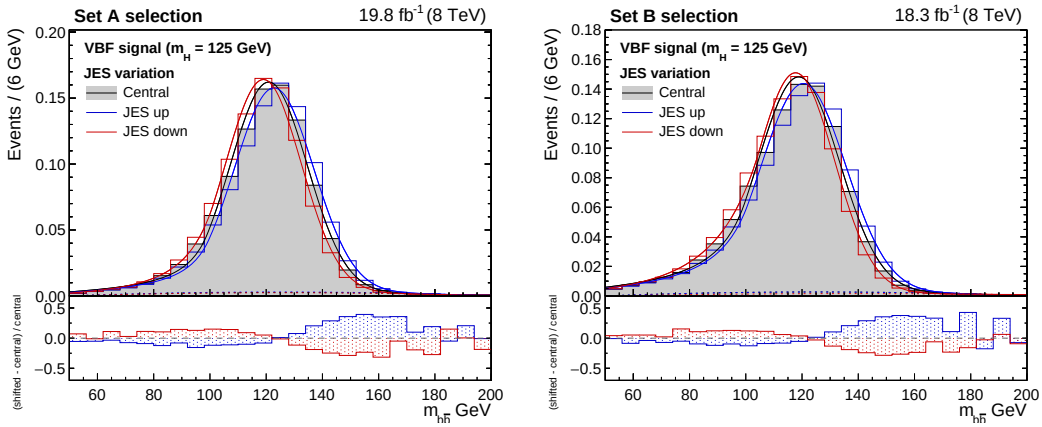


Figure 9.30: Effect of the jet energy scale uncertainty on the shape of the b-quark pair invariant mass for the Set A (left) and Set B (right) selections, for the VBF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

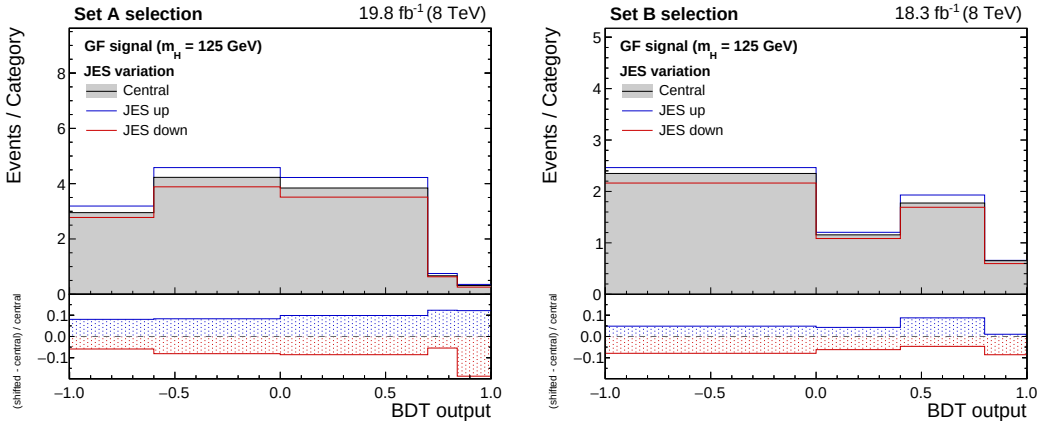


Figure 9.31: Jet energy scale uncertainty per category for the Set A (left) and Set B (right) selections, for the GF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

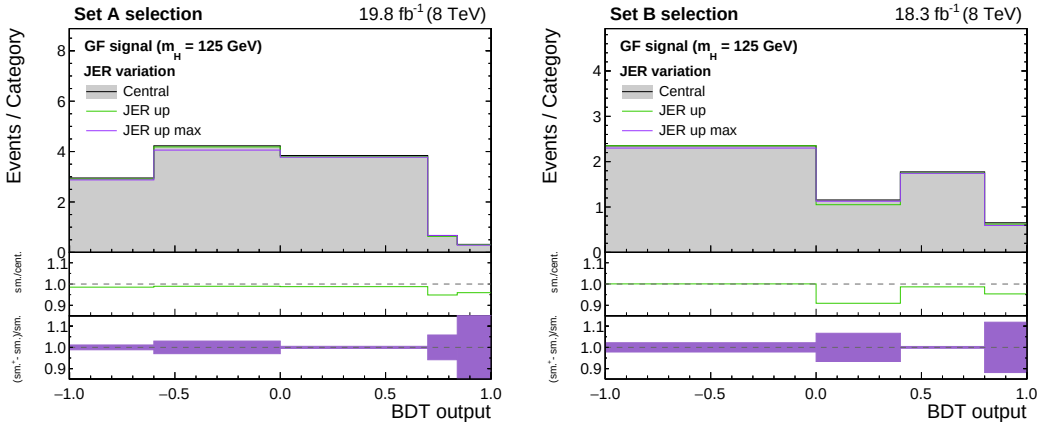


Figure 9.32: Jet energy resolution uncertainty per category for the Set A (left) and Set B (right) selections, for the GF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

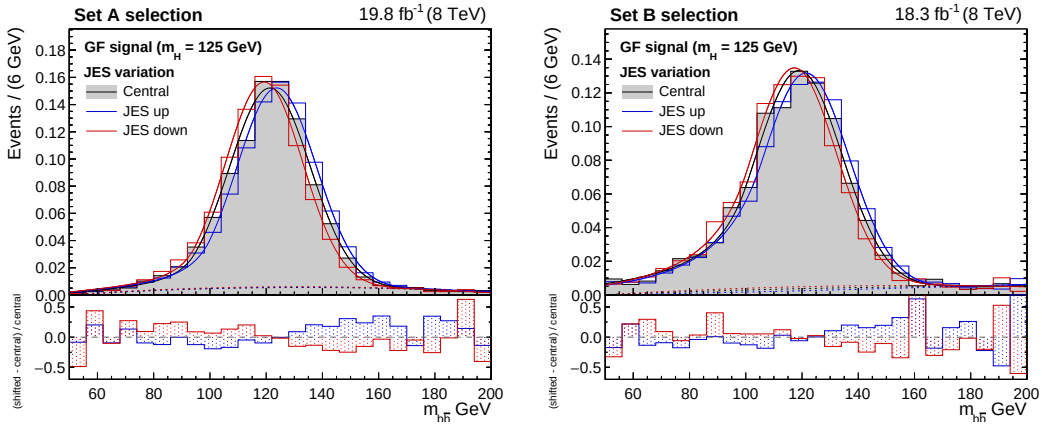


Figure 9.33: Effect of the jet energy scale uncertainty on the shape of the b-quark pair invariant mass for the Set A (left) and Set B (right) selections, for the GF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

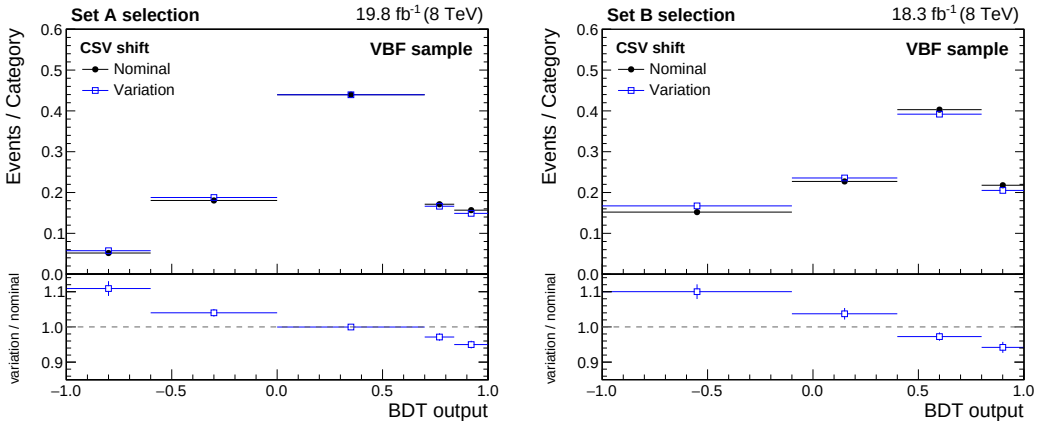


Figure 9.34: Effect of the CSV discriminant reshaping on the BDT output, per event category, for the Set A (left) and Set B (right) selections, for the VBF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

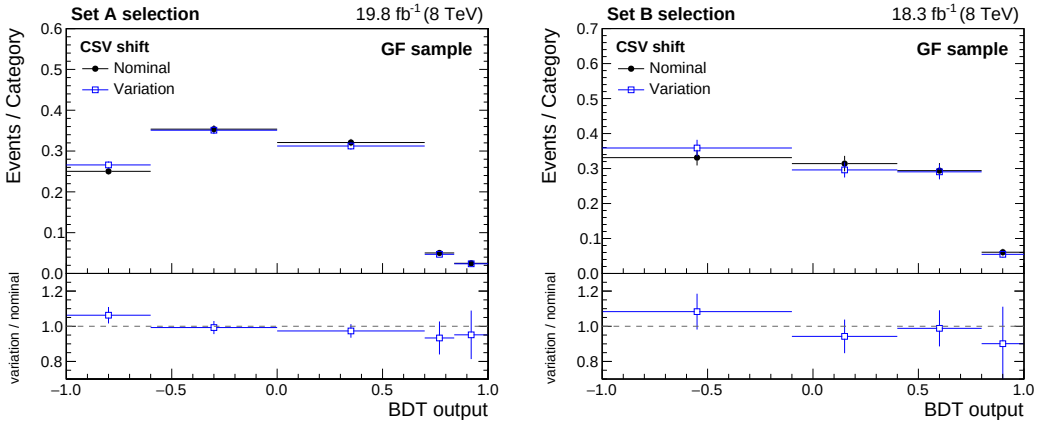


Figure 9.35: Effect of the CSV discriminant reshaping on the BDT output, per event category, for the Set A (left) and Set B (right) selections, for the GF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

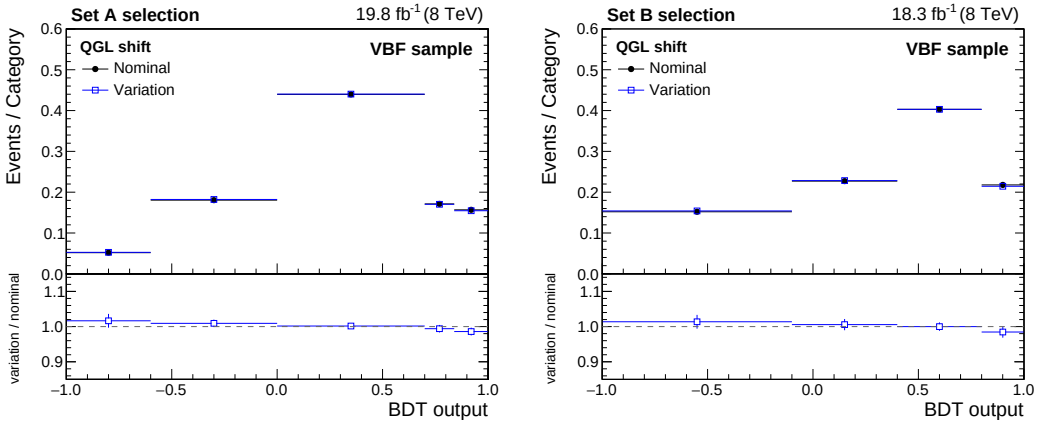


Figure 9.36: Effect of the QGL discriminant reshaping on the BDT output, per event category, for the Set A (left) and Set B (right) selections, for the VBF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

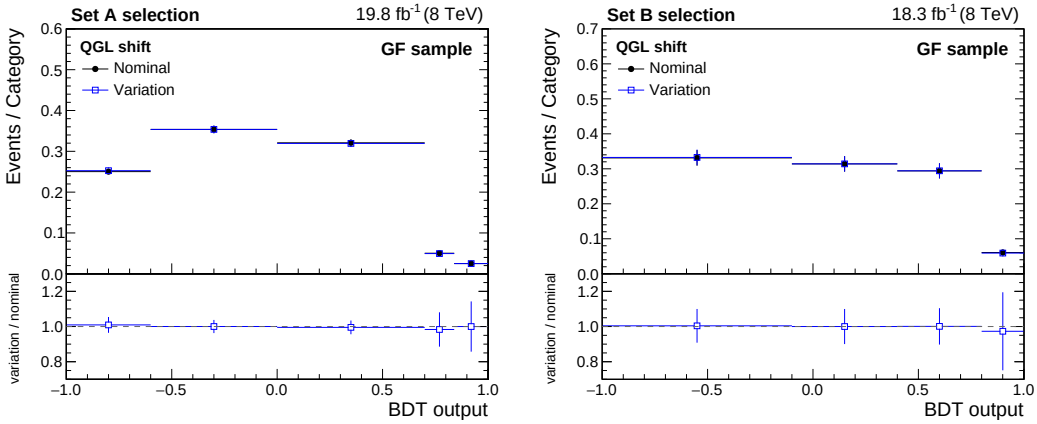


Figure 9.37: Effect of the QGL discriminant reshaping on the BDT output, per event category, for the Set A (left) and Set B (right) selections, for the GF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

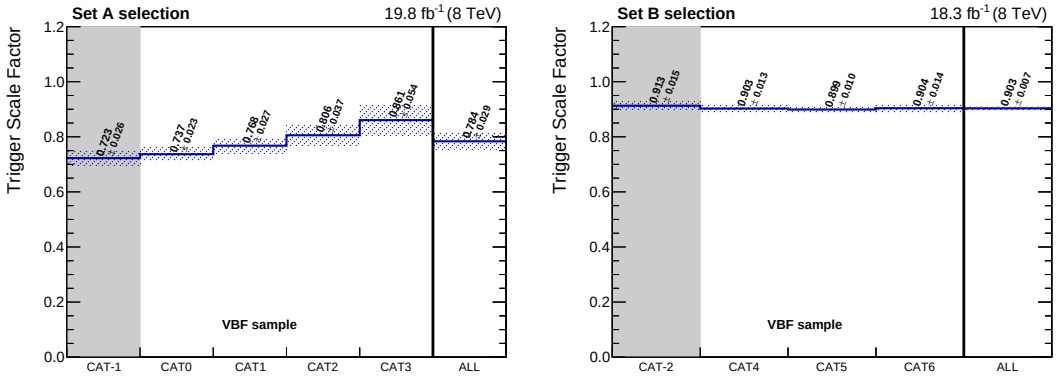


Figure 9.38: Trigger efficiency scale factor distributions convoluted with the signal distribution, per category, for the VBF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

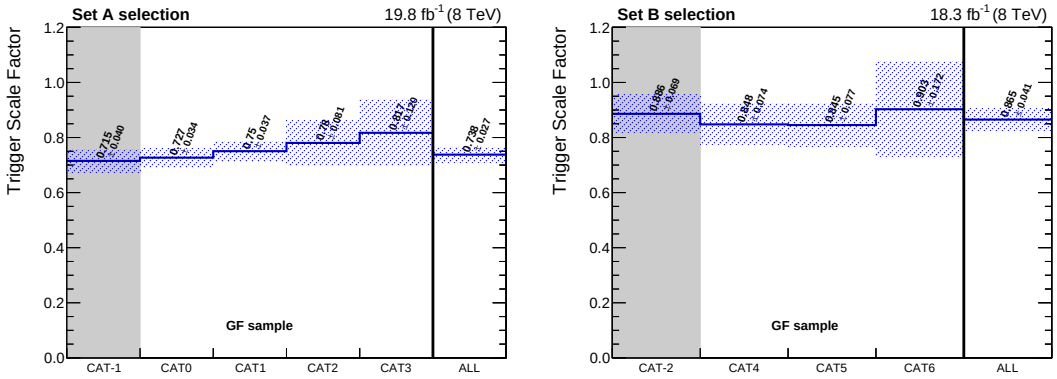


Figure 9.39: Trigger efficiency scale factor distributions convoluted with the signal distribution, per category, for the GF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

9.5.3 Theory uncertainties

The theory uncertainties affect all simulated samples, including the signal samples. All uncertainties are determined separately per event category of the Set A and Set B selections, and for the VBF and GF $H \to b\bar{b}$ signal.

- **Cross section:** the production cross section or global scale uncertainty is taken from the handbook of LHC Higgs cross sections [143]. The values for $m_H = 125$ GeV are 2.8% for VBF production, and 8% for GF production.
- **Underlying event:** the uncertainty due to the underlying event and parton shower modelling is computed using an alternate showering model: the original hard processes generated using the Powheg Monte Carlo generator, were reshowered using the Pythia 8 generator instead. The result is compared to the original Pythia 6 version, and is shown in Figs. 9.40 and 9.41.
- **PDFs and α_S:** the uncertainty due to PDFs and α_S can be split into a global effect and a residual effect specific to the phase space used. The global uncertainties are taken from the handbook of LHC cross sections [143]. The residual effect is evaluated according to the PDF4LHC recommendations [144, 145]: uncertainties are computed covering the CT10, MSTW2008 and NNPDF2.3 PDFs, and the total uncertainty is taken as the envelope around all contributions, as shown in Figs. 9.42 and 9.43.
- **Scale variation:** the effect of the factorisation and renormalisation scale uncertainties on the acceptance was evaluated by introducing 0.5-2 times variations to the scales, using the Powheg Monte Carlo generator.

All these systematic uncertainties are considered in the global fit, the results of which are presented in the following chapter 10.

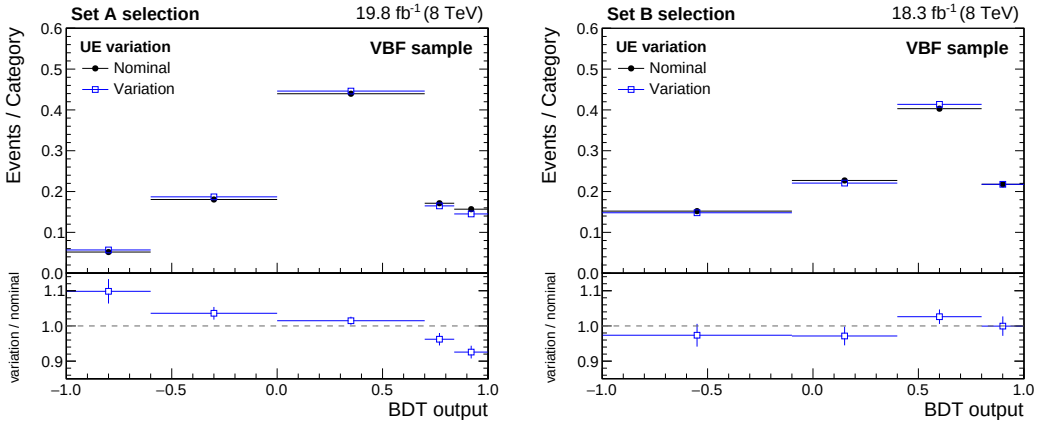


Figure 9.40: Effect of the underlying event variation on the BDT output, per event category for the Set A (left) and Set B (right) selections, for the VBF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

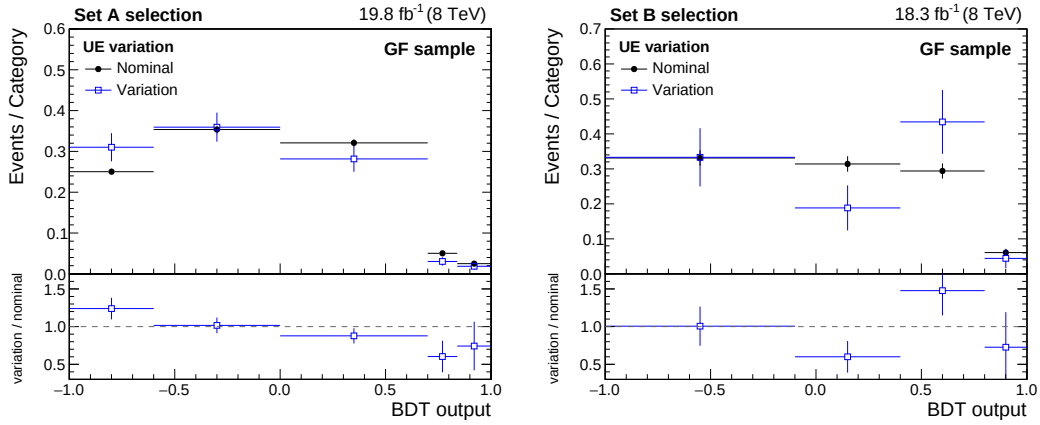


Figure 9.41: Effect of the underlying event variation on the BDT output, per event category for the Set A (left) and Set B (right) selections, for the GF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

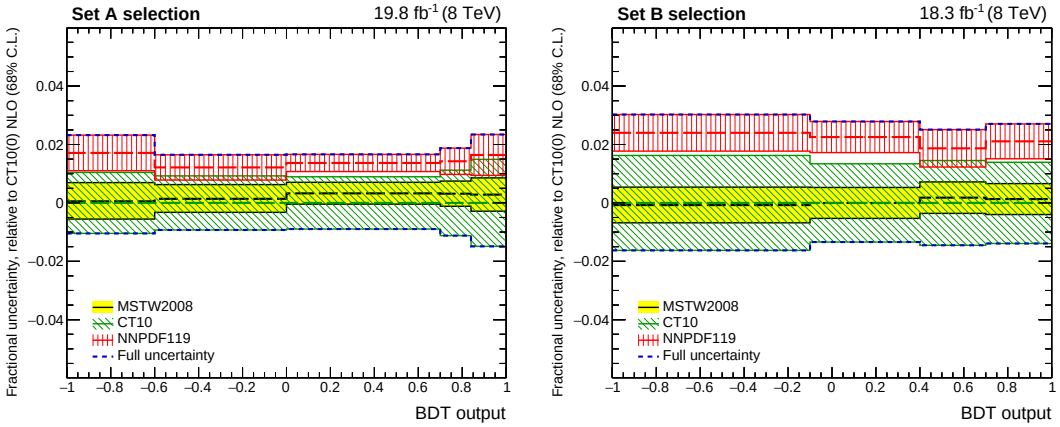


Figure 9.42: Fractional PDF uncertainty per category for the Set A (left) and Set B (right) selections, for the VBF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

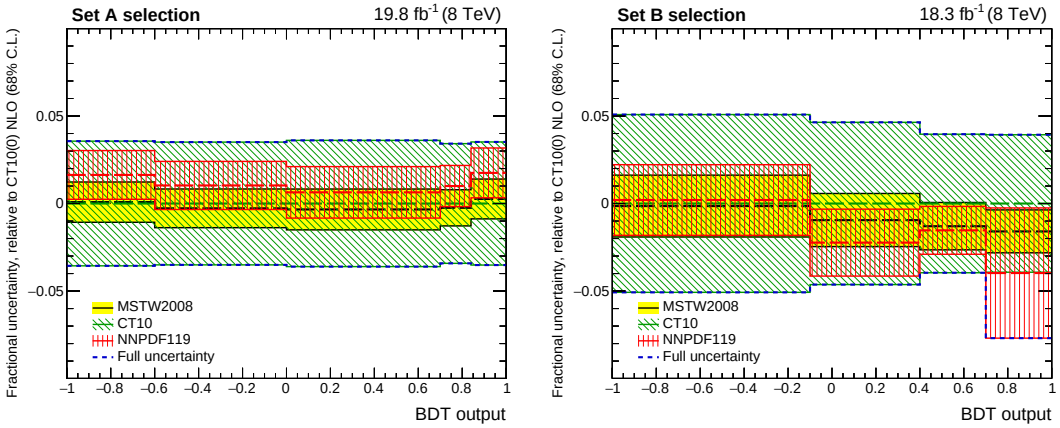


Figure 9.43: Fractional PDF uncertainty per category for the Set A (left) and Set B (right) selections, for the GF $H \to b\bar{b}$ simulated sample with $m_H = 125$ GeV.

Chapter 10

Results

10.1 Results of the Higgs boson fit

A binned maximum likelihood fit is performed simultaneously in all event categories of the Set A and Set B selections, testing both the background only and the background plus signal hypotheses. The full dataset, recorded at a centre-of-mass energy $\sqrt{s} = 8$ TeV, corresponds to an integrated luminosity of 19.8 fb^{-1}. For the global fit, the full fit model is used as described in section 9.4.5, including all systematic uncertainties discussed in section 9.5 as nuisance parameters. The global fit is repeated for each of the five Higgs boson mass hypotheses (115-135 GeV). In this chapter, the results are presented for the main mass point, $m_H = 125$ GeV; matching figures for the remaining mass points are given in appendix C. The fitted distributions, including the curves for the QCD background, the total background, and the total signal plus background, are shown in Figs. 10.1a and 10.1b for Set A and Set B; also the background-subtracted results are shown.

Limits on the signal strength are derived using the modified CL_S method; the results are listed in Table 10.1 and shown in Fig. 10.2. An excess is observed with respect to the expected values, for all mass points. The observed limit ranges from 5.5 to 5.8 between the lowest and the highest mass point, while the expected limit ranges between 2.2 and 3.7. The injected limit, i.e. the expected limit in the presence of a standard model Higgs boson with $m_H = 125$ GeV, is also given.

Table 10.1: Summary of the upper limits on the signal cross section times branching ratio, the significance, the p-value, and the best fitted signal strength, in units of σ_{SM}, for the Higgs boson fit.

m_H (GeV)	Exp. limit	Obs. limit	Inj. limit	Fitted μ
115	2.16	4.99	2.95	$2.81^{+1.25}_{-1.10}$
120	2.26	5.26	3.04	$2.88^{+1.36}_{-1.19}$
125	2.55	5.48	3.33	$2.79^{+1.53}_{-1.35}$
130	2.82	5.04	3.59	$2.17^{+1.61}_{-1.44}$
135	3.73	5.82	4.43	$2.17^{+2.01}_{-1.83}$
m_H (GeV)	Exp. sign.	Obs. sign.	Exp. p-value	Obs. p-value
115	1.00	2.67	0.159	0.0037
120	0.95	2.56	0.171	0.0052
125	0.83	2.18	0.203	0.0146
130	0.74	1.56	0.229	0.0558
135	0.56	1.23	0.286	0.111

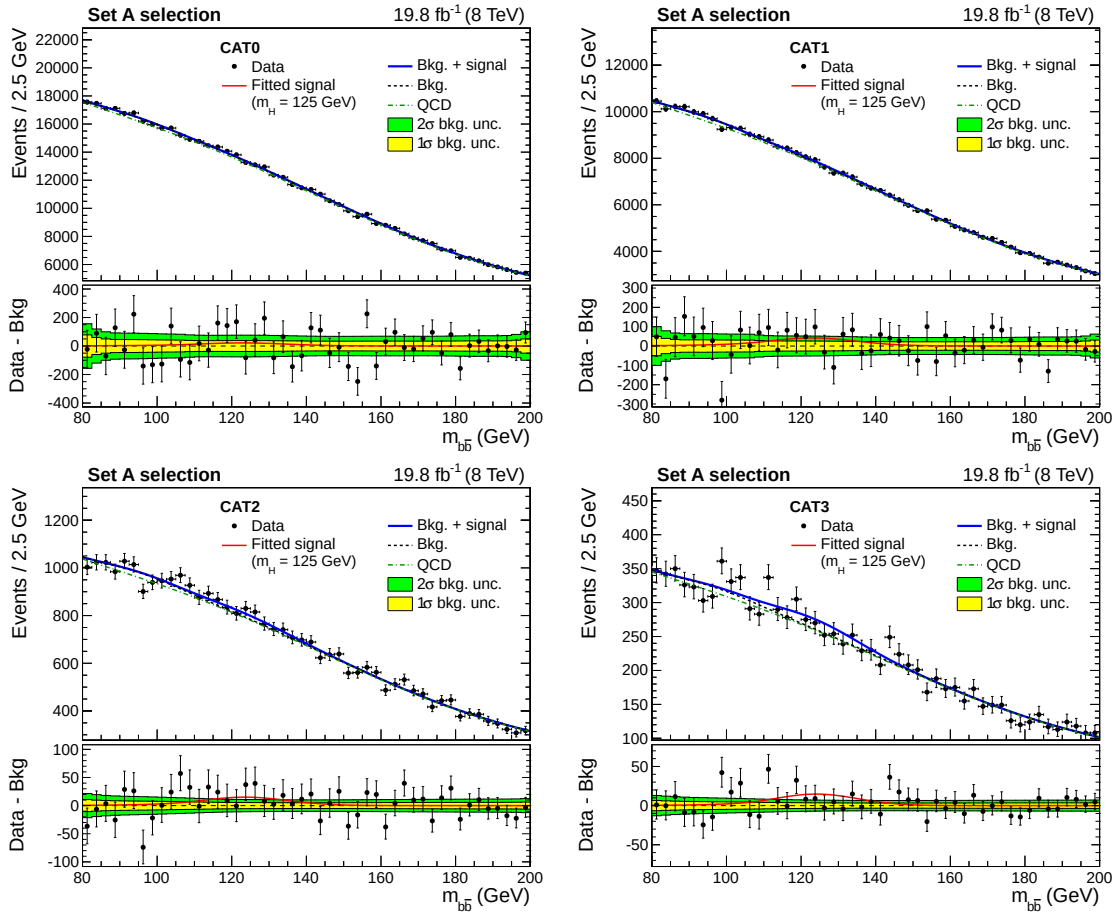


Figure 10.1a: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Set A and Set B selections, for a VBF $H \to b\bar{b}$ signal with $m_H = 125$ GeV. The top panels show the data (black markers), as well as the fitted curves: QCD background (green dash-dotted), total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

The result corresponds to an observed significance of 2.2 standard deviations, for an expected significance of 0.8 standard deviations. The significance is illustrated in Fig. 10.3, which shows the expected and observed likelihood profile of the signal strength for the $m_H = 125$ GeV Higgs boson mass hypothesis. The significances and matching p-values for all mass hypotheses are shown in Fig. 10.4.

