



Université Catholique de Louvain
Secteur des Sciences et Technologies
Institut de Recherche en Mathématiques et Physique
Centre for Cosmology, Particle Physics and Phenomenology

First measurement of the associated production of a Z boson with b jets at the LHC

Doctoral dissertation presented by

Ludivine CÉARD

in fulfillment of the requirement for the degree of Doctor in Sciences

Pr. Christophe DELAERE (<i>Advisor</i>)	UCL, Belgium
Pr. Jean-Marc GÉRARD (<i>Chairman</i>)	UCL, Belgium
Pr. Ernesto MIGLIORE	Università di Torino, Italy
Pr. Didar DOBUR	UGent, Belgium
Pr. Vincent LEMAÎTRE	UCL, Belgium
Pr. Christophe RINGEVAL	UCL, Belgium

June 8th, 2015

Nay, answer me: stand, and unfold yourself.

Shakespeare, from Hamlet

Remerciements

Durant la rédaction de ma thèse j'ai souvent pensé au moment où finalement j'écirai les remerciements. Celui-ci étant arrivé la tâche est plus compliquée que ce que j'imaginai. Durant ces années de thèse j'ai interagi avec de nombreuses personnes, qui grâce à leur passion, leur aide, leurs encouragements m'ont permis de finaliser ce travail.

Je voudrais en premier lieu remercier mon superviseur Christophe Delaere, qui a cru en moi à un moment clef de mon parcours et qui m'a offert sa confiance et la chance de travailler avec lui. Durant tout le parcours de ma thèse son aide, ses suggestions et les discussions que nous avons eues m'ont à chaque fois stimulée. Il m'a toujours poussée à me dépasser mais en veillant à ce que je conserve mon indépendance et que je puisse m'approprier pleinement mes travaux. Peu d'étudiants connaissent la chance de développer un code avec leur promoteur, ou celle d'écrire une note d'analyse à six mains. Ou l'opportunité de participer à une analyse de la première ligne de code à la dernière typo de l'article avant publication. Encore moins sont allés faire de l'escalade ou de la spéléologie avec leur promoteur. Pour tout cela je tiens à le remercier.

Je tiens aussi à remercier les Professeurs Jean-Marc Gérard, Ernesto Migliore, Didar Dobur, Vincent Lemaître et Christophe Ringeval d'avoir accepté de faire partie de mon jury. Leurs commentaires et les discussions que nous avons eues, ont été extrêmement utiles dans l'écriture du document final.

Je dois également remercier tous mes collaborateurs, Tristan avec qui l'analyse a débuté, Arnaud avec qui le PatTuple et MadGraph étaient plus drôles et qui m'a soutenue par skype plus tard quand il le fallait. Roberto mon colloc de bureau qui répondait à mes questions entre deux meetings, et Adrien C. sans qui je serais probablement toujours en train de débbugger mon code.

En plus de mes partenaires directs, CP3 est rempli de personnes sans qui ma thèse n'aurait pas été la même. Je veux donc dire merci à Lucia, Claude, Gauthier, Brieuc pour leur bonne humeur, leurs oreilles attentives et les discussions autour d'un café (ou un sacco de sushis). Merci à Antonio d'avoir vaincu l'application doctorale. Une mention spéciale pour les productions MadGraph et les discussions phéno. pour Michele, Alex M., Marco, Pierre et Fede. En prime la découverte de la meilleure pizzeria de Bruxelles avec Fede.

Mais aussi pour les séances de psychologie et de conduite improvisées avec Laurent F. et Bob, et tous les bons moments, et les bonnes bières, au bureau, à Louvain ou à Bruxelles. Je dois une grande reconnaissance à l'équipe informatique, Jérôme, Pavel et Juan qui m'ont tiré de nombreux mauvais pas.

Et bien sûr un grand merci à Ginette qui m'a aidée dans la jungle administrative pendant toutes ses années et à qui j'ai confié la liste exhaustive de tous mes petits problèmes.

J'ai eu la chance de passer la première année de ma thèse au CERN, où je dois saluer toute l'ancienne et actuelle équipe Tracker: Gaëlle, Gordon, Tomas ... J'ai passé toute cette année en compagnie d'amis formidables avec qui j'ai vécu mille aventures sur terre, sur neige, dans l'eau et dans l'air. Dani mon 'jeune' colloc, Ro, Marco, Albi, Klo, Kia, Mario, Maciek, Tuukka, Octavio, Valeria, Silvia et tous les autres! Chacun peut se targuer de m'avoir hébergé moi et/ou mes affaires.

De retour à Bruxelles je ne savais pas encore que j'allais passer de tout aussi bons moments avec des amis souvent confrères: Aidan, Giuseppe et Laurent T. qui m'ont accueillie à bras (et chauffage) ouverts dans leur bureau quand j'en avait besoin. Mais bien sûr aussi Maëlle, Adriana, Erik, Sabrina, Adrien T. et Alexandre L. qui a décortiqué l'unfolding avec moi à mains nues.

Au CERN Simon m'a apporté toute son aide et son enthousiasme pour cet unfolding. Je ne serais probablement jamais venue à bout de cette partie en solo de mon travail si il ne m'avait pas sommé d'arrêter de pinailler.

Merci également à Olivier qui est toujours là en cas de question de physique existentielle et urgente, mais aussi pour son aide dans l'éducation du tigre.

Et merci à mon amie de toujours (j'arrondis 26 ans à toujours), Amandine, qui a sans cesse cru en moi, qui a eu tous mes numéros de téléphone et à qui je dois un power point spécial sur ma thèse. Elle m'a tellement écoutée en parler pendant des heures qu'à présent elle utilise le mot 'plot' comme une vraie physicienne.

Enfin je remercie toute ma famille, pour leur support sans faille tout au long de ma vie et encore plus durant la thèse. Mes parents qui ont acceptés tous mes choix farfelus, mes grands-parents qui regardent les émissions sur le CERN, ma soeur et mon frère qui ont été derrière moi à chaque étape. Malgré la distance nous restons terriblement proches, et ils ont même eu le courage de venir me rendre visite dans cette nordique et glaciale contrée qu'est la Belgique.

Mais si il y a une personne qui mérite une médaille pour avoir traversé cette thèse avec moi, c'est bien Thomas. Non seulement il a brillamment réussi sa thèse, mais il m'a soutenue et écoutée tout au long de la mienne. Il a cru en moi à chaque instant et m'a interdit d'échouer. Les weekends au labo, les nuits de rédaction, le stress, le doute, tout ce qui aurait pu nous faire craquer nous a rendu plus forts. Danke für alles.

Abstract

The start of the LHC in 2009 opened a new era in particle physics. The collisions, produced at an energy and a rate unequalled, allowing to probe energy sectors and rare processes so far inaccessible. In summer 2012 the announcement of the discovery a new boson, presenting the properties of the long-sought Higgs boson was confirming the last prediction of the particle physics theoretical framework called the standard model. This theory, even though highly predictive and never defeated experimentally, is known to have limitations and lacks explanation for several physics observations.

The production of a Z boson in association with b jets is a process for which different theoretical predictions from the standard model differ, depending on the treatment of the b jets in the calculation (so-called 4- and 5-flavour schemes). The measurement of the Z plus b jets cross section and the study of its kinematics constitute then a strong QCD test. The Z plus at least two b jets process constitutes the main background for studies of the newly discovered boson, as well as for searches in theories beyond the standard model, presenting the same final state.

In this thesis the cross section measurement is done both for Z plus exactly one b jet and Z plus at least two b jets, together giving the result for Z plus at least one b jet. The data used, corresponding to an integrated luminosity of 5 fb^{-1} , were recorded by the CMS detector from proton-proton collisions at a centre-of-mass energy of 7 TeV at the LHC. The event selection is led by the care to optimise the signal over background ratio and to select robust well identified objects. b-tagging techniques are used for the selection of the b jets. A special attention is given to the background estimation, partially data-driven when possible. To yield a cross section at the particle level, all the efficiencies of reconstruction and selection and the detector effects are estimated taking into account the migrations in terms of b jet multiplicity. The results are compared with several theoretical predictions coming from calculations at leading order or next-to-leading order and in the two schemes available.

Important kinematic variables shown at the reconstructed level are also deconvoluted at the generator level. For this, two different unfolding techniques are explained and used: the Singular Value Decomposition and the Bayesian unfolding. The analysis selection was implemented in the RIVET framework, so that the differential cross sections can be compared to any generator.

Contents

Introduction	1
1 Physics at the LHC and Motivations for Z plus b jets Studies	5
1.1 The Standard Model	5
1.1.1 Building Blocks of the Matter	5
1.1.2 Fundamental Interactions	10
1.2 Physics at the LHC	15
1.2.1 A Collision Event	15
1.2.1.1 Parton Distribution Functions and Hard Process	15
1.2.1.2 Parton Shower	20
1.2.1.3 Matching Parton Showers and Matrix Elements	21
1.2.1.4 Hadronisation	23
1.2.1.5 Decay	24
1.2.1.6 Underlying Events	25
1.2.2 b-initiated Processes: the Z plus b jets Associated Pro- duction Case	26
1.2.2.1 Production	26
1.2.2.2 4-flavour and 5-flavour Schemes	28

1.2.2.3	Background for Searches	29
2	The Experimental Setup	33
2.1	The Large Hadron Collider	34
2.1.1	The CERN Accelerator Complex	35
2.1.2	The LHC Machine	36
2.1.3	Luminosity	38
2.1.4	LHC Run 1	39
2.1.5	LHC Schedule	41
2.2	The Compact Muon Solenoid	43
2.2.1	General CMS Layout	43
2.2.2	Coordinate System	44
2.2.3	Magnet System	45
2.2.4	Subdetectors	46
2.2.4.1	Tracker	46
2.2.4.2	Electromagnetic Calorimeter	48
2.2.4.3	Hadronic Calorimeter	50
2.2.4.4	Muon System	51
2.2.5	Trigger System	52
3	Measurement of the Associative Production Cross Section of a Z Boson with b jets	55
3.1	Analysis Strategy	56
3.2	Event Reconstruction and Selection	57
3.2.1	Datasets and Simulation Samples	57
3.2.2	Leptons	60
3.2.3	Jets and b-tagging	67
3.2.4	Missing Transverse Energy: MET	70

3.2.5	Summary of the Selection	73
3.2.6	Correction Factors	74
3.2.6.1	Pileup Reweighting	74
3.2.6.2	Lepton Scale Factors	75
3.2.6.3	b-tagging Scale Factors	80
3.2.7	Data and Simulation Yields Comparison	83
3.2.8	Data and MC Comparison of Kinematic Variables	84
3.3	Backgrounds	86
3.3.1	Background due to $t\bar{t}$ production	86
3.3.2	b-purity	91
3.3.2.1	Strategy: Per-jet Purity vs. Per-event Purity	91
3.3.2.2	Estimation of Jet Fractions with Templates Fit	92
3.3.2.3	Expected b-jet Content and Comparison with Data	95
3.3.2.4	Full Calculation of the Z+2b Purity	99
3.3.2.5	Cross Checks	100
3.3.3	Other Backgrounds: ZZ and ZH	102
3.3.4	Summary of the Backgrounds	103
3.4	Efficiencies and Migrations	104
3.4.1	Generator Level and Reconstructed Level Definitions . .	105
3.4.2	MET significance Efficiency	106
3.4.3	b-tagging Efficiency	106
3.4.4	Lepton Selection Efficiency	107
3.4.5	Reconstruction Efficiency	108
3.4.6	Lepton Acceptance	109
3.4.7	Final Matrix	110
3.5	Systematic Uncertainties	111
3.6	Cross Sections Results	115

4	Kinematic Observables and Unfolding	121
4.1	Interesting Kinematic Observables	121
4.2	Unfolding	126
4.2.1	Unfolding Methods	126
4.2.1.1	Bayesian Unfolding	127
4.2.1.2	Singular Value Decomposition	130
4.2.2	Performing the Unfolding	137
4.2.2.1	Construction of the Response Matrix	139
4.2.2.2	Background Subtraction	141
4.2.2.3	Choice of Method and Regularisation Parameter	144
4.3	Generators Comparisons	153
4.3.1	RIVET	153
4.3.2	Comparison at Generator Level	155
4.3.3	Comparison with Leading Order Generators	156
4.3.4	Comparison with Next-to-Leading Order Generator	158
4.3.5	Results	162
4.3.5.1	Transverse momentum of the Z boson	163
4.3.5.2	Transverse momentum of the leading b jet	164
4.3.5.3	Transverse momentum of the subleading b jet	165
4.3.5.4	Transverse momentum of the bb system	166
4.3.5.5	Invariant mass of the bb system	167
4.3.5.6	Invariant mass of the Zbb system	168
4.3.5.7	Difference ΔR between the two b jets	169
4.3.5.8	Azimuthal angular difference between the Z and the bb system	170
4.4	Summary	171
	Conclusions	173

A	Yields	177
B	Lepton Scale Factors	179
C	Unfolding Results for all the Kinematic Observables	185
D	Generator Comparisons for all the Kinematic Observables	211
	Bibliography	285

Introduction

The quest for knowledge about our surroundings and the universe is almost as old as mankind. In the context of centuries of science, particle physics sits as a relatively new field whose beginning lies at the end of the 19th century. This discipline took on the challenge to answer the biggest questions about the very composition of everything at the smallest scale and the description of all known physics phenomena. After several decades of discoveries and elaboration of concepts, its achievements were united in a global theory called the standard model (SM). This model tells us that in spite of the great variety of the nature, all matter can be described in terms of the same relatively small number of elementary constituents and their interactions via fundamental forces.

The standard model has been incredibly successful, with the observations of all its predicted particles and properties over the years. The discoveries were achieved in experiments always more powerful and technologically advanced than the previous ones, up to the actual Large Hadron Collider (LHC) at CERN¹, which produces the highest-energy man-made particle collisions. One of the reasons for building the LHC, the search for the last-to-be-observed standard model fundamental particle, the Higgs boson, was crowned with success with the announcement of the discovery of a new scalar boson in July 2012. So far this boson is showing all the properties of the SM Higgs boson. However, in spite of its accomplishments the standard model has limitations. On one hand limitations about answering all the physics observations and giving a complete picture of all fundamental forces, including the gravitation, which gave rise to other models. And on the other hand, limitations in the calculation accuracy or the need for measured parameters.

¹European Organization for Nuclear Research. The name CERN is derived from the French acronym for Conseil Européen pour la Recherche Nucléaire

In this context, the study of the associated production of a Z boson with one, or two or more b jets takes the verification of the SM, as well as the search for new physics, a step forward. Indeed, both the cross section measurement and the description of the kinematic variables provide an important test of Quantum Chromodynamics (QCD), for which the theoretical predictions for this process vary. The predictions are related to the treatment of the b quark in b-initiated processes, and the process is also sensitive to the b quark density in the proton, a so far unmeasured quantity. The process of the production of a Z in association with b jets also acts as one of the main backgrounds for the study of the vector boson associated production of the SM Higgs boson decaying to bottom quarks. For this channel, evidences have been found in data without claim for discovery, and it is still under investigation and measurement. Beyond the standard model, the Z plus b jets process is also a background in the search for new particles in certain supersymmetric or supersymmetric-like models.

The LHC provides the first set of events where a Z is produced together with b jets in proton-proton collisions, which are dominated by the gluon-gluon contribution. Previous measurements were performed at the Tevatron collider at Fermilab but arising from proton-antiproton collisions, thus dominated by another type of diagram (quark-antiquark). Besides, these former measurements were only presenting results as a ratio with respect to the inclusive Z cross section or the Z plus jets cross section. With an unprecedented number of Z plus b jets events selected (ten times higher than at the Tevatron), thanks to the high luminosity of the LHC, the first measurement of the associative production cross section of a Z Boson with b jets will be presented. The high luminosity allows even to perform separately the first cross section measurement for the production of a Z plus at least two b jets.

In physics, one wants to give the measurement of a process as it happens in nature, detached from the measurement tool effects, which is not 100% efficient and can further create distortions. This is done in the goal to be able to compare the measurement with the theory, but also to reproduce it with another experimental setting. This philosophy brought the cross section, as well as the variables of the Z plus b jets process, to be unfolded to their ‘truth’ level, via different techniques.

The thesis will be structured in the following way. Chapter 1 presents the particles of the standard model and the fundamental forces ruling their interactions. Furthermore, the theoretical notions involved in a proton-proton collision will be evoked, as well as their simulations to build templates to be compared to data. The signatures of the Z plus b jets process and the motivations for its

study, will then be developed.

Chapter 2 introduces the experimental facility that allows us to observe such events. The LHC, the complex machine responsible for the production of the data, will be generally described. A more detailed picture will be given of the Compact Muon Solenoid (CMS), one of the LHC experiments, and the detector from which the events are reconstructed and selected.

Chapter 3 presents in detail the cross section measurements performed for the associated production of a Z boson with either, exactly one b jet, or at least two b jets. The objects and events selection, the scale factor estimations, the background estimations and validations and the corrections for the detector effects will be exposed, before to reach the final results and the comparison with the theoretical predictions.

Finally, the unfolding of the differential cross sections is explained and performed in Chapter 4, followed by a comparison with different theoretical predictions.

Chapter 1

Physics at the LHC and Motivations for Z plus b jets Studies

1.1 The Standard Model

1.1.1 Building Blocks of the Matter

The matter around us, as well as the matter constituting us appears very varied, different and changing. But the entire description of matter can be in fact reduced to few elementary constituents and fundamental forces mediating their interactions. The theoretical framework summarising this knowledge, started to be elaborated from the sixties, is called the standard model.

A long way towards the infinitesimally small, which could start in the 5th century B.C., with Empedocles's model of the fundamental structure of the matter in four ultimate elements: water, earth, fire and air [1]. In spite of its elegance, this model was wrong, and 24 centuries later Mendeleev gave a new answer: a periodic table of over 100 chemical elements. But this was far from being the end of the path to answer about the composition of the matter. Nowadays knowledge of the elementary particles constituting the matter fits

in Table 1.1¹: 12 fundamental particles, the smallest building blocks with no substructure. The way to this modern conception of matter, included in the standard model, spreads over more than two centuries. It benefited from the contributions of so many scientists, and went through so many steps, experiments and reasoning, that it would go far beyond the scope of this thesis to claim to name them all. However, one can remind some of the key predictions and discoveries, as well as some important concepts.

Table 1.1: *The standard model elementary fermionic particles. In each box the symbol of the fermion, its name, its mass (top), its charge in units of elementary charge (middle) and its spin (bottom) are given. The masses are given here for illustration, the exact values with their uncertainties can be found in [2].*

		Generation					
		I		II		III	
Quarks	u	2.3 MeV		C	1.28 GeV	t	173 GeV
		$\frac{2}{3}$			$\frac{2}{3}$		$\frac{2}{3}$
	up	$\frac{1}{2}$		charm	$\frac{1}{2}$	top	$\frac{1}{2}$
	d	4.8 MeV		S	95 MeV	b	4.18 GeV
$-\frac{1}{3}$			$-\frac{1}{3}$		$-\frac{1}{3}$		
	down	$\frac{1}{2}$		strange	$\frac{1}{2}$	bottom	$\frac{1}{2}$
Leptons	e	511 keV		μ	106 MeV	τ	1.78 GeV
		-1			-1		-1
	electron	$\frac{1}{2}$		muon	$\frac{1}{2}$	tau	$\frac{1}{2}$
	ν_e	<2 eV		ν_μ	<0.19 MeV	ν_τ	<18.2 MeV
0			0		0		
	electron	$\frac{1}{2}$		muon	$\frac{1}{2}$	tau	$\frac{1}{2}$
	neutrino			neutrino		neutrino	

In Mendeleev's 1871 periodic table, the proliferation of elements and the systematic organisation, suggested a substructure. A first evidence-based theory of the atom had been developed in the early 19th century by Dalton, and stated that elements react in multiples of discrete units, the atoms. But it was the discovery of the electron in 1896 by Thomson that introduced the first probe for the atom substructure [3]. By measuring the mass of cathode rays, Thom-

¹The two fundamental constants of relativistic quantum mechanics are the Planck constant, h ($\hbar = \frac{h}{2\pi} = 1.055 \times 10^{-34}$ J·s), and the velocity of light in vacuum, c ($c = 2.998 \times 10^8$ m·s⁻¹). In this thesis the choice has been made to use a system of units in which $\hbar$ is one unit of action and c is one unit of velocity. So $\hbar = c = 1$ and the mass (m), momentum (mc) and energy (mc^2) will be all given in terms of GeV (or other appropriate factor of eV).

son showed that they were made of particles, but much lighter ones than the lightest atom, hydrogen. Therefore, they could not be atoms but new particles: the first subatomic particle had been discovered, which was later named electron, carrying the negative electric charge.

In 1911, after bombarding a metal foil with alpha particles (helium nuclei), and observing their deflection, Rutherford proposed that the positive charge of the atom is concentrated in a tiny nucleus at the centre of the atom [4]. In 1917, Rutherford proved that the hydrogen nucleus is present in other nuclei. The hydrogen atom was known to be the simplest and lightest atom, and its nucleus was given in 1920 the name of proton.

The existence of the neutron was not discovered until 1932 by Chadwick [5]. He demonstrated that the unusual long-range radiation, unaffected by electric fields, thought first to be γ rays, were in fact consisting of uncharged particles with about the same mass as the proton.

The model of the atom was given: a positively charged nucleus and negatively charged electrons. The nucleus being composed of two types of smaller particles: the neutrons carrying no charge and the proton carrying a positive electric charge. To ensure the electric neutrality of the atom there are as many electrons as protons. A chemical element is then defined as consisting of a single type of atom distinguished by its atomic number, which is the number of protons in its nucleus. The protons and neutrons are bound together with a strong or nuclear force to form the nucleus, which is bound with electron(s) via the electromagnetic force.

From the quantum mechanics point of view, electrons, protons and neutrons follow the Fermi–Dirac statistic and have a half-integer spin, which is an intrinsic form of angular momentum, discovered in 1922 with the Stern–Gerlach experiment [6] and integrated as quantum number by Pauli. The Fermi–Dirac statistic involves the obedience to the Pauli exclusion principle, which states that two identical fermions cannot occupy the same quantum state simultaneously. The conversion of neutrons into protons is done by the weak interaction, responsible for the radioactive β decay of nuclei, and in order to keep energy, momentum and spin conserved, Pauli introduced in 1930 a new particle, neutral and undetected. This particle was later called the electron neutrino [7] and was detected in 1956 [8].

The Leptons

The electron and the neutrino form a separate group of particles, elementary since without a substructure, and called leptons. Each represent one ‘flavour’ of leptons. The discovery in 1936, in cosmic rays, of the muon [9], a heavier version of the electron with the same charge and behaving in a similar way, brought another flavour to the lepton family. In 1962, at the Brookhaven National

Laboratory, it was showed that more than one type of neutrino exists, by first detecting interactions of the muon neutrino [10, 11] by shooting an accelerated proton beam on thick targets. When the third type of lepton, the tau, was discovered in 1975 at the Stanford Linear Accelerator Center (SLAC) [12], it was expected to have an associated neutrino too. The first evidence for this tau neutrino came from the observation of missing energy and momentum in tau decays (analogous to the beta decay), in 2000 at Fermilab [13].

The lepton group consists now of six flavours arranged in three generations. In addition, according to Dirac's theory each particle should have its own anti-particle characterised with the same mass but opposite charge. The anti-leptons take the name of the leptons preceded with the prefix 'anti-'. However, the anti-electron is more commonly called positron. The question whether the electrically neutral neutrinos are their own antiparticles (Majorana particles) or not (Dirac particles) remains open.

The Quarks

The protons and the neutrons are also not the only ones of their kind, but contrarily to the six leptons, they turned out to be just the lightest particles in a very large spectrum of strongly interacting fermion states, called baryons. A similarly high number of integer spin (0 or 1) particles called mesons have also been observed. The pion is the lightest meson and was discovered in 1947 from atmospheric cosmic rays [14], following the prediction by Yukawa [15]. All these particles, mesons and baryons, that undergo strong interactions are collectively called hadrons.

This proliferation suggested also a substructure of the hadrons. The quark model was independently proposed by the physicists Gell-Mann [16] and Zweig [17] in 1964. This model states that the hadrons are composed of smaller elementary particles, called quarks. The quarks are fermions with a spin $1/2$. The baryons, including protons and neutrons, are made of three quarks. The proton should be made of two so-called up quarks with electrical charge $+2/3$ and a so-called down quark with charge $-1/3$. For the neutrons the numbers of up and down quarks are reversed. Just like for the leptons, each quark has an anti-quark of the same mass but with opposite charge. The mesons are then formed by the combination of a quark and an anti-quark.

From the experimental side, it was in 1968, that deep inelastic scattering experiments at the SLAC showed that the proton contained much smaller, point-like objects. Therefore, confirming that the proton was not an elementary particle [18, 19]. The objects that were observed at SLAC would later be identified as up and down quarks, two different quark flavours.

Another flavour was predicted by Gell-Mann and Zweig's three-quark model:

the strange quark. Its existence was indirectly validated by the SLAC's scattering experiments, and provides an explanation for the kaon and pion hadrons discovered in cosmic rays.

Soon after, in 1970, the existence of a new charm quark was predicted in the Glashow–Iliopoulos–Maiani (GIM) mechanism [20]. This mechanism explained why the expressions that change the flavour of a fermion current without altering its electric charge are suppressed, and thus correspond to the experimental fact that such phenomena have never been observed. Charm quarks were produced almost simultaneously by two teams in 1974, one at SLAC and the other at the Brookhaven National Laboratory. The charm quarks were observed bound with charm antiquarks in the so-called J/Ψ meson.

The number of supposed quark flavours grew to the current six in 1973, when Kobayashi and Maskawa noted that the experimental observation of charge and parity (CP) violation [21] could be explained if there was another pair of quarks. Harari was the first to coin the terms top and bottom for the additional quarks [22]. And in 1977, the bottom quark was observed at Fermilab, via the observation of an Upsilon meson (Υ) formed from a bottom quark and its antiparticle [23]. After this success the existence of the top quark was expected, since without it the bottom quark would have been without a partner. However, it was not until 1995 that the top quark was finally observed by the CDF [24] and DØ [25] collaborations at Fermilab. Its mass was much larger than expected, almost as large as that of a gold atom, making it the heaviest of all the fundamental particles.

When implementing the quark scheme for all hadrons, a new problem to respect the Pauli principle appears. Let us take the example of the $\Delta^{++} = uuu$: discovered in 1951, the uuu configuration matches the properties of this baryon, but constitutes a completely symmetric ground state of identical fermions. In 1964, Greenberg introduced the notion of colour charge, a new property or quantum number for quarks, to explain how they can coexist inside some hadrons in otherwise identical quantum states. Quarks then come in three primary colours: red (R), green (G) and blue (B). The antiquarks are assigned the complementary colours, $\bar{R}$, $\bar{G}$, $\bar{B}$. Furthermore, all particle states observed in nature must be colourless or 'white', leading to a unique set of combinations of quarks: equal mixture of RGB, equal mixture of $\bar{R}\bar{G}\bar{B}$, and equal mixture of colour and complementary colour $R\bar{R}$, $G\bar{G}$, $B\bar{B}$. These possibilities correspond to the particle states observed in nature. The colour quantum is hidden in the sense that quarks themselves are coloured and thus not observable alone.

1.1.2 Fundamental Interactions

In order for all the fundamental particles to interact with each other, forces have to act between them. The electromagnetic and the strong forces have already been evoked. In total there are four fundamental forces, three of them mediated by particles with an integer spin, called bosons, following the Bose-Einstein statistics.

The electromagnetic force was the first known of these fundamental interactions. Electromagnetism is the force that acts between electrically charged particles. This phenomenon includes the electrostatic force acting between charged particles at rest, and the combined effect of electric and magnetic forces acting between charged particles moving relatively to each other. Electromagnetism is infinite-ranged and is carried by a massless boson called photon.

It was in 1873 that Maxwell unified the preceding developments on electricity and magnetism, into a single theory and discovered the electromagnetic nature of light. In classical electromagnetism, the electromagnetic field obeys a set of equations known as Maxwell's equations, and the electromagnetic force is given by the Lorentz force law. But this classical electromagnetism was difficult to reconcile with classical mechanics. In 1905, Einstein solved this problem with the introduction of the special relativity, which replaces classical kinematics with a new theory of kinematics that is compatible with classical electromagnetism [26]. Moreover, Einstein undermined the very foundations of classical electromagnetism. He was inspired by Planck's 'quanta' idea and posited that light could exist in discrete particle-like quantities, taking latter the name of photons. In this theory of the photoelectric effect, the light is no longer considered a continuous wave, and can explain the work led by Planck who already showed that hot objects emit electromagnetic radiation in discrete packets, the quanta [27]. Planck's and Einstein's theories were progenitors of quantum mechanics, which, when formulated in 1925, necessitated the invention of a quantum theory of electromagnetism. This theory, completed in the 1940s-1950s, describing the electromagnetic interactions of electrons and photons, is known as quantum electrodynamics (QED), and was the prototype of quantum field theory (QFT).

The strong interaction is the force that holds quarks together to form protons, neutrons, and other hadron particles. The colour charge assigned to the quarks is the charge of this strong field, mediated by quanta called gluons. These massless particles are exchanged between the quarks and antiquarks and must be coloured themselves. They are in fact bicoloured objects which exist in eight different states, and can interact with each other. By analogy with

QED, a quantised theory of the colour, the quantum chromodynamics, can be constructed. It is also a renormalisable gauge theory, but with eight gluons instead of one photon, that can interact with themselves, making it a ‘non-Abelian’ theory. The quarks and the gluons are globally called partons.

The strength of the strong force (approximately 100 times stronger than electromagnetism at its effective distance of 10^{-15} m), is such that the energy of quarks being separated is high enough to produce new massive particles. Thus, if hadrons are struck by high-energy particles, they give rise to new hadrons instead of emitting freely moving radiations (gluons). This property of the strong force is called colour confinement, and it prevents the free ‘emission’ of strong force: instead, in practice, jets of massive colour neutral particles are observed.

The strong interaction is also responsible for the binding of protons and neutrons together to form atoms. In this context it is called the nuclear force (or residual strong force). In this case, it is the residue of the strong interaction between the quarks that make up the protons and neutrons. It can be compared to the Van der Waals interaction in electromagnetism, which can make neutral molecule interact via high order dipole-dipole effect. As such, the residual strong interaction obeys a quite different distance-dependent behaviour between nucleons, from when it is acting to bind quarks within nucleons. The binding energy, that is partly released upon breakup of a nucleus, is related to the residual strong force and is used in nuclear power.

To compare the behaviour of the electromagnetic and the strong force one can perform a *charge screening* and note the drastically different effects on QED and QCD. In QFT the electron is not just a point charge particle but it can emit a photon subsequently annihilating into an e^+e^- pair and so on. The original electron is seen as surrounded by e^+e^- pairs, the positive charges being closer to the electron and the negative charges further away, screening the negative charge of the original electron. When moving a positive test charge closer to the electron, the cloud of positrons is penetrated, and the closer one comes to the electron the higher is the charge measured. In QED the measured charge depends on the distance at which one is probing the electron, as illustrated in Figure 1.1 (a) from [28].

The same calculation can be carried out for the colour charge, except that in this case the gluons, themselves carrying a colour charge, introduce new configurations involving gluons splitting into gluon pairs. They spread the colour charge of the quarks, resulting in, for example, a red charge to be preferentially surrounded by other red charges. When one moves a test charge closer to the original red quark, the cloud of red charges is penetrated and the measured red

charge decreases. This behaviour is illustrated in Figure 1.1 (b) and is called ‘asymptotic freedom’, which means that two red quarks with a reduced strength colour field behave essentially as free particles for very small separation. The ‘colour confinement’ is also seen as the colour charge tends to infinity when the quarks are separated, making the creation of new heavy particles possible.