The best fitted signal strength is $\mu = \sigma/\sigma_{SM} = 2.8^{+0.53}_{-0.35}$. This is compatible with the standard model expectation of $\mu = 1$ at the 8% level, as is illustrated in Fig. 10.5. To determine the compatibility, pseudo-experiments are generated and fitted, and the integrated probability is computed.

Figure 10.6 then shows the post-fit nuisance pulls and uncertainties for the global fit in all categories, defined as (post-fit value - pre-fit value) / pre-fit value.

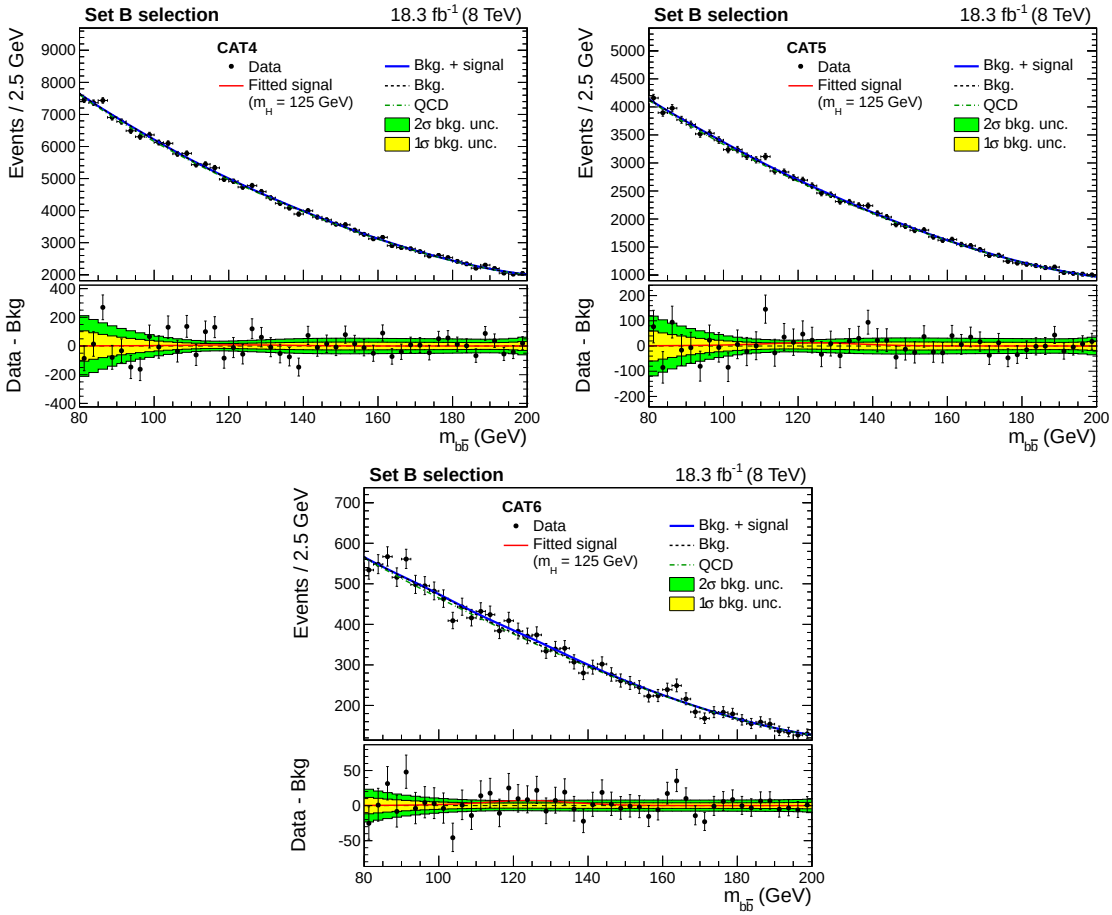


Figure 10.1b: Continued from Fig. 10.1a

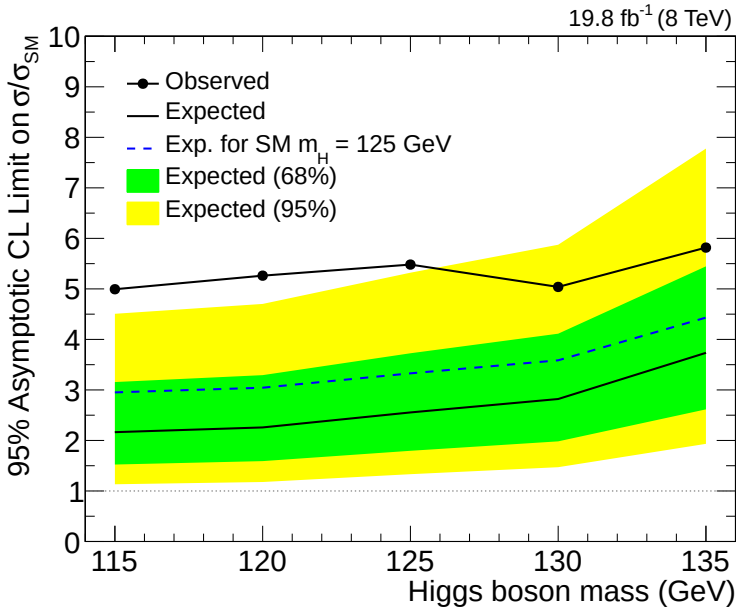


Figure 10.2: Expected and observed 95% asymptotic confidence level limits on the signal cross section times branching ratio, in units of the SM expected cross section, as a function of the Higgs boson mass, including all event categories. The limits in the presence of a SM Higgs boson with mass $m_H = 125$ GeV are indicated by the dashed curve.

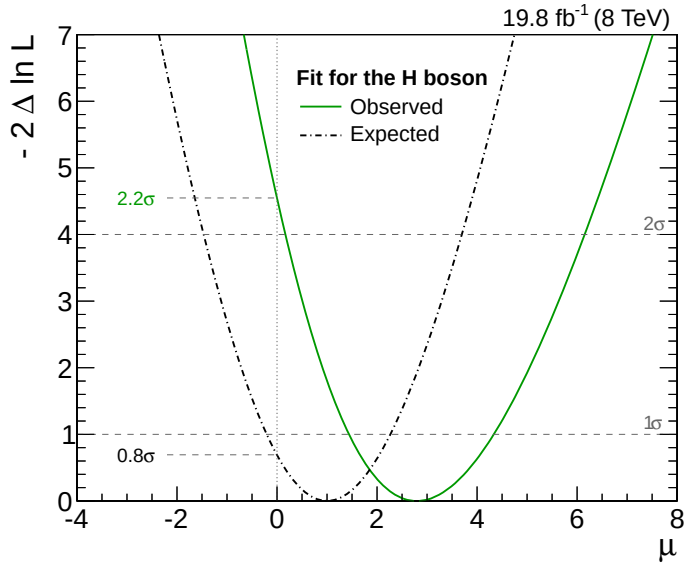


Figure 10.3: Expected and observed likelihood profile of signal strength $\mu = \sigma/\sigma_{\mathrm{SM}}$, for Higgs boson mass hypothesis $m_H = 125$ GeV.

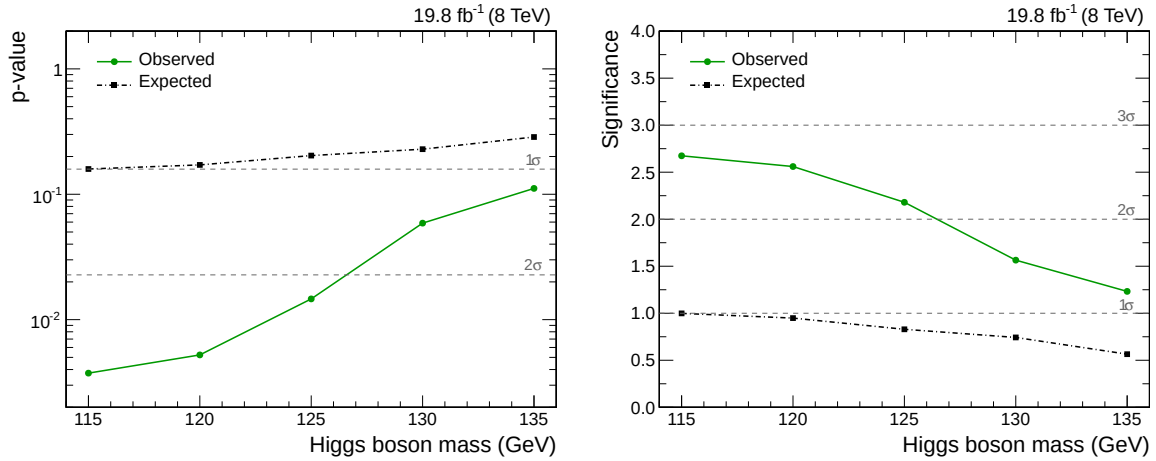


Figure 10.4: Expected and observed p-value (left) and significance (right), for Higgs boson mass hypothesis $m_H = 125$ GeV.

Further tests are executed probing the contribution of all individual event categories. Figure 10.7 shows the signal strength for each category. The global fit result is largely constrained by CAT2, CAT3 and CAT6, the most signal-like categories of both the Set A and Set B selections. In addition, the limits on the signal strength were derived excluding one or more categories. Figure 10.8 shows the upper limit per dataset, with either Set A or Set B excluded from the fit. The same is done excluding one category at a time in Fig. 10.9; reference categories CAT0 and CAT4 cannot be excluded, since they are required for the derivation of a QCD template.

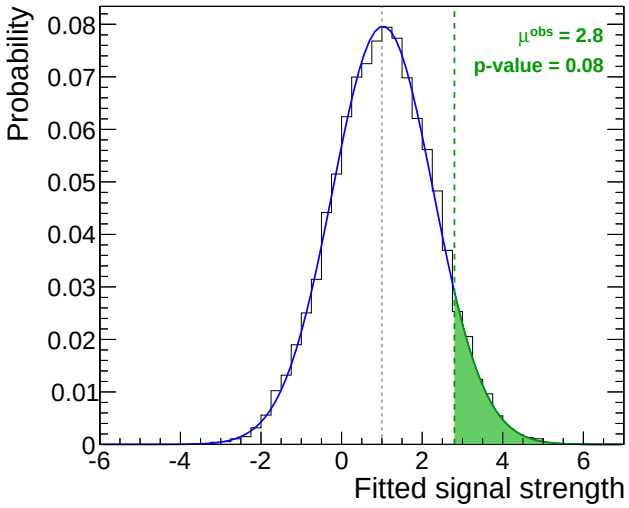


Figure 10.5: Probability distributions of the fitted signal strength of SM pseudo-experiments, for Higgs boson mass hypothesis $m_H = 125$ GeV. The integrated probability of an observation $\mu > 2.8$ is 8%.

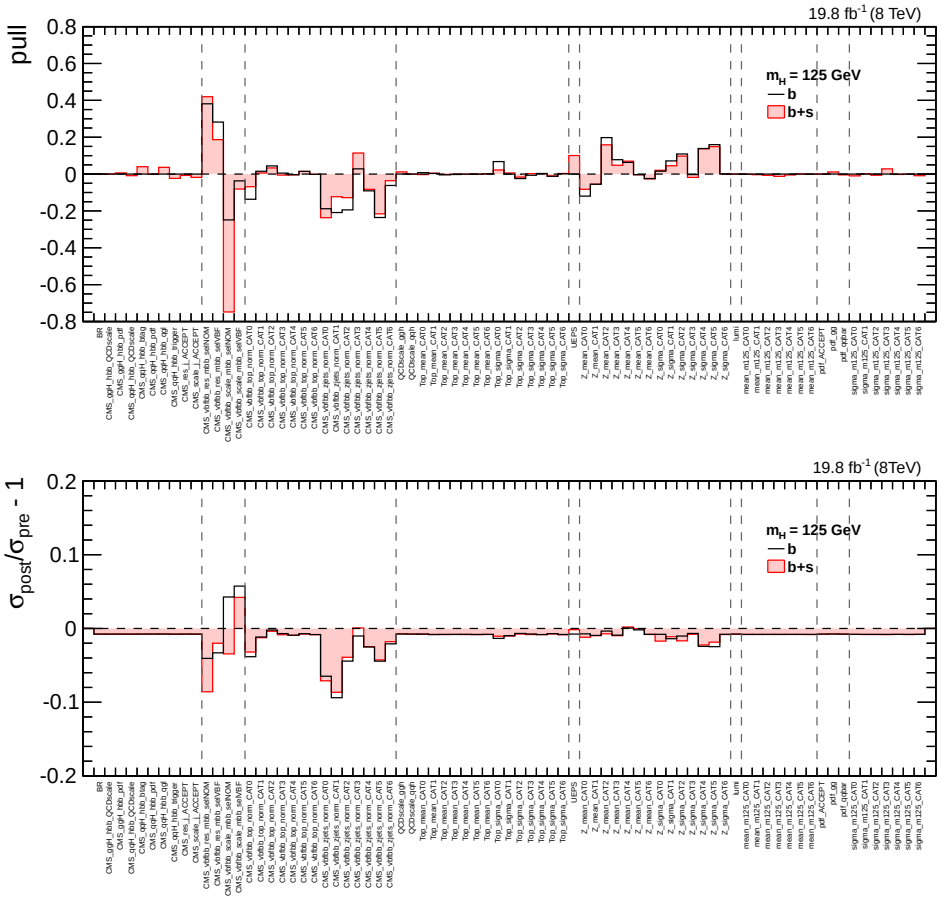


Figure 10.6: Post-fit nuisance pulls (top) and uncertainties (bottom) for the simultaneous fit of the b-quark pair invariant mass distribution in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 125$ GeV.

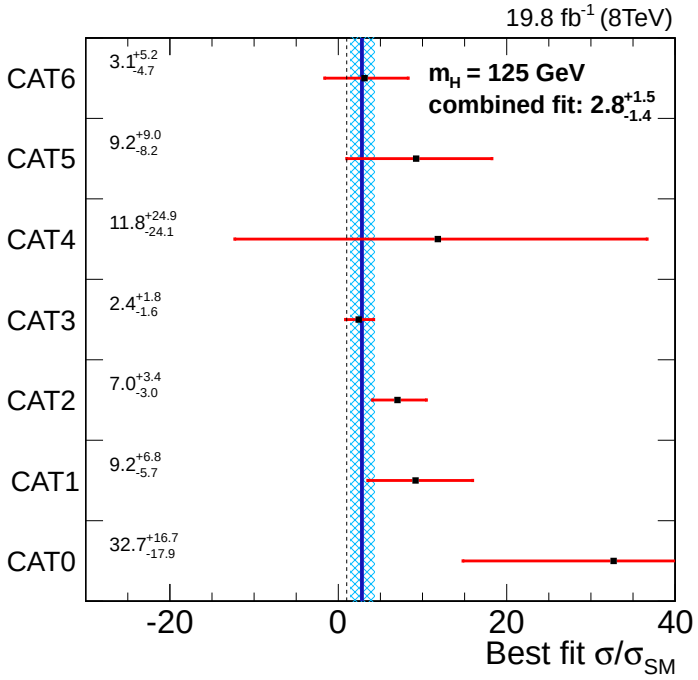


Figure 10.7: Fitted signal strength for each individual category and for all categories combined, in units of the SM expected cross section, $\mu = \sigma/\sigma_{\mathrm{SM}}$, for Higgs boson mass hypothesis $\mathrm{m_H} = 125$ GeV.

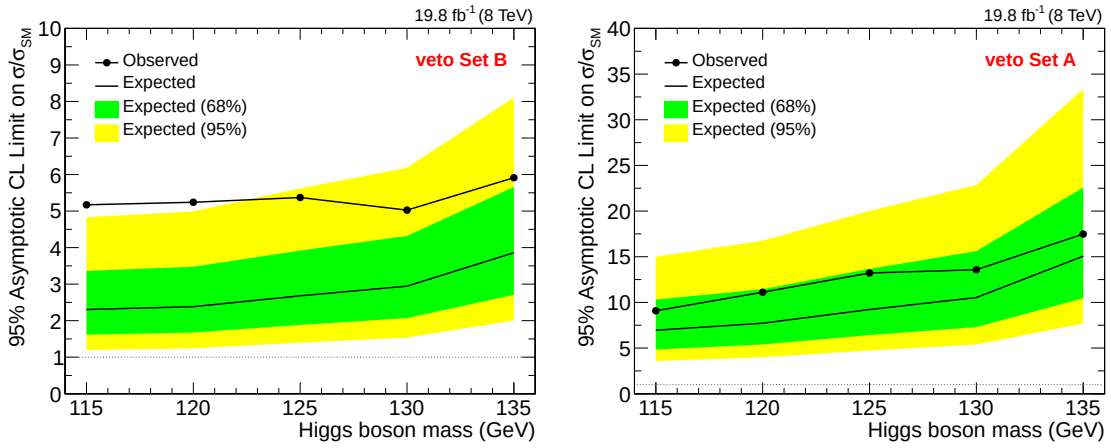


Figure 10.8: Expected and observed 95% confidence level limits on the signal cross section times branching ratio, in units of the SM expected cross section, as a function of the Higgs boson mass, using only the categories of Set A (CAT0-3, left) or Set B (CAT4-6, right).

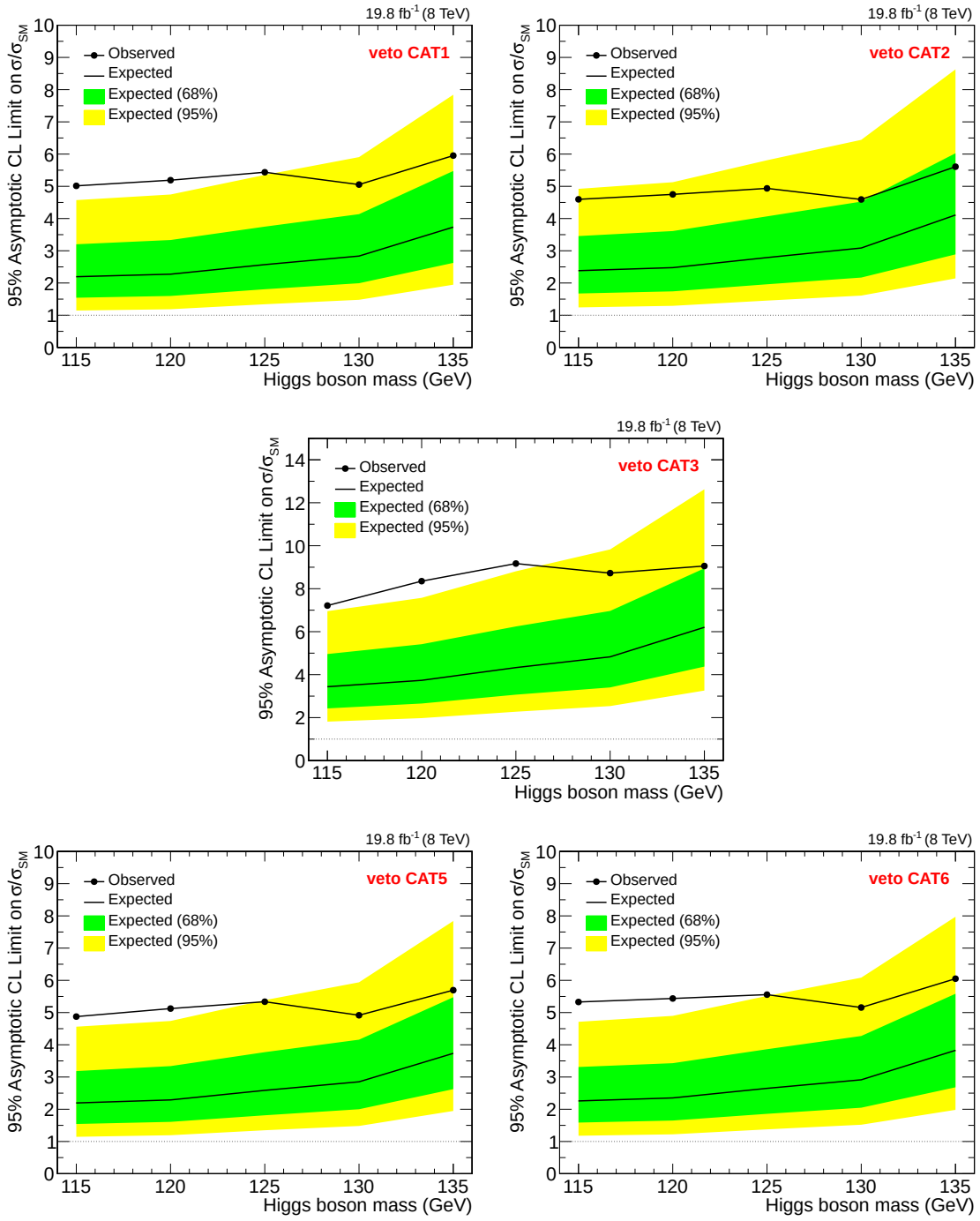


Figure 10.9: Expected and observed 95% confidence level limits on the signal cross section times branching ratio, in units of the SM expected cross section, as a function of the Higgs boson mass, excluding one category at a time: CAT1 (top left), CAT2 (top right), CAT3 (centre), CAT5 (bottom left) and CAT6 (bottom right). Reference categories CAT0 and CAT4 cannot be excluded.

10.2 Combination with other $H \to b\bar{b}$ results

Next to the VBF $H \to b\bar{b}$ analysis described in this thesis, searches have also been performed by the CMS collaboration for the $H \to b\bar{b}$ signal in the VH [120] and $t\bar{t}H$ [121, 122] production modes. The results of all three searches are combined by performing one global fit, which takes into account the correlations between the systematic uncertainties of the individual analyses. For the fit the data selections of each analysis were verified to be exclusive, such that no data is double-counted.

Before performing the combined fit, the fits of each individual analysis are repeated as a cross check. The results are listed in Table 10.2 for the $m_H = 125$ GeV mass point, and are fully compatible with the original results. The expected significance indicates that the new VBF $H \to b\bar{b}$ result is expected to be the second largest contribution to the combination, after the VH result, and before the $t\bar{t}H$ result.

Table 10.2: Summary of the upper limits on the signal cross section times branching ratio, the significance, and the best fitted signal strength, in units of σ_{SM}, for the individual channels of the $H \to b\bar{b}$ combination.

Channel	Exp. limit	Obs. limit	Exp. Sign.	Obs. Sign.	Fitted μ
$VH \to b\bar{b}$	0.85	1.68	2.52	2.08	$0.89^{+0.46}_{-0.44}$
$t\bar{t}H \to b\bar{b}$	3.48	4.09	0.58	0.37	$0.67^{+1.82}_{-1.81}$
VBF $H \to b\bar{b}$	2.55	5.48	0.83	2.18	$2.79^{+1.53}_{-1.35}$
Combined	0.78	1.77	2.70	2.55	$1.04^{+0.44}_{-0.42}$

The results of the combined fit are then summarised in Table 10.3. The fit yields a best-fitted signal strength $\mu = \sigma/\sigma_{SM} = 1.04^{+0.44}_{-0.42}$, compatible with the standard model expectation of $\mu = 1$. The expected and observed limits on the cross section times branching ratio are included in Table 10.3 and shown in Fig. 10.10. The observed limit ranges from 1.5 to 3.2 between the lowest and the highest mass point (115-135 GeV), while the expected limit ranges between 0.7 and 1.4.

The observed significance of the combined result is 2.6 standard deviations, for an expected significance of 2.7 standard deviations. Figure 10.11 shows the expected and observed likelihood profiles of the signal strength, for all channels of the $H \to b\bar{b}$ combination separately and the combined result itself. The significances and matching p-values of the combined result are given for all mass points in Fig. 10.12.

Table 10.3: Summary of the upper limits on the signal cross section times branching ratio, the significance, the p-value, and the best fitted signal strength, in units of σ_{SM}, for the $H \to b\bar{b}$ combination.

m_H (GeV)	Exp. limit	Obs. limit	Fitted μ
115	0.66	1.45	$0.83^{+0.38}_{-0.35}$
120	0.72	1.73	$1.06^{+0.43}_{-0.44}$
125	0.78	1.77	$1.04^{+0.44}_{-0.42}$
130	1.01	2.72	$1.72^{+0.58}_{-0.56}$
135	1.35	3.17	$1.89^{+0.96}_{-0.89}$

m_H (GeV)	Exp. sign.	Obs. sign.	Exp. p-value	Obs. p-value
115	3.01	2.37	0.00129	0.00901
120	2.93	2.74	0.00128	0.00312
125	2.70	2.57	0.00346	0.00513
130	2.03	3.21	0.0214	0.000655
135	1.68	2.54	0.0465	0.00561

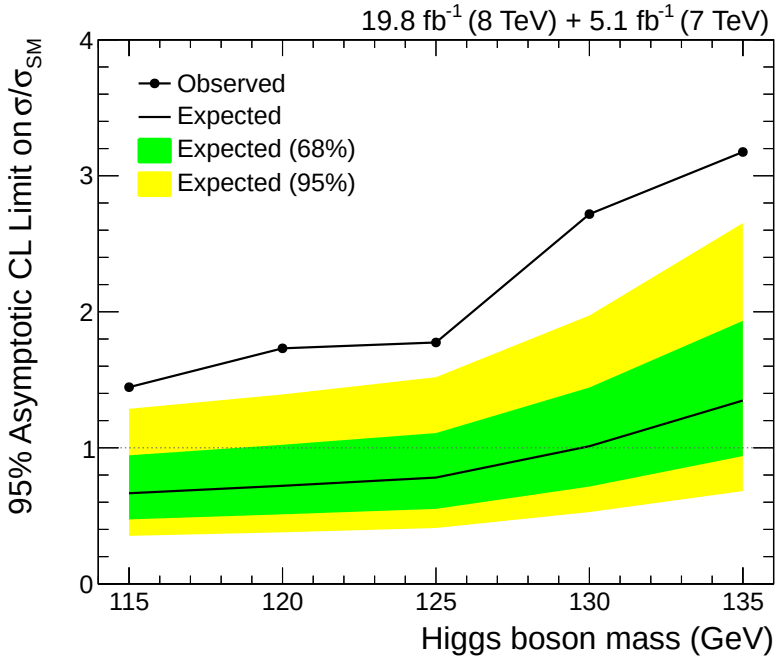


Figure 10.10: Expected and observed 95% confidence level limits on the signal cross section times branching ratio, in units of the SM expected cross section, as a function of the Higgs boson mass, for the $H \to b\bar{b}$ combination.

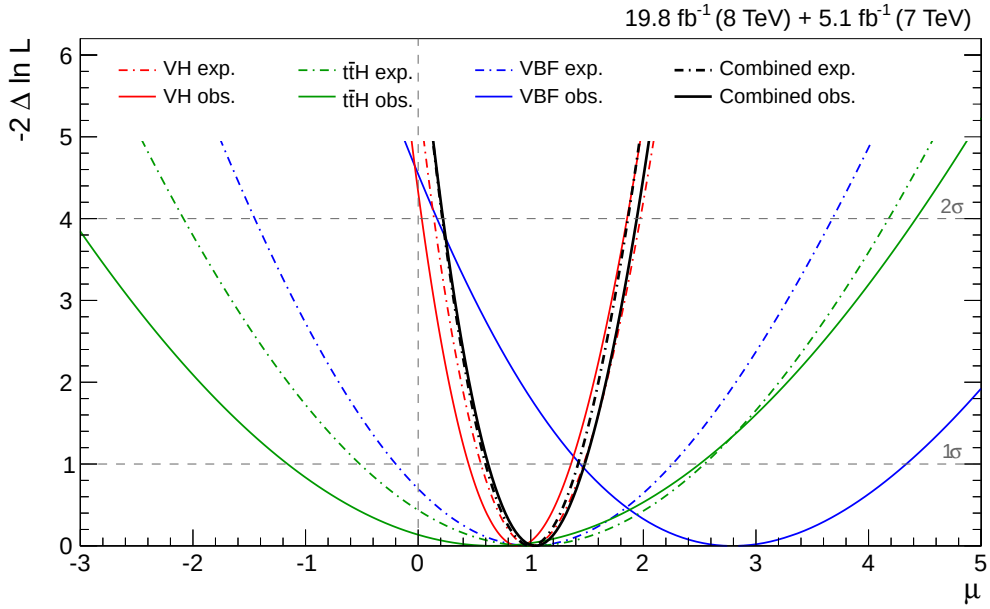


Figure 10.11: Expected and observed likelihood profile of signal strength $\mu = \sigma/\sigma_{\mathrm{SM}}$, for Higgs boson mass hypothesis $m_H = 125$ GeV, for the $H \to b\bar{b}$ combination.