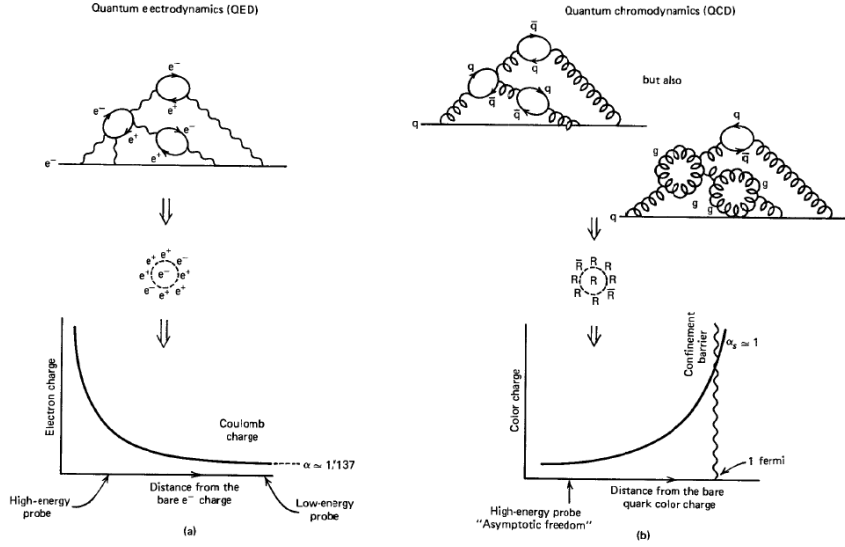


Figure 1.1: Screening of the electric (a) and colour (b) charge in quantum field theory [28].

The weak interaction is the mechanism responsible for the weak force or weak nuclear force. It is mediated by two charged bosons $W^\pm$ and a neutral boson Z , which are emitted and absorbed. They are heavy (of the order of 100 GeV) and therefore short-lived. The range of the interaction is very short (of the order of 10^{-17} m). The weak field couples with equal force to quarks and leptons, but with a strength of several orders of magnitude less than the electromagnetic force, or than the strong force. Historically, the development of the weak interaction theory as we know it nowadays ranges from the β rays problem in 1914, to the observation of the W and Z bosons at the Super Proton Synchrotron at CERN in 1983. In the late 19th century the β rays posed the problem of having a continuous spectrum, contrarily to the other known α and γ rays. Besides, in 1920 Rutherford had described the nucleus as consisting of protons and electrons, which was causing a problem of behaviour of some atoms supposed to be fermions but behaving like bosons. The predictions of Pauli led

to the discovery of the neutron and the prediction of the neutrino, the β particle being then an electron or positron. In 1933, Fermi proposed the first theory of the weak interaction, known as Fermi's interaction [29]. He suggested that the β decay could be explained by a four-fermion interaction, involving a contact force with no range, which is now known to not be exactly correct, and better described as a non-contact force field having a very small finite range. From this corner stone the building of the theory involves many brilliant physicists over years of discoveries. One can mention some essential steps like the prediction of the violation of the reflection symmetry and, thus, parity violation, in 1956 by Lee and Yang [30], which led Wu to observe it experimentally one year later, in the disintegration of cobalt atoms with the spin oriented in a magnetic field [31]. The V-A theory from Feynman, Gell-Mann, Sudarshan and Marshak in 1958 introduces the vector bosons as mediators of the force [32, 33]. The weak neutral currents, inducing flavour-changing processes, were only predicted in 1973 by Glashow, Salam and Weinberg and confirmed the same year, in a neutrino experiment in the Gargamelle bubble chamber at CERN [34]. The existence of the W and Z bosons was not directly confirmed until 1983 [35, 36, 37, 38].

The standard model of particle physics describes the electromagnetic interaction and the weak interaction as two different aspects of the electroweak interaction, the theory of which was developed around 1968 by Glashow, Salam and Weinberg.

The Brout-Englert-Higgs (BEH) mechanism provides an explanation for the presence of three massive gauge bosons, carriers of the weak interaction, and the massless photon of the electromagnetic interaction. The BEH field was proposed in 1964 independently by Brout and Englert on one side [39], and Higgs on the other side [40], and is mediated by the so-called Higgs boson. The masses of the fermions arise from the Yukawa interaction that predicts their couplings with the Higgs boson, the coupling strength being proportional to the mass of the particle. The Higgs boson was the last predicted particle of the SM not to have been observed, and its existence was only confirmed in July 2012 at CERN [41, 42].

The fourth force of the nature, missing in this list, is probably the most known: **the gravity**. This force, also called gravitation, describes how all physical bodies attract each other and acquire weight. However, even if dominant at an astronomical scale (electrostatic attraction being not relevant for large celestial bodies which are seen as with a net electric charge of zero), it is negligible on a microscopic scale where elementary particle physics happens. Even if acting in principle on all particles, it could not be integrated in the SM due to the exclusive long-range nature of its interaction. Newton's law of universal gravitation

in 1687 was a good approximation of the behaviour of gravitation. It postulates that the gravitational force of two bodies of mass is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. But our present-day understanding of gravitation stems from Einstein's general theory of relativity of 1915, a more accurate description of gravitation in terms of the geometry of space-time [43]. Gravity is described there as a consequence of the curvature of space-time.

Merging general relativity and quantum mechanics (or quantum field theory) into a more general theory of quantum gravity is an area of active research. In some beyond the standard model theories it is hypothesised that gravitation is mediated by a massless spin-2 particle called the graviton.

All standard model bosons are summarised in Table 1.2.

Table 1.2: *The standard model elementary bosons. In each box the symbol of the boson, its name, its mass (top), its charge in unit of elementary charge (middle) and its spin (bottom) are given. The masses are given here for illustration, the exact values with their uncertainties can be found in [2], and in [44] for the Higgs boson.*

* *The mass of the gluon is theoretical.*

Fundamental Force	Boson	
Electromagnetic	γ	0 eV
		0
	photon	1
Strong	g	0 eV *
		0
	gluon	1
Weak	$W^{\pm}$	80.4 GeV
		± 1
	W boson	1
	Z^0	91.2 GeV
		0
	Z boson	1
	H^0	125.09 GeV
		0
	Higgs boson	0

1.2 Physics at the LHC

This section is dedicated to the description of what is happening in a collision of accelerated protons. The discussion is oriented towards the events produced at the Large Hadron Collider at CERN in France and Switzerland. Even though the acceleration and detection of the products of collisions will be discussed in the Chapter 2, we set up here the theoretical concepts involved in these collisions, coming from the partons properties and actions of forces.

The complex procedure of the collision can in fact be simulated, based on a theory, by various programs, relying on the Monte Carlo integration method. The theory can be the standard model or alternatively beyond the SM theories. A Monte Carlo event generator is a program able to reproduce the particles resulting from the collision with the same probability of production as in the nature. This event by event simulation of the collisions will represent our theoretical benchmark for later comparison with the results obtained with data from experiments. It is thus essential to understand the details and the eventual differences of the simulations, to give an interpretation of the comparison.

The description can be separated in a number of sequential steps listed in Figure 1.2. The figure displays this process from the initial interaction between protons, down to the decay of long-lived particles in the experiment. Such a factorised approach allows the fine tuning of each individual step.

1.2.1 A Collision Event

1.2.1.1 Parton Distribution Functions and Hard Process

As seen before, the proton is made of two up quarks and one down quark with the colour charges R, G, B. Gluons are constantly exchanged between quarks through a virtual emission and absorption process, during which a colour exchange occurs. These individual quarks are called valence quarks and provide only about 1% of the mass-energy of the proton; most of the mass-energy of the proton being in the form of the strong force field energy. Since gluons carry colour charge, they themselves are able to emit and absorb other gluons, but also to split into a virtual quark-antiquark pair that annihilates shortly afterwards into another gluon within the proton. This results in a constant flux of gluon splits and creations, commonly known as ‘the sea’. The quarks populating it are called sea quarks and can be of any flavour.

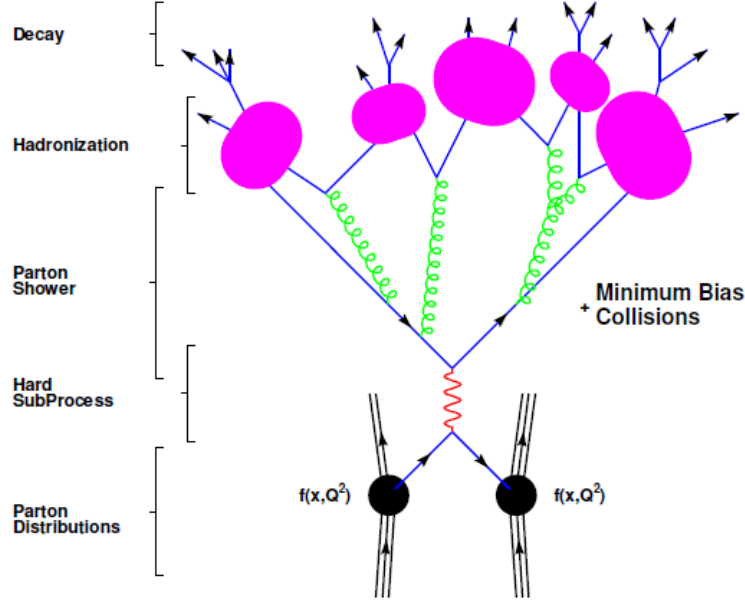


Figure 1.2: Overview of the event generation procedure.

The interaction between two incoming accelerated protons is often soft and elastic and will not lead to the creation of a Z boson for example. Only a hard fundamental interaction between the partons of the two protons will result in the process we are interested in.

When two protons collide, each parton takes a certain fraction x of the initial proton energy and will interact with a parton of the opposite proton. The probability density $f_p(x, Q^2)$ to find a parton p , with the fraction x of the longitudinal proton momentum in the proton-proton centre of mass frame depends on Q^2 , where Q is the four-momentum transfer between the partons of the collision, or the energy scale of the hard interaction. The probability to find a gluon or a quark inside a hadron depends on the scale at which one is probing it. The parton distribution function (PDF) describes this probability. PDFs cannot be calculated from first principles and need to be measured from experiments. Currently, they are essentially extracted from the study of lepton-hadron collisions such as those provided by HERA [45, 46]. At hadron colliders, some constraints are also obtained by the study of the Drell–Yan process [47] or by the measurement of the W^+W^- asymmetry [48]. Various sets of PDFs exist and are provided by different collaborations such as CTEQ [49, 50], MSTW [51] and NNPDF [52].

An example of parton distribution functions is shown in Figure 1.3 for two different Q^2 scales. For low value of Q^2 the proton can be described only with its three valence quarks, but at higher Q^2 the sea quarks and gluons can be probed as well and must be added to the description of the proton.

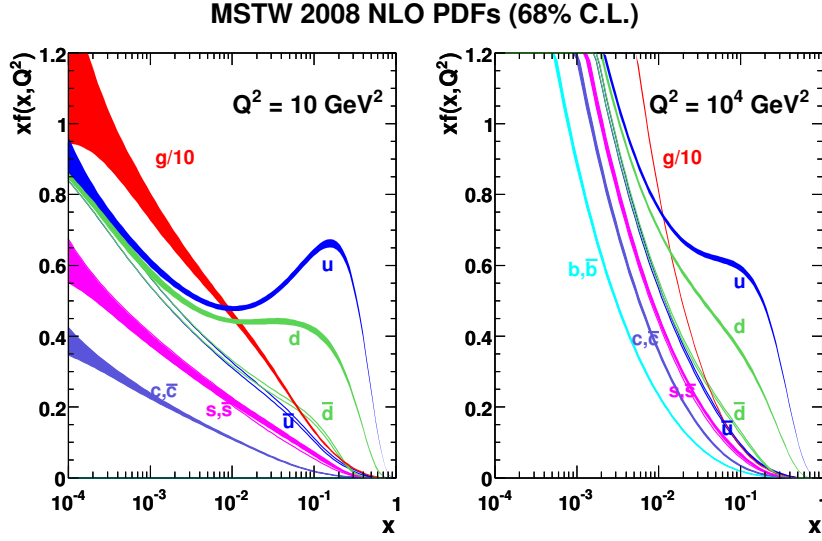


Figure 1.3: PDFs for the different partons in a proton, including the one standard deviation (1σ) uncertainty bands, with the MSTW 2008 NLO parametrization, for two different energy scales Q^2 [51].

To probe physics at a certain energy scale, the value for Q^2 has to be taken in the range of the invariant mass M^2 of the system, with $M = \sqrt{x_1 x_2 s}$, where s is the squared centre of mass energy of the colliding protons, and x_1 and x_2 are the momentum fractions taken by each of the interacting partons. In Figure 1.4 another PDFs example is given with a Q^2 chosen at the scale of the Z boson mass (91.18 GeV [2]), where the gluons and the valence quarks are dominating.

The PDF measured at a given Q^2 can be extrapolated at another energy using the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equation [53, 54, 55].

The total cross section for hadronic process can then be computed by separating the short distance physics, i.e. hard scattering described by perturbative QCD, from the long distance scattering, i.e. the kinematic of the proton described by the PDFs:

$$\sigma_{pp \rightarrow X} = \sum_{p,p'} \int_0^1 dx_1 dx_2 f_p(x_1, \mu_F^2) f_{p'}(x_2, \mu_F^2) \hat{\sigma}(x_1, x_2, \mu_F, \alpha_s(\mu_R)) \quad (1.1)$$

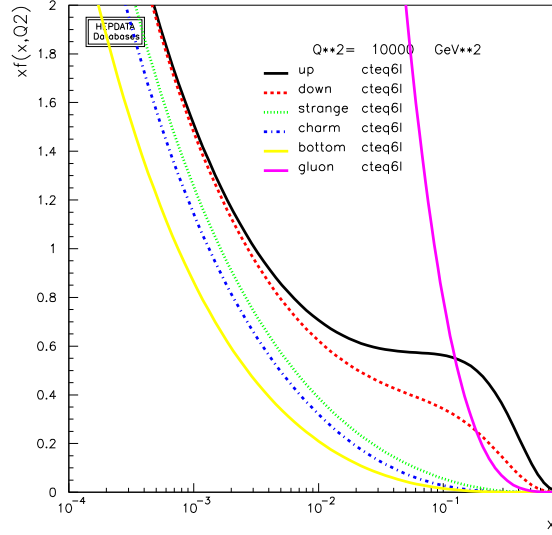


Figure 1.4: PDF for the different partons in a proton, obtained with the CTEQ6L parametrization. Q^2 is chosen for physics studies at the Z boson mass scale. The plot was generated with the tool from the HepData project [56].

This is the *factorisation theorem* which relates the total cross section to the partonic one. μ_R is the scale at which the strong coupling constant α_s is measured and is called the renormalisation scale. μ_F is the mathematical scale at which the decoupling of short and long range physics happens and is called the factorisation scale. The hard scattering can then be calculated at a given order for a given process and convoluted with the PDFs that are universal functions, not depending on the process under consideration. The partonic cross section at the n -th order in terms of the strong coupling constant α_S is given by:

$$\hat{\sigma} = \alpha_S^k \sum_{i=0}^n C_i \alpha_S^i. \quad (1.2)$$

The coefficients C_i depend on the kinematic of the process but also on the renormalisation and factorisation scales and can be calculated in orders of α_S . If $i = 0$ the process is said to be calculated at leading order (LO), if $i = 1$ at next-to-leading order (NLO), if $i = 2$ at next-to-next-to-leading order (NNLO), etc. The mathematical dependence of C_i to μ_F and μ_R is of the form of $\log\left(\frac{\mu}{Q}\right)$ where Q is the typical scale of the process under consideration (for a Z boson production that would be its mass). If one could calculate the partonic cross section at infinite order the scales terms would cancel each other; but in order

to calculate it at a certain order, the terms of higher orders must remain small, and this is only possible if the $\log\left(\frac{\mu}{Q}\right)$ stay small. Since a choice has to be made for the values of these scales, the most reasonable is typically $\mu_F = \mu_R = Q$.

The description of the DGLAP equation and renormalisation group equations go beyond the scope of this manuscript, but were evoked to introduce the renormalisation and factorisation scales that will be discussed later on for some of the simulations.

The partonic cross section can be expressed in terms of the probability amplitude $|\mathcal{M}|^2$ of the hard scattering for a specific process between these partons. This amplitude is derived from the elements of the scattering matrix, that relates the initial state and the final state of a physical system undergoing a scattering process. The perturbative calculation of the scattering matrix results to a Feynman amplitude which contains parton level informations about couplings between particles, intermediate resonances in the process and the spin correlations. The calculation of this hard scattering process probability amplitude is called matrix element (ME) calculation.

To simulate the first part of the process involving the PDFs description, one can use general-purpose programs like PYTHIA [57, 58] or HERWIG [59]. These are complete event generation chains, meaning that in addition to the hard process they also take care of the steps that will be described in the next sections: parton showering, hadronisation and underlying event description. For the matrix element calculation these generators only consider the leading order. The radiative corrections being not explicitly calculated but rather treated in the parton shower. As it will be discussed in the next section, the parton shower describes well low Q^2 radiations but fails to accurately describe the hard radiations. This results in a limited performance for PYTHIA and HERWIG to describe events with additional hard radiations (jets) in the final state.

A more accurate approach is the one followed by MADGRAPH/MADEVENT [60, 61]. MADGRAPH is a matrix element generator that can generate the leading order Feynman diagrams for any given process based on the standard model, or any other user-defined model. It allows to calculate exactly the Feynman amplitudes associated to a process, taking into account for instance the possible interferences. Here, up to five extra partons are included in the matrix element. The integration over a phase space gives an estimation of the cross section, and the generation of events based on the diagrams produced by MADGRAPH, is taken care of by MADEVENT. The event description from MADGRAPH already gives a better description than the other LO generators because it includes the real corrections.

To take into account the virtual loop corrections as well, and have a complete NLO calculation, one can use the most recent version of MADGRAPH: MADGRAPH5_aMC@NLO [62]. In this version the order of the QCD corrections can be chosen to be LO or NLO. When chosen to be at NLO it will be said to be generated with aMC@NLO in the rest of the thesis. The parton showering will still be needed on top of the ME generation, to describe softer radiations.

Other programs, like PowHeg [63] or MC@NLO [64, 65], provide a complete NLO calculation including both real and virtual loop corrections. They are therefore limited to the radiation of one extra parton. Where MADGRAPH produces real corrections to the LO calculation for up to five additional partons, this kind of NLO generators must rely on the parton shower approximation beyond one additional parton. However, it accurately predicts inclusive observables to NLO accuracy like inclusive cross sections.

1.2.1.2 Parton Shower

Incoming partons involved in the initial state, and outgoing partons involved in the final state of the hard process, are subject to radiations of quarks and gluons. Incorporation in the matrix element should be done for hard radiations and is possible but the current generators are limited to a certain number of partons in the final state. Moreover, the ME calculation leads to divergences for collinear or soft radiation if one does not consider both real and virtual corrections.

To form a parton shower, successive branchings are performed on the partons until the energy drops below an energy scale where $\alpha_s \sim 1$, where the non-perturbative regime is reached and where hadronisation will take over. When a parton a branches into the partons b with a fraction z of the energy of a , and c with an energy fraction $1 - z$, the branching probability is given by:

$$d\mathcal{P}_{a \rightarrow bc} = \frac{\alpha_s}{2\pi} \frac{dQ^2}{Q^2} \mathcal{P}_{a \rightarrow bc}(z) dz \exp \left(\underbrace{- \sum_{b,c} \int_{Q^2}^{Q_{max}^2} \frac{dQ'^2}{Q'^2} \int \frac{\alpha_s}{2\pi} \frac{dQ^2}{Q^2} \mathcal{P}_{a \rightarrow bc}(z') dz'}_{\text{Sudakov form factor}} \right) \quad (1.3)$$

In equation 1.3 one gets rid of the divergences, i.e. when $z \rightarrow 0$ (soft divergence) or when $Q^2 \rightarrow 0$ (collinear divergence), by the implementation of a cut-off scale Q_{cutoff}^2 of the order of 1 GeV^2 at which hadronisation occurs. To prevent the branching ratio to be larger than 1, a Sudakov form factor is added [66].

This factor represents the probability $\mathcal{P}_{a \rightarrow a}(Q^2, Q_{max}^2)$ for a parton a to evolve without radiating from a virtuality Q_{max}^2 to a virtuality Q^2 . In the case of final state radiation (FSR), for a given event, a random number $\mathcal{R}$ is drawn from a uniform distribution in range [0–1] and a branching is simulated at a Q^2 value such that $\mathcal{P}_{a \rightarrow a}(Q^2, Q_{max}^2) = \mathcal{R}$ if $Q_{max}^2 > Q_{cutoff}^2$. This showering is then repeated for the daughter partons until reaching the Q_{cutoff}^2 scale. The case of initial state radiation (ISR) is more complex since the incoming protons have a structure. The ISR is done backwards and the sign of the virtualities considered will be negative. It is important to note that the parton showering can be evolved with other variables than the virtualities evoked so far. In PYTHIA for example the evolution can also be performed in terms of the squared transverse momentum $p_T^2 \approx z(1-z)m^2$.

1.2.1.3 Matching Parton Showers and Matrix Elements

In the previous two sections, the higher order matrix elements and parton showers were introduced as two possibilities to describe the ISR and FSR in an event. On one hand the ME can describe a $2 \rightarrow 3$ particles process with full NLO accuracy, and provide real corrections to LO diagrams, with the generation of up to four additional partons in the case of a Z boson production. On the other hand the parton showering has no constrain on the multiplicity of the parton and can describe soft and collinear emission without divergences. The two approaches are in fact not alternative but complementary, and should be combined to give an optimal description of what happens in nature and is seen in the data. When doing this merging a double counting of radiation might occur since an (n+1)-jet event can be created in two ways: either with the ME generator calculating an (n+1)-parton hard interaction, which when passed on to the parton shower, can produce a radiation in a cascade, observed as a jet; or from an n-parton event, when a sufficiently hard additional parton is generated in the parton shower leading also to an (n+1)-jet event. To avoid this double counting different jet merging (also called matching) techniques exist. The principle is to divide the phase space in two different regions, characterised by the hardness of the QCD radiations. The region with soft and collinear emissions will be described by parton showering, while the region with harder emissions will be taken care of by the ME generator. The scale chosen to separate the two regions is generically called the matching scale Q_{match} . The following discussion will be restricted to MADGRAPH interfaced with PYTHIA that are used to generate the signal events for this thesis.

The MLM matching approach

The MLM matching regroups the k_T -MLM and the shower- k_T schemes. Both use the transverse momentum measure (called k_T in jet algorithms) to determine the Q_{match} and are based on the original MLM method [67, 68]. The k_T -MLM is the default matching scheme in PYTHIA and the one used for the event generation in this thesis. It works as follows:

- The matrix elements events are generated in MADGRAPH with additional light ISR partons (for example Z+0,1,2,3,4 partons), distant from each other by Q_{cut}^{ME} . The events with ME partons below the scale Q_{match} should be rejected after parton shower. However, since the cut will be performed on the combined radiation from the shower, a smearing due to the size of the jet across the matching scale may happen. The rejection cut Q_{cut}^{ME} will then be taken smaller than Q_{match} .
- A reweighting by $\alpha_s(k_T)$ at each vertex is done, in order to reproduce what is performed in the shower algorithm. This allows a smoother transition between the two regions of phase-space.
- The parton showering is performed with PYTHIA, by default the p_T -ordered showers will be used.
- The partons are clustered in jets using the k_T algorithm [69].
- Starting from the hardest parton the algorithm looks for the closest jet. The jet is said to be matched to the parton if at smaller distance than a maximum determined by Q_{match} . Once matched, the jet is removed from the list to avoid double matching, and the matching continues to the next hardest parton.
- For events with a multiplicity below the highest multiplicity (for example Z+0,1,2,3 partons), all jets must match a parton and inversely. For the highest multiplicity (for example Z+4 partons), the partons should all have matched jets and the rest of the jets can remain unmatched, describing soft radiations.

The shower- k_T scheme works in the same sequence, with the difference of the rejection criteria which are: $Q_{hardest}^{PS} > Q_{match}$ for lower-multiplicity events and $Q_{hardest}^{PS} > Q_{softest}^{ME}$ for the highest multiplicity events. This scheme gives equivalent results to the k_T -MLM with the advantage to have $Q_{match} = Q_{cut}^{ME}$.

The philosophy of the MLM matching lies on rejection while other methods use a reweighting of the events, like e.g. the CKKW matching [70], using the probability to do not have additional radiation above Q_{match} .

The (a)MC@NLO approach

The aMC@NLO package attempts to automate the MC@NLO matching procedure [71]. In this scheme, first the full NLO matrix element computation including (n+1)-parton corrections is performed. Then an analytic calculation is made to estimate how a first branching in the parton shower of an n-parton event would migrate it to the (n+1)-parton phase space. This estimation can be used as a correction term (the shower subtraction term) to the (n+1)-parton matrix element calculation, to remove the migration from the n-parton events. After this the two separated population samples, (n+1)- and n-parton are passed through a parton shower program and finally added. The correction term is therefore dependent on the type of parton shower program, which needs to be known before the generation of events. Moreover, the n and (n+1) phase space contributions are finite but not necessarily positive definite, which results in a set of events with a negative weight.

1.2.1.4 Hadronisation

After a certain number of successive splittings of the partons in the parton showering, the energies drop to the already mentioned cut-off scale, called Λ_{QCD} , where α_S becomes close to 1. Below this scale, the phase space is non-perturbative and the coloured partons created from the previous steps will recombine to form colourless hadrons. This phenomenon is called hadronisation and is currently described only through phenomenological models. The model used in PYTHIA, which takes care of the hadronisation after the parton shower, is the so-called Lund string model [72]. It is based on the assumption of linear confinement in QCD, and describes the partons as connected to each other by a string. When two coloured partons separate, the potential energy of the strong colour field, represented by the string, increases. When the energy is high enough the string can break creating a new colour neutral quark-antiquark pair also bounded by a string, shorter (or with less potential energy) this time. The ‘string breaking’ process will continue until only on mass shell hadrons remain. The probability for a parton to give rise to a specific type of hadron, in function of the string momentum fraction is described by the fragmentation functions. These empirical functions cannot be calculated from theory and are fitted to match the experimental measurements [2]. After fragmentation the remaining mesons and baryons are not all stable and will decay to daughter particles. These, in turn, decay themselves if they are not stable.

1.2.1.5 Decay

The decay step is again taken care of by general-purpose programs like PYTHIA. Since the b quark will play a key role in the rest of this thesis, the review of its hadronisation process will illustrate this section. In general it is very complex to identify the flavour of a quark from the experimental observations. The b quark provides, through its hadronisation process, some characteristics which allow one to recognise its signature. In the hadronisation chain the b and $\bar{b}$ quarks can regroup with lighter quarks and antiquarks to form the so-called B hadrons. The most often produced B hadrons are listed in Table 1.3.

Table 1.3: *The most prominent b-flavoured hadrons with their production fraction (b-hadronisation fraction) and their lifetime [2]. The quark content is also given for the B mesons and the list of b-baryons is given for the b-baryon admixture.*

Hadron	Content	Production fraction (%)	Lifetime ($\times 10^{-12} s$)
B^+ / B^-	$u\bar{b} / \bar{u}b$	40.2 ± 0.7	1.638 ± 0.004
$B^0 / \bar{B}^0$	$d\bar{b} / \bar{d}b$	40.2 ± 0.7	1.519 ± 0.005
$B_s^0 / \bar{B}_s^0$	$s\bar{b} / \bar{s}b$	10.5 ± 0.6	1.512 ± 0.007
b-baryons ADMIXTURE	$\Lambda_b, \Xi_b, \Sigma_b, \Omega_b$	9.2 ± 1.5	1.449 ± 0.015

It is interesting to notice that the total B hadron admixture presents an average lifetime τ of $1.529 \pm 0.009 \times 10^{-12}$ seconds [2]. This lifetime is equivalent to a typical travelled distance, $c \cdot \tau$, of approximately 0.46 mm before the decay of the B hadron, in a non relativistic case. In the case of a relativistic B hadron like the ones produced in a collision at LHC, the travelled distance is getting larger with the hadron momentum. Following the formula $d_{\text{flight}} = \gamma \cdot c \cdot \tau = \sqrt{1 + (\frac{p}{m_0})^2} \cdot c \cdot \tau$, where γ is the Lorentz factor expressed in terms of the hadron momentum p , and its mass at rest m_0 , one can estimate the range of values for the B hadron distance of flight. The rest mass being about 5 GeV and assuming the momentum to be comprised between 1 GeV and 1000 GeV, the distance of flight will range between approximately 0.5 mm and 9.5 cm. In all cases, this is translated into the observation of an additional vertex, displaced with respect to the primary vertex of the hard interaction. This secondary vertex will be used in the experiment to identify observed jets of particles as coming from a b quark decay, as will be shown in Section 3.2.3.

In PYTHIA the B hadron decay is based on the spectator model, in which the non-b quark of the hadron does not take part in the decay. This spectator

quark will however enter the flavour composition of the daughter particles. The b quark decays the most often to a virtual W^* boson and a c quark, due to the domination of the probability of a transition from a b quark to another c quark, over the transition to a u quark [2]. The W^* boson will subsequently decay to a lepton and its associated neutrino or to a quark-antiquark pair. All remaining quarks will themselves hadronise (in PYTHIA according to the model described in the previous section), and the particles daughters will also, depending on their nature, hadronise or decay. This ‘chain’ reaction results in a jet of hadrons as the observable for a quark.

1.2.1.6 Underlying Events

In addition to the hard interaction that engages one of the partons of the incoming protons, interactions between the rest of the partons, called the proton remnants, also take place. Since one parton is missing in each proton, the remnants are coloured partons which will be subjects to the hadronisation. These beam remnants consist thus of hadrons carrying a small transverse momentum and are emitted very forward. But remnant partons can also interact in addition to the hard interaction and this effect is called multi-parton interactions (MPIs) [73]. Beam remnant together with MPIs are labelled as underlying events. These events cannot be described from first principles and phenomenological models are tuned via the measurement of the multiplicity of low transverse momentum tracks. The tuning is implemented in the programs taking care of the parton shower and the hadronisation; in PYTHIA it has been cross checked to be in agreement with the data to which the simulations will be compared to [74].

1.2.2 b-initiated Processes: the Z plus b jets Associated Production Case

Processes with third generation quarks (b and t quarks) play a key role in proton-proton (p-p) collisions. Their peculiar signature allows to identify them better than light flavoured ones. The displaced secondary vertex was already mentioned as being a clear b quark signature. Such processes also receive special theory attention. In the case of the associative production of a Z boson with one or several b jets, the interest lies both on the production mode and on the final state.

1.2.2.1 Production

The production of a dilepton pair from p-p collisions via a Z boson is done by the Drell–Yan process (DY). This process describes the production of a pair of oppositely charged leptons of same flavour, from the interaction of a quark with an anti-quark. The interaction between the two fermion pairs is mediated by a Z boson or by a virtual photon γ^* .

In theory there is an indifferent selection of a Z or a γ^* . The Z/ γ^* interference contribution is considered to be part of the measured signal, even if the Z component is very large in the mass region considered. One could always write Z/ γ^* but for simplicity of notation Z will be used in the rest of this thesis.

The production of Z+b-quark(s) in p-p collisions can originate from gluon-gluon (gg) or quark-antiquark (q $\bar{q}$) interactions. From the PDF distributions of Figures 1.3 and 1.4 it is easy to notice that the gg contribution is dominating at the LHC where for a centre-of-mass energy of 7 or 8 TeV and a Q of about 100 GeV the x fraction will be relatively small, compared to the Tevatron case. Indeed, at the Tevatron, the centre-of-mass energy was about 2 TeV for the same Q and thus a higher x fraction for the partons involved, which makes the q $\bar{q}$ interaction more important [75, 76, 77].

The production of a Z and only one b quark can be approached by two different calculation schemes. The first one is to consider the initial gluons splitting in a collinear b $\bar{b}$ pair, which is an important source at the LHC, resulting in only one b quark observed at high transverse momentum. In this case it is convenient to think of the splitting in one b quark remaining at low p_T while the other one takes part in the hard interaction. In other words, the b quark PDF distribution is taken into account at the ME level. The leading order diagrams are presented in Figure 1.5 and the calculations are performed at NLO with the mass of the b set to zero. Such a scheme is called the 5-flavour (5F) scheme.



Figure 1.5: *LO diagrams of the associated production of a Z boson and a single high- p_T b quark ($Q = b$) in the 5F scheme.*

The $gb \rightarrow Zb$ process can then be used to provide information on the b quark content of the proton. Currently the b distribution function is derived perturbatively from the measured gluon distribution [78] but no direct measurement has been performed yet.

An alternative way to get a $Z+b$ final state is to consider the gluon splitting at parton shower level, with one heavy quark emitted collinearly and thus not observed. This approach can apply to the $gg \rightarrow Zb\bar{b}$ and the $q\bar{q} \rightarrow Zb\bar{b}$ production, as leading order process. Some corresponding diagrams are shown in Figure 1.6, and the calculations are performed with the mass of the b quark non null. This scheme is called the 4-flavour (4F) scheme.