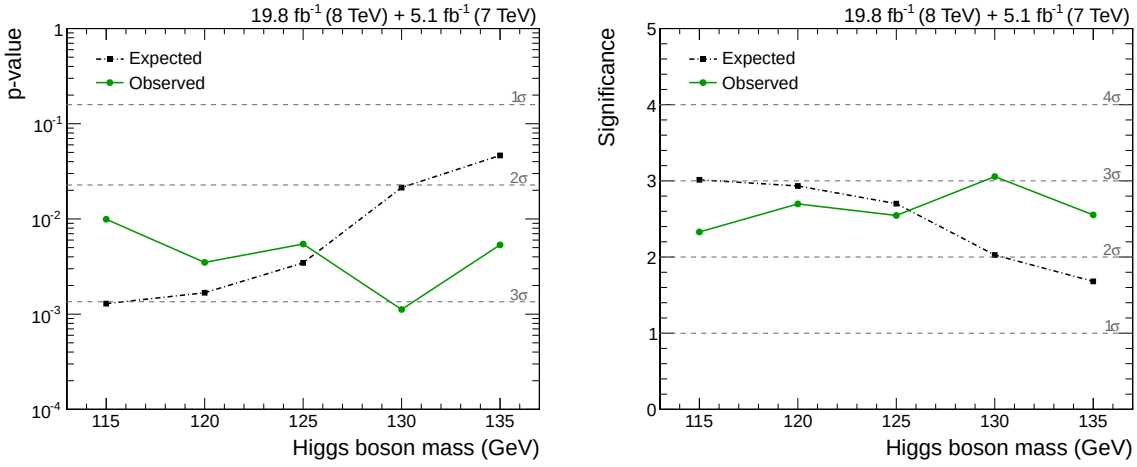


Figure 10.12: Expected and observed p-value (left) and significance (right), for Higgs boson mass hypothesis $m_H = 125$ GeV, for the $H \to b\bar{b}$ combination.

Chapter 11

Summary and Outlook

The analysis presented in this thesis is the search for the standard model Higgs boson, produced by vector boson fusion, and decaying to bottom quarks [116]. The search was executed using collision data from the LHC experiment, collected by the CMS detector in 2012, during LHC Run I, at a centre-of-mass energy $\sqrt{s} = 8$ TeV. The dataset corresponds to an integrated luminosity of 19.8 (nominal) + 18.3 (parked) fb^{-1}.

The standard model Higgs boson was first observed by CMS in the $\gamma\gamma$, ZZ and WW decay channels. Search have been performed in the $\text{b}\bar{\text{b}}$ decay channel as well, using the VH and $\text{t}\bar{\text{t}}\text{H}$ production modes, but so far no conclusive $\text{H} \to \text{b}\bar{\text{b}}$ observation has been made. While it has a much larger production cross section times branching ratio, the sensitivity of the VBF $\text{H} \to \text{b}\bar{\text{b}}$ analysis is lower than that of the other $\text{H} \to \text{b}\bar{\text{b}}$ channels, due to the fact that it considers a fully hadronic final state. Still, an $\text{H} \to \text{b}\bar{\text{b}}$ analysis in the VBF production mode has the potential of adding the required sensitivity to build towards an observation of the Higgs boson decaying to fermions. It is also the first all hadronic final state Higgs boson search at the LHC.

The largest challenge of the VBF $\text{H} \to \text{b}\bar{\text{b}}$ analysis is the very large QCD background to the all hadronic final state. The analysis strategy consists of three stages. First, data is selected online using dedicated triggers, exploiting the specific properties of the VBF mechanism as well as using b-tagging; offline, an additional set of preselection criteria is applied, chosen according to the measured trigger efficiency. Secondly, additional variables are constructed, building on properties of signal and background components; a large set of input variables showing good discriminating power are then used to train a multivariate signal versus background discriminant. This last step is done avoiding any correlation to the b-quark pair invariant mass, such that in the third and final stage, a fit of the invariant mass spectrum can be made.

I have been part of the VBF $\text{H} \to \text{b}\bar{\text{b}}$ group since the start of the analysis in 2011, at the beginning of my PhD. A first rough version of the analysis was completed in 2013, but it took until mid 2015 before the second, optimised version was published. At all times, I was involved in writing the documentation, presenting the results to the wider collaboration, and getting the paper ready for publication. I contributed to the development of the dedicated VBF $\text{H} \to \text{b}\bar{\text{b}}$ trigger, and was responsible for evaluating trigger efficiencies and determining data vs. simulation trigger scale factors once data taking had started. The trigger efficiency studies were also relevant in the optimisation of the offline preselection criteria. After my work on the trigger, I validated the selected datasets by constructing data versus simulation comparisons. In the second half of my PhD, I focused on the global fit of the invariant mass spectrum, the third stage of the analysis. During the development stages of the fit framework, I worked on the derivation and optimisation of fit templates for the signal and background components, the

implementation of transfer functions for the QCD background templates. In addition, I was involved in the validation of the fit procedure, more specifically in the bias studies that were performed for the polynomial order of the QCD background template and the transfer functions. Once the fit framework was set up, I evaluated various systematic uncertainties which were to be considered in the global fit, including the uncertainties on the jet energy scale and resolution, the uncertainties due to the trigger, and the PDF uncertainties. Once all inputs were ready, I performed the simultaneous binned maximum likelihood fits of the invariant mass spectrum, derived upper limits on the production cross section times branching ratio and determined the significance of the fit results. Finally, I was involved in the $H \to b\bar{b}$ combination analysis: I studied the systematics of the three contributing analyses to identify correlations, and prepared input configurations for the combined global fit.

The VBF $H \to b\bar{b}$ analysis yields a best-fitted signal strength $\mu = 2.8^{+1.5}_{-1.4}$, with an observed significance of 2.2 standard deviations (the expected significance is 0.8 standard deviations). The compatibility of such an observation with the standard model expectation of $\mu = 1$ is 8%. Upper limits were set on the production cross section times branching ratio, in the 115-135 GeV mass range. An excess is observed for all mass points: the expected limits range between 2.2 and 3.7, while the observed limits range between 5.0 and 5.8. The significance of this result is not large enough to claim an observation of the Higgs boson in the VBF $H \to b\bar{b}$ channel.

The best-fitted signal strength for the $H \to b\bar{b}$ combination, combining the VH, $t\bar{t}H$ and VBF production modes, is 1.0 ± 0.4, with observed (expected) significance of 2.6 (2.7) standard deviations.

More recently, also a combination of the $b\bar{b} + \tau\tau$ results was performed, which finds a best-fitted signal strength $\mu = 0.9 + 0.2$, with an observed (expected) significance of 4.2 (4.7) standard deviations [146]. With this combination, the CMS collaboration has provided the first tentative proof of the Higgs boson decaying to fermions.

Looking forward, the search for the Higgs boson decaying to $b\bar{b}$ in the VBF production mode will be continued at the LHC. The additional data at higher centre-of-mass values, should make a 5σ observation of the Higgs boson decay to fermions possible in the near future. In general, many studies performed during LHC Run II will aim to measure in detail the properties of the observed boson: spin, parity, width (lifetime), and cross section. After that, and also already at the same time, the way lies open for searches for additional Higgs bosons, Dark Matter, and more exotic New Physics.

Appendix A

Event interpretation discriminant

In the implementation of the b-likelihood event interpretation discriminant, a coding mistake led to the wrong values being filled for the b-tag and pseudorapidity rank of the jets. Effectively, only two of the four inputs to the discriminant had real merit, the other two were insignificant. As a result, the gain obtained from the use of the event discriminant, was smaller than it could have been. The result with the suboptimal discriminant however, was still better than it would have been had the simple b-tag ordering interpretation been used. For completeness, the event discriminant input variables and performance with the corrected implementation is shown in Figs. A.1-A.4. Due to time limitations, it is not feasible to also evaluate the effect of this change on the entire analysis chain.

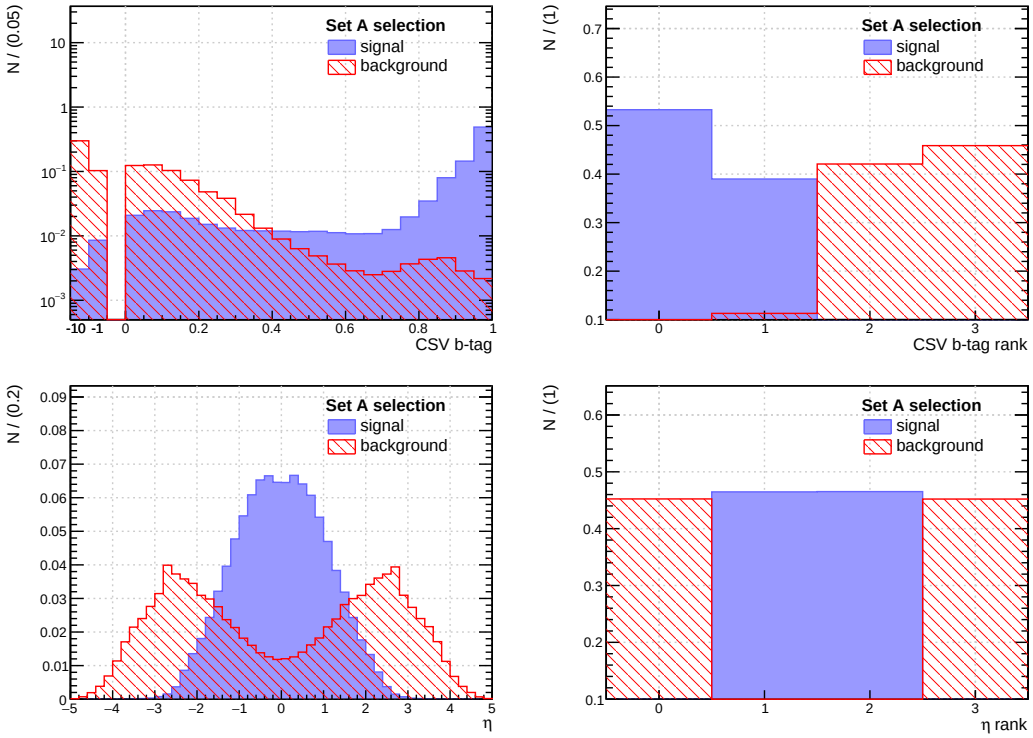


Figure A.1: Input variables for the event interpretation b-likelihood discrimininant, for Set A (after trigger selection). Signal consists of jets matched to b-partons, background contains all other jets; using the VBF $H \to b\bar{b}$ signal sample for $m_H = 125$ GeV. (corrected)

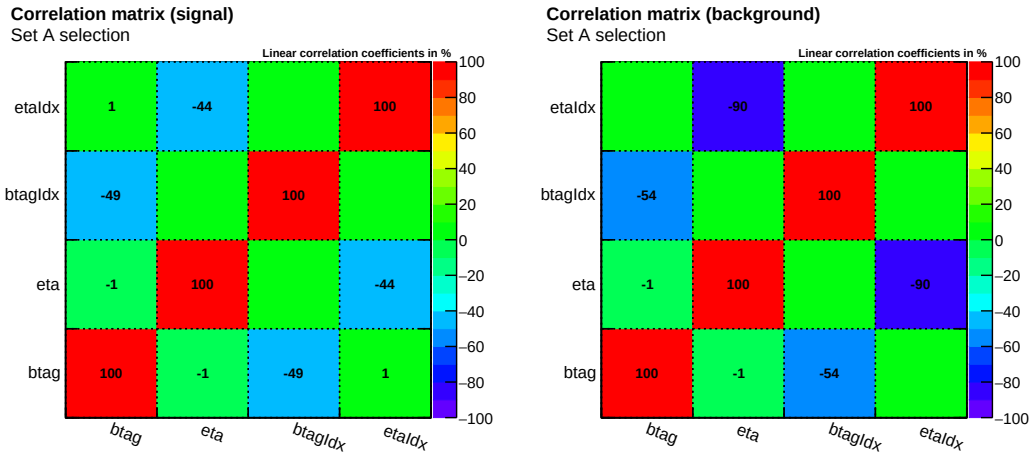


Figure A.2: Linear correlation coefficients between the event interpretation b-likelihood discriminant, for Set A (after trigger selection). (corrected)

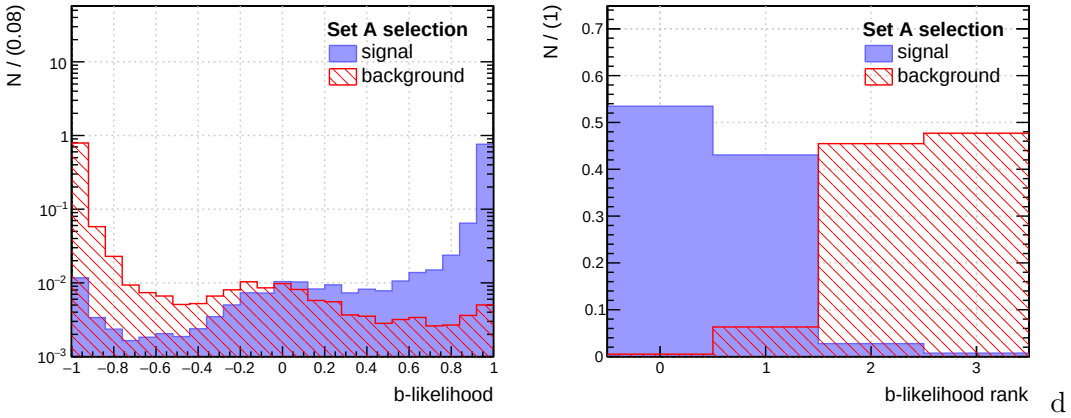


Figure A.3: Output of the event interpretation b-likelihood discriminant, for Set A (after trigger selection). (corrected)

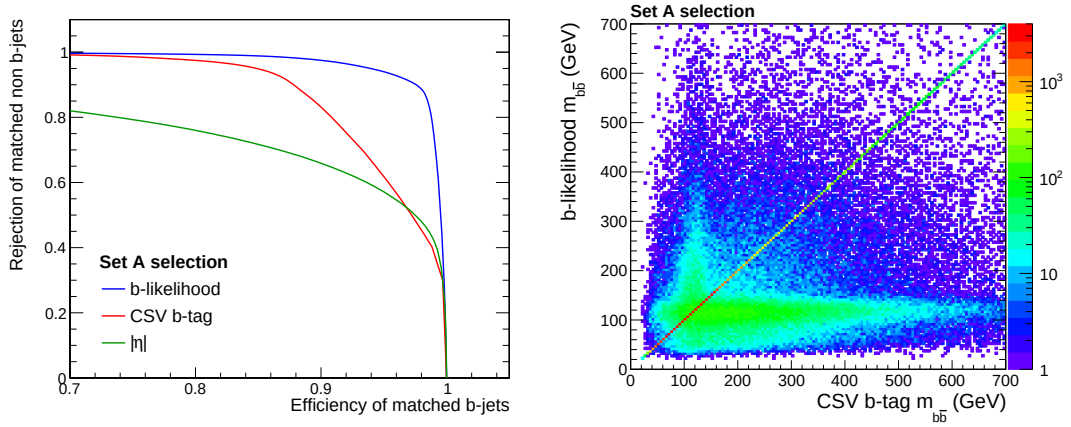


Figure A.4: Background rejection vs. signal efficiency (left) and Comparison of b-quark pair invariant mass $m_{b\bar{b}}$ in the CSV b-tag interpretation and the b-likelihood interpretation (right), for Set A (after trigger selection). (corrected)

Appendix B

Data vs. simulation comparisons: top control sample

As mentioned in section 8.3, an additional set of comparison plots is constructed for a phase space region where the QCD background contribution is greatly suppressed, allowing the validation of simulation samples without needing to take into account the leading order limitations of the QCD simulation sample. Especially for important distributions such as the multivariate signal vs. background discriminant output and the b-quark pair invariant mass, having a validation in a second phase space is interesting.

The chosen alternative selections (Top A and Top B) ask for exactly one lepton (with isolation $<$ 0.15, tight lepton ID, $p_T > 20$ GeV and $|\eta| < 2.4$), $E_T^{miss} > 45$ GeV, and respectively the standard Set A and Set B selection. This leads to a $t\bar{t}$ dominated sample, with some contribution from single-top events. QCD multijet events are suppressed by the lepton and E_T^{miss} requirements, W(lν)+jets events are suppressed by the b-tagging and jet multiplicity requirements, and Z(ll)+jets events are suppressed by the veto on additional leptons and the E_T^{miss} requirements.

B.1 Set A selection (top control region)

Figures B.1-B.7 show the comparisons for the Top A selection. The following groups of variables are included:

- qq$'$ system properties (Fig. B.1)
- $b\bar{b}$ system properties (Fig. B.2)
- QGL discriminant output (Fig. B.3)
- b-likelihood scores (Fig. B.4)
- CSV b-tag values (Fig. B.5)
- hadronic activity (Fig. B.6)
- multivariate discriminants (Fig. B.14)

The agreement observed between data and simulation is good in all distributions, except for the ones related to the additional hadronic activity (Fig. B.6). This can largely be attributed to the choice of the underlying event model in the simulation, which doesn't describe the additional hadronic activity in the $t\bar{t}$ sample very well. The observed activity is lower than expected, which makes the data look more signal-like. This is confirmed by the multivariate discriminant output distribution (Fig. B.7).

B.2 Set B selection (top control region)

Figures B.8-B.14 show the comparisons for the Top B selection. The following groups of variables are included:

- qq' system properties (Fig. B.8)
- $b\bar{b}$ system properties (Fig. B.9)
- QGL discriminant output (Fig. B.10)
- b-likelihood scores (Fig. B.11)
- CSV b-tag values (Fig. B.12)
- hadronic activity (Fig. B.13)
- multivariate discriminants (Fig. B.14)

Also in the top control region the same conclusions can be drawn for the Top B selection comparisons as for the Top B selection ones. The overall agreement between data vs. background simulation is good, except in the additional hadronic activity distributions (Fig. B.13).

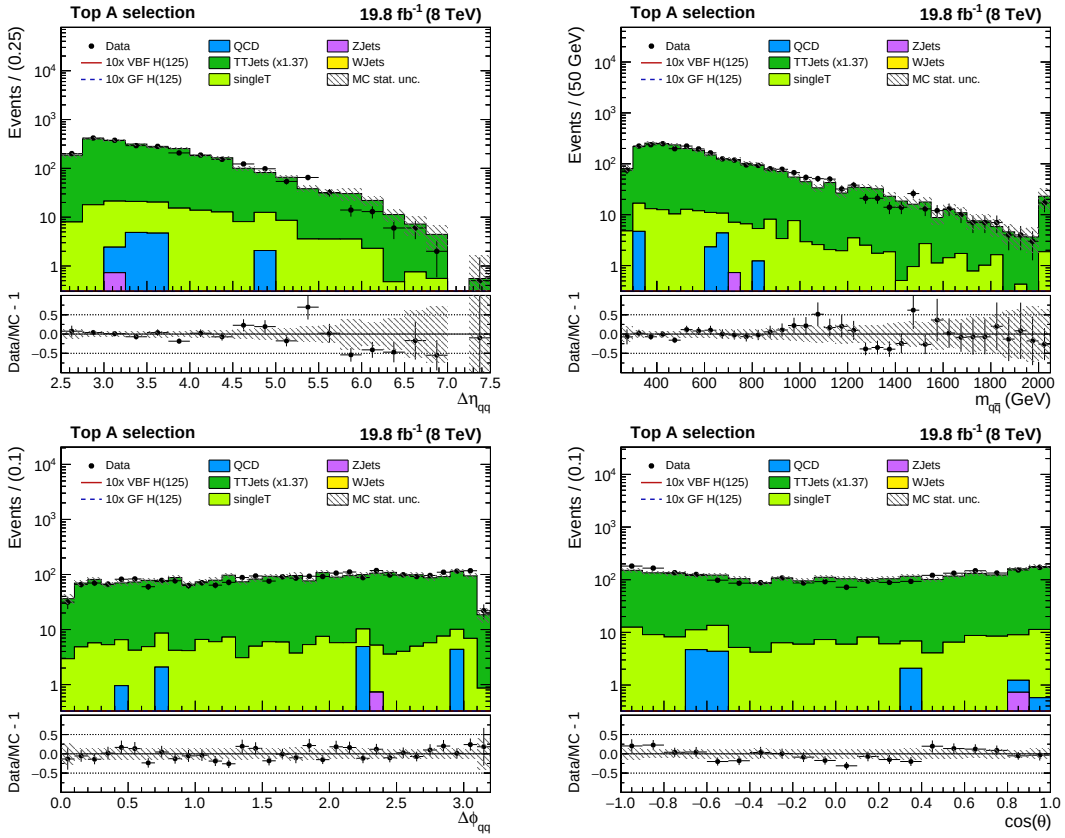


Figure B.1: Kinematic properties of the qq′ system (pseudorapidity separation, invariant mass, angular separation, and angle with the b$\bar{\text{b}}$ system, for the Top A selection.

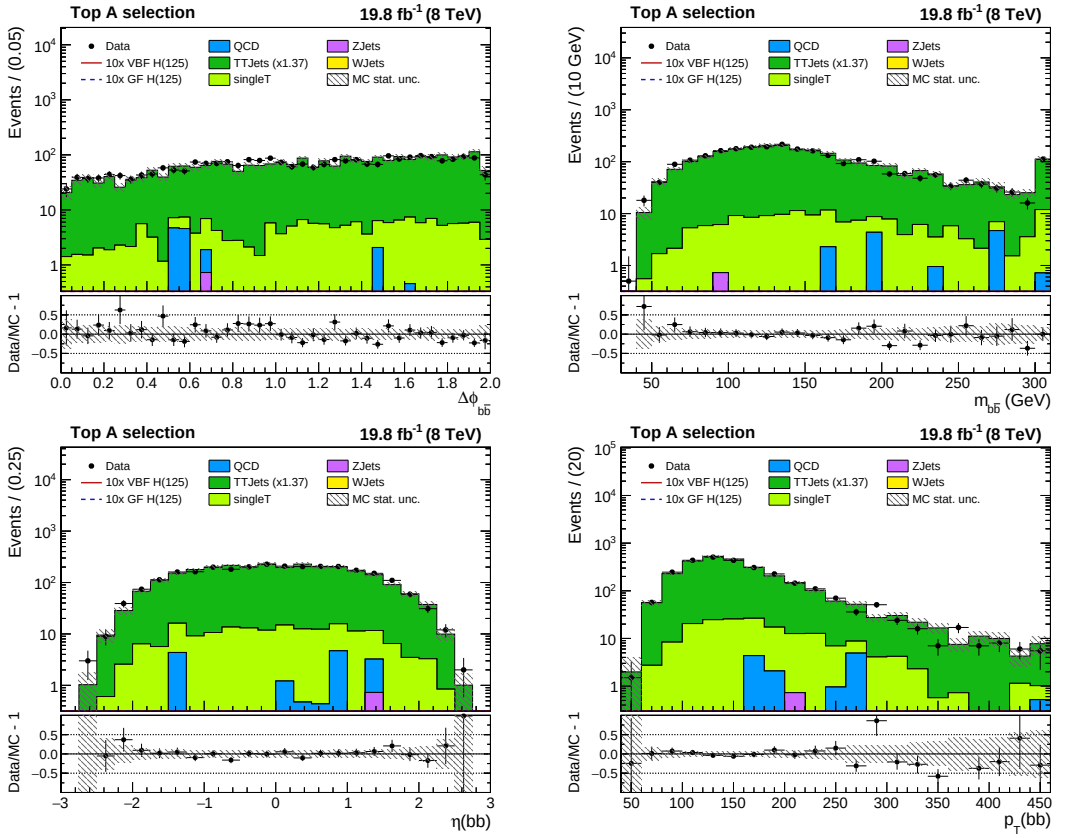


Figure B.2: Kinematic properties of the b$\bar{\text{b}}$ system (angular separation, invariant mass, pseudorapity, and transverse momentum), for the Top A selection.

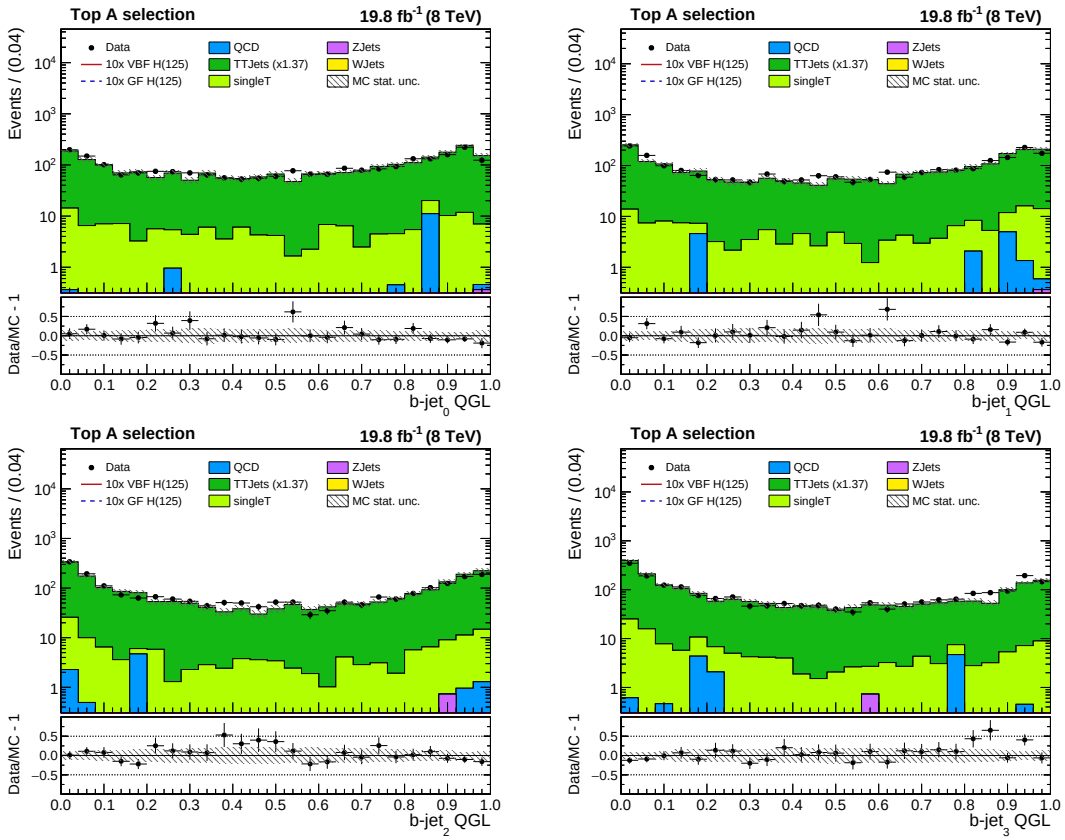


Figure B.3: Distribution of the quark-gluon likehood discriminant value of the jets, ordered by b-likelihood value, for the Top A selection.

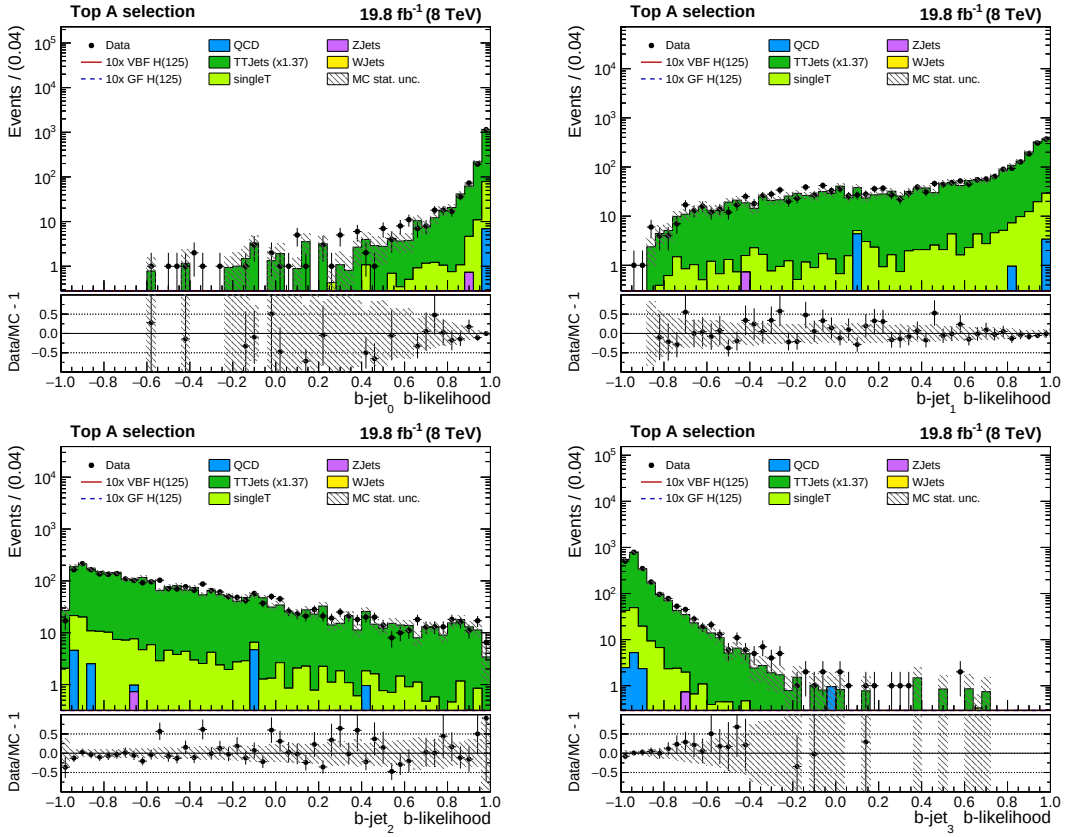


Figure B.4: Distributions of the b-likelihood discriminant values of the jets, ordered by b-likelihood discriminant value, for the Top A selection.