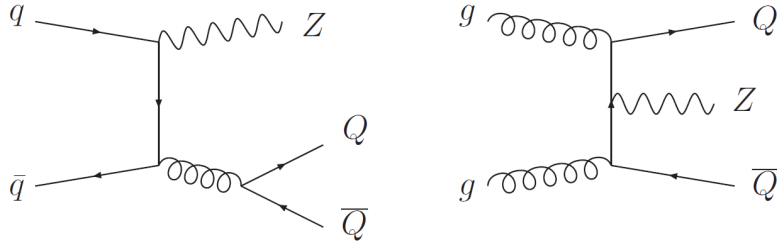


Figure 1.6: *LO diagrams contributing to the associated production of a Z boson and two b quarks ($Q = b$). The Zb final state can be yielded if one of the b quarks is emitted collinearly. In the case of $q\bar{q}$ interaction, the Z boson can also be radiated off the final-state quark. The 4F and 5F scheme can be attached to the diagrams (with the possibility of $q = b$ for 5F).*

The leading order processes for $Z+2$ jets, with one or more jets arising from a b quark, are also given either by $q\bar{q} \rightarrow Zb\bar{b}$ shown in Figure 1.6, this time with both b jets observed, or from $bq'(g) \rightarrow Zbq'(g)$ where q' can be a light

(Z+1b+1j) or a heavy b quark (Z+2b). Diagrams for the later can be found in Figure 1.7 for the Z+1b+1j case.

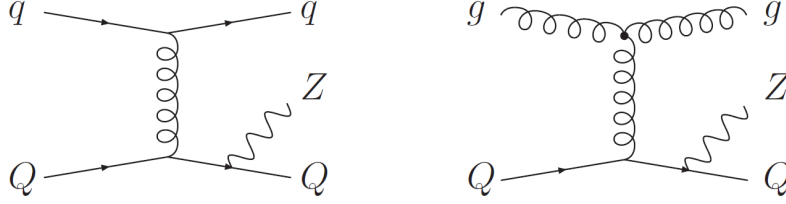


Figure 1.7: *LO diagrams contributing to the associated production of a Z boson and two high- p_T jets, one of which arises from a b quark ($Q = b$), in the 5F scheme. In the case of Qq interaction, the Z boson can also be radiated off the initial-state quark.*

When correcting the processes at the next-to-leading order, one has to add the contribution of diagrams with one extra real radiation and diagrams with a virtual correction via a loop.

1.2.2.2 4-flavour and 5-flavour Schemes

The two different schemes, 4-flavour and 5-flavour, describing QCD processes that involve b quarks such as the Z+b('s) production, are also implemented in simulation. In the 4F scheme b quarks are implemented only in the final state and are considered massive. They do not enter the composition of the PDFs which will take into account only the four lightest flavours of quarks (u,d,c,s) and the gluons. In the 5F scheme on the contrary, b quarks are included in the PDFs description, and thus in the calculations, and are considered massless.

For the 4F scheme to be reliable, the b quark mass has to be assumed to be of the same order as the other hard scales in the process. It can be not the case if the p_T of the b quark itself is very high, or if a weak boson is present in the final state, which is the case for Z+b('s). When a cross section is characterised by two different scales, perturbative calculations display $\log \frac{Q^2}{m_b^2}$ and $\left(\frac{Q^2}{m_b^2}\right)^n$. Here large logarithms of type $\log \frac{Q^2}{m_b^2}$, of initial state (IS) or final state (FS) nature, appear and can spoil the accuracy of a fixed order calculation. If the calculation could be performed at infinite order those logarithms would cancel out, but at NLO for example the expansion might diverge.

In these cases with a scale higher than m_b , the 5F approach could be favoured, since the consideration of $m_b = 0$ leads to the resummation of the IS large logarithms into the PDF of the b, and the FS ones into perturbative fragmentation functions. Another advantage lies in the massless scheme calculations, very simplified, due to the reduction of external legs and of scales in the processes. The cross section can thus be calculated to NNLO.

In the massive 4F scheme, computation is more complex due to the higher multiplicity. But its advantage is the full kinematic description of the heavy quark at LO, which becomes even more accurate via a NLO correction. The implementation in the parton shower is more straightforward than for the 5F scheme.

The use of the two schemes is in fact complementary, as it was shown that the disadvantage of the 4F, namely the possible badly-converging perturbative expansion due to large logarithms, hardly appears in hadron collisions [79]. For many observables the effects of the resummation are small and even more negligible than some other approximations involved in fixed order calculation.

To infinite order in perturbation theory the two schemes are defined such as to be exactly identical. But at finite order differences can appear. The comparison of the two different schemes, with real data, for both the calculation of the cross section and the kinematic variables description, sets an example as precision test of perturbative QCD.

1.2.2.3 Background for Searches

At the time this thesis started the Higgs boson remained the last particle predicted by the SM which had not been observed experimentally. The decay of a Higgs particle is dominated by the $b\bar{b}$ final state for a hypothetical mass of the Higgs smaller than about 130 GeV, as can be seen on the top left plot of the Figure 1.8.

However, a direct observation would be very difficult since the signature of the $H \rightarrow b\bar{b}$ would be drowned in the many other processes that can produce jets of hadrons coming from a b quark. These processes creating one or multiple jets of hadrons from quarks are gathered under the name QCD multijet events. The high rate of the QCD production makes the detection of a $H \rightarrow b\bar{b}$ signature within them almost impossible. One can get an idea of the dominance of the production of jets at the LHC with the bottom plot of Figure 1.8, where the cross section of jets production is at least five orders of magnitude higher than a SM Higgs boson production.

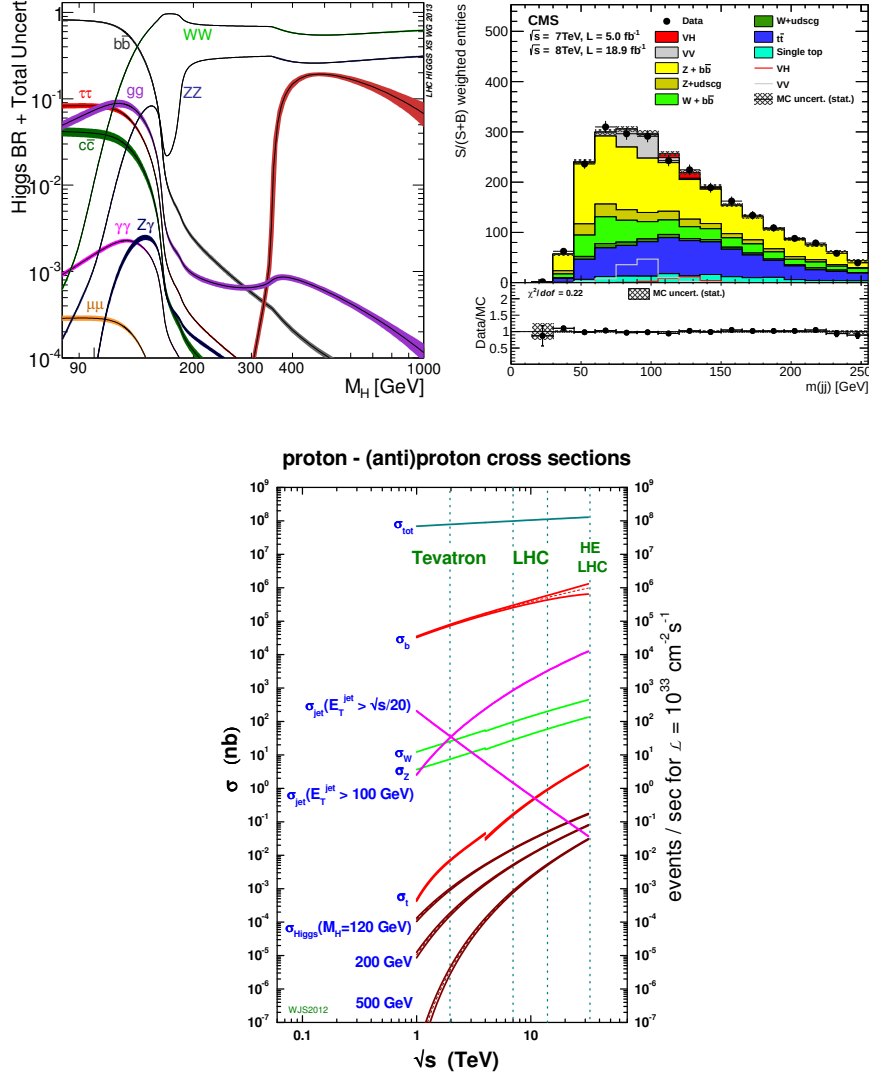


Figure 1.8: Top left: standard model Higgs boson decay branching ratios and total uncertainties [80]. Top right: Weighted dijets invariant mass distribution, combined for all channels for the vector boson associated production of a Higgs in two b jets. The expected signal used corresponds to the production of the SM Higgs boson with a mass of 125 GeV [81]. Bottom: Production cross section of the standard model processes as a function of the p-p (or p- $\bar{p}$) centre-of-mass energy [82].

To search for such a Higgs boson decay, one can require its production in association with a vector boson, Z or W, called the $VH(b\bar{b})$ channel. A corresponding Feynman diagram can be found in Figure 1.9. The vector bosons are characterised by a clear signature of isolated lepton(s) and will drastically reduce the QCD background, which by definition does not include such leptons.

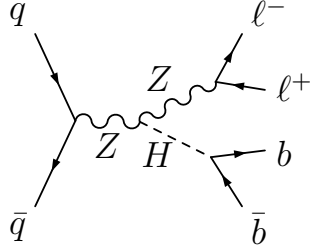


Figure 1.9: Feynman diagram of the $ZH(b\bar{b})$ channel.

By looking at the main backgrounds contributing to the search in the $VH(b\bar{b})$ channel, on the top right plot in Figure 1.8, one can see without surprise that the dominant one comes from the associated production of a Z and a $b\bar{b}$ pair, called hereafter the Zbb process. It presents the same final state and thus is irreducible. So the knowledge of the characteristics of this Zbb process was directly motivated by the search for this SM last predicted particle.

However, the discovery of a new boson in July 2012 by the ATLAS and CMS collaborations [41, 42] did not include this channel. The new particle was found to have a mass m_H near 125 GeV [44] and properties compatible with those of the standard model Higgs boson. Significant signals have been observed in channels where the boson decays to gg , ZZ , or WW , which confirm its interaction with the bosons of the SM. In the fermionic sector, evidences have been accumulated, but the observation of a five σ deviation² signal in one of the $H \rightarrow b\bar{b}$ and $H \rightarrow \tau\bar{\tau}$ decay channels, remains important to establish the new boson as the SM one.

The vector boson associated channel is still actively probed, and to date no evidence of signal has been found [81], as one can notice in Figure 1.8 (top right). The Zbb process remains the dominating background, making still interesting to know as much as possible about its features.

²A 5 σ deviation is equivalent to a likelihood that the signal results of a fluctuation of less than one chance in 3.5 million. The standard in particle physics for a discovery.

Moreover, one can also quote the Zbb process as background for searches in some beyond the standard model theories. Indeed, in spite of the successes of the SM predictions verified by experimental evidence, the observation of this new particle on one hand vindicates the reliability of the SM, and on the other hand, reemphasises the weaknesses of this model. Weaknesses which consist in unanswered phenomena like the hierarchy problem, which manifests itself as the large quantum corrections to the Higgs boson mass, or the observed matter-antimatter asymmetry in the universe.

Some alternatives are proposed in the form of an extended scalar sector with the introduction of more than one Higgs doublets. These theories are generally called Two Higgs Doublet Models (2HDM) [83] and present five physical Higgs bosons: two neutral scalars (h , H), one neutral pseudo-scalar (A), and two charged scalars (H^+ , H^-). A known special case is the Minimal Supersymmetric Standard Model (MSSM) [84], which predicts amongst others, the decay of a heavy pseudo-scalar particle to a Z and the SM Higgs boson, subsequently decaying to a $b\bar{b}$ pair: $A \rightarrow Z(\ell\ell)h(b\bar{b})$. For the search of such a decay the Zbb process is an important background.

In a more general way the 2HDM models predict the decay: $A \rightarrow Z(\ell\ell)H(b\bar{b})$, which also results in a Zbb final state. Another type of model comes from imposing a twisted custodial symmetry to the general 2HDM [85, 86]. In this inverted hierarchy model, a new heavy particle would decay to a Z and a light pseudo-scalar new particle, itself decaying to a b -quark pair: $H \rightarrow Z(h)A(b\bar{b})$. The Zbb process represents clearly an important background for all these decays. The Feynman diagram for a generic search in both the general and the inverted 2HDM is presented in Figure 1.10.

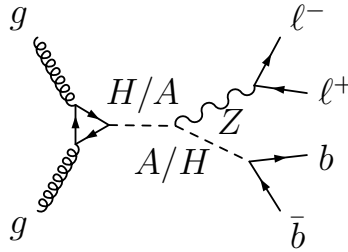


Figure 1.10: Feynman diagram of the $H/A \rightarrow ZA/H(b\bar{b})$ processes predicted in the 2HDM theories.

Chapter 2

The Experimental Setup

Since the proposition of the standard model, a huge scientific effort has been raised to verify its predictions. This goes by the observation of the particles the SM predicts. Among the SM elementary particles, only the electrons, the photons and the neutrinos are stable and can be observed freely. However, since the neutrinos interact weakly with matter, they need special very large detectors to be observed. Some others like the muons can be intercepted from a natural source of particle production, the cosmic rays. To observe the other particles, quarks hidden in hadrons and bosons of high masses, all having a very short life time, the solution is to create them and collect signs of their decay products as they interact with the matter of a detector. For this purpose particles colliders, of always increasing energy, have been built, in association with detectors, with always better performing technologies.

This effort led to important breakthroughs in modern physics: one can mention the discovery of the Z and W bosons at the Super Proton Synchrotron (SPS) [35, 36, 37, 38] followed by the verification of their properties at the Large Electron Positron (LEP) collider [87, 88], both at CERN, and the discovery of the top quark at the Tevatron at Fermilab [24, 25]. The effort culminated with the LHC at CERN, which is currently the most powerful accelerator ever build. It was designed to find, or alternatively exclude, the last unobserved particle predicted by the SM, the Higgs boson. Thanks to its great centre of mass energy, and its high rate of collisions, searches for alternative models could also be performed, as well as the precise measurement of low cross section SM processes.

This chapter will describe the experimental setup that led to the collection of data which were used for the results of this thesis. First, the LHC will be briefly described as the source of the data, then the Compact Muon Solenoid (CMS) detector will be reviewed as responsible for the complex collection and reconstruction of the data.

2.1 The Large Hadron Collider

The Large Hadron Collider [89] is currently the most recent and the most energetic hadron collider in the world, situated at CERN at the border of France and Switzerland, close to Geneva. It is designed to accelerate protons (p) or heavy ions (HI) and collide them in p-p collisions, HI-HI collisions or asymmetric p-HI collisions, respectively. The machine lies 100 m underground, in a tunnel of 26.7 km circumference, which has previously hosted the LEP, shutdown in 2000. The LHC construction was decided in 1994 and finished 15 years later. The first beam was recorded in 2008 and the first collisions took place in March 2010, dully recorded by the four main experiments: ATLAS [90], ALICE [91], CMS [92] and LHCb [93].

The LHC was designed to accelerate protons (ions) up to an energy of 7 TeV (2.76 TeV per nucleon). The present thesis deals with data obtained from proton-proton collisions and thus the description will be focused on this specific type of collisions. The choice of proton collisions at the LHC, rather than electron-positron collisions like it was the case for the LEP, is explained by the synchrotron radiation phenomenon. When accelerated in a curved path, charged particles emit an electromagnetic radiation. For relativistic charged particles whose path is bent by a magnetic field the emission is called synchrotron radiation and its radiated power is proportional to $\left(\frac{E}{Rm}\right)^4$, where E is the energy of the particle, m its mass and R the radius of the ring. This results in an emission 10^{12} less important for protons compared to electrons, which explains why for the same ring radius the energy reached at LEP was ‘only’ 104.5 GeV while the LHC can manage to accelerate protons up to 7 TeV. The choice of proton-proton (p-p) over proton-antiproton (p- $\bar{p}$) collisions, like it was the case at the Tevatron, is explained by the ease to produce protons compared to antiprotons, which allows for a higher number of collisions per second.