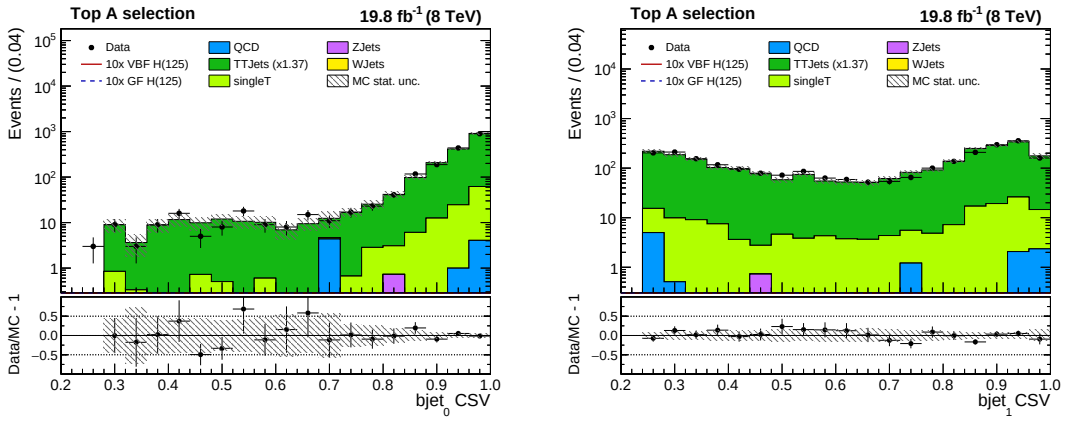


Figure B.5: CSV b-tag value distributions of the two leading b-jets, ordered by CSV b-tag value, for the Top A selection.

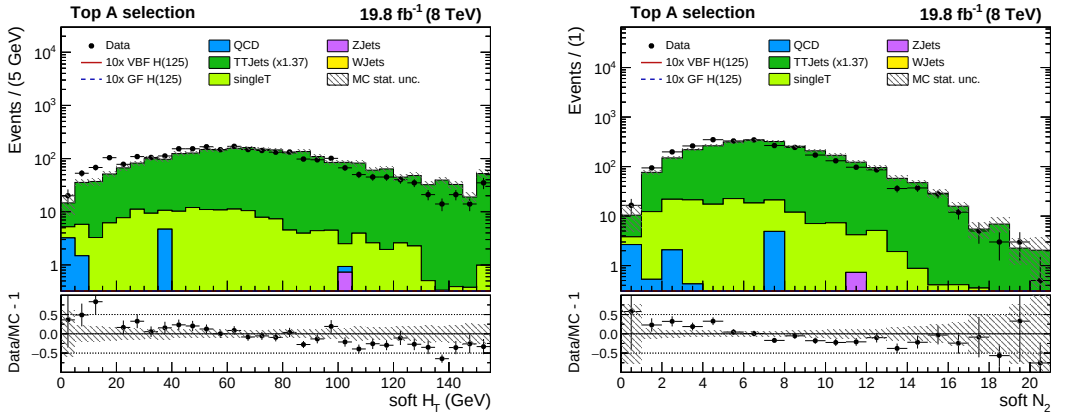


Figure B.6: Soft track activity momentum sum and multiplicity distributions, for the Top A selection.

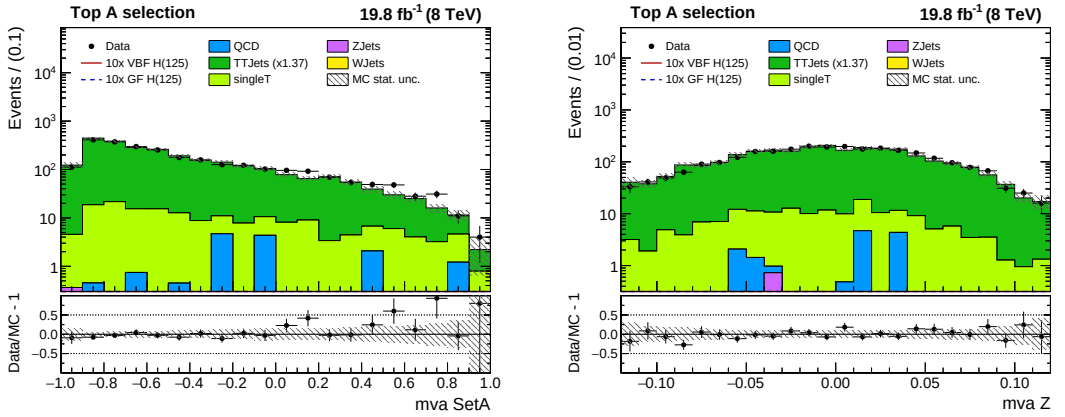


Figure B.7: Multivariate discriminant output values (BDT and Fisher discriminant), for the Top A selection.

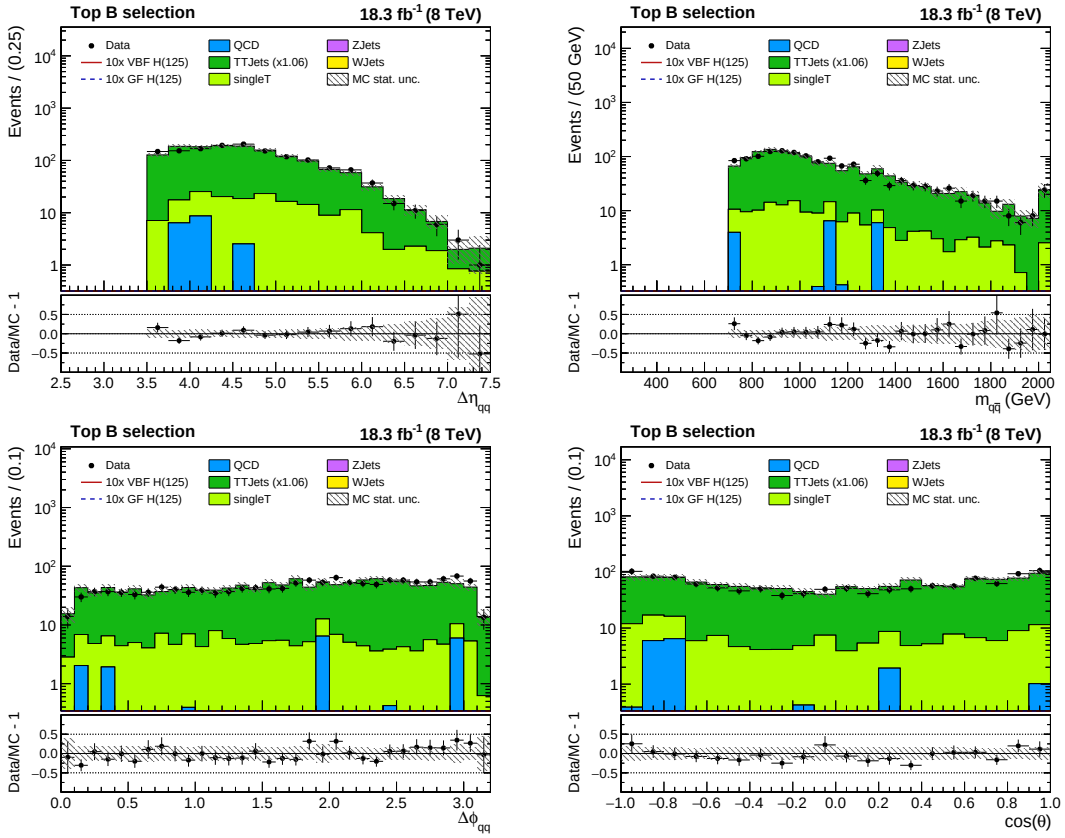


Figure B.8: Kinematic properties of the qq′ system (pseudorapidity separation, invariant mass, angular separation, and angle with the $b\bar{b}$ system, for the Top B selection.

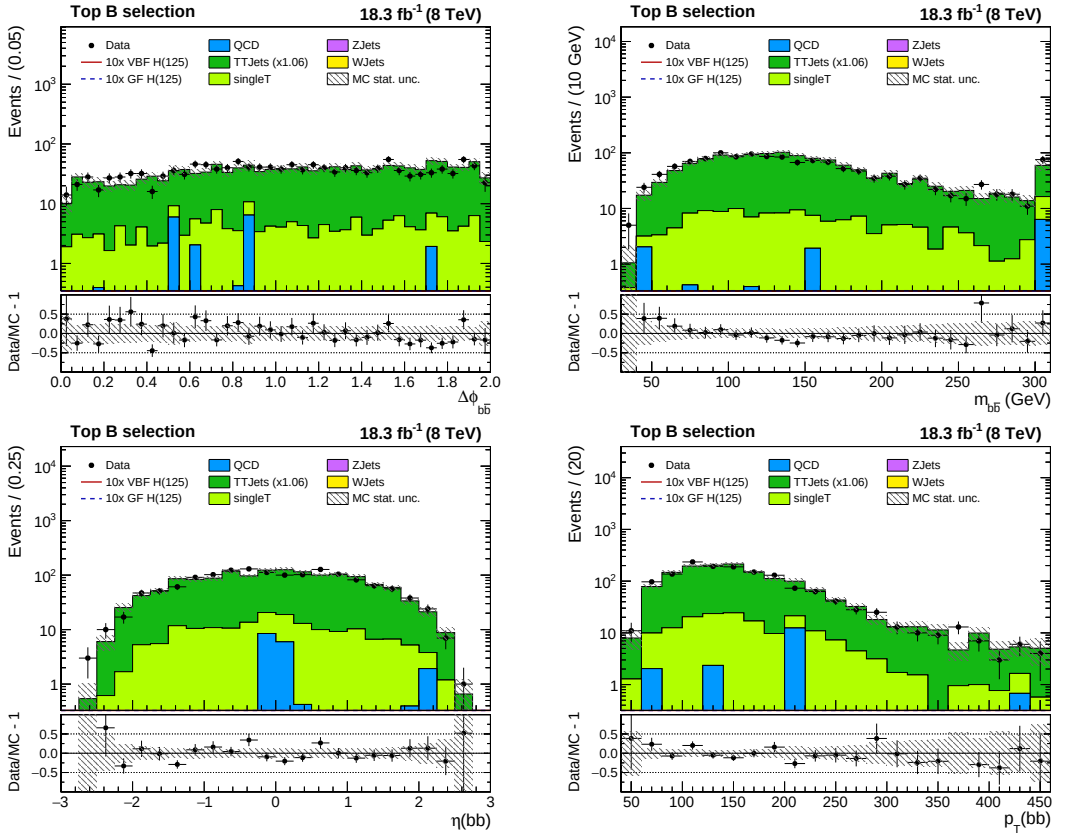


Figure B.9: Kinematic properties of the $b\bar{b}$ system (angular separation, invariant mass, pseudorapity, and transverse momentum), for the Top B selection.

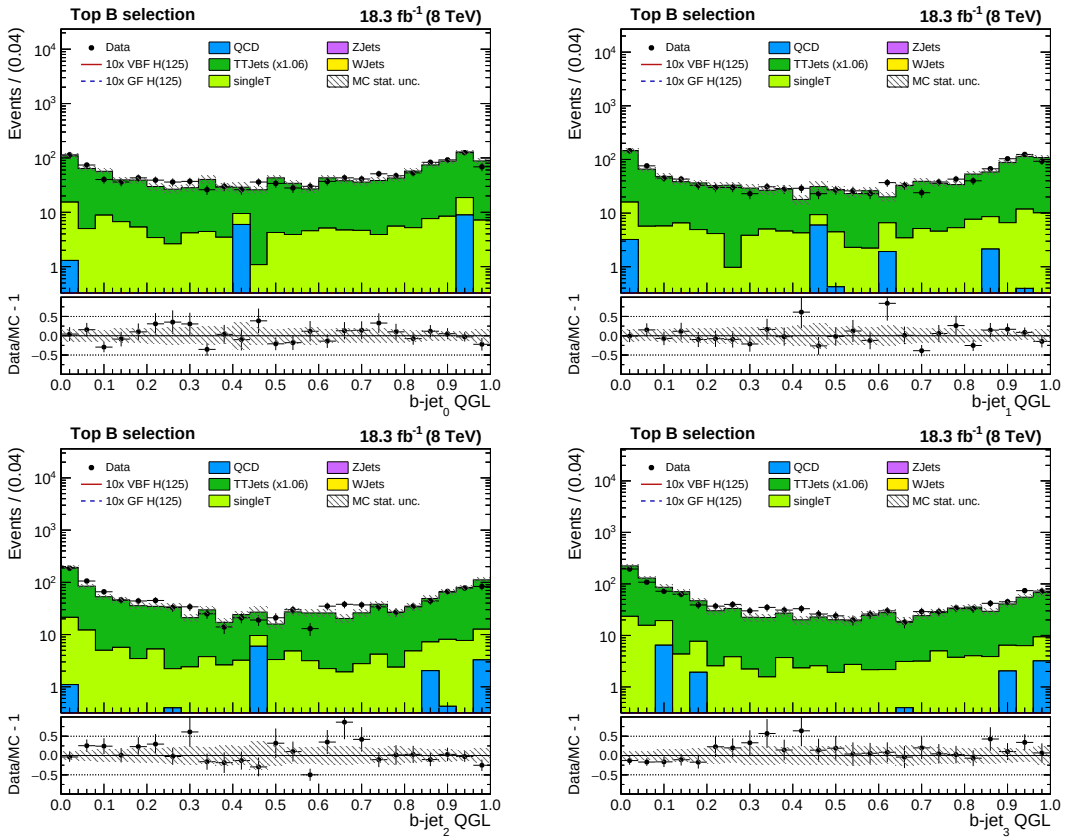


Figure B.10: Distribution of the quark-gluon likehood discriminant value of the jets, ordered by b-likelihood value, for the Top B selection.

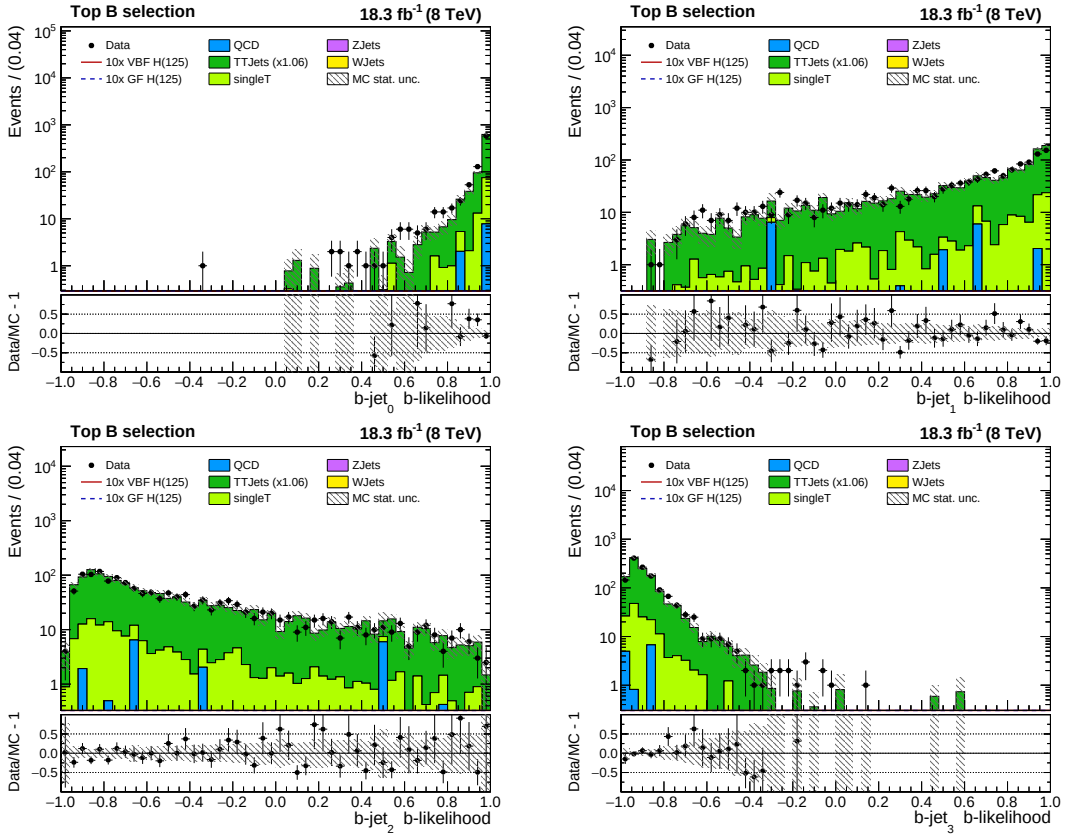


Figure B.11: Distributions of the b-likelihood discriminant values of the jets, ordered by b-likelihood discriminant value, for the Top B selection.

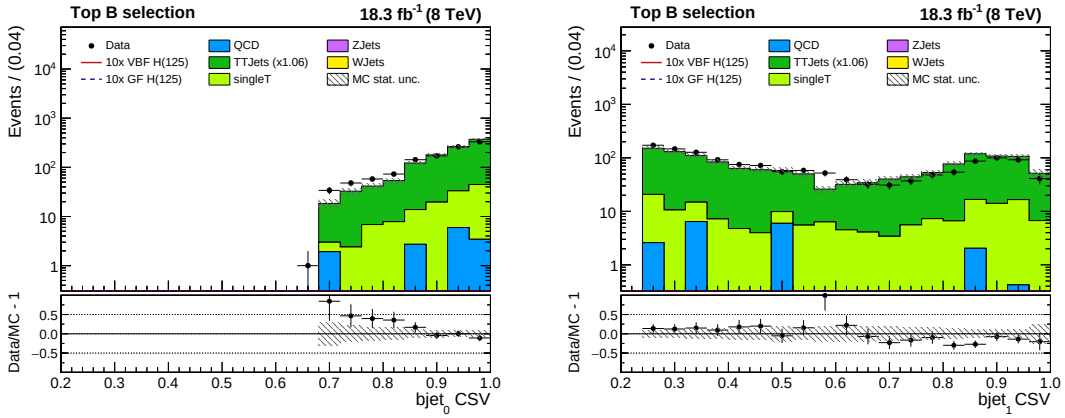


Figure B.12: CSV b-tag value distributions of the two leading b-jets, ordered by CSV b-tag value, for the Top B selection.

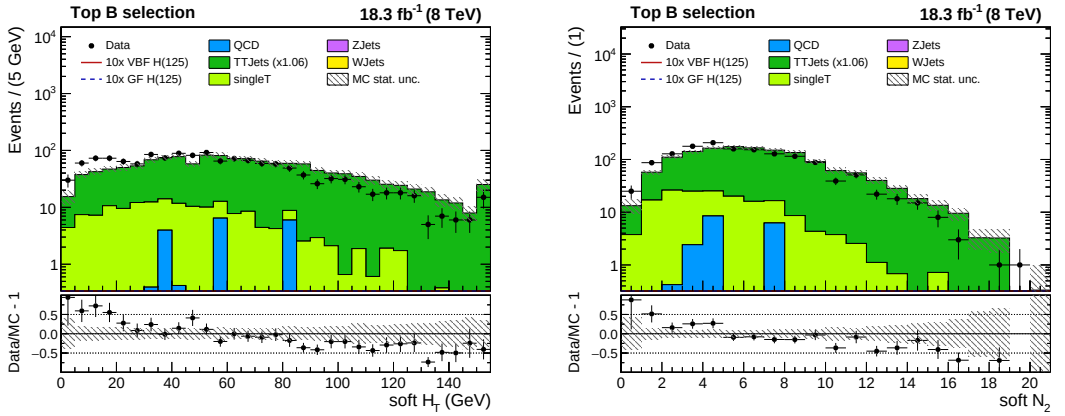


Figure B.13: Soft track activity momentum sum and multiplicity distributions, for the Top B selection.

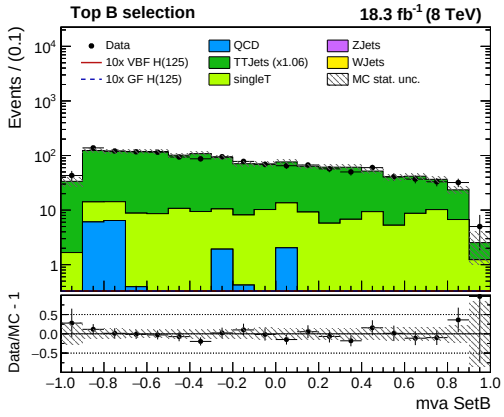


Figure B.14: Multivariate discriminant output value (BDT), for the Top B selection.

Appendix C

Alternate Higgs boson mass hypotheses in the VBF $H \to b\bar{b}$ search

The global fit for the VBF $H \to b\bar{b}$ analysis was performed for five different Higgs boson mass hypotheses, $m_H = 115$-135 GeV, in steps of 5 GeV. The signal templates and results for the main mass point, $m_H = 125$ GeV, were given in section 9.4.4 and section 10.1 respectively.

The first section of this appendix lists the signal templates for the remaining mass points: Figs. C.1 to C.8.

In the second section, the results for the remaining mass points are given. Figures C.9a to C.12b show the fitted distributions, Figs. C.13 to C.16 summarise the nuisance parameters, and Figs. C.17 and C.18 illustrate the contributions of the individual categories to each fit.

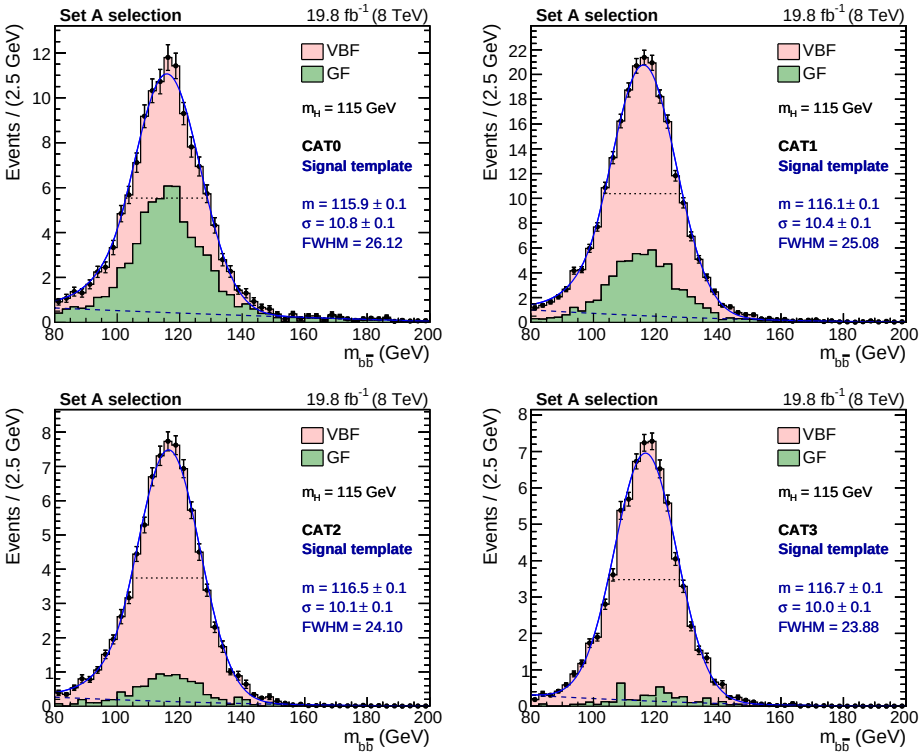


Figure C.1: Templates for the b-quark pair invariant mass distribution for the Set A selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 115$ GeV.

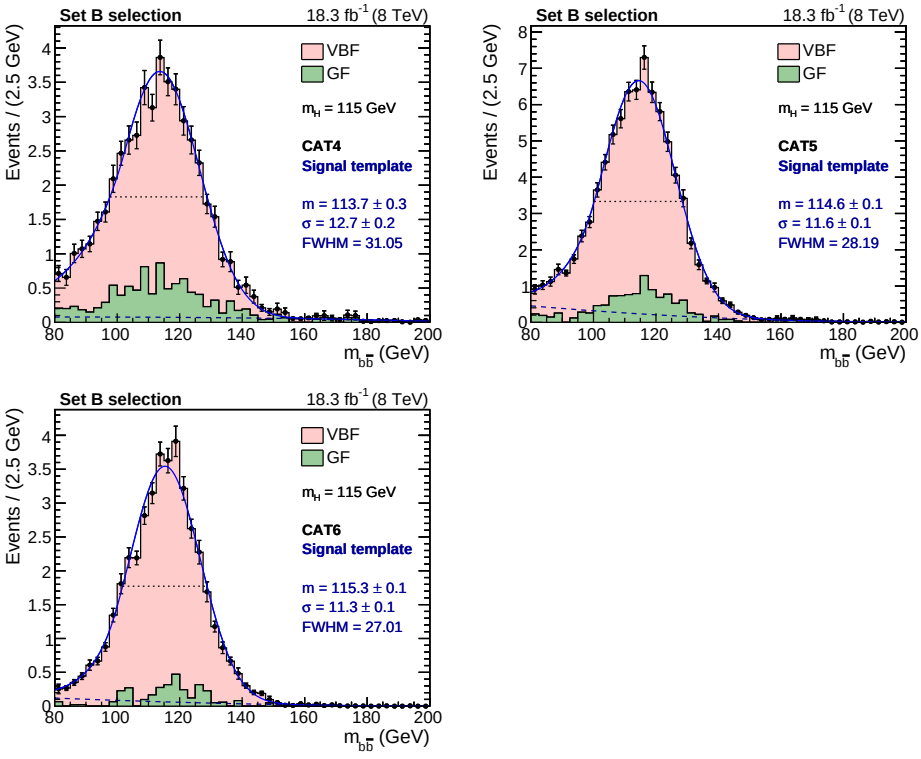


Figure C.2: Templates for the b-quark pair invariant mass distribution for the Set B selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 115$ GeV.

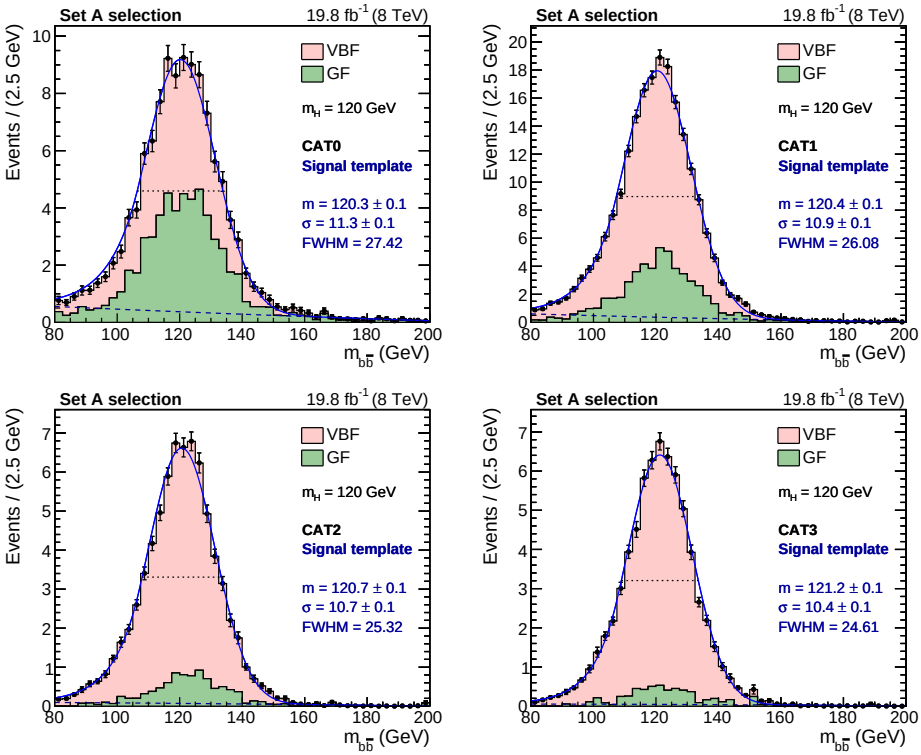


Figure C.3: Templates for the b-quark pair invariant mass distribution for the Set A selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 120$ GeV.