2.1.1 The CERN Accelerator Complex

To achieve the collision energy the protons are accelerated in several steps, that involve a large complex of several accelerators which are chained to each other, so that the output beam of one of them can be used as input to the next one to be further accelerated. First, the protons are produced by ionising hydrogen atoms and then accelerated to 50 MeV in the LINAC2 linear accelerator, from which they exit as a pulsed beam. They are then injected into the Proton Synchrotron Booster (PSB), made of four synchrotron rings stacked on each other, to reach an energy of 1.4 GeV. The next injection is made in the Proton Synchrotron (PS) which is a 628 m circumference storage ring started in 1959, that now accelerate the protons up to 25 GeV, before to pass them to the Super Proton Synchrotron (SPS). In the PS the beam takes its final bunch structure with a determined bunch spacing. The 6.9 km circumference SPS will accelerate the proton bunches further to 450 GeV. The protons are then finally injected into the LHC to reach their maximal energy, designed to be 7 TeV. Two beams are actually circulating in the LHC in opposite direction, and in order to maximise the probability of collision the beams are squeezed in the transverse plane and eventually collided in four points within the four main experiments. The complete acceleration phase takes between one and two hours. The total time protons are circulated in the LHC is called a fill and can last around ten hours, before the beams are dumped.

For the acceleration of heavy ions, a similar chain takes place. It starts with an highly purified lead sample, heated up and producing lead vapour, subsequently ionised to get ions Pb^{29+} . Those ions are accelerated in the LINAC3 linear accelerator, and passed through a carbon foil, to end up as ions Pb^{54+} with an energy per nucleon of 4.2 MeV/u. The ions are then injected in the Low Energy Ion Ring (LEIR), that splits the beam into four bunches and accelerates them up to 72 MeV/u. The next steps are then the same as for the protons: injection and acceleration in the PS and then in the SPS to reach energies of 5.9 GeV/u and then 177 GeV/u, respectively, and the final injection in the LHC to reach their final energy, designed to be 1.38 TeV/u.

The complete chain of accelerators can be found in Figure 2.1.

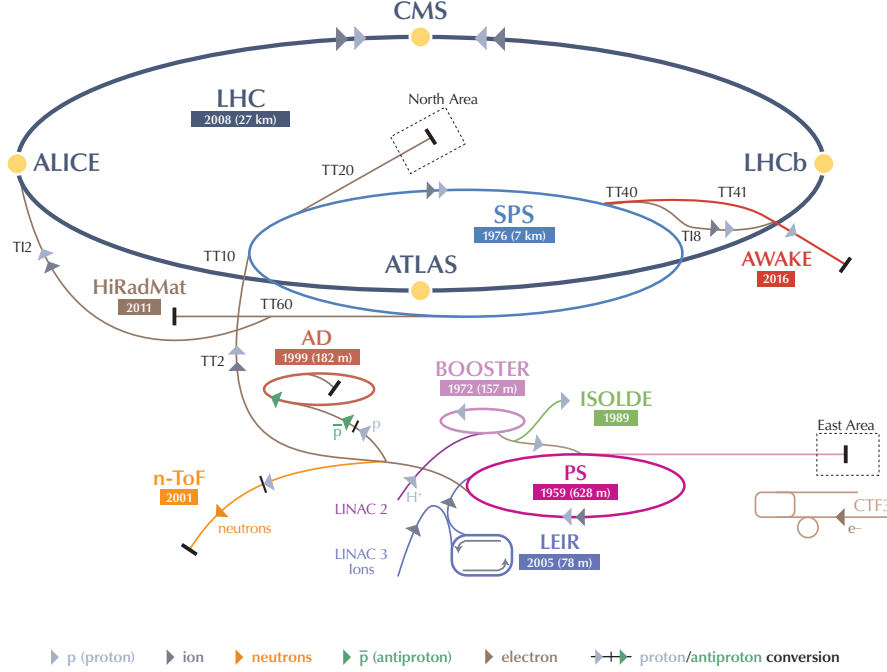


Figure 2.1: The CERN accelerator complex [94].

2.1.2 The LHC Machine

The LHC is divided in eight arcs and eight straight sections. In the arcs the beam is bent by superconducting dipoles, and in the straight sections can be found, either the equipment needed for the functioning of the accelerator, or the interaction points (IP) where the two beams are collided inside the four main experiments. Figure 2.2 sketches the general layout of the two LHC rings, with the division in octants which start from the middle of an arc and end in the middle of the following arc, including a complete straight section. To accelerate the beams and compensate for the synchrotron radiation losses, eight radio frequency cavities (RF) are used per beam, situated in the straight interaction region of the Octant 4 (IR4). The cavities are superconducting and operate at 400 MHz. The beam dump system is situated in IR6 and designed to make a fast extraction of the beams out of the two rings into one tunnel each leading to a beam dump. The beam dump is a cylinder of graphite composite, eight meters long and one meter in diameter, which is encased in concrete.

Finally the sections IR3 and IR7 are dedicated to beam cleaning in order to absorb part of the primary beam halo and of the secondary radiation.

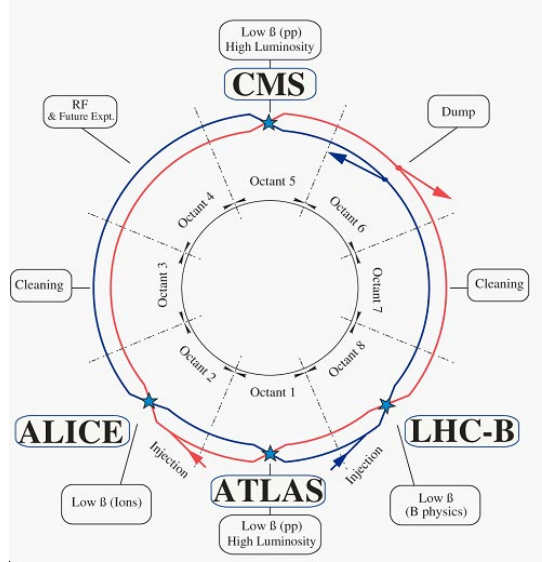


Figure 2.2: Schematic view of the LHC ring. Beam 1 is clockwise (red) while beam 2 is anticlockwise (blue). They collide in four interaction points corresponding to the four main experiments [89].

Two counter-rotating beams with particles of the same charge, would need in principle two opposite magnetic fields to be guided along the accelerator. To achieve this without using one accelerator for each beam, special twin-beam-pipe magnets were designed to guide each beam. Dipoles and quadrupoles of the LHC present such a feature, as can be seen in Figure 2.3 displaying the cross section of a dipole. The 1232 15 m long dipoles which bend the beams, and the 392 5-7 m long quadrupoles which focus the beams, are using the superconductor Niobium-titanium (NbTi). Cooled down to 1.9 K with superfluid helium, the superconducting magnets can produce a magnetic field up to 8.33 T, which is the high value required to steer the design 7 TeV beams. In addition, a number of higher order magnets correct the trajectories of off momentum particles, compensate for the beam imperfections, or squeeze the beam before the collisions.

To prevent the particles in the beam to interact with gas molecules that could be found in the beam pipes, these are pumped down to an ultrahigh vacuum, presenting a pressure as low as 10^{-10} mbar at room temperature (approximately

10^{-13} times the atmospheric pressure). Two other vacuum systems exist: one for the insulation of the cryomagnets and one for the insulation of the helium distribution.

LHC DIPOLE : STANDARD CROSS-SECTION

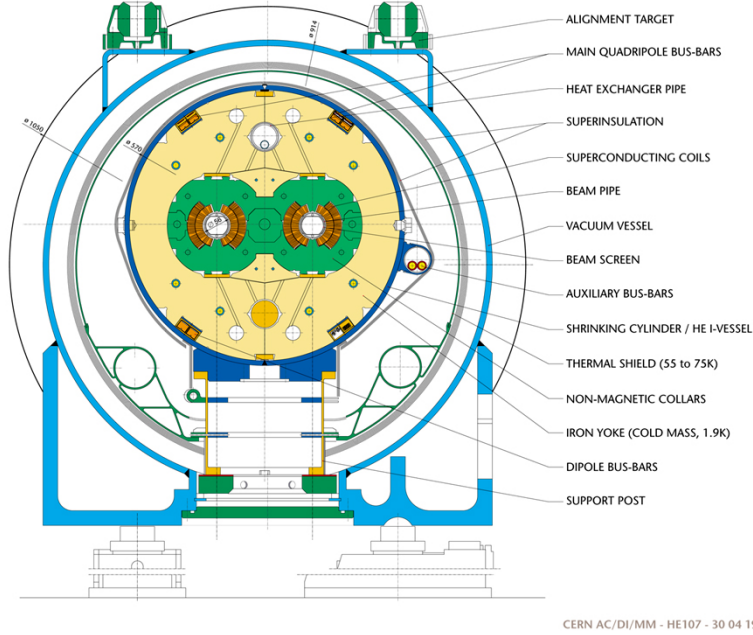


Figure 2.3: Cross section of a superconducting dipole of the LHC [89].

2.1.3 Luminosity

The most important characteristic of an accelerator together with its energy, is its instantaneous luminosity $\mathcal{L}$. It defines the number $\mathcal{N}$ of events that will be produced for a given process, with a given cross section σ , according to:

$$\mathcal{N} = \sigma \mathcal{L} . \quad (2.1)$$

A high luminosity allows to observe rare processes defined by a small cross section. Its accumulation or integration directly represents the data ‘statistics’ or quantity of events.

The luminosity is machine dependent and defined as:

$$\mathcal{L} = n_b \frac{N_1 N_2}{\sigma_x \sigma_y} f_{rev} , \quad (2.2)$$

where n_b is the number of bunches in one beam, N_1 and N_2 the numbers of particles in each of the colliding bunches, σ_x and σ_y the transverse width in x and y of the beams, and f_{rev} the revolution frequency of one bunch. So in order to increase the luminosity one can increase the number of particles in a bunch or reduce the interaction area $\sigma_x\sigma_y$ by focusing the beam.

In the case of the LHC the number and type of particles in each beam are the same: $N_1 = N_2 = N_b$. Since the interaction area is a parameter difficult to measure directly, the luminosity can be rewritten in terms of beam parameters easier to access:

$$\mathcal{L} = n_b N_b^2 f_{rev} \frac{\gamma}{4\pi\epsilon_n\beta^*} F, \quad (2.3)$$

where γ is the relativistic gamma factor, ϵ_n the normalised transverse beam emittance from the injectors, β^* is the so-called beta function at the collision point, and F is the geometric luminosity reduction factor due to the small crossing angle of the two beams at the interaction point (IP).

With each collision the number of protons in a bunch decreases, and thus the instantaneous luminosity. After a certain number of collisions the luminosity is too degraded and it is more efficient to abort the fill, dump the beam and refill the machine with new protons. The integrated luminosity is the luminosity integrated over the time. The curve of this integrated luminosity with respect to the time can be found for the run 1 conditions in Figure 2.4 in the next section.

A high luminosity, so a high rate of collisions, comes at a price: for one bunch crossing several interactions can happen and are referred to as in-time pileup. Each interaction forms a primary vertex, from which all the particles created originate. The number of those primary vertices per bunch crossing is machine related, since it depends on the number of bunches n_b , and of number of protons per bunch N_b , as well as the size of the interaction area. In addition, if the bunch spacing is small, some interactions coming from another bunch crossing, happening before or after, can be seen and are referred to as out-of-time pileup. All these interactions happening besides the hard interaction one wants to observe are entangled and called generically pileup.

2.1.4 LHC Run 1

All the components of the LHC have been designed for the accelerator to reach a nominal energy of 7 TeV per proton beam, creating p-p collisions of 14 TeV centre-of-mass energy. However, up to now this energy has not been reached

because of concerns about the stability of the electrical interconnections between the dipoles. This worry has its origin in the start of the LHC: the 10th of September 2008 proton beams were circulated successfully in the LHC but nine days later a faulty electrical connection led a magnet quench to induce the rupture of a liquid helium enclosure, causing an explosive release of the helium in the tunnel, damaging 50 magnets and contaminating the vacuum pipe. After 14 months of repairs, protons beams were circulated again in November 2009 at an energy of 450 GeV and 1.18 TeV, and the first collisions between two beams of 3.5 TeV energy took place on the 30th of March 2010. These collisions mark the beginning of the first LHC data taking period stretching from 2010 to 2013, called LHC run 1. In 2010 and 2011 the LHC was operated with protons beams of 3.5 TeV energy (centre-of-mass energy of $\sqrt{s} = 7$ TeV) and in 2012 the beams were collided at 4 TeV ($\sqrt{s} = 8$ TeV). Between November and December 2011 heavy ion collisions took place with a centre-of-mass energy reaching 2.76 TeV/u. Asymmetric proton-heavy ion collisions were recorded between January and February 2013 with $\sqrt{s} = 5.02$ TeV/u.

For the luminosity related parameters, the values during run 1 are also different from the design ones. Table 2.1 lists all the parameters entering the instantaneous luminosity Formula 2.3, for the proton-proton collisions data taking eras of 2010, 2011 and 2012 separately, and compares them with the design values. One can mention some parameters; the bunch spacing for example, which was designed to be 25 ns, corresponding to a spatial separation of about 7.5 m between the bunches. But, at best, only every second bunch slot was filled in the latest periods of run 1, resulting in a bunch spacing of 50 ns. The same goes for the number of colliding bunches which was designed to be 2808 for p-p collisions, and 592 for HI-HI collisions, and at maximum reached the numbers of 1380 and 356, respectively. On the contrary the number of protons per bunch superseded the design value from the first collisions in 2010. An indication of the pileup is given by the expression of the mean number of interactions per bunch crossing over the concerned period.

Within these run conditions of the LHC, the CMS detector measured the integrated luminosity, and the display of its cumulative value can be seen in Figure 2.4 (left) for the years 2010, 2011 and 2012 for p-p collisions. The total integrated luminosities delivered by the LHC added up to 44.2 pb⁻¹, 6.1 fb⁻¹ and 23.3 fb⁻¹, respectively.

The work presented in this thesis is using the data recorded by the CMS detector in 2011. However a look on the first 2010 data with the nascent analysis was given in the first six months of the thesis but will not be treated here. One has to mention that a small but significant difference exists between the

Table 2.1: *Machine parameters of the LHC during the three p-p collision periods of run 1: 2010, 2011, 2012, and the design values [95]. The average pileup is also given referring to the mean number of interactions per bunch crossing.*

Parameters	2010	2011	2012	Design value
Beam energy (TeV)	3.50	3.50	4	7
Bunch spacing (ns)	150	75/50	50	25
n_b	368	1380	1380	2808
N_b	1.2×10^{11}	1.45×10^{11}	1.7×10^{11}	1.15×10^{11}
f_{rev} (Hz)	11245	11245	11245	11245
ϵ_n (mm·mrad)	≈ 2.0	≈ 2.4	≈ 2.5	3.75
β^* (m)	3.5	1.0	0.6	0.55
F (m)	0.84	0.84	0.84	0.76
Peak $\mathcal{L}$ ($\text{cm}^{-2}\text{s}^{-1}$)	2.1×10^{32}	3.7×10^{33}	7.7×10^{33}	1.0×10^{34}
$\langle \frac{N_{\text{interaction}}}{\text{bunch crossing}} \rangle$	1	[5–9]	21	~ 25

integrated luminosity delivered by the LHC and the one recorded by the CMS detector. It corresponds to the time periods for which the LHC was producing collisions while CMS could not record them for different technical reasons. For the 2011 data a total difference of about 10% was accounted, leading to a total data sample of 5.55 fb^{-1} . These integrated luminosities (delivered by the LHC and recorded by CMS) can be observed versus time in Figure 2.4 (right) for the year 2011.

2.1.5 LHC Schedule

The LHC was designed to run over many years, with a life made of run periods, short shutdowns at the end of every year and long shutdowns every couple of years. After run 1, the Long Shutdown 1 (LS1) started in February 2013 and will be over very soon. As these lines are written, beam tests have already been performed and the start of run 2 is foreseen for May or June 2015. During the LS1 the LHC has been upgraded to produce higher centre-of-mass energy collisions and finally reach the nominal 14 TeV. In 2015, the p-p collisions are planned to take place with beams of 6.5 TeV energy and a bunch spacing of

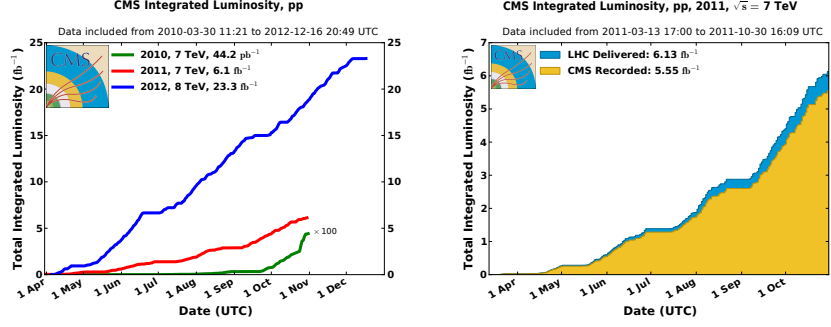


Figure 2.4: Evolution of the integrated luminosity versus time delivered by the LHC, as measured by CMS, for the first three years of running of the LHC (left). Evolution of the LHC delivered luminosity and the CMS recorded luminosity during the 2011 proton-proton collision campaign (right) [96].