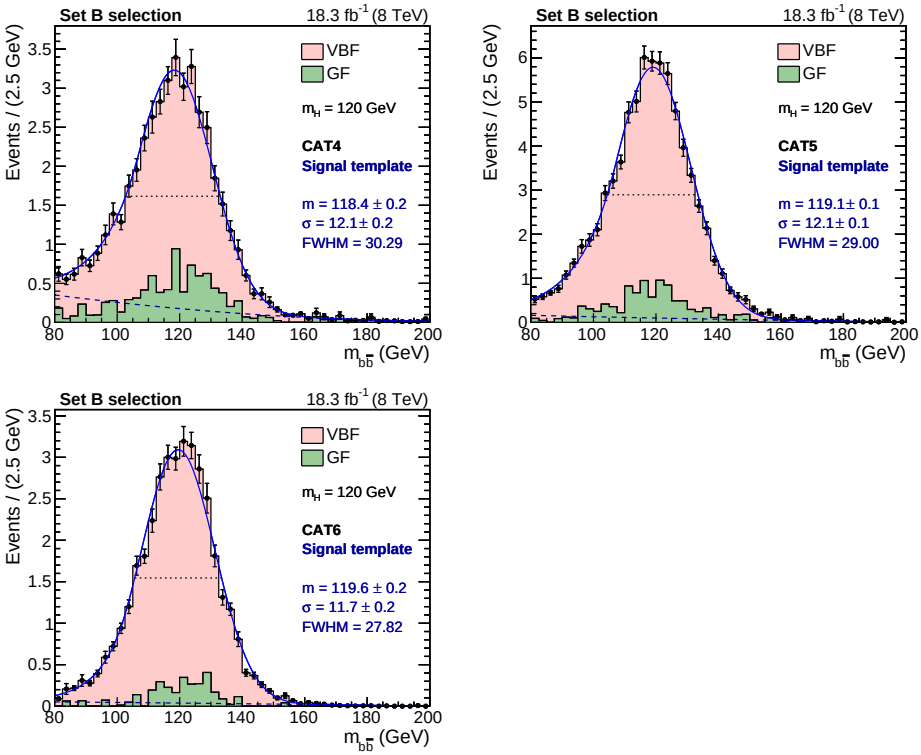


Figure C.4: Templates for the b-quark pair invariant mass distribution for the Set B selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 120$ GeV.

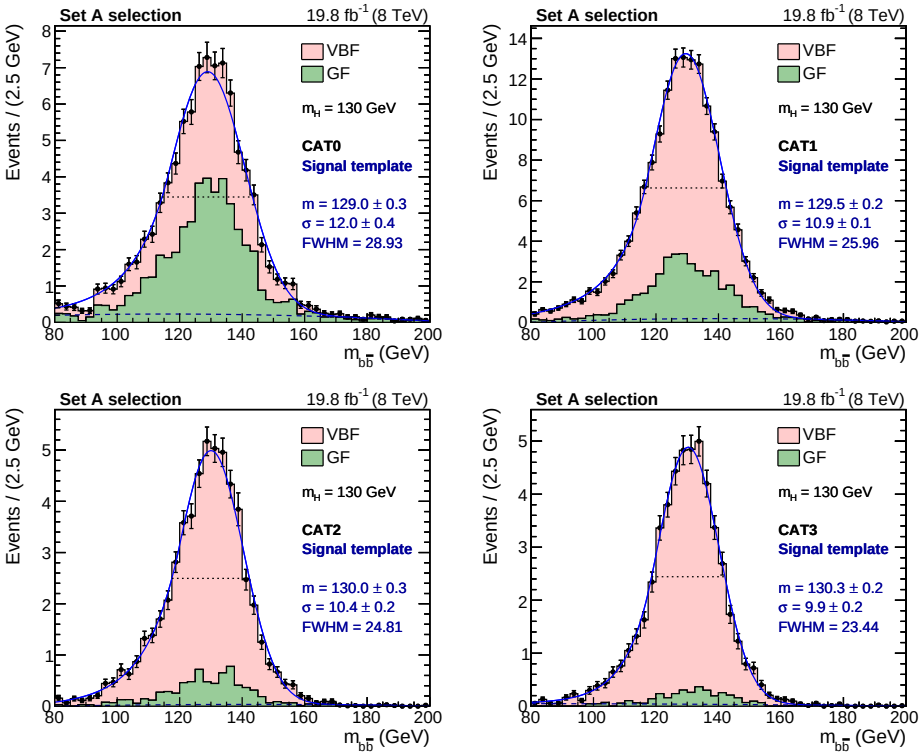


Figure C.5: Templates for the b-quark pair invariant mass distribution for the Set A selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 130$ GeV.

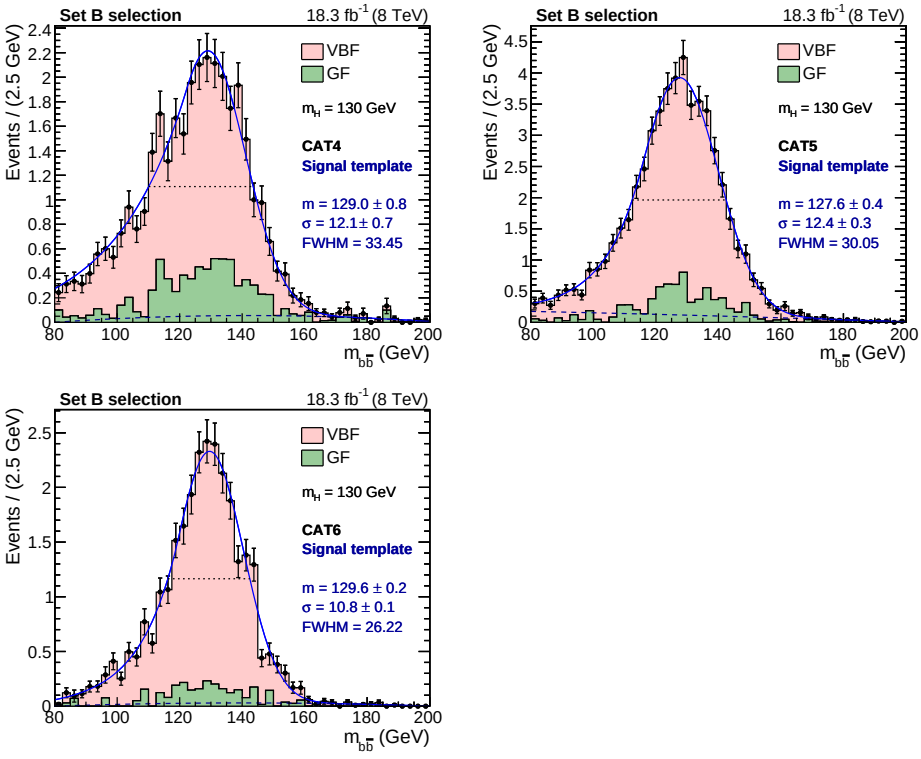


Figure C.6: Templates for the b-quark pair invariant mass distribution for the Set B selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 130$ GeV.

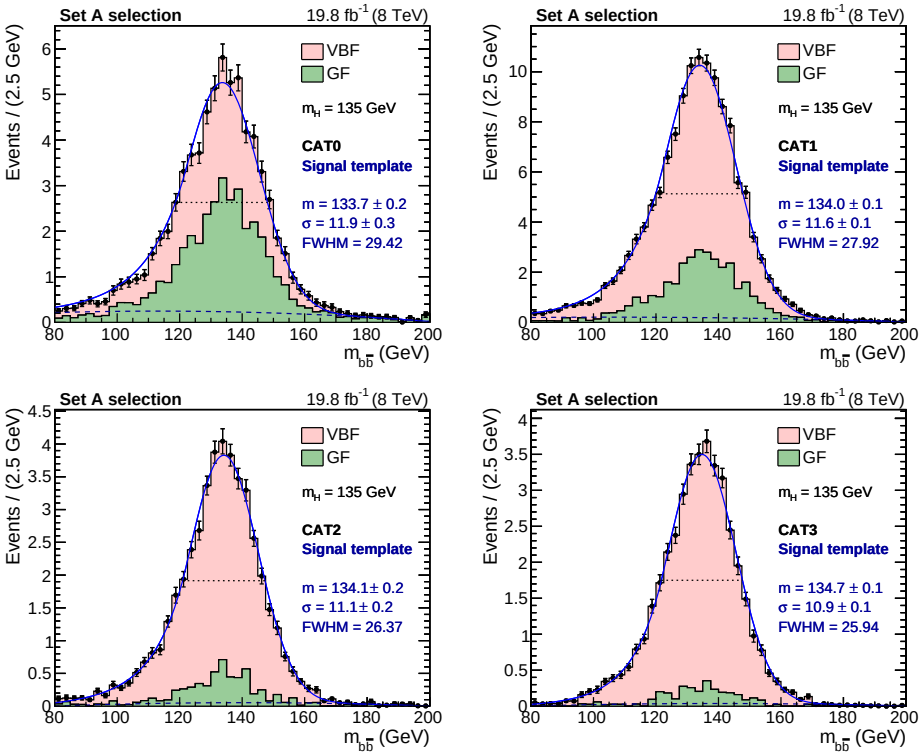


Figure C.7: Templates for the b-quark pair invariant mass distribution for the Set A selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 135$ GeV.

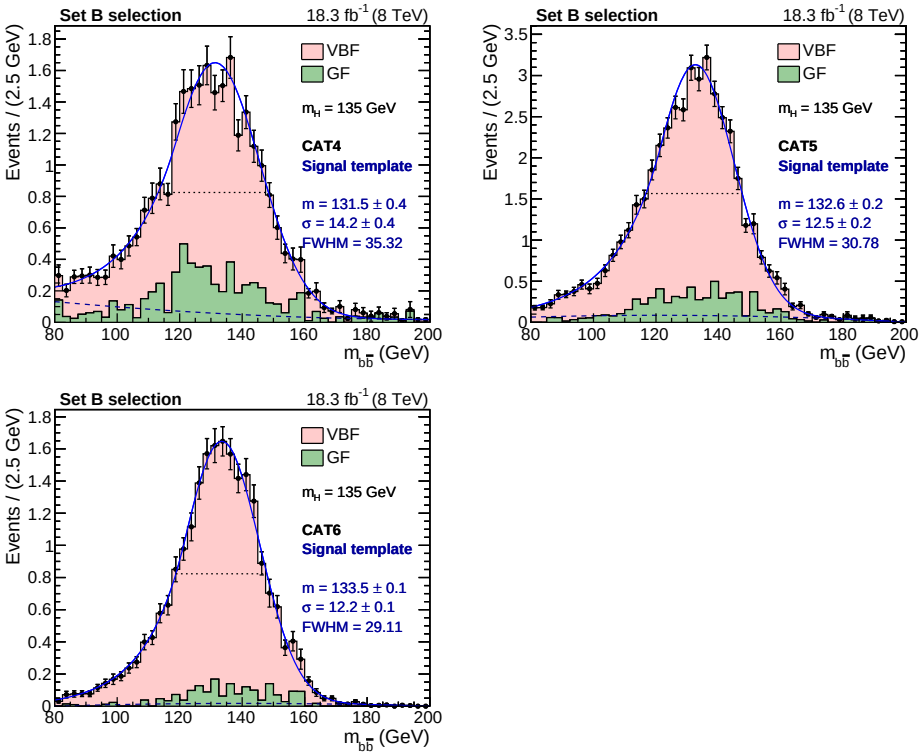


Figure C.8: Templates for the b-quark pair invariant mass distribution for the Set B selection, for the VBF and GF $H \to b\bar{b}$ signal at $m_H = 135$ GeV.

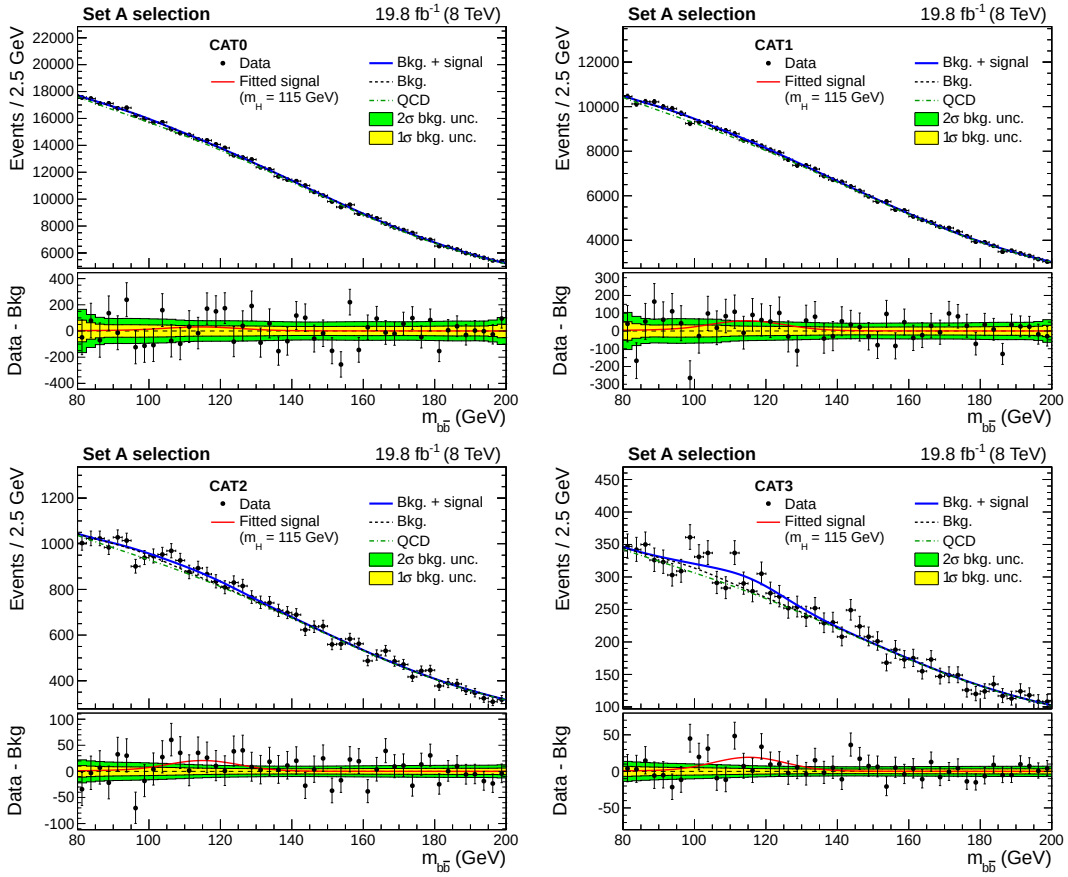


Figure C.9a: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 115$ GeV. The top panels show the data (black markers), as well as the fitted curves: QCD background (green dash-dotted), total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

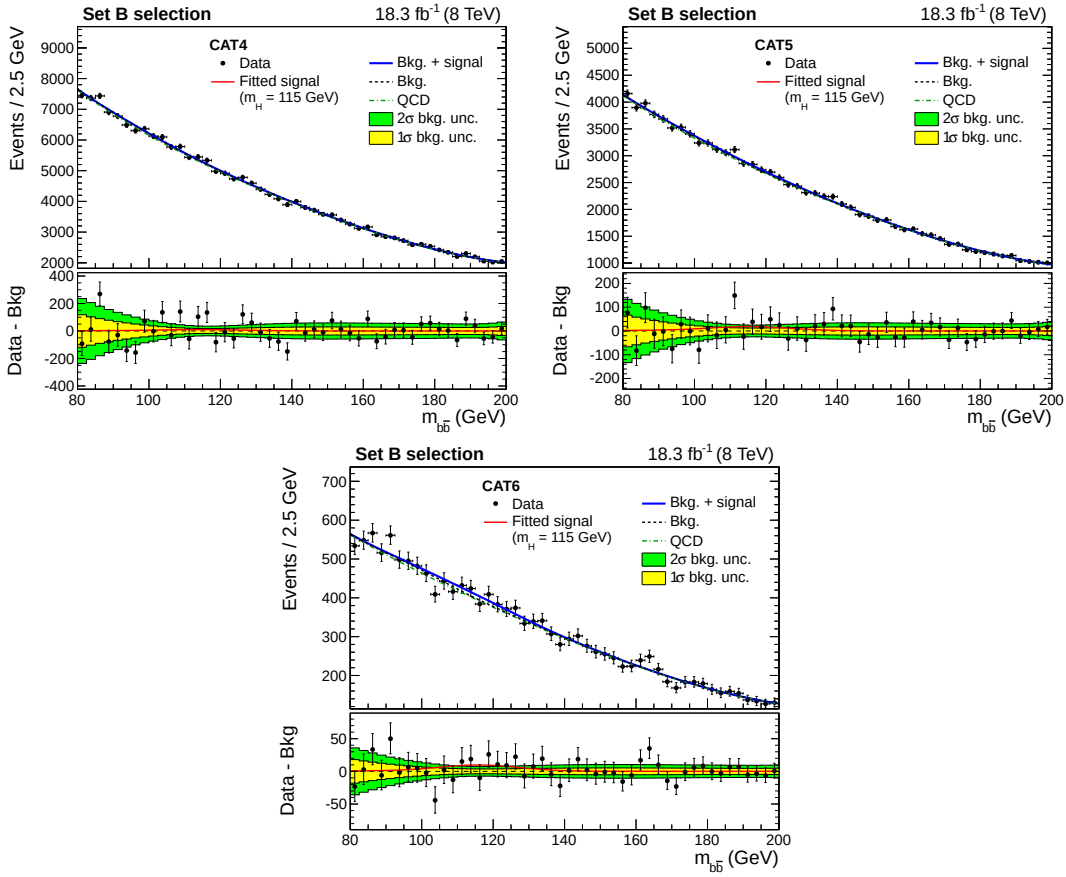


Figure C.9b: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 115$ GeV. The top panels show the data (black markers), as well as the fitted curves: QCD background (green dash-dotted), total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

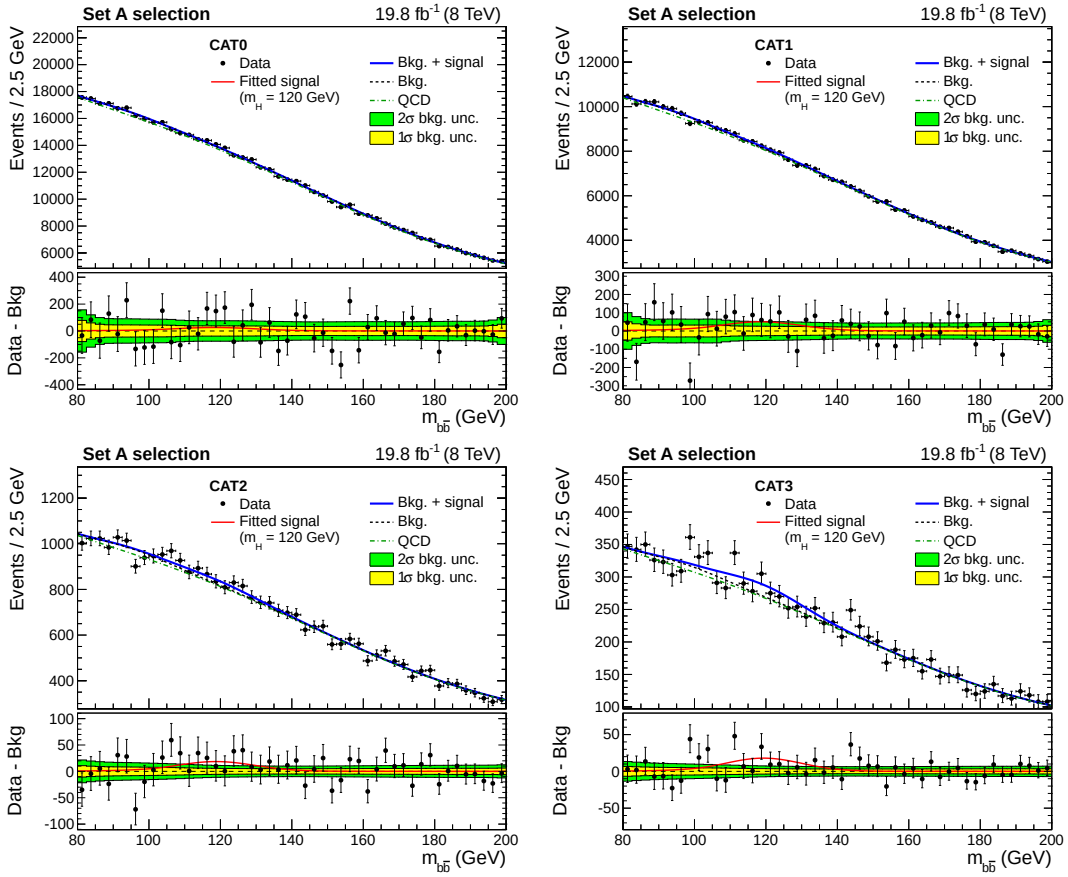


Figure C.10a: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 120$ GeV. The top panels show the data (black markers), as well as the fitted curves: QCD background (green dash-dotted), total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

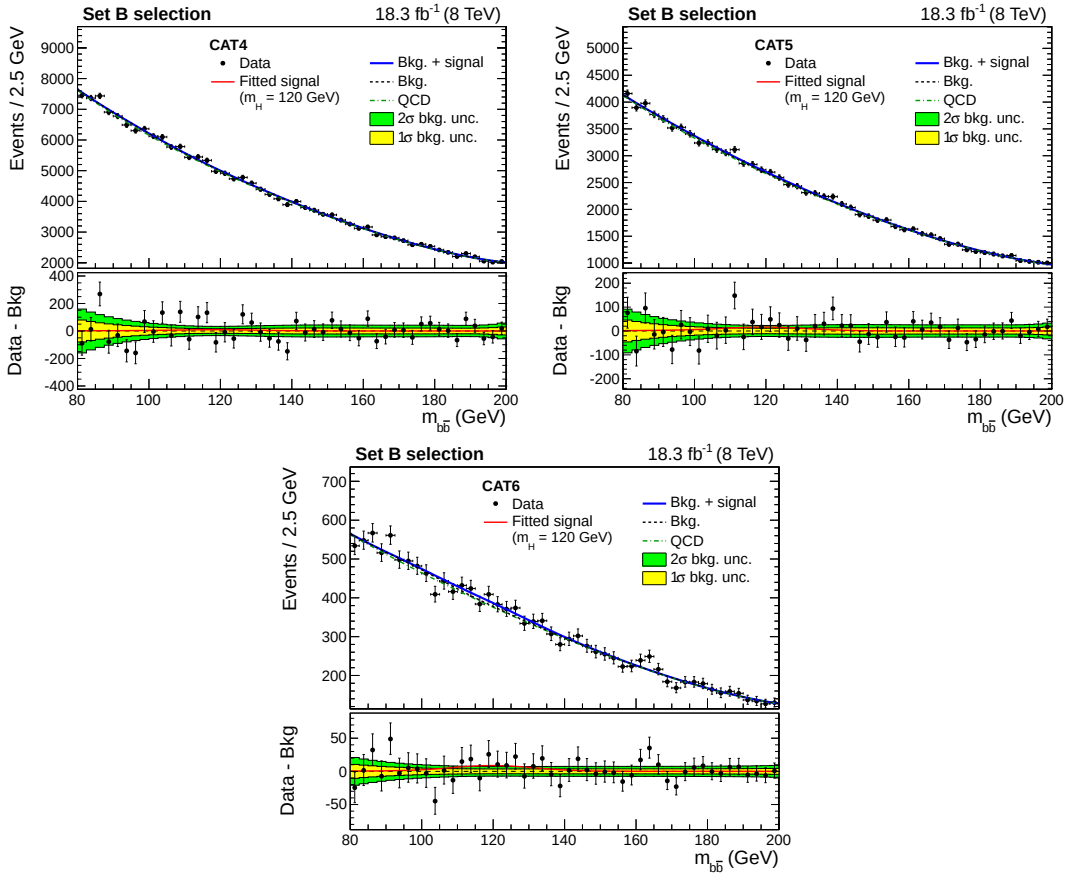


Figure C.10b: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 120$ GeV. The top panels show the data (black markers), as well as the fitted curves: QCD background (green dash-dotted), total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

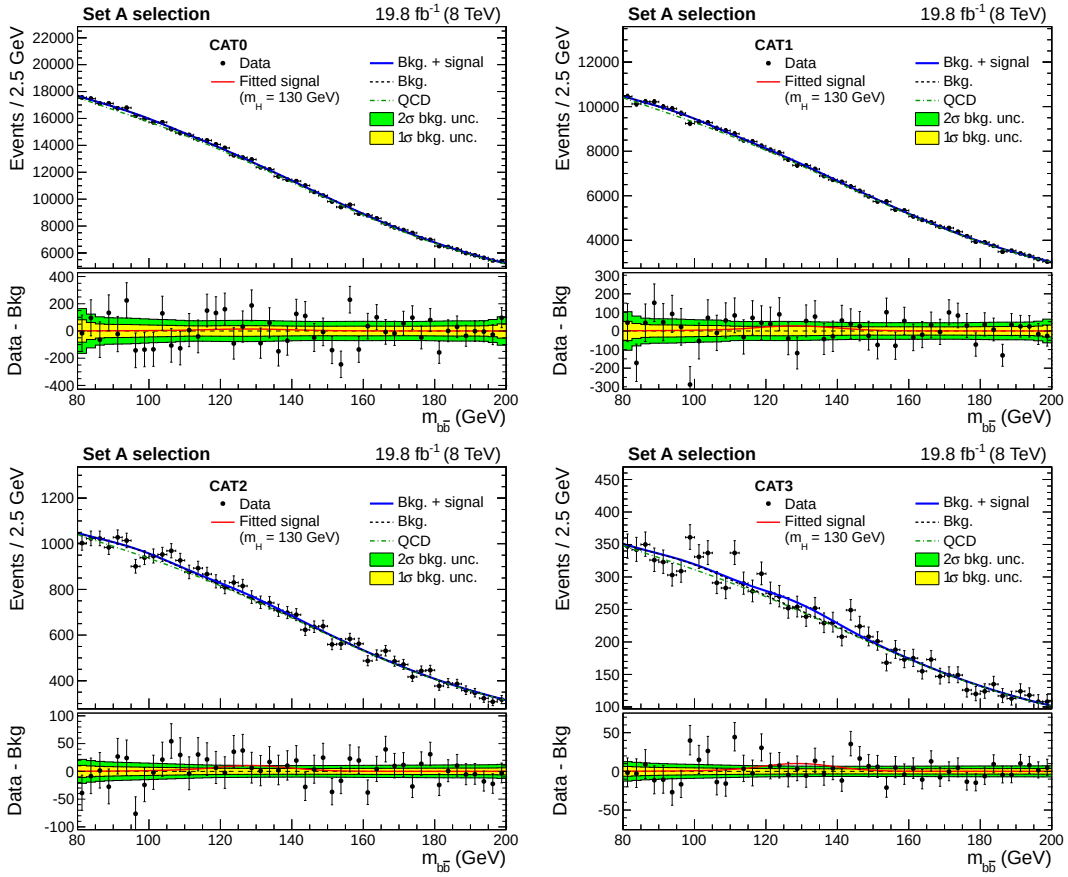


Figure C.11a: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 130$ GeV. The top panels show the data (black markers), as well as the fitted curves: QCD background (green dash-dotted), total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

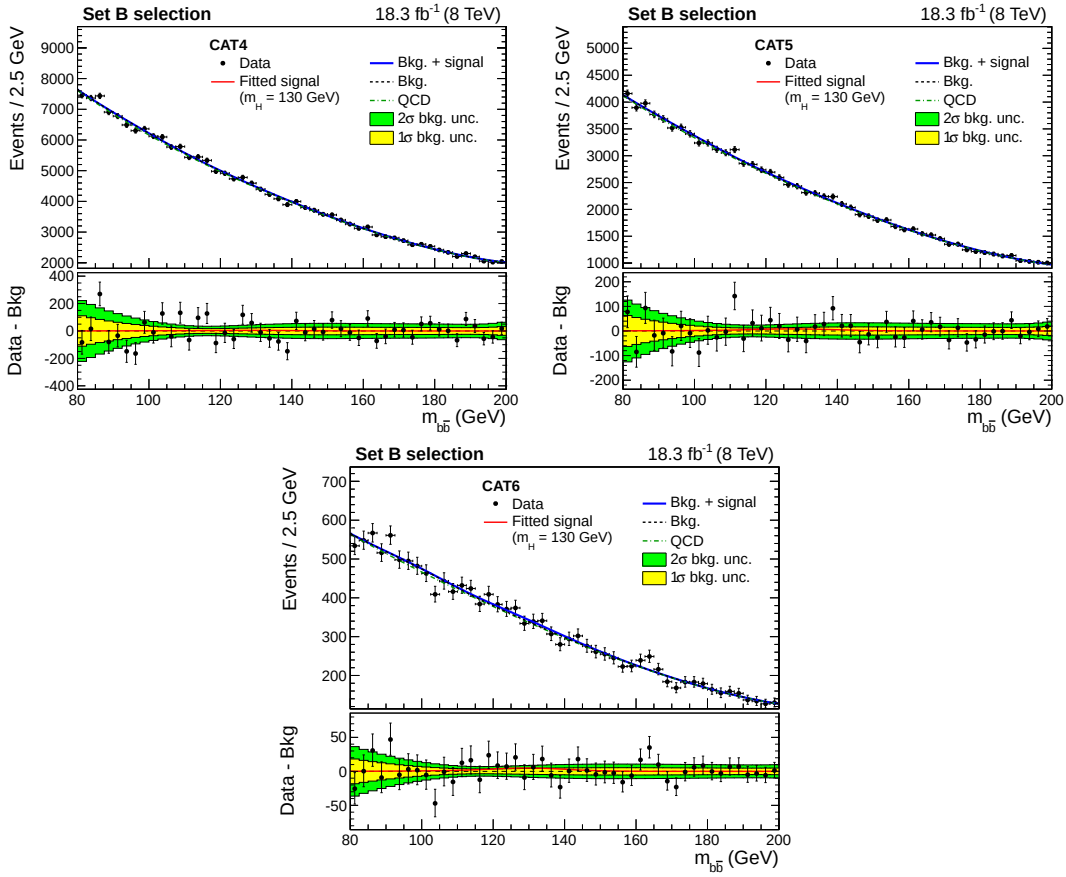


Figure C.11b: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 130$ GeV. The top panels show the data (black markers), as well as the fitted curves: QCD background (green dash-dotted), total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

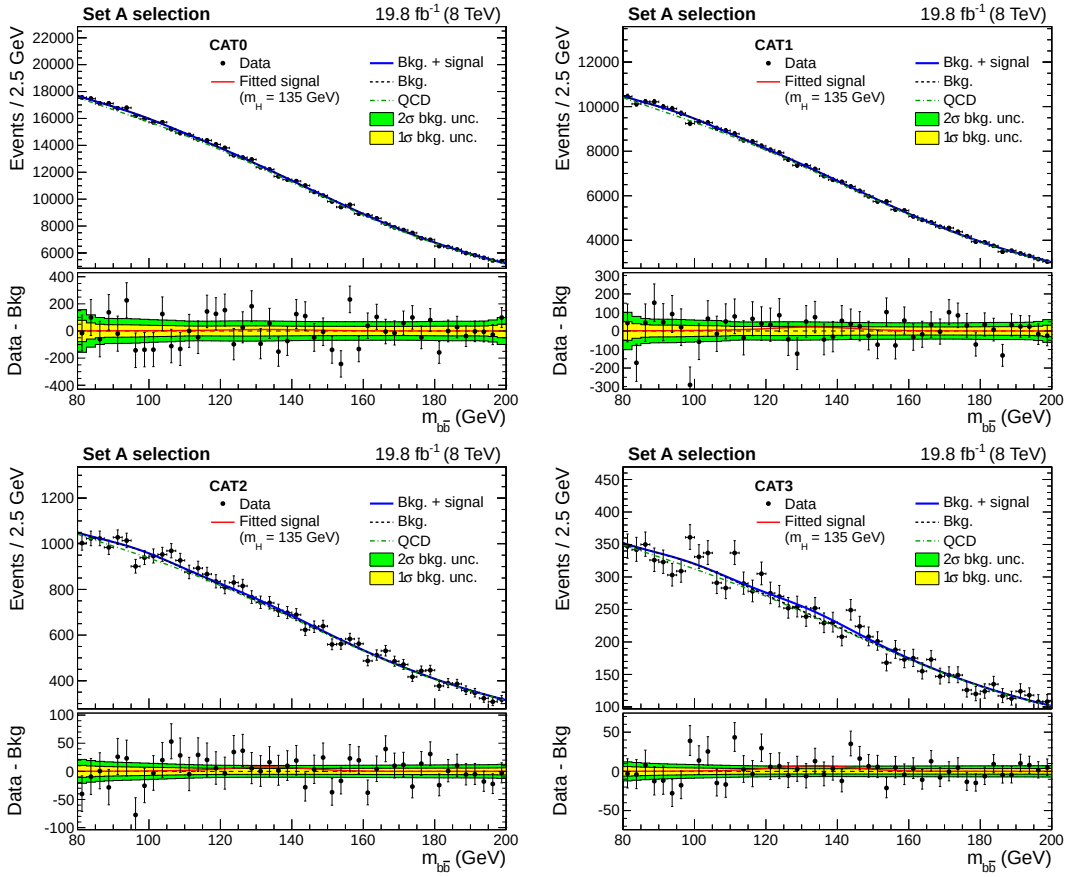


Figure C.12a: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 135$ GeV. The top panels show the data (black markers), as well as the fitted curves: QCD background (green dash-dotted), total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

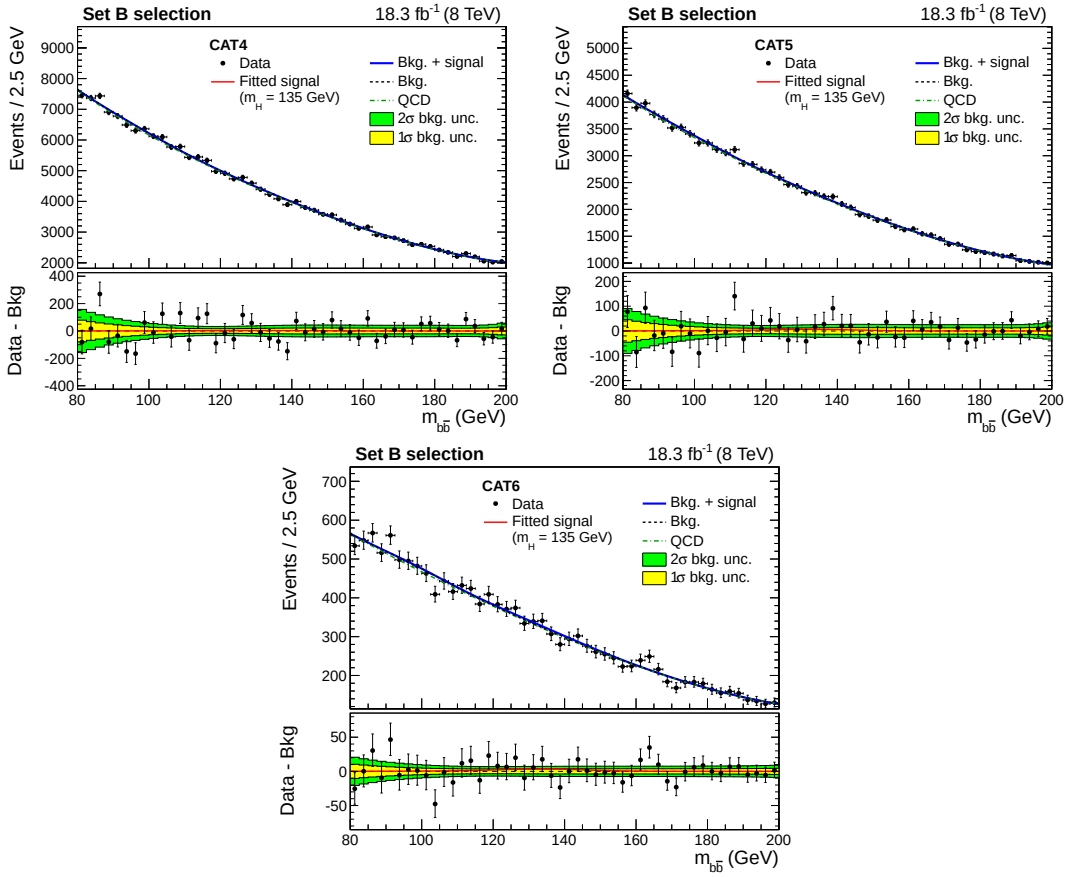


Figure C.12b: Simultaneous fit of the signal plus background models to the b-quark pair invariant mass distributions in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 135$ GeV. The top panels show the data (black markers), as well as the fitted curves: QCD background (green dash-dotted), total background (black dashed) and total signal plus background (blue solid). The bottom panels show the background-subtracted result, with one and two standard deviation uncertainty bands (yellow and green bands), and the fitted signal (red solid).

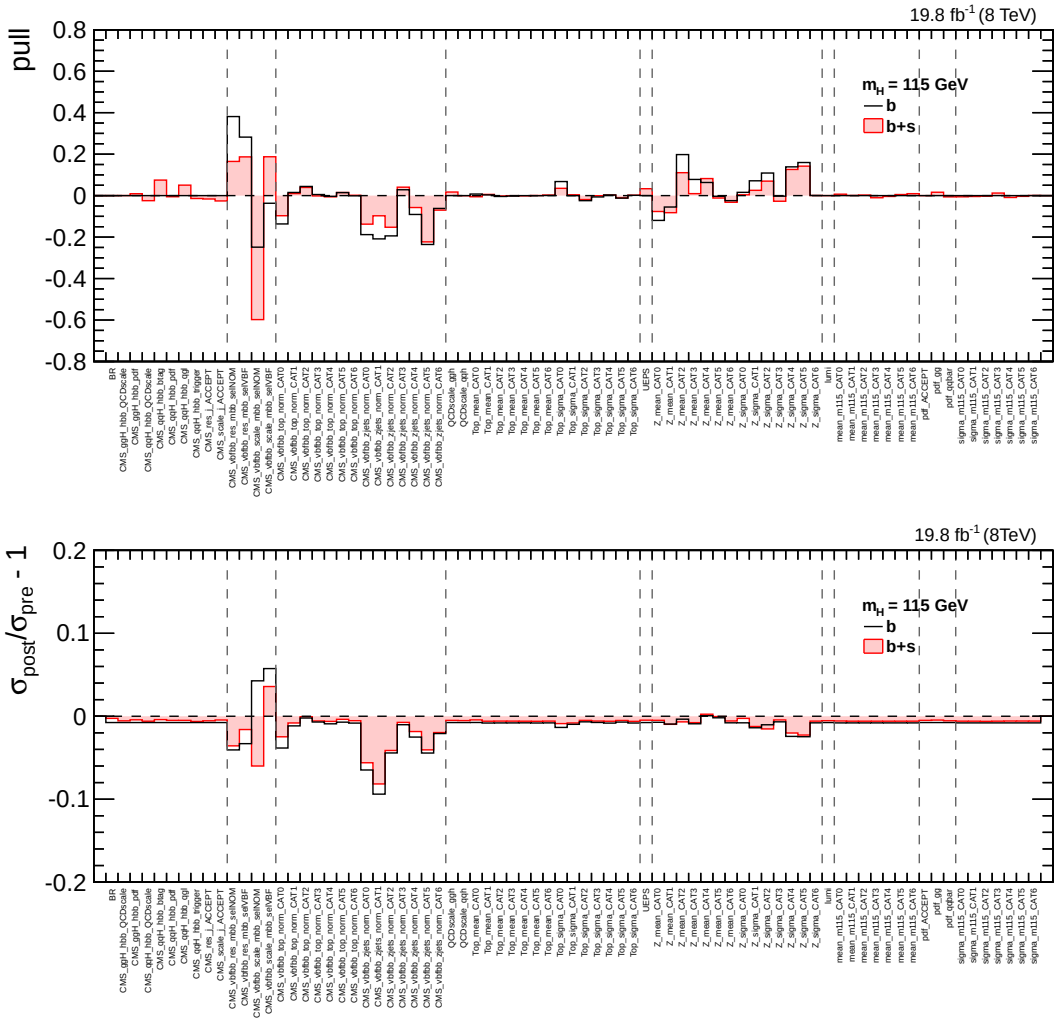


Figure C.13: Post-fit nuisance pulls (top) and uncertainties (bottom) for the simultaneous fit of the b-quark pair invariant mass distribution in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 115$ GeV.

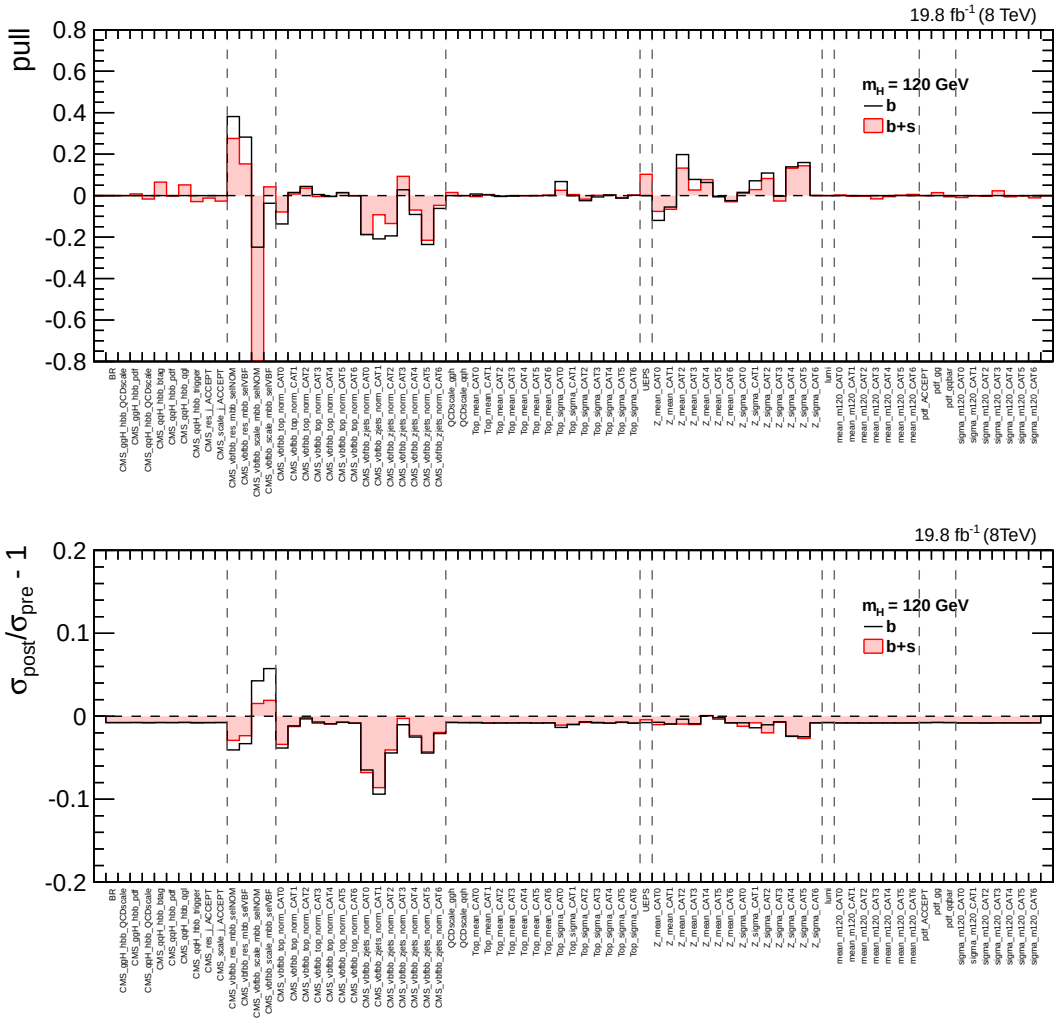


Figure C.14: Post-fit nuisance pulls (top) and uncertainties (bottom) for the simultaneous fit of the b-quark pair invariant mass distribution in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 120$ GeV.

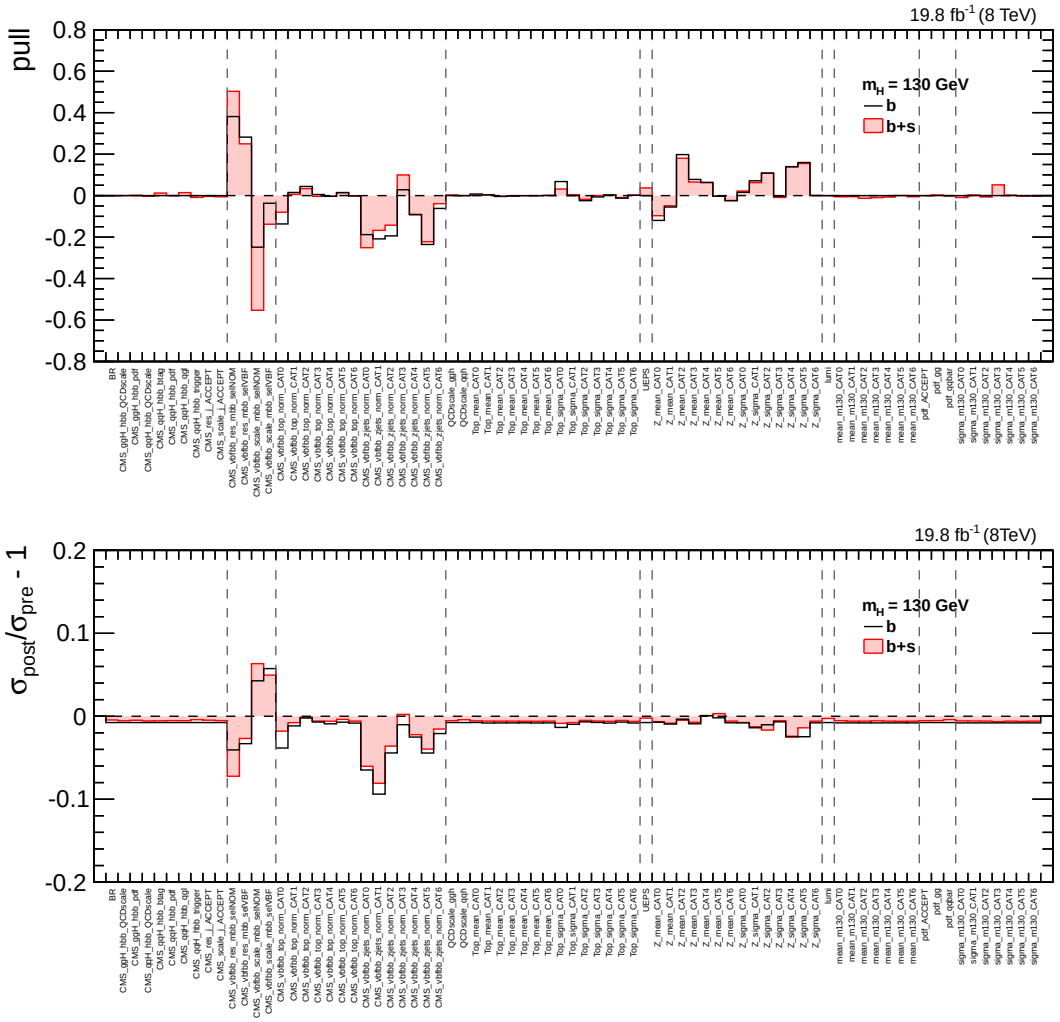


Figure C.15: Post-fit nuisance pulls (top) and uncertainties (bottom) for the simultaneous fit of the b-quark pair invariant mass distribution in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 130$ GeV.

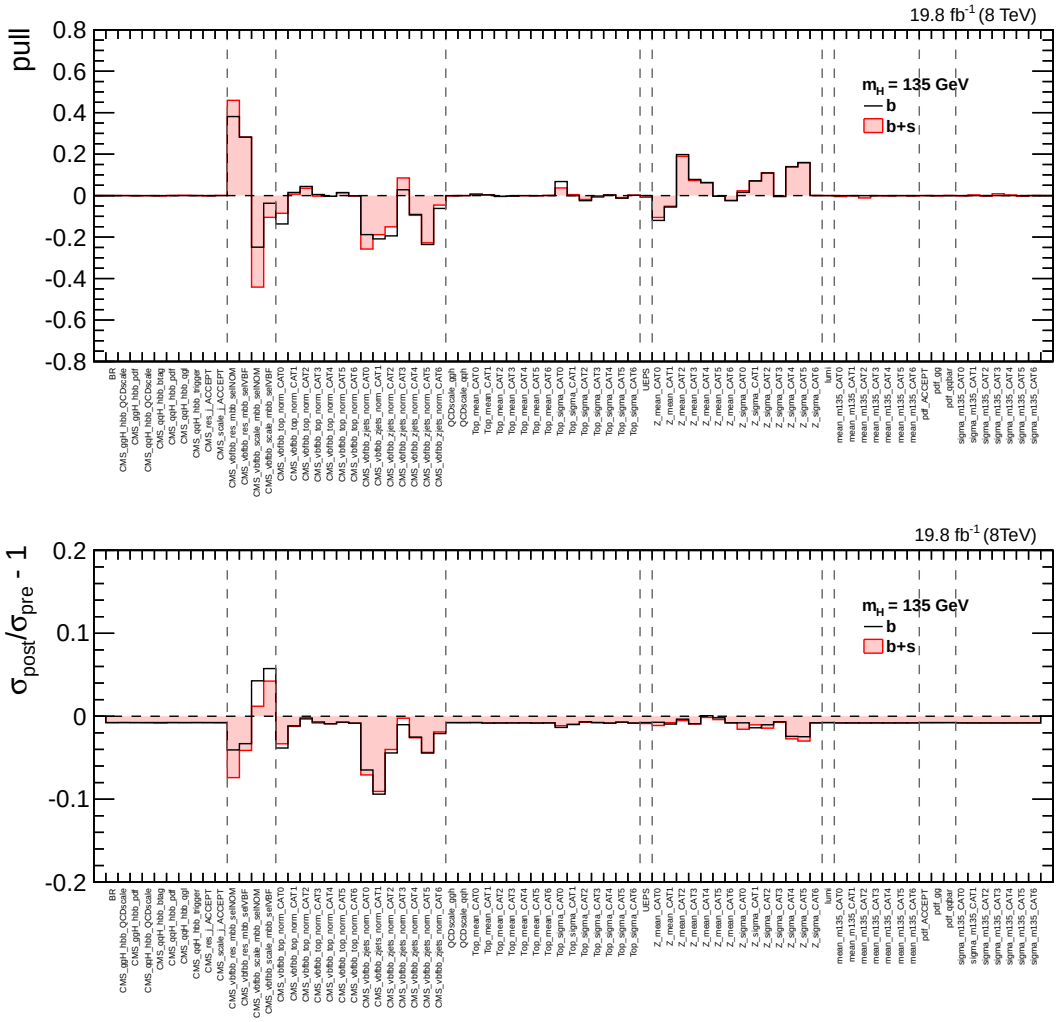


Figure C.16: Post-fit nuisance pulls (top) and uncertainties (bottom) for the simultaneous fit of the b-quark pair invariant mass distribution in all categories of the Set A and Set B selections, for Higgs boson mass hypothesis $m_H = 135$ GeV.

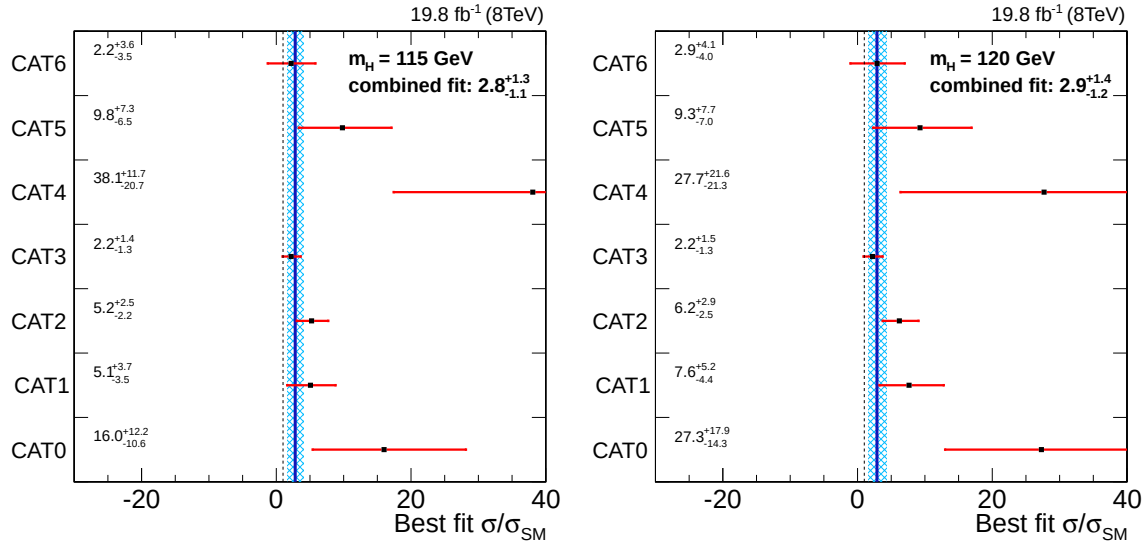


Figure C.17: Fitted signal strength for each individual category and for all categories combined, in units of the SM expected cross section, $\mu = \sigma/\sigma_{\mathrm{SM}}$, for Higgs boson mass hypotheses $m_H = 115$ GeV (left) and $m_H = 120$ GeV (right).

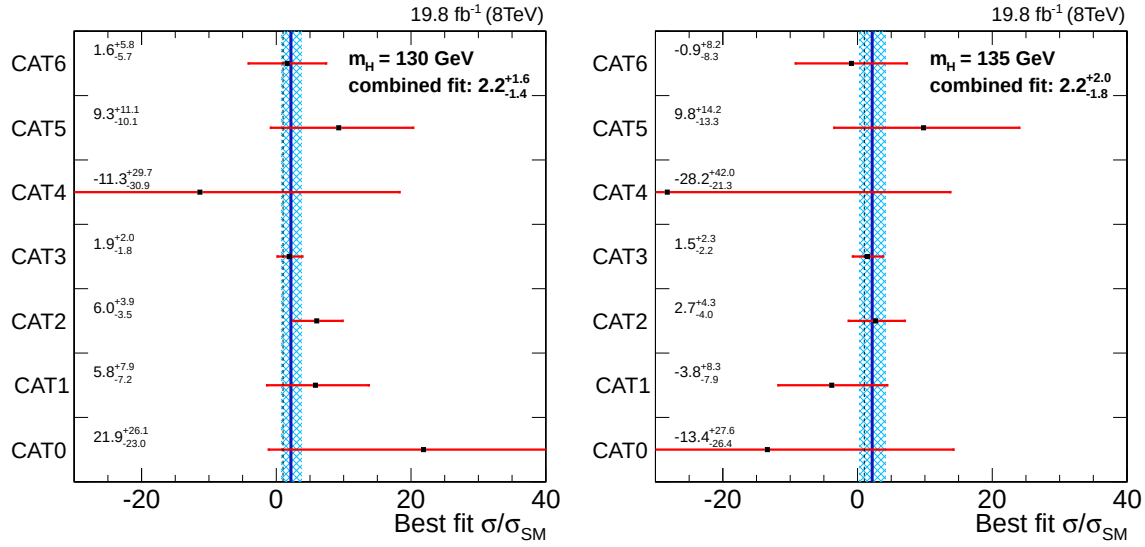


Figure C.18: Fitted signal strength for each individual category and for all categories combined, in units of the SM expected cross section, $\mu = \sigma/\sigma_{\mathrm{SM}}$, for Higgs boson mass hypotheses $m_H = 130$ GeV (left) and $m_H = 135$ GeV (right).