25 ns, afterwards the beam energy is planned to be increased to 7 TeV. The expectation is to collect about 100 fb^{-1} of p-p collision data before the next long shutdown in 2018. After LS2, a third run is planned for 2020 and is expected to deliver 300 fb^{-1} of data with 14 TeV p-p collision until 2022, which corresponds to the end of the LHC phase 1. The latter will be followed by a two year major upgrade of the machine which will become the High Luminosity Large Hadron Collider (HL-LHC). The HL-LHC is expected to produce an instantaneous luminosity five times greater than the one at the LHC, and to collect about 3000 fb^{-1} of 14 TeV p-p collisions before the end of the program planned for 2035 [97]. A timeline in Table 2.2 summarises all the running and shutdown periods until 2029.

In synchronisation with the LHC shutdowns the detectors of all its experiments are being upgraded as well, to deal with the always higher luminosity environment, and to replace parts that will be damaged by high radiation.

Table 2.2: Schedule of the LHC starting from run 1 in 2010 and up to 2029. For the runs the collisions centre-of-mass energy, the peak instantaneous luminosity and the integrated luminosity are specified.

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Run 1 7-8 TeV, $0.7 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, 25 fb^{-1}			LS1		Run 2 13-14 TeV, $1.6 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, 150 fb^{-1}			LS2 Phase-I install.	
2020	2021	2022	2023	2024	2025	2026	2027	2028	2029
Run 3 14 TeV, $2.3 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, 350 fb^{-1}			LS3 Phase-II install.		Run 4 14 TeV, $5.7 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, 1000 fb^{-1}			LS4	

2.2 The Compact Muon Solenoid

CMS stands for Compact Muon Solenoid. The CMS detector is located at the point 5 (P5) of the LHC ring, in a cavern 100 m underground. It is one of the two general-purpose detectors of the LHC, the ATLAS detector being the other one.

2.2.1 General CMS Layout

The CMS detector presents the traditional multi-layered design of high energy detectors in collider experiments. It has a cylindrical shape with a total length of 28.7 m and a diameter of 15.0 m. The main feature of the detector is its superconducting solenoid magnet, able to produce a magnetic field of 3.8 Tesla in order to bend the trajectories of the charged particles produced in the collisions. The magnet legitimises both the ‘solenoid’ and the ‘compact’ labels, since the high field allows the detector to track the particles within a restricted distance, if we compare it for example to the ATLAS detector with its length of 46 m and 12 m radius. The weight of the CMS detector is very high: 14000 tonnes (the double of ATLAS), making it the heaviest detector ever built in a collider experiment. The ‘muon’ label comes from its great performance in the muon reconstruction.

The detector can be seen as divided in two parts: the central cylindrical one, the barrel, which is then closed with the disk-shaped endcap sections on both ends. This design is covering most of the 4π solid angle.

The subdetectors embedded in the 6 m diameter bore of the magnet, and thus inside the magnetic field parallel to the beam axis, are, starting from the interaction point:

- the inner tracking system where the trajectory of charged particles, ionising the medium, can be reconstructed.
- the electromagnetic calorimeter (ECAL), absorbing and measuring the energy of photons and electrons, mostly.
- the hadronic calorimeter (HCAL), absorbing and measuring the energy of hadrons.

Outside the solenoid the muon system can be found, whose layers are sandwiched with the ones of the steel return yoke for the magnetic field. Figure 2.5 displays a sketch of a three dimensional view of the detector, in its configuration during the 2011 and 2012 data taking eras.

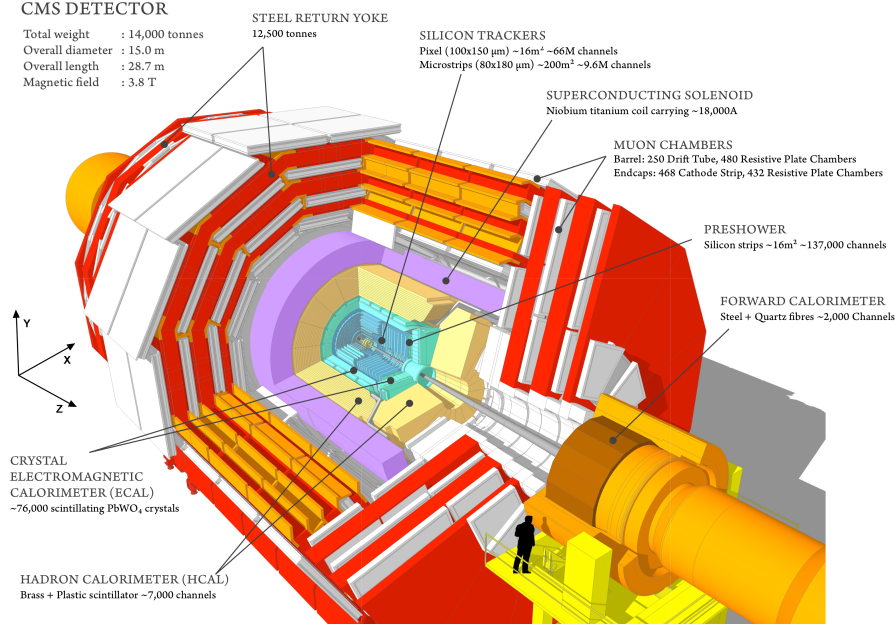


Figure 2.5: Schematic three dimensional view of the CMS detector [98].

2.2.2 Coordinate System

CMS uses a right-handed coordinate system whose origin is the nominal interaction point. The x -axis points toward the centre of the LHC, the y -axis points vertically upwards, perpendicular to the LHC plane, and the z -axis points horizontally along the anti-clockwise beam direction. The particles produced in collisions could fly in any direction, leading to the use of spherical coordinates to precise a position in the detector. The radial coordinate r is the distance in the x - y plane, the azimuthal angle ϕ is the angle measured in the x - y plane from the x -axis (ranging 0 to 2π), and finally the polar angle θ is measured also in the x - y plane but from the z -axis (ranging 0 to π). The θ angle is more conveniently transformed in the pseudorapidity η defined as:

$$\eta = -\ln \left[\tan \left(\frac{\theta}{2} \right) \right], \quad (2.4)$$

which ranges from zero in the x - y plane to positive and negative infinity, respectively toward the positive and negative z -axis. The region of the detector is considered forward when close to the z -axis so for values of $|\eta| \gtrsim 3$.

The transverse energy (E_T) and momentum (p_T) are defined from the x and y

coordinates.

One can also define the distance between two objects by $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

2.2.3 Magnet System

The superconducting magnet is a 12.5 m long solenoid, made of a four layer winding of NbTi reinforced with aluminium, reaching a total weight of 220 tonnes. It is housed in a coil cryostat of about 40 m³, that maintains an operating temperature of 4.5 K with liquid helium. The magnet was designed to operate at 4 T but the field was kept at 3.8 T during run 1 of the LHC to increase the safety margin and prolong the magnet lifetime. That field allows to measure the momentum of high energy particles with a resolution better than 10% for a $p_T \sim 1$ TeV. To reach 3.8 T, a 18 kA current circulates in the cables, inducing a stored energy of 2.5 GJ.

The return yoke outside the solenoid is composed of five barrel wheels and two endcaps comprising three disks each, interleaved with the muon chambers. The central wheel holds the cryostat coil. The whole return yoke system is made of steel and weights an impressive total of 12 500 tonnes. A simulation of the magnetic field intensity in the different regions of the CMS detector, as well as the field lines, can be observed in Figure 2.6.

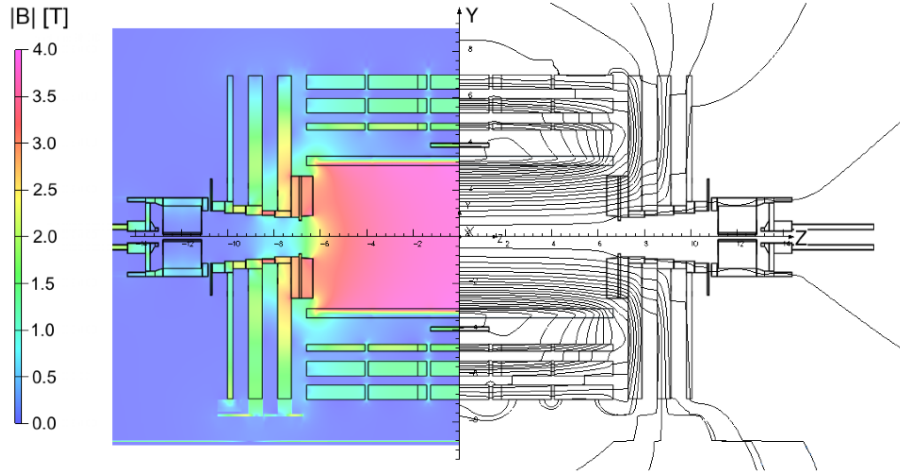


Figure 2.6: Simulation of the magnetic field magnitude (left) and of the magnetic field lines (right) in the y - z plane of the CMS detector, for a central field of 3.8 T [99].

2.2.4 Subdetectors

2.2.4.1 Tracker

The inner tracking system lies the closest around the interaction point (IP). Its purpose is to measure precisely and efficiently the trajectories of the particles bent by the magnetic field. The challenge was to deal with the pileup with a high granularity and a fast response, while keeping the amount of material as low as possible to avoid parasitic interactions, and to resist to radiation. The answer was found in building the tracker entirely with silicon detectors. It is divided in the inner pixel detector, and the surrounding silicon strip tracker, reaching a total length of 5.8 m and a diameter of 2.5 m. A general view of the tracker and its subsystems can be found in Figure 2.7.

The pixel detector consists of three 53 cm long barrel-shaped layers, of radii 4.4, 7.3 and 10.2 cm, and is closed by two endcap disks which extend the tracker coverage up to $|\eta| = 2.5$. The barrel and forward systems are respectively called BPix and FPix. The detector covers an area of about 1 m^2 and has 66 million pixels, each of them of size $100 \times 150 \mu\text{m}^2$ in the $r\text{-}\phi \times z$ directions. For each charged particle it provides three high precision space points. The passing charge creates electron-hole pairs in the silicon and the created electrons will move to the readout surface of the pixel due to the high voltage applied. Because of the magnetic field, these electrons undergo a Lorentz drift in the $\vec{E} \times \vec{B}$ direction, and spread the collected signal charge over more than one pixel. A charge interpolation of the analogue pulse height being read out, allows a spacial resolution of 15 (20) μm in the $r\text{-}\phi$ (z) direction for the BPix. The FPix detectors are tilted at 20° to induce charge-spreading and achieve 15 μm resolution in both directions.

As seen in Section 1.2.1.5, the displaced secondary vertex, signature of a b quark hadronising in a B meson, can lie between around 0.5 mm and 9.5 cm from the primary vertex, depending on the B meson momentum. So for a b quark produced at the interaction point, its secondary vertex will be contained in the pixel detector. The fine pixel spatial resolution is responsible for a small impact parameter resolution, important for the secondary vertex resolution.

The silicon strip (SiStrip) tracker is made of ten layers of microstrip detectors and is situated around the pixel detector. It is composed of four different subsystems: the Tracker Inner Barrel (TIB) made of four barrel layers extending to a radius of 55 cm and the Tracker Inner Disk (TID) made of three disks at each end in the forward region. The Tracker Outer Barrel (TOB) extends, with six barrel layers, to a radius of 116 cm and to $|z| = 118 \text{ cm}$. Above $|z| = 124 \text{ cm}$ and up to $|\eta| = 2.5$, the Tracker EndCaps (TEC+ and TEC- according to their

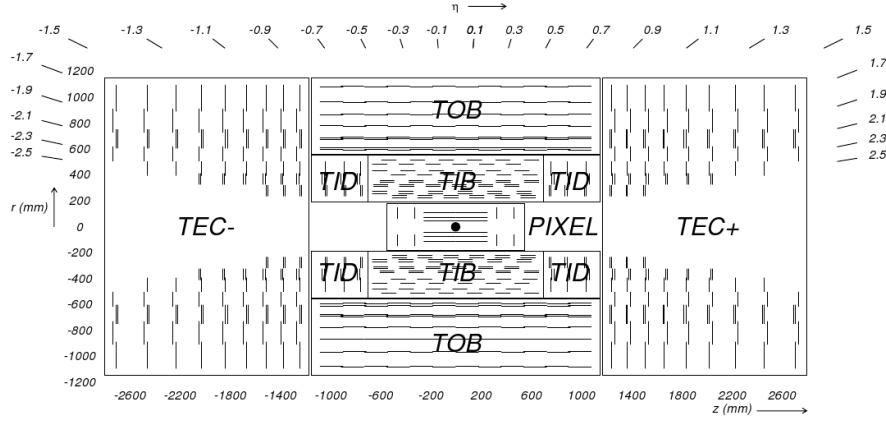


Figure 2.7: Schematic sectional view of the CMS tracker. Detector modules are represented by a line. Double lines indicate back-to-back stereo-modules [92].

location along the z -axis) are made of nine disks each. The SiStrip tracker, the largest ever built, totals up 9.3 millions of strips on a 198 m^2 active silicon area. The pitch of the strip ranges from $80\text{ }\mu\text{m}$ in the innermost layers of the TIB, to $183\text{ }\mu\text{m}$ in the four first layers of the TOB. For the disks, the pitch varies between 97 and $184\text{ }\mu\text{m}$. The disks thickness is $320\text{ }\mu\text{m}$ for TIB, TID, and the first four inner disks of the TEC, and $500\text{ }\mu\text{m}$ for the TOB and the outer rings of the TEC.

The sensors have their strips parallel to the beam axis in the barrels and radial in the disks, and can therefore measure only one coordinate. But the presence of a second microstrip detector on some modules, mounted back-to-back with a stereo angle of 100 mrad with respect to the first one, provides more constraining measurements for the tracks reconstruction fit. These stereo-modules giving two hits are shown in Figure 2.7 as a double line.

A charged particle, e.g. a muon, produced at the IP and flying through the complete tracker with a $|\eta|$ smaller than 2.4, well inside the acceptance of the tracker, should leave between 10 and 14 hits in the complete tracker, depending on its η . Such a muon can then be reasonably assumed to leave in general three hits in the pixel detector, and eight hits in the silicon strip tracker layers.

The material budget for the whole tracker is shown in Figure 2.8 and is comprised between 0.4 and 2 radiation lengths.

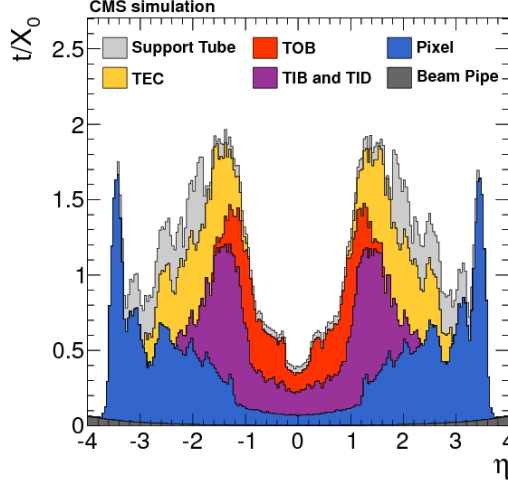


Figure 2.8: Tracker material budget per radiation length as a function of the pseudorapidity for the different subdetectors [100].

2.2.4.2 Electromagnetic Calorimeter

The electromagnetic calorimeter positioned outside the tracking volume, has the purpose to measure the energy of light electromagnetically interacting particles, i.e. mainly electrons and photons. It relies on the electromagnetic shower phenomenon, happening for high energy electrons entering matter. They will emit photons via bremsstrahlung in the ECAL material, themselves converting to electron-positron pairs. The repetition of the process resulting in a cascade of electrons, positrons and photons. The exited atoms of the material emits photons, that can be collected by a photodetector. The amount of light produced is proportional to the initial energy of the incident particle, allowing its measurement. The idea is to be able to contain the whole shower in a small volume by choosing an appropriate material.

The ECAL is made of lead tungstate (PbWO_4) crystals and is described as a nearly hermetic, homogeneous calorimeter. The high density (8.28 g/cm^3), short radiation length (0.89 cm) and small Molière radius¹ (2.2 cm) of these crystals allow a fine granularity for a compact calorimeter, that can fit inside the solenoid bore of CMS. The scintillation decay time is such that 80% of the light is emitted in 25 ns , which is the smallest bunch crossing window the LHC

¹Radius of a cylinder containing on average 90% of the shower energy.

has been designed for.

The ECAL can be subdivided in three parts as shown in Figure 2.9: the barrel, the endcap and the preshower. The ECAL barrel covers the range $|\eta| < 1.479$ and possesses 61200 crystals of dimensions $22 \times 22 \times 230$ mm, weighting a total of 67.4 tonnes. These crystals are mounted with their surface facing the interaction region but with a 3° tilt in η and ϕ , to avoid to have particles crossing the gap in between adjacent crystals and be missed. The ECAL endcaps add 14648 crystals of dimension $28.6 \times 28.6 \times 220$ mm, extending the ECAL coverage to $|\eta| = 3.0$. The ECAL preShower (ES) detectors are installed in the forward area ($1.653 < |\eta| < 2.6$) and are made of lead radiators to initiate electromagnetic cascades, interleaved with silicon sensors of surface 63×63 mm². However, in the transition region between barrel and endcaps, for $1.479 < |\eta| < 1.566$, the measurement is extremely degraded, because of the tracker services that have to pass through the ECAL. On the edges of the barrel and preshower, a margin of two crystals is taken for precision measurement, which results in a gap region for $1.4442 < |\eta| < 1.566$, and a maximal value $|\eta| < 2.5$, which will be taken into account for physics analysis. The ES are necessary because of the high rate of pions in this region which decay in two very collinear photons. This detector improves the identification of the electrons and the spacial resolution of the electrons and photons. The energy resolution of the ECAL is expressed as $\frac{\sigma}{E}$ and is estimated to 1–2% for $|\eta| < 1.1$, increasing up to 4% with the pseudorapidity [101].

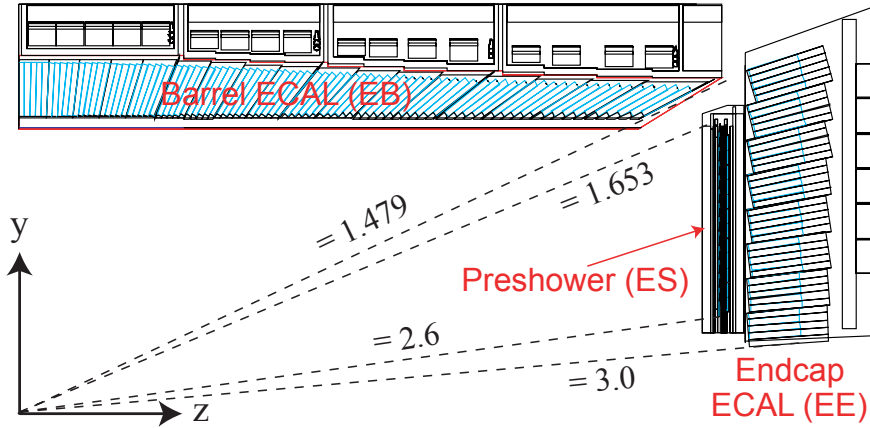


Figure 2.9: Longitudinal cross section of a quadrant of the ECAL detector with the positions of its different components [102].

2.2.4.3 Hadronic Calorimeter

The hadronic calorimeter of CMS was designed to measure the hadron jets that will shower in its material. The HCAL is a sampling calorimeter, in which the material that produces the particle shower is distinct from the material that measures the deposited energy. Plates of brass as absorber alternate with plastic scintillator tiles as active medium.

The HCAL central part, the barrel (HB), extend to $|\eta| < 1.4$ and is made of 16 absorber plates. The scintillator plates are segmented in 32 regions in η and 36 regions in ϕ , resulting in 1152 towers of dimensions: $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$. The innermost and outermost plates are made of stainless steel for structural strength. The barrel lies between the ECAL and the magnet coil, which constrains its size and thus the amount of material used: the number of interaction lengths varies between 5 and 10, which can be too small to contain the whole hadronic shower. The HCAL is extended outside the solenoid by a tail catcher outer calorimeter (HO) recovering the rest of the shower. The HO uses the solenoid as an additional absorber, and is made of two sensitive layers, placed on each side of the first layer of the iron return yoke. It allows to extend the HCAL radiation lengths to at least 12, and $|\eta|$ to about 1.3.

The HCAL endcaps (HE) cover the region $1.3 < |\eta| < 3$, the region $1.3 < |\eta| < 1.4$ being shared between HB and HE. 34% of the final state particles will be found in the HE, and as it is placed at the end of the solenoidal magnet its material must be non-magnetic. Brass is again a good choice, and the outermost layer is made of stainless steel to attach it to the muon endcap yoke. The granularity of the HE is $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ for $|\eta| < 1.6$ and $\Delta\eta \times \Delta\phi \approx 0.17 \times 0.17$ for $|\eta| \geq 1.6$.

An additional forward calorimeter (HF) is placed at 11.2 m of the IP, extending the $|\eta|$ coverage of the HCAL from 3.0 to 5.2. This region will undergo the highest particle flux and energy deposit per collision. The radiation hardness of the material was thus crucial, leading to the choice of Cherenkov-light detector quartz fibres as the active medium and steel as absorber. Over 1000 km of fibres are used in the HF, which are characterised by an important attenuation length. The HF is also used for the real-time monitoring of the instantaneous luminosity.

Figure 2.10 shows the position of the four components of the HCAL within a quadrant of the the CMS detector.

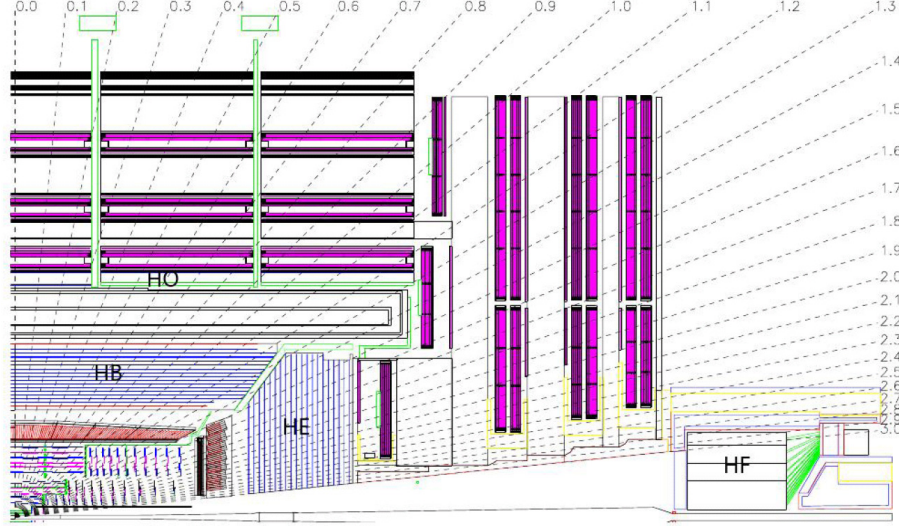


Figure 2.10: Schematic sectional view of a quadrant of the CMS detector with the different HCAL subdetectors named [92].

2.2.4.4 Muon System

The muon system, the outermost subdetector of CMS, surrounding the solenoid coil, is of central importance for CMS as indicated by its middle name. The muon is the only known charged particle that will leave the HCAL without being stopped and with none or very little energy loss in all the subdetectors mentioned above. It is the only one that will interact with the muon system. This system is composed of three different types of muon chambers, which are gaseous detectors, interleaved with the iron yoke layers.

- Drift tubes (DTs)

The DTs are placed in the barrel covering the range $|\eta| < 1.2$. They are filled with a mixture of Ar and CO₂ gas and they have a length of 2 m and a cross section of 4.3×1.3 cm. The three stations closest to the IP have eight chambers with their wires aligned to the beam pipe to give a measurement in the r - ϕ bending plane of the track, and four others with their wires perpendicular to the beam pipe to measure the z coordinate. The last and fourth station only measures the r - ϕ coordinate. Each chamber is made of four layers of drift cells. This technology is suited for the central part of the detector where the muon flux is lower and

the magnetic field uniform. They exhibit a very good single-hit spatial resolution of around $100\,\mu\text{m}$ and an angular resolution of $1\,\text{mrad}$.

- Cathode Strip Chambers (CSCs)

The CSCs are placed in the forward region of the endcaps, $0.9 < |\eta| < 2.4$, where the magnetic field and muon rate are higher. Four stations were installed in each endcap. They use a mixture of Ar, CO₂ and CF₄ and provide a measurement in the r - ϕ plane with their strips pointing radially outward. These chambers also present a good single-hit position resolution of typically $200\,\mu\text{m}$ with an angular resolution of the order of $10\,\text{mrad}$.

- Resistive Plate Chambers (RPCs)

The RPCs are installed both in the barrel and endcaps, extending approximately to $|\eta| < 1.6$. This accompaniment of both the DTs and the CSCs is explained by the fact that their time response for the triggering of an event containing a muon is comparable to the bunch spacing, making it too slow for the high pileup environment of the LHC. The solution is found with the RPCs, which presents a time resolution of $1\,\text{ns}$. Their spacial resolution is however much coarser than the other technology and thus they alone cannot be used.

The positions of all the muon chamber types within CMS are given in Figure 2.11. The determination of the position of a muon track is done up to $|\eta| = 2.4$, from information from the muon chambers alone or from the combination of the muon chambers and the tracking system information. For a muon of transverse momentum below $200\,\text{GeV}$, the resolution is mainly driven by the tracker. Typically it is around 1% (2%) in the barrel (endcaps) for a $50\,\text{GeV}$ muon. For $1\,\text{TeV}$ muons the resolution gives the best performance by using the combined information and reach 5% (10%) in the barrel (endcaps) [99].

2.2.5 Trigger System

Under its nominal design conditions, the LHC can deliver a bunch crossing rate of $40\,\text{MHz}$. It is impossible to read, store and reconstruct such a huge amount of data. To select the events of interest, characterised by particles of high transverse momentum, a sophisticated trigger system exists. It is a two-steps system, composed of a level-1 (L1) trigger that first reduces the rate to $100\,\text{kHz}$, and of a high-level trigger (HLT) which finally lowers it further to $400\,\text{Hz}$, before to store the events on disk.

The L1 is build from programmable hardware and situated in the service and

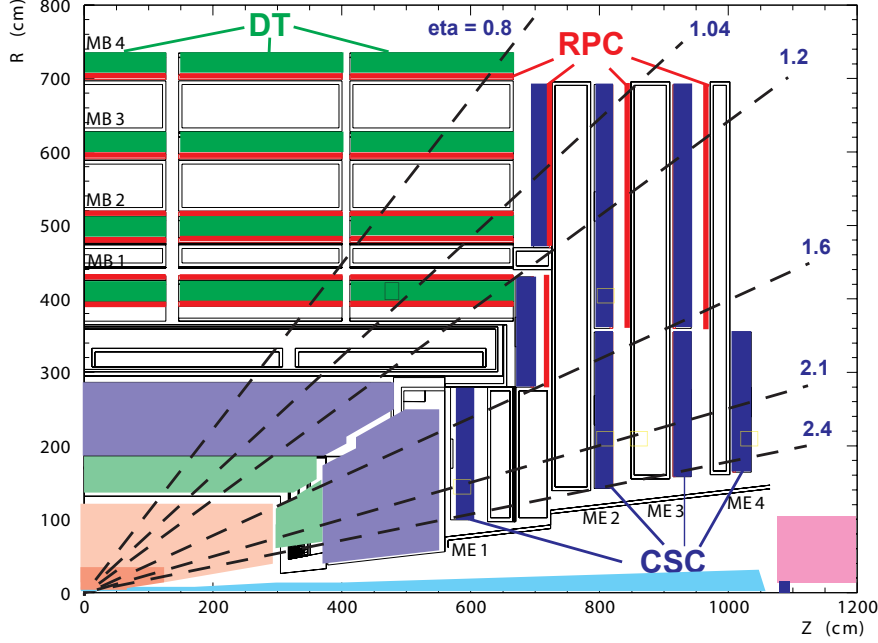


Figure 2.11: Schematic sectional view of a quadrant of the CMS detector with the different muon subdetectors highlighted [92].

experiment caverns underground. It has a requested latency time of $3.2\ \mu\text{s}$ which leads to the constrain to take into account only information from the calorimeters and the muon chambers, since a track reconstruction would take too much time. During this decision time, the data information is kept in a buffer memory installed next to the detector, which can host 128 collisions. To classify and rank the event candidate as interesting or not, energy, momentum and quality of the objects are extracted from the low resolution data read out of the subdetectors. The information are gathered first regionally and then globally for the calorimeters on one side and the muon system on the other side. Eventually the global trigger will take the final decision, with up to 128 trigger algorithms that can be executed in parallel. The events surviving the L1 trigger will be passed to the HLT.

The HLT is situated in the surface building and is a CPU farm linked to the data acquisition system and running a special part of the CMS software. The filter farm consists of more than 1000 processors that have access to all the read out information, and will classify the events according to ‘paths’. Each processor works on one event at the time with an average decision time of 100 ms,

which allows to use more complicated reconstruction algorithms. However, to save computation time, each of the approximately 400 trigger paths of the HLT is built from a succession of filter modules and reconstruction modules. The reconstruction modules are only started if the previous filter module quality condition was fulfilled, and the reconstruction taking the most time, like the track reconstruction, are performed at the end of the path. An event can be accepted by several paths.

If the acceptance rate of some path is too high, due to, for example low thresholds on the transverse momentum of the particles, they can be programmed to be executed only once every given number of events. The trigger are said to be prescaled with a value corresponding to the given number and decrease the rate for that path by the same factor. These triggers are of particular importance for the study of the pileup, which requires a low biased selection of events. Several prescale values are predefined for a trigger path, and can change during a CMS run to keep an optimal global trigger rate.

Chapter 3

Measurement of the Associative Production Cross Section of a Z Boson with b jets

This work aims at measuring the cross sections of the associative production of a Z boson and exactly one b jet, and of a Z boson and at least two b jets. The cross section of the production of a Z boson with at least one b jet will be simply deduced from the two other measurements. An estimation of the ratio of the selection Z+b-jets over the selection Z+jets will also be given. This work has been published in spring 2014 [103] and constitutes the state-of-the-art for that signature. A first measurement of the associative process of a Z boson with at least one b jet, had already been published in 2013 [104] on the first half of the present dataset. While the philosophy of the precision measurement corrected to the generator level cross section, as well as the selection and processes involved have been essentially the same, several changes have greatly improved the analysis. Important changes with respect to this first measurement will be mentioned at the end of this chapter.

The two working points (WP) of selection in the following analysis are:

- Z+1b: selection of a Z boson and exactly one b jet
- Z+2b: selection of a Z boson and at least two b jets

3.1 Analysis Strategy

The main steps of the analysis, described in details in the next sections, are the following:

- CMS reconstruction of all the objects in the final state.
- Preselection of the events in both data and simulations.
 - Normalisation of the simulations to the data luminosity using the theoretical cross sections.
 - Reweighting of the simulations to reproduce the selection efficiencies observed in data.
 - * Efficiencies evaluated from data and Monte Carlo samples independent of the ones of the analysis, with such methods as *Tag and Probe*.
 - * The reweighted simulations will be used later to estimate the selection efficiencies and the acceptance effects.
- Comparison data/Monte Carlo to exclude any major problem on one side or the other.
- Optimisation of the selection to reduce the contamination from the backgrounds while keeping a good signal selection efficiency.
- Estimation of the remaining background contributions with almost fully data driven methods. In particular the estimation of the purity of the sample in terms of b content.
- Calculation of the number of background-free selected events for the reconstructed level.
- Extraction of the cross section at the particle level by intermediate of correction matrices for the selection efficiencies and the effects of the detector, derived from the simulation.

The important point in the extraction of the cross section is that, if the estimation of the backgrounds