Bibliography

[1] S. Glashow, *Partial Symmetries of Weak Interactions,* Nucl.Phys. **22**, pp. 579–588 (1961), doi:10.1016/0029-5582(61)90469-2. (i, 3, 259)

[2] S. Weinberg, *A Model of Leptons,* Phys.Rev.Lett. **19**, pp. 1264–1266 (1967), doi:10.1103/PhysRevLett.19.1264. (i, 3, 259)

[3] A. Salam, *Weak and Electromagnetic Interactions,* Proceedings of the Eighth Nobel Symposium **C680519**, pp. 367–377 (1968). (i, 3, 259)

[4] H. Fritzsch, M. Gell-Mann, and H. Leutwyler, *Advantages of the Color Octet Gluon Picture,* Phys.Lett. **B47**, pp. 365–368 (1973), doi:10.1016/0370-2693(73)90625-4. (i, 3, 259)

[5] Y. Nambu and M. Han, *Three triplets, paraquarks, and colored quarks,* Phys.Rev. **D10**, pp. 674–683 (1974), doi:10.1103/PhysRevD.10.674. (i, 3, 259)

[6] D. Gross and F. Wilczek, *Asymptotically Free Gauge Theories. 1,* Phys.Rev. **D8**, pp. 3633–3652 (1973), doi:10.1103/PhysRevD.8.3633. (i, 3, 259)

[7] H. D. Politzer, *Asymptotic Freedom: An Approach to Strong Interactions,* Phys.Rept. **14**, pp. 129–180 (1974), doi:10.1016/0370-1573(74)90014-3. (i, 3, 259)

[8] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons,* Phys.Rev.Lett. **13**, pp. 321–323 (1964), doi:10.1103/PhysRevLett.13.321. (i, 3, 259)

[9] P. W. Higgs, *Broken symmetries, massless particles and gauge fields,* Phys.Lett. **12**, pp. 132–133 (1964), doi:10.1016/0031-9163(64)91136-9. (i, 3, 259)

[10] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons,* Phys.Rev.Lett. **13**, pp. 508–509 (1964), doi:10.1103/PhysRevLett.13.508. (i, 3, 259)

[11] G. Guralnik, C. Hagen, and T. Kibble, *Global Conservation Laws and Massless Particles,* Phys.Rev.Lett. **13**, pp. 585–587 (1964), doi:10.1103/PhysRevLett.13.585. (i, 3, 259)

[12] M. E. Peskin and D. V. Schroeder, *An Introduction to quantum field theory*, Addison-Wesley (1995), ISBN-9780201503975. (3, 13)

[13] F. Halzen and A. D. Martin, *Quarks and Leptons: An Introductory Course in Modern Particle Physics*, Wiley (1984), ISBN-9780471887416. (3)

[14] C. P. Burgess and G. D. Moore, *The standard model: A primer*, Cambridge University Press (2006), ISBN-9780521860369. (13, 19, 28)

[15] LHC Higgs Cross Section Working Group, *LHC Higgs Cross Section Working Group Picture Gallery,* https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGCrossSectionsFigures (April 2015). (19, 20)

[16] R. K. Ellis, W. J. Stirling, and B. Webber, *QCD and collider physics*, Cambridge University Press (2003), ISBN-9780521545891. (21, 55, 57)

[17] N. Cabibbo, L. Maiani, G. Parisi, and R. Petronzio, *Bounds on the Fermions and Higgs Boson Masses in Grand Unified Theories,* Nucl.Phys. **B158**, pp. 295–305 (1979), doi: 10.1016/0550-3213(79)90167-6. (22)

[18] J. Ellis, J. Espinosa, G. Giudice, A. Hoecker, and A. Riotto, *The Probable Fate of the Standard Model,* Phys.Lett. **B679**, pp. 369–375 (2009), doi: 10.1016/j.physletb.2009.07.054, arXiv:0906.0954. (22)

[19] ATLAS collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC,* Phys. Lett. **B716**, pp. 1–29 (2012), doi: 10.1016/j.physletb.2012.08.020, arXiv:1207.7214. (22, 27, 28)

[20] CMS collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC,* Phys. Lett. **B716**, pp. 30–61 (2012), doi: 10.1016/j.physletb.2012.08.021, arXiv:1207.7235. (22, 27, 28)

[21] ATLAS collaboration, CMS collaboration, *Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments,* ATLAS-HIGG-2014-14, CMS-HIG-14-042, CERN-PH-EP-2015-075 (2015), arXiv:1503.07589. (22, 28, 29)

[22] S. Dittmaier and M. Schumacher, *The Higgs Boson in the Standard Model - From LEP to LHC: Expectations, Searches, and Discovery of a Candidate,* Prog.Part.Nucl.Phys. **70**, pp. 1–54 (2013), doi: 10.1016/j.ppnp.2013.02.001, arXiv:1211.4828. (23)

[23] R. Assmann, M. Lamont, and S. Myers, *A brief history of the LEP collider,* Nucl. Phys. Proc. Suppl. **109B**, pp. 17–31 (2002), doi: 10.1016/S0920-5632(02)90005-8. (23)

[24] LEP Working Group for Higgs boson searches, *Search for the standard model Higgs boson at LEP,* Phys.Lett. **B565**, pp. 61–75 (2003), doi: 10.1016/S0370-2693(03)00614-2, arXiv:hep-ex/0306033. (23)

[25] CDF collaboration, D0 collaboration, *Tevatron Legacy,* Nuovo Cim. **C035N3**, pp. 181–186 (2012), doi: 10.1393/ncc/i2012-11243-4, arXiv:1202.6196. (23)

[26] Tevatron New Physics Higgs Working Group, *Updated Combination of CDF and D0 Searches for Standard Model Higgs Boson Production with up to 10.0 fb^{-1} of Data,* FERMILAB-CONF-12-318-E, CDF-NOTE-10884, D0-NOTE-6348 (2012), arXiv:1207.0449. (23, 24, 25)

[27] ATLAS collaboration, *Combined search for the Standard Model Higgs boson using up to 4.9 fb^{-1} of pp collision data at $\sqrt{s} = 7$ TeV with the ATLAS detector at the LHC,* Phys.Lett. **B710**, pp. 49–66 (2012), doi: 10.1016/j.physletb.2012.02.044, arXiv:1202.1408. (24)

[28] CMS collaboration, *Combined results of searches for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV,* Phys.Lett. **B710**, pp. 26–48 (2012), doi:10.1016/j.physletb.2012.02.064, arXiv:1202.1488. (24)

[29] ALEPH, DELPHI, L3, OPAL, LEP Electroweak Working Group, SLD Heavy Flavor Group, *A Combination of preliminary electroweak measurements and constraints on the standard model,* SLAC-R-643, CERN-EP-2002-091, LEPEWWG-2002-02, ALEPH-2002-042-PHYSIC-2002-018, DELPHI-2002-098-PHYS-927, CERN-L3-NOTE-2788, OPAL-PR-370 (2002), arXiv:hep-ex/0212036. (25)

[30] Gfitter collaboration, *Status of the global electroweak fit of the Standard Model,* PoS **EPS-HEP2009**, p. 366 (2009), arXiv:0909.0961. (25)

[31] Gfitter collaboration, *The global electroweak fit and constraints on new physics with Gfitter,* PoS **ICHEP2010**, p. 404 (2010), arXiv:1010.5678. (25)

[32] Gfitter collaboration, *Updated Status of the Global Electroweak Fit and Constraints on New Physics,* Eur.Phys.J. **C72**, p. 2003 (2012), doi: 10.1140/epjc/s10052-012-2003-4, arXiv:1107.0975. (25, 26)

[33] Gfitter collaboration, *The Electroweak Fit of the Standard Model after the Discovery of a New Boson at the LHC,* Eur.Phys.J. **C72**, p. 2205 (2012), doi: 10.1140/epjc/s10052-012-2205-9, arXiv:1209.2716. (25)

[34] Gfitter collaboration, *The global electroweak fit at NNLO and prospects for the LHC and ILC,* Eur.Phys.J. **C74**, p. 3046 (2014), doi:10.1140/epjc/s10052-014-3046-5, arXiv:1407.3792. (25)

[35] Gfitter collaboration, *Gfitter home page,* http://project-gfitter.web.cern.ch/project-gfitter/Standard_Model (March 2015). (26)

[36] ATLAS collaboration, *Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment,* ATLAS-CONF-2015-007, ATLAS-COM-CONF-2015-011 (2015), cds:2002212. (28, 29, 30)

[37] CMS collaboration, *Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV,* CMS-HIG-14-009, CERN-PH-EP-2014-288 (2014), arXiv:1412.8662. (28, 29, 30)

[38] CMS collaboration, *Limits on the Higgs boson lifetime and width from its decay to four charged leptons,* Phys. Rev. **D92**, no. 7 p. 072010 (2015), doi:10.1103/PhysRevD.92.072010, arXiv:1507.06656. (28)

[39] CMS, *Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV,* Phys. Rev. **D92**, no. 1 p. 012004 (2015), doi: 10.1103/PhysRevD.92.012004, arXiv:1411.3441. (28, 30)

[40] ATLAS collaboration, *Study of the spin and parity of the Higgs boson in HVV decays with the ATLAS detector,* ATLAS-CONF-2015-008 (2015), cds:2002414. (28)

[41] L. Evans and P. Bryant, *LHC Machine,* JINST **3**, p. S08001 (2008), doi:10.1088/1748-0221/3/08/S08001. (33)

[42] T. Binoth, C. Buttar, P. J. Clark, and E. W. N. Glover, *Proceedings, 65th Scottish Universities Summer School in Physics: LHC Physics (SUSSP65),* CRC Pr. (2012), http://www.crcpress.com/product/isbn/9781439837702. (33)

[43] O. S. Brüning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole, and P. Proudlock, *LHC Design Report*, CERN (2004), cds:782076. (33)

[44] CERN, *CERN's Accelerator Complex,* cds:1621583 (April 2015). (34)

[45] N. Cartiglia, *Measurement of the proton-proton total, elastic, inelastic and diffractive cross sections at 2, 7, 8 and 57 TeV,* (2013), arXiv:1305.6131. (34)

[46] CMS collaboration, *Multi-year collision plots and annual charts of luminosity and pile-up,* https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults (June 2015). (35, 36)

[47] ATLAS collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider,* JINST **3**, p. S08003 (2008), doi:10.1088/1748-0221/3/08/S08003. (35)

[48] CMS collaboration, *The CMS experiment at the CERN LHC,* JINST **3**, p. S08004 (2008), doi:10.1088/1748-0221/3/08/S08004. (35, 38, 41, 42, 43, 44, 46, 48, 49)

[49] LHCb collaboration, *The LHCb Detector at the LHC,* JINST **3**, p. S08005 (2008), doi:10.1088/1748-0221/3/08/S08005. (35)

[50] ALICE collaboration, *The ALICE experiment at the CERN LHC,* JINST **3**, p. S08002 (2008), doi:10.1088/1748-0221/3/08/S08002. (35)

[51] LHCf collaboration, *The LHCf detector at the CERN Large Hadron Collider,* JINST **3**, p. S08006 (2008), doi:10.1088/1748-0221/3/08/S08006. (35)

[52] TOTEM collaboration, *The TOTEM experiment at the CERN Large Hadron Collider,* JINST **3**, p. S08007 (2008), doi:10.1088/1748-0221/3/08/S08007. (35)

[53] MoEDAL collaboration, *Technical Design Report of the MoEDAL Experiment,* CERN-LHCC-2009-006, MoEDAL-TDR-001 (2009), cds:1181486. (35)

[54] CMS collaboration, *CMS physics: Technical design report, volume I,* CERN-LHCC-2006-001, CMS-TDR-008-1 (2006), cds:922757. (38)

[55] D. Barney, *Cms slice diagram,* https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=5697 (June 2015). (39)

[56] *CMS: The tracker. Technical design report* (1998), CERN-LHCC-98-06, CMS-TDR-5, cds:368412. (40)

[57] CMS collaboration, *Description and performance of track and primary-vertex reconstruction with the CMS tracker,* JINST **9**, no. 10 p. P10009 (2014), doi:10.1088/1748-0221/9/10/P10009, arXiv:1405.6569. (40, 65, 67, 68)

[58] C. W. Fabjan and F. Gianotti, *Calorimetry for particle physics,* Rev. Mod. Phys. **75**, pp. 1243–1286 (2003), cds:6922522. (42)

[59] CMS collaboration, *CMS: The electromagnetic calorimeter. Technical design report,* CERN-LHCC-97-33, CMS-TDR-4 (1997), cds:349375. (43)

[60] CMS collaboration, *Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in pp Collisions at* $\sqrt{s} = 7$ *TeV,* JINST **8**, p. P09009 (2013), doi:10.1088/1748-0221/8/09/P09009, arXiv:1306.2016. (43)

[61] CMS collaboration, *CMS: The hadron calorimeter. Technical design report,* CERN-LHCC-97-31 (1997), cds:357153. (44)

[62] CMS collaboration, *Performance of the CMS Hadron Calorimeter with Cosmic Ray Muons and LHC Beam Data,* JINST **5**, p. T03012 (2010), doi:10.1088/1748-0221/5/03/T03012, arXiv:0911.4991. (44)

[63] CMS collaboration, *CMS: The muon system. Technical design report,* CERN-LHCC-97-32 (1997), cds:343814. (46)

[64] CMS collaboration, *The performance of the CMS muon detector in proton-proton collisions at sqrt(s) = 7 TeV at the LHC,* JINST **8**, p. P11002 (2013), doi:10.1088/1748-0221/8/11/P11002, arXiv:1306.6905. (46, 47)

[65] CMS collaboration, *CMS: The TriDAS project. Technical design report, Vol. 1: The trigger systems,* CERN-LHCC-2000-038 (2000), cds:706847. (49)

[66] CMS collaboration, *CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger,* CERN-LHCC-2002-026 (2002), cds:578006. (49)

[67] R. Brun and F. Rademakers, *ROOT: An object oriented data analysis framework,* Nucl. Instrum. Meth. **A389**, pp. 81–86 (1997), doi:10.1016/S0168-9002(97)00048-X. (50)

[68] A. Buckley *et al.*, *General-purpose event generators for LHC physics,* Phys. Rept. **504**, pp. 145–233 (2011), doi:10.1016/j.physrep.2011.03.005, arXiv:1101.2599. (53, 56, 58)

[69] F. Krauss, *QCD and Monte Carlo tools,* CERN-Fermilab Hadron Collider Physics Summer Schools (2007), https://indico.cern.ch/event/6238/other-view?view=cdsagenda. (54, 56)

[70] A. Buckley, J. Ferrando, S. Lloyd, K. Nordström, B. Page, M. Rüfenacht, M. Schönherr, and G. Watt, *LHAPDF6: parton density access in the LHC precision era,* Eur. Phys. J. **C75**, no. 3 p. 132 (2015), doi:10.1140/epjc/s10052-015-3318-8, arXiv:1412.7420. (55)

[71] L. Lönnblad, *Monte Carlo Techniques and Physics,* Terascale Monte Carlo School (2008), https://indico.desy.de/conferenceOtherViews.py?view=standard&confId=708. (56)

[72] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani, *Matching matrix elements and shower evolution for top-quark production in hadronic collisions,* JHEP **01**, p. 013 (2007), doi:10.1088/1126-6708/2007/01/013, arXiv:hep-ph/0611129. (58, 61)

[73] S. Catani, F. Krauss, R. Kuhn, and B. R. Webber, *QCD matrix elements + parton showers,* JHEP **11**, p. 063 (2001), doi:10.1088/1126-6708/2001/11/063, arXiv:hep-ph/0109231. (58, 60)

[74] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand, *Parton Fragmentation and String Dynamics,* Phys. Rept. **97**, pp. 31–145 (1983), doi:10.1016/0370-1573(83)90080-7. (58)

[75] B. R. Webber, *A QCD Model for Jet Fragmentation Including Soft Gluon Interference,* Nucl. Phys. **B238**, p. 492 (1984), doi:10.1016/0550-3213(84)90333-X. (58)

[76] JADE collaboration, *Particle Distribution in Three Jet Events Produced by e+ e- Annihilation,* Z. Phys. **C21**, p. 37 (1983), doi:10.1007/BF01648774. (58)

[77] CDF collaboration, *Evidence for color coherence in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV,* Phys. Rev. **D50**, pp. 5562–5579 (1994), doi:10.1103/PhysRevD.50.5562. (58)

[78] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *PYTHIA 6.4 Physics and Manual,* JHEP **05**, p. 026 (2006), doi:10.1088/1126-6708/2006/05/026, arXiv:hep-ph/0603175. (60, 94)

[79] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1,* Comput. Phys. Commun. **178**, pp. 852–867 (2008), doi:10.1016/j.cpc.2008.01.036, arXiv:0710.3820. (60)

[80] M. Bahr *et al.*, *Herwig++ Physics and Manual,* Eur. Phys. J. **C58**, pp. 639–707 (2008), doi:10.1140/epjc/s10052-008-0798-9, arXiv:0803.0883. (60)

[81] J. Bellm *et al.*, *Herwig++ 2.7 Release Note,* IPPP-13-88, MCNET-13-15, DCPT-13-176, DESY-13-186, KA-TP-31-2013, –ZU-TH-23-13 (2013), arXiv:1310.6877. (60)

[82] T. Gleisberg, S. Hoeche, F. Krauss, M. Schönherr, S. Schumann, F. Siegert, and J. Winter, *Event generation with SHERPA 1.1,* JHEP **02**, p. 007 (2009), doi:10.1088/1126-6708/2009/02/007, arXiv:0811.4622. (60)

[83] F. Maltoni and T. Stelzer, *MadEvent: Automatic event generation with MadGraph,* JHEP **02**, p. 027 (2003), doi:10.1088/1126-6708/2003/02/027, arXiv:hep-ph/0208156. (61)

[84] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, *MadGraph 5 : Going Beyond,* JHEP **06**, p. 128 (2011), doi:10.1007/JHEP06(2011)128, arXiv:1106.0522. (61, 94)

[85] C. Oleari, *The POWHEG-BOX,* Nucl. Phys. Proc. Suppl. **205-206**, pp. 36–41 (2010), doi:10.1016/j.nuclphysbps.2010.08.016, arXiv:1007.3893. (61)

[86] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms,* JHEP **11**, p. 040 (2004), doi:10.1088/1126-6708/2004/11/040, arXiv:hep-ph/0409146. (61, 94)

[87] P. Golonka, B. Kersevan, T. Pierzchala, E. Richter-Was, Z. Was, and M. Worek, *The Tauola photos F environment for the TAUOLA and PHOTOS packages: Release. 2.,* Comput. Phys. Commun. **174**, pp. 818–835 (2006), doi:10.1016/j.cpc.2005.12.018, arXiv:hep-ph/0312240. (61, 94)

[88] GEANT4, *GEANT4: A Simulation toolkit,* Nucl. Instrum. Meth. **A506**, pp. 250–303 (2003), doi:10.1016/S0168-9002(03)01368-8. (61)

[89] CMS collaboration, *Tracking and Primary Vertex Results in First 7 TeV Collisions,* CMS-PAS-TRK-10-005 (2010), cds:1279383. (65, 66)

[90] CMS collaboration, *CMS Tracking Performance Results from early LHC Operation,* Eur.Phys.J. **C70**, pp. 1165–1192 (2010), doi:10.1140/epjc/s10052-010-1491-3, arXiv:1007.1988. (65, 125)

[91] P. Billoir, *Progressive track recognition with a Kalman like fitting procedure,* Comput. Phys. Commun. **57**, pp. 390–394 (1989), doi:10.1016/0010-4655(89)90249-X. (65)

[92] K. Rose, *Deterministic annealing for clustering, compression, classification, regression, and related optimization problems,* Proceedings of the IEEE **86**, no. 11 pp. 2210–2239 (1998), doi:10.1109/5.726788. (68)

[93] R. Frühwirth, W. Waltenberger, and P. Vanlaer, *Adaptive vertex fitting,* J. Phys. **G34**, p. N343 (2007), doi:10.1088/0954-3899/34/12/N01. (68)

[94] G. P. Salam and G. Soyez, *A Practical Seedless Infrared-Safe Cone jet algorithm,* JHEP **0705**, p. 086 (2007), doi:10.1088/1126-6708/2007/05/086, arXiv:0704.0292. (69)

[95] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti-k(t) jet clustering algorithm,* JHEP **0804**, p. 063 (2008), doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189. (69)

[96] CMS collaboration, *Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS,* JINST **6**, p. P11002 (2011), doi: 10.1088/1748-0221/6/11/P11002, arXiv:1107.4277. (70, 73, 172)

[97] CMS collaboration, *Jet Performance in pp Collisions at 7 TeV,* CMS-PAS-JME-10-003 (2010), cds:1279362. (70)

[98] CMS collaboration, *Commissioning of TrackJets in pp Collisions at 7 TeV,* CMS-PAS-JME-10-006 (2010), cds:1275133. (70)

[99] CMS collaboration, *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET,* CMS-PAS-PFT-09-001 (2009), cds:1194487. (70, 72)

[100] CMS collaboration, *Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector,* CMS-PAS-PFT-10-001 (2010), cds:1247373. (70)

[101] CMS collaboration, *Jet Energy Scale and Resolution in the CMS Experiment,* CMS-JME-13-004 (2015). To be submitted to JINST. (73, 74, 75, 76, 96, 121)

[102] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual,* Eur. Phys. J. **C72**, p. 1896 (2012), doi:10.1140/epjc/s10052-012-1896-2, arXiv:1111.6097. (74)

[103] D. Bertolini, P. Harris, M. Low, and N. Tran, *Pileup Per Particle Identification,* JHEP **10**, p. 59 (2014), doi:10.1007/JHEP10(2014)059, arXiv:1407.6013. (74)

[104] M. Cacciari, G. P. Salam, and G. Soyez, *SoftKiller, a particle-level pileup removal method,* Eur. Phys. J. **C75**, no. 2 p. 59 (2015), doi:10.1140/epjc/s10052-015-3267-2, arXiv:1407.0408. (74)

[105] P. Berta, M. Spousta, D. W. Miller, and R. Leitner, *Particle-level pileup subtraction for jets and jet shapes,* JHEP **06**, p. 092 (2014), doi:10.1007/JHEP06(2014)092, arXiv:1403.3108. (74)

[106] CMS collaboration, *Pileup Jet Identification,* CMS-PAS-JME-13-005 (2013), cds:1581583. (74)

[107] CMS collaboration, *Identification of b-quark jets with the CMS experiment,* JINST **8**, p. P04013 (2013), doi:10.1088/1748-0221/8/04/P04013, arXiv:1211.4462. (77, 78)

[108] K. Konishi, A. Ukawa, and G. Veneziano, *A Simple Algorithm for QCD Jets,* Phys. Lett. **B78**, p. 243 (1978), doi:10.1016/0370-2693(78)90015-1. (79)

[109] CMS collaboration, *Performance of quark/gluon discrimination in 8 TeV pp data,* CMS-PAS-JME-13-002 (2013), cds:1599732. (79)

[110] T. Cornelis, *Measurement of the electroweak production of a Z boson in association with two jets with the CMS detector,* PhD thesis Universiteit Antwerpen, 2015. (79, 80)

[111] CMS collaboration, *Measurement of electroweak production of two jets in association with a Z boson in proton-proton collisions at* $\sqrt{s} = 8$ *TeV,* Eur. Phys. J. **C75**, no. 2 p. 66 (2015) , doi:10.1140/epjc/s10052-014-3232-5, arXiv:1410.3153. (80)

[112] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. Meridiani, I. Puljak, C. Rovelli, R. Salerno, and Y. Sirois, *Electron reconstruction in CMS,* Eur. Phys. J. **C49**, pp. 1099–1116 (2007), doi:10.1140/epjc/s10052-006-0175-5. (80)

[113] CMS collaboration, *Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at* $\sqrt{s}$ *= 8 TeV,* JINST **10**, no. 06 p. P06005 (2015), doi:10.1088/1748-0221/10/06/P06005, arXiv:1502.02701. (80)

[114] CMS collaboration, *Performance of CMS muon reconstruction in pp collision events at* $\sqrt{s} = 7$ *TeV,* JINST **7**, p. P10002 (2012), doi:10.1088/1748-0221/7/10/P10002, arXiv:1206.4071. (81)

[115] CMS collaboration, *Performance of Missing Transverse Momentum Reconstruction Algorithms in Proton-Proton Collisions at p s = 8 TeV with the CMS Detector,* CMS-PAS-JME-12-002 (2012), cds:1543527. (81)

[116] CMS, *Search for the standard model Higgs boson produced through vector boson fusion and decaying to* $b\bar{b}$*,* Phys. Rev. **D92**, no. 3 p. 032008 (2015), doi:10.1103/PhysRevD.92.032008, arXiv:1506.01010. (85, 193, 260)

[117] CMS collaboration, *Measurement of the properties of a Higgs boson in the four-lepton final state,* Phys. Rev. **D89**, no. 9 p. 092007 (2014), doi:10.1103/PhysRevD.89.092007, arXiv:1312.5353. (85)

[118] CMS collaboration, *Observation of the diphoton decay of the Higgs boson and measurement of its properties,* Eur. Phys. J. **C74**, no. 10 p. 3076 (2014), doi:10.1140/epjc/s10052-014-3076-z, arXiv:1407.0558. (85)

[119] CMS collaboration, *Measurement of Higgs boson production and properties in the WW decay channel with leptonic final states,* JHEP **01**, p. 096 (2014), doi:10.1007/JHEP01(2014)096, arXiv:1312.1129. (85)

[120] CMS collaboration, *Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks,* Phys. Rev. **D89**, no. 1 p. 012003 (2014), doi:10.1103/PhysRevD.89.012003, arXiv:1310.3687. (85, 188)

[121] CMS collaboration, *Search for the associated production of the Higgs boson with a top-quark pair,* JHEP **09**, p. 087 (2014), doi:10.1007/JHEP09(2014)087,10.1007/JHEP10(2014)106, arXiv:1408.1682. [Erratum: JHEP10,106(2014)]. (85, 188)

[122] CMS collaboration, *Search for a Standard Model Higgs Boson Produced in Association with a Top-Quark Pair and Decaying to Bottom Quarks Using a Matrix Element Method,* Eur. Phys. J. **C75**, no. 6 p. 251 (2015), doi:10.1140/epjc/s10052-015-3454-1, arXiv:1502.02485. (85, 188)

[123] CMS collaboration, *Evidence for the 125 GeV Higgs boson decaying to a pair of* τ *leptons,* JHEP **05**, p. 104 (2014), doi:10.1007/JHEP05(2014)104, arXiv:1401.5041. (85)

[124] CMS collaboration, *Evidence for the direct decay of the 125 GeV Higgs boson to fermions,* Nature Phys. **10**, pp. 557–560 (2014), doi:10.1038/nphys3005, arXiv:1401.6527. (85)

[125] R. Field, *Early LHC Underlying Event Data - Findings and Surprises,* in Hadron collider physics. Proceedings, 22nd Conference, HCP 2010, Toronto, Canada, August 23-27, 2010 (2010), arXiv:1010.3558. (94)

[126] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, P. M. Nadolsky, and W. K. Tung, *New generation of parton distributions with uncertainties from global QCD analysis,* JHEP **07**, p. 012 (2002), doi:10.1088/1126-6708/2002/07/012, arXiv:hep-ph/0201195. (94)

[127] H.-L. Lai, M. Guzzi, J. Huston, Z. Li, P. M. Nadolsky, J. Pumplin, and C. P. Yuan, *New parton distributions for collider physics,* Phys. Rev. **D82**, p. 074024 (2010), doi: 10.1103/PhysRevD.82.074024, arXiv:1007.2241. (94)

[128] R. Gavin, Y. Li, F. Petriello, and S. Quackenbush, *FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order,* Comput. Phys. Commun. **182**, pp. 2388–2403 (2011), doi:10.1016/j.cpc.2011.06.008, arXiv:1011.3540. (94)

[129] Y. Li and F. Petriello, *Combining QCD and electroweak corrections to dilepton production in FEWZ,* Phys. Rev. **D86**, p. 094034 (2012), doi:10.1103/PhysRevD.86.094034, arXiv:1208.5967. (94)

[130] R. Gavin, Y. Li, F. Petriello, and S. Quackenbush, *W Physics at the LHC with FEWZ 2.1,* Comput. Phys. Commun. **184**, pp. 208–214 (2013), doi: 10.1016/j.cpc.2012.09.005, arXiv:1201.5896. (94)

[131] M. Czakon, P. Fiedler, and A. Mitov, *Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$,* Phys. Rev. Lett. **110**, p. 252004 (2013), doi:10.1103/PhysRevLett.110.252004, arXiv:1303.6254. (94)

[132] N. Kidonakis, *Differential and total cross sections for top pair and single top production,* in Proceedings, 20th International Workshop on Deep-Inelastic Scattering and Related Subjects (DIS 2012) pp. 831–834 (2012), arXiv:1205.3453. (94)

[133] A. Hoecker *et al.*, *TMVA - Toolkit for Multivariate Data Analysis,* PoS **ACAT**, p. 040 (2007), arXiv:physics/0703039. (114, 121, 127, 149)

[134] K. I. Joy, *Bernstein polynomials,* Department of Computer Science, University of California, Davis, http://idav.ucdavis.edu/education/CAGDNotes/Bernstein-Polynomials.pdf. (144)

[135] ATLAS collaboration, *Measurement of the cross section of high transverse momentum $Z \to b\bar{b}$ production in proton–proton collisions at $\sqrt{s} = 8TeV$ with the ATLAS Detector,* Phys. Lett. **B738**, pp. 25–43 (2014), doi:10.1016/j.physletb.2014.09.020, arXiv:1404.7042. (144)

[136] M. Oreglia, *A Study of the Reactions $\psi' \to \gamma\gamma\psi$,* PhD thesis SLAC, 1980. (144)

[137] CMS collaboration, *Cms higgs analysis combined limit package (combine),* https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideHiggsAnalysisCombinedLimit (December 2015). (protected page). (144)

[138] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling,* eConf **C0303241**, p. MOLT007 (2003), arXiv:physics/0306116. (144, 145)

[139] ATLAS Collaboration, CMS Collaboration, LHC Higgs Combination Group, *Procedure for the LHC Higgs boson search combination in Summer 2011,* (August 2011), cds:1379837. (145)

[140] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics,* Eur. Phys. J. **C71**, p. 1554 (2011), doi:10.1140/epjc/s10052-011-1554-0, 10.1140/epjc/s10052-013-2501-z, arXiv:1007.1727. [Erratum: Eur. Phys. J.C73,2501(2013)]. (145, 146)

[141] T. Junk, *Confidence level computation for combining searches with small statistics,* Nucl. Instrum. Meth. **A434**, pp. 435–443 (1999), doi:10.1016/S0168-9002(99)00498-2, arXiv:hep-ex/9902006. (145)

[142] A. L. Read, *Presentation of search results: The CL(s) technique,* J. Phys. **G28**, pp. 2693–2704 (2002), doi:10.1088/0954-3899/28/10/313. (145, 147)

[143] LHC Higgs Cross Section Working Group, *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties,* CERN-2013-004, FERMILAB-CONF-13-667-T (2013), doi:10.5170/CERN-2013-004, arXiv:1307.1347. (173, 178)

[144] M. Botje *et al.*, *The PDF4LHC Working Group Interim Recommendations,* (2011), arXiv:1101.0538. (178)

[145] S. Alekhin *et al.*, *The PDF4LHC Working Group Interim Report,* (2011), arXiv:1101.0536. (178)

[146] CMS collaboration, *Search for the standard model higgs boson produced through vector boson fusion and decaying to $b\bar{b}$; additional figures,* http://cms-results.web.cern.ch/cms-results/public-results/publications/HIG-14-004/index.html (December 2015). (194)

List of abbreviations

ALICE	A Large Ion Collider Experiment 34
ATLAS	A Toroidal LHC ApparatuS i
$b\bar{b}$	Bottom/anti-bottom quark pair iii
BDT	Boosted Decision Tree 121
BEH	Brout-Englert-Higgs i
Calo	Calorimeter- 70
CERN	European Organization for Nuclear Research 17
CHS	Charged Hadron Subtraction 73
CL	Confidence Level 23
CMS	Compact Muon Solenoid i
CMSSW	CMS software 50
CSV	Combined Secondary Vertex 77
CTF	Combinatorial Track Finder 65
E_T^{miss}	Missing transverse energy 39
ECAL	Electromagnetic CALorimeter 38
eV	Electron volt i
η	Pseudorapidity 39
Fermilab	Fermi National Accelerator Laboratory 23
FSR	Final state radiation, cfr. parton shower. 57
$\gamma\gamma$	Photon pair 19
GF	Gluon-gluon fusion (production) 17
GWS	Glashow-Weinberg-Salam 3
$H \rightarrow b\bar{b}$	Higgs boson decaying to a bottom/anti-bottom quark pair ii
HCAL	Hadron CALorimeter 38
HE	High Efficiency 77
Higgs	Higgs boson i
HLT	High-Level Trigger 48
HP	High Purity 77
ISR	Initial state radiation, cfr. parton shower. 57

itPU In-time pile-up 35

JEC Jet Energy Calibration 73
JER Jet Energy Resolution 76
JES Jet Energy Scale 152
JP Jet probability 77
JPT Jet-plus-track 70

L1 Level-1 trigger 48
LEP Large Electron–Positron collider 23
LHC Large Hadron Collider i
LHCb Large Hadron Collider beauty 34
LHCf Large Hadron Collider forward 35
LO Leading Order 17

m_H Higgs boson mass 17
MB Minimum Bias: events triggered with a very general trigger, requiring just some activity in the event, nothing more restrictive 60
MC Monte Carlo: numerical method using pseudo-random numbers. 53
MoEDAL Monopole and Exotics Detector At the LHC 35
$\mu\mu$ Mu lepton pair 28

NLO Next-to-Leading Order 55
NNLO Next-to-Next-to-Leading Order 55

ootPU Out-of-time pile-up 35

PDF Parton distribution function 55
PF Particle Flow 70
pile-up The presence of additional proton-proton interactions in a collision event 35
PS Proton Synchtrotron 33
p_T Transverse momentum 40
PV Primary Vertex 40

Q Charge, generator of $U(1)_{EM}$, gauge group of electromagnetic interactions 5
QCD Quantum Chromodynamics 3
QED Quantum Electrodynamics 5
QFT Quantum Field Theory 3
QGL Quark Gluon Likelihood 79

ROOT Software package for data analysis 50

SM Standard Model i

SPS Super Proton Synchtrotron 33

$\sqrt{s}$ Centre-of-mass energy 17

SSB Spontaneous Symmetry Breaking 6

SSC Superconducting Super Collider 23

SSV Simple Secondary Vertex 77

SV Secondary Vertex 40

$t\bar{t}$ Top/anti-top quark pair 18

$t\bar{t}H$ Associated (production) with a top/anti-top quark pair 18

$\tau\tau$ Tau lepton pair 19

TC Track Counting 77

TCHE Track Counting High Efficiency 77

TCHP Track Counting High Purity 77

Tevatron Tevatron 23

TOTEM TOTal Elastic and diffractive cross section Measurement 35

UE Underlying Event: combination of shower activity, hadronisation and secondary interactions in collision events. 60

VBF Vector Boson Fusion (production) ii

VH Associated (production) with a vector boson 17

WW W boson pair 19

ZB Zero Bias: events triggered at random, without any trigger requirements 68

ZZ Z boson pair 19

List of figures

List of tables

Samenvatting

Onze huidige kennis van de fundamentele deeltjes en krachten voorkomend in de natuur, is vervat in het standaardmodel van de deeltjesfysica [1–7]. Een groot aantal theoretische en experimentele bijdragen, gedaan over vele jaren, hebben samen geleid tot de ontwikkeling van dit model in de jaren '60 en '70. Nog meer metingen hebben in tussentijd ingestaan voor de validatie ervan. Het standaardmodel bevat fermionen, deeltjes met halfwaardige spin, die de bouwstenen van alle materie vormen; en ijkbosonen, deeltjes met heeltallige spin, die instaan voor het dragen van de krachten. Drie van de vier gekende fundamentele wisselwerkingen worden beschreven door het standaardmodel: de elektromagnetische kracht, de zwakke kernkracht, en de sterke kernkracht; de zwaartekracht is niet inbegrepen. Het standaardmodel is een ijktheorie: de formulering van de theorie reflecteert de symmetrieën die kunnen worden waargenomen in de natuur. Een gevolg hiervan is dat het model, in haar meest simpele vorm, vereist dat alle deeltjes massaloos zouden zijn. Om het bestaan van massieve deeltjes te verklaren, wordt het Brout-Englert-Higgs mechanisme (BEH) [8–11] verwerkt in het standaardmodel.

Het BEH mechanisme werd onafhankelijk voorgesteld door verscheidene onderzoekers in 1964. Het beschrijft hoe deeltjes, die schijnbaar massaloos moeten zijn, massief kunnen worden. Er wordt een veld gepostuleerd – het BEH veld – alsook een daarmee geassocieerd boson – het Higgsboson. Dit veld genereert spontane symmetriebreking, en de koppeling van deeltjes met het veld zorgt ervoor dat ze massa verkrijgen. De massa van het Higgsboson zelf, een belangrijke parameter in het standaardmodel, kan niet theoretisch worden voorspeld. In de plaats daarvan moet ze experimenteel gemeten worden.

Gedurende vele jaren, sinds de voorspelling in 1964, zijn wetenschappers actief op zoek geweest naar het Higgsboson. Zoektochten werden uitgevoerd bij de Large Electron-Positron versneller in CERN[1] (1989–2000), bij de Tevatron in Fermilab (1987–2011), en meest recent bij de Large Hadron Collider (LHC) in CERN (2009–). De LHC is vandaag de dag de grootste en meest krachtige deeltjesversneller in de wereld. Hij werd ontworpen en gebouwd tussen 1984 en 2009, en is bedoeld om protonen te doen botsen bij massamiddelpuntsenergieën van 14 TeV. Een van de belangrijkste onderzoeksdoelstellingen van de LHC was het vinden van het Higgsboson. Sinds 2009 worden er bij de LHC botsingen uitgevoerd bij massamiddelpuntsenergieën van 7, 8 en 13 TeV. Het bestaan van het Higgsboson kon niet worden aangetoond, totdat in 2012, op 4 juli, zowel de ATLAS als de CMS collaboraties hun resultaten presenteerden en de ontdekking van een nieuw deeltje aankondigden. De eigenschappen van dit deeltje komen overeen met die van het standaardmodel Higgsboson. Sinds 2012 zijn verdere experimenten uitgevoerd, bedoeld om de karakteristieken van het deeltje verder de verifieëren; totnogtoe komen alle meting overeen met de voorspellingen van het standaardmodel.

[1] Europese Raad voor Kernonderzoek

Higgsbosonen kunnen op allerhande verschillende manieren worden geproduceerd en waargenomen bij botsingsexperimenten. In deze thesis stel ik een zoektocht naar het standaardmodel Higgsboson voor, waarin gekeken wordt naar Higgsbosonproductie via vector boson fusie (VBF), en Higgsbosonverval naar twee bottom quarks ($H \to b\bar{b}$) [116]. De data gebruikt voor deze analyse werd opgenomen met de CMS detector bij de LHC in 2012, tijdens Run I. De massamiddelpuntsenergie van de botsingen was op dat moment 8 TeV, en het experiment verzamelde 19.8 fb^{-1} (nominaal) + 18.3 fb^{-1} (uitgesteld) aan data.

Het standaardmodel Higgsboson werd voor het eerst waargenomen door CMS in de vervalkanalen naar $\gamma\gamma$, ZZ en WW. Zoektochten in het vervalkanaal naar twee bottom quarks werden ook reeds uitgevoerd door CMS, specifiek voor de gevallen waar de Higgsbosonen geproduceerd worden samen met een vector boson, of samen met twee top quarks. De waarnemingen hebben echter nog onvoldoende statistische significantie. De werkzame doorsnede van de Higgsbosonproductie via vector boson fusie is groter dan deze van de twee bovenvernoemde productiemechanismen, maar de gevoeligheid van een analyse in dit kanaal is lager. Dit komt doordat de eindtoestand van het VBF $H \to b\bar{b}$ proces volledig uit hadronen bestaat, en dat een analyse ervan alsdusdanig meer hinder ondervindt van een achtergrondsignaal veroorzaakt door QCD multijet interacties. Hoedanook, kan een analyse in het VBF kanaal wel bijdragen tot de gezamenlijke gevoeligheid van de $H \to b\bar{b}$ metingen, en mogelijkerwijze de resultaten voldoende verbeteren om te kunnen spreken van een waarneming van het Higgsboson vervallend naar fermionen. De VBF $H \to b\bar{b}$ analysis is ook de eerste meting bij de LHC van een Higgsbosonverval met een volledig hadronische eindtoestand.

De grootste uitdaging voor de VBF $H \to b\bar{b}$ analyse is het enorme QCD achtergrondsignaal. De strategie van het onderzoek bestaat uit drie delen: eerst wordt een selectie van de data uitgevoerd, zowel online (tijdens het experiment) als offline (achteraf). De online selectiecriteria spelen in op de eigenschappen van het VBF productiemechanisme, en op het gebruik van b-tagging. Offline worden aangepaste criteria gebruikt, geoptimaliseerd op basis van de waargenomen efficiëntie van de online selectie. Vervolgens worden extra variabelen gedefinieerd, gebaseerd op eigenschappen van het VBF $H \to b\bar{b}$ signaal. Een set van variabelen die goed het verschil tussen signaal en achtergrond weergeven, wordt uiteindelijk gecombineerd in een multivariante discriminant. Bij het uitvoeren van deze stap wordt vermeden variabelen te gebruiken die een correlatie vertonen met de invariante massa van het bottom quarkpaar. Zodoende kan tenslotte in de laatste stap een fit worden gemaakt van het massaspectrum, zonder rekening te moeten houden met vervormingen van het spectrum ten gevolge van de vorige stap.

Ik ben lid geweest van de VBF $H \to b\bar{b}$ onderzoeksgroep sinds het begin van de analyse in 2011, ongeveer tegelijkertijd met de start van mijn doctoraat. Een eerste ruwe versie van het onderzoek werd gepubliceerd in 2013, maar het was pas midden 2015 dat de tweede, verbeterde versie gepubliceerd werd. Gedurende de hele periode sinds 2011 was ik betrokken bij het schrijven van documentatie voor de analyse, het presenteren van de resultaten in de ruimere onderzoeksgroep, en het voorbereiden van het artikel voor publicatie. In 2011 en 2012 heb ik

vooral bijgedragen aan de ontwikkeling van de online selectie, de trigger. Ik stond naast de ontwikkeling van de selectie zelf ook in voor het bestuderen van de selectie-efficiëntie en het afleiden van correctiefactoren tussen experimentele data en simulatie. De studies uitgevoerd in deze context waren ook van belang voor de optimalisatie van de offline selectiecriteria. In de tweede fase van de analyse, in 2013, heb ik gewerkt aan de validatie van de data door het maken van vergelijkingen tussen experimentele- en simulatiedata. Gedurende de tweede helft van mijn doctoraat, in 2014 en 2015, was ik verantwoordelijk voor het op punt stellen en uitvoeren va de fit van het invariante massaspectrum van het bottom quarkpaar, en de statistische interpretatie van de resultaten. Tegelijkertijd was ik ook betrokken bij het onderzoek naar systematische onzekerheden die een rol spelen in deze analyse. Helemaal aan het einde van het onderzoek, heb ik mee ingestaan voor het combineren van de VBF $H \to b\bar{b}$ resultaten en deze van de eerder uitgevoerde $H \to b\bar{b}$ metingen.

De VBF $H \to b\bar{b}$ analyse neemt een verhoging waar van het gemeten signaal, in vergelijking met de door het standaardmodel voorspelde waarde. De statistische significantie van de resultaten is echter onvoldoende om te kunnen spreken van een Higgsbosonobservatie in het VBF kanaal. Ook in combinatie met de andere $H \to b\bar{b}$ resultaten blijft de significantie beneden de in de deeltjesfysica algemeen aangenomen drempelwaarde vereist om te kunnen spreken van een observatie. Recent werd ook nog een combinatie gemaakt van de $H \to b\bar{b}$ resultaten en deze uit het vervalkanaal naar twee tau leptonen ($H \to \tau\tau$). Met een significantie van de waarneming van 4.2 standaardafwijkingen, levert de CMS collaboratie het eerste voorzichtige bewijs voor het verval van het Higgsboson naar fermionen.

In de toekomst zal bij de LHC de zoektocht naar het Higgsboson, geproduceerd via het VBF mechanisme en vervallend naar bottom quarks, verdergezet worden. Meer data, bij hogere massamiddelpuntsenergieën, moet het mogelijk maken om in de nabije toekomst een observatie te doen van het Higgsbosonverval naar fermionen, met een significantie van vijf standaardafwijkingen of meer. In het algemeen zullen veel studies uitgevoerd tijdens Run II van de LHC als doel hebben de eigenschappen van het geobserveerde boson in detail te meten: de spin, de pariteit, de vervalbreedte (of levensduur), en de productie werkzame doornede. Daarna, maar ook al tegelijkertijd, ligt de weg open voor zoektochten naar andere Higgsbosonen, donkere materie, en meer exotische Nieuwe Fysica.

Curriculum Vitae

Personal Details

Name	Sara Caroline Alderweireldt	Address	Rozenlaan 14
Date of Birth	4th April 1988		2970 Schilde
Place of Birth	Wilrijk, Belgium		Belgium
Nationality	Belgian		

Education

2011–2016 **PhD in Experimental Particle Physics**, *University of Antwerp.*

Thesis title: Search for the Standard Model Higgs boson produced via vector boson fusion and decaying to bottom quarks

Promoter: Prof. Dr. N. van Remortel

2009–2011 **Master of Science in Physics** (with greatest honours), *University of Antwerp.*

Thesis title: Determination of PDF4MC parton densities from HERA data using the Pythia Monte Carlo generator

Promoters: Prof. Dr. P. Van Mechelen & Prof. Dr. H. Jung

2006–2009 **Bachelor of Science in Physics** (with greatest honours), *University of Antwerp.*

Thesis title: (experimental) Simulation study of the trigger rate for collisions with Drell-Yan lepton pairs in the CMS-CASTOR calorimeter

Promoters: Prof. Dr. P. Van Mechelen

Thesis title: (theoretical) Quantitative boundary layer characterisation for a $SrTiO_3/La_{2/3}Sr_{1/3}MnO_3/SrTiO_3$ multilayer structure using high-angle annular dark field scanning transmission electron microscopy

Promoters: Prof. Dr. S. Bals, Dr. S. Van Aert & Dr. J. Verbeeck

Funding

PhD Fellowship of the Research Foundation – Flanders (FWO), at the University of Antwerp, 4 years, from Oct 2011.

Post-doc Postdoctoral researcher for the Foundation for Fundamental Research on Matter (FOM), at Nikhef institute for Subatomic Physics, 3 years, from Jan 2016.

Research Experience

Higgs analysis My main area of expertise is Higgs Search, the topic of my PhD thesis being the search for the Standard Model Higgs boson produced by vector-boson fusion and decaying to b-quarks (VBF H→$b\bar{b}$). My experience in this field can be subdivided into a few categories:

- **Trigger and event selection**: At the start of the analysis in 2011, I was involved in the development of a dedicated VBF trigger, which was active during the full 2012 data-taking. Subsequently, as soon as data became available and again at different milestones during the runs, I performed trigger efficiency studies which first led to the optimisation of the trigger and then also allowed the determination of ideal offline preselection cuts for the analysis and the creation of data versus Monte Carlo correction factors.
- **Template creation and fitting**: The approach of the VBF H→$b\bar{b}$ analysis is to select events using dedicated triggers and preselections, then to make use of multivariate discriminators (independent of the $b\bar{b}$ invariant mass) to maximise signal versus background discrimination, and finally to fit the $b\bar{b}$ invariant mass spectrum looking for a mass bump. This last step requires templates for the different signal and background components. I worked on extracting these templates from Monte Carlo samples and building full fit models for the final fit using the RooFit package.
- **Systematic uncertainties**: I actively contributed to obtaining and evaluating a full list of both theoretical and experimental uncertainties for the VBF H→$b\bar{b}$ analysis, studying jet energy scale and resolution systematics, PDF uncertainties and trigger uncertainties.
- **Global fit and limit setting**: Bringing together all the components of the previous analysis steps, I learnt how to use the CMS combine toolkit to perform the global fit for the analysis and to calculate the limit and significance values.
- **Channel combination**: A combination was performed of the CMS H→$b\bar{b}$ results in the VH, $t\bar{t}$H and VBF channels. I studied the systematic uncertainties of the three analyses and the possible correlations, and prepared input configurations for the combination global fit.

Monte Carlo generation During my bachelor's and master's theses I was introduced to the use of Monte Carlo generators, looking mainly at Pythia 8 but also at other mainstream generators where needed. I used the LHAPDF, Rivet and Professor toolkits, first during my master's thesis when comparing theoretical predictions to Hera data, and later on during my PhD, in the context of the ridge project I worked on early in 2012, as well as when working on improving Pythia 8 tunes using LHC data during 2013.

Shift work I performed shift work for the CMS Experiment,
- online DQM CMS control room shifts
- offline CMS workflow monitoring shifts.

Hardware As a CERN summer student in 2010, I developed a monitoring tool with a graphical user interface to allow continuous monitoring of a set of high-voltage power supplies used in the Detector Control System of the ALICE experiment. In addition I gained experience with various tabletop electronics experiments as an assistant in the practical sessions of the electronics course at my university.

Skills

Programming I have experience with software development on Unix/Linux, Mac and Windows systems, using the bash, C, C++, Java, FORTRAN, php and Python programming languages and packages such as combine, CMSSW, Geant, LaTeX, OpenMP, Professor, Pythia, Rivet, ROOT and VTK

Languages Dutch (native speaker), English (full professional working proficiency), French (limited working proficiency), Italian (limited working proficiency), German (notions)

Summer Schools

- **MCNet and Terascale Alliance Monte Carlo School 2013**,
  University of Göttingen, 5th-9th August 2013.
- **BND Graduate School 2012**,
  University of Bonn, 21nd September - 2nd October 2012.
- **BND Graduate School 2011**,
  University of Nijmegen, 19th-30th September 2011.
- **CERN Summer Student Programme 2010**,
  CERN, 21st June - 3rd September 2010.

Research Stays

- **Collaboration with P. Skands on implementing a ridge model in the Pythia 8 Monte Carlo generator**,
  CERN, 13th January - 15th February 2012.

Conference and Workshop Talks and Posters

- **Study of Higgs bosons decaying to bottom quarks at CMS** (talk),
  SUSY 2015, *Lake Tahoe, US*, 24 August 2015.
- **Search for the Standard Model Higgs boson produced by vector boson fusion and decaying to bottom quarks** (poster),
  Lepton Photon 2015, *Ljubljana, Slovenia*, 17-22 August 2015.
- **Search for the Standard Model Higgs boson produced by vector boson fusion and decaying to bottom quarks** (poster),
  EPS-HEP 2015, *Vienna, Austria*, 22-29 July 2015.
- **VBF scalar to bb analysis** (talk),
  IAP meeting, *UCL, Louvain-la-Neuve, Belgium*, 19 June 2015.
- **CMS VBF Hbb analysis preapproval (CMS-HIG-14-004)** (talk),
  CMS HIG PAG, *CERN, Switzerland*, 9 September 2014.
- **CMS VBF Hbb analysis status** (talk),
  Scalar Search and Study in Belgium, *VUB, Brussels, Belgium*, 25 January 2014.
- **Vector Boson Fusion leading to Higgs production and subsequent decay in bottom quarks** (poster),
  EPS-HEP 2013, *Stockholm, Sweden*, 18-24 July 2013.
- **CMS VBF Hbb analysis update** (talk),
  IAP meeting, *UGhent, Ghent, Belgium*, 14 June 2013.
- **CMS VBF Hbb analysis approval (CMS-HIG-13-0.05)** (talk),
  CMS HIG central approval session, *CERN, Switzerland*, 10 May 2013.
- **CMS VBF Hbb analysis status** (talk),
  Higgs Search in Belgium, *ULB, Louvain-la-Neuve, Belgium*, 23 November 2012.
- **Obtaining the CMS ridge effect with multiparton interactions** (talk),
  MPI@LHC 2011, *DESY, Hamburg (Germany)*, 21 November 2011.

Teaching Experience

2014-2015
2013-2014
2012-2013 **Advanced C++ programming for physicists: teaching assistant supervising programming and problem solving tasks**,
University of Antwerp, undergraduate level, 48 contact hours/year.

2013-2014
2011-2012 **Electronics: teaching assistant supervising lab work and problem sets**,
University of Antwerp, undergraduate level, 24 contact hours/year.

2013-2014
2012-2013 **RooT introductory course**,
University of Antwerp, undergraduate level, 8 contact hours/year.

2012-2015 **Bachelor's thesis support for six undergraduate students**,
University of Antwerp.

Publications with significant contribution

- **Search for the standard model Higgs boson produced by vector-boson fusion and decaying to bottom quarks** – CMS-HIG-14-004.
  CMS collaboration, Phys. Rev. D92 (2015) 032008,
  doi:10.1103/PhysRevD.92.032008, arXiv:hep-ex/1506.01010
- **Higgs to bb in the VBF channel** – CMS-PAS-HIG-13-011.
  CMS collaboration (2013), CMS Physics Analysis Summary.
- **Proceedings, EPS-HEP 2013** – C13-07-18 (2013) 128, PoS.
  Contribution: Vector Boson Fusion leading to Higgs production and subsequent decay in bottom quarks, *S. Alderweireldt* (2013), inspire:1290947.
- **Proceedings, MPI@LHC 2011** – doi:10.3204/DESY-PROC-2012-03 (2012) 33.
  Contribution: Obtaining the CMS Ridge effect with Multiparton Interactions, *S. Alderweireldt & P. Van Mechelen* (2012), arXiv:hep-ph/1203.2048.

Publications as part of the CMS collaboration

247 publications, source: inspire-hep

Research Interests

Having been involved in particle physics as a Bachelor's, Master's and PhD student since 2008, I have had the opportunity to work with the CMS experiment at LHC throughout various interesting periods: the startup, Run 1, the Higgs discovery, and most recently Long Shutdown 1. As a PhD student I worked on Higgs analysis, performing the search for the Standard Model Higgs boson produced in vector-boson fusion and decaying to b-quarks (VBF $H \to b\bar{b}$). My contributions were focused on trigger development and efficiency studies, multivariate methods for event interpretation and signal versus background discrimination, fit model building in the context of a mass spectrum fit, systematic uncertainty evaluation, and limit setting. I gained practical experience in handling large data volumes and in software development. Over time, I experienced working within small as well as large working groups, and grew more proficient at communicating on both technical and informational levels.

Now at the end of LS1, a bright future continues to lie ahead for CERN and the LHC. Run 2 will open new research avenues by almost doubling the collision energy, but will also bring a multitude of theoretical, technological and computational challenges. There will be many interesting possibilities for immediate data analyses, while at the same time, we need to look beyond the next few years; with Long Shutdown 2, Run 3, HL-LHC and future experiments.

I am interested initially in two very broad areas of research in data analysis at LHC: Higgs precision measurements and continued searches on one hand, and Beyond the Standard Model searches including Supersymmetry and Dark Matter on the other hand. I am most drawn towards contributing in the field of Higgs Physics. The list of precision measurements to be performed is long: e.g. narrowing down the mass and the width of the boson, probing its spin and couplings – to vector bosons, fermions and itself – and testing various model parameters in fits. In parallel, opportunities are opening up to study rarer decays and more difficult channels, such as $H\rightarrow \mu\mu$ or $H\rightarrow Z\gamma$; or to perform searches for multiple Higgs. In line with my experience as a PhD student, it would be interesting to continue the VBF $H\rightarrow b\bar{b}$ search, which can contribute further to the measurement of the coupling of Higgs to fermions. Additionally, as the $t\bar{t}H$ production mode is the one with the highest cross section increase, I would like to participate in Run 2 studies probing the more sensitive decay channels in this mode. Finally, it would be rewarding to be able to take my knowledge of multi-(b)jet final states, and to apply it in searches for multiple Higgs or other BSM topologies with many jets.

Crucial to the success of any of the analyses however, is that we continuously rethink, retune and optimise our tools and methods. This will be especially relevant with the move from 8 to 13 TeV. For Run 2 the higher pileup conditions will pose a challenge for tracking, reconstruction and identification. In keeping with my PhD work, I am keen to participate in studies concerning jet tagging: b-tagging improvements, both at trigger level and offline; top tagging; quark/gluon discrimination; etc. Another topic of interest is trigger development, analysis specific as well as general, and in particular concerning topologies with many (tagable) jets.

Looking towards the future, at the front of hardware development, I would like to devote some of my time to projects concerning forward detector performance, mainly in tracking but also in calorimetry. Performance in this region will be a significant element in the possibilities for tagging and identification, which in turn is essential in many new physics searches, as well as for precision measurements. It also relates to my background working with VBF topologies, as studies with VBF or vector-boson scattering will benefit from improved forward coverage.

The last physics topic I discuss is Monte Carlo generation. Input from this field is very important in the workings of many analyses. Next to using classic generators which have widely proven their worth, it is necessary that newer methods are validated in analysis, alongside accepted ones. Additionally, the experimental community can make effective contributions to the performance and development of generators by providing data and helping to tune model parameters to the experimental setup at hand.

Finally, I want to be an advocate of good communication and participation in different research areas. We are all team players and only through interacting can the different communities effectively consolidate their efforts. For the same reasons, I'm very open to organising outreach projects and taking time to teach the next generation of scientists.

Acknowledgements

The past four years have been, in many ways, a very interesting adventure.

I would like to take this opportunity to extend my warmest thanks to all people who have made it possible for me to undertake this PhD project, and helped make it a wonderful experience.

First of all I'd like to thank Nick van Remortel, my promotor, who provided the opportunity for me to do a PhD in Antwerp, and introduced me to experimental Higgs boson research. You always identified nicely challenging goals, and allowed me the freedom to achieve them independently. Thanks for the guidance and support throughout these past four years, for sharing your experience, and for encouraging me to try new things.

I am also indebted to Sandra Van Aert, Michiel Wouters, Barbara Clerbaux and Ivo van Vulpen, for accepting to be in my PhD jury.

Next, I'm grateful to Xavier Janssen, whom I shared an office with for four years. Thanks for helping me get started, by introducing me to the people and software of CMS; for all the advice, technical and other; and for the many insights and ideas which led to nice solutions and results. In particular, thank you for your patience, and your wit.

Furthermore, I want to sincerely thank Sten Luyckx, who started his PhD at the same time I did, for being a great office mate and friend. We shared many learning experiences, and you helped me out massively on many occasions. Being able to discuss things with you has always been very efficient. Thanks for making the atmosphere in the office light, but being serious at all the right times.
In the same sense, a great thank you to Jasper Lauwers, who isn't easy to get to know, but proved to be a brilliant office mate as well, and filled Sten's shoes magnificently. You added various truly nice insights to my understanding of my own analysis, sometimes with simple offhand comments. Keep making them!

Then to all other members of the particle physics group, past and present, thank you for making it a great place to work, and for many pleasant lunches and coffee breaks.
In particular to Pierre van Mechelen, Benoît Roland and Sunil Bansal, for being the adult voices, and providing the answers to many questions. To Tom Mertens, Pieter Taels and Frederik Van der Veken, for drawing me into discussions and entertaining my questions about theoretical particle physics – the dark side. To Alex Van Spilbeeck, Hans Van Haevermaet, Tom Cornelis and Merijn Van de Klundert, for being great colleagues and friends, and guaranteeing fun times. Finally to Sarah Van Mierlo, for always being there to help out, solving all kinds of problems, and immediately understanding things that others did not.

I also send an enthusiastic thank you to Kostas Kousouris and Paolo Azzurri, and the other members of the VBF $H \to b\bar{b}$ group. You allowed me to learn and to grow, giving me a nudge here and there when I didn't always believe I could do something. Thank you for the many hours spent putting this analysis together and obtaining a nice result.
At the same time, thank you to all my other colleagues and friends at CERN. You all help make it a unique research environment, which I hope to be part of for a whole lot longer.

A very special thanks then goes to Ben Smart. We had many interesting discussions, and you always managed to add a view from a different perspective, which often provided the start to a solution. We shared many precious moments, in and outside physics, and all around Europe; these past few years wouldn't have been anywhere near as nice an experience without you. Thank you!

Thank you also to Senne, Anne, and Marjolijn, for being the most amazing friends I can think of, for reminding me that there is a life outside of particle physics, and for sharing their very own particular kinds wisdom. Thanks for being there and making my spare time enjoyable!

Finally, I wholeheartedly thank my parents and grandparents, for always encouraging me to pursue my interests and dreams, for making me aspire to do well, and for their unconditional support. Thank you also, for putting up with my temper, which flared occasionally when work was difficult. I couldn't have done this without you!

To all of you, thank you very much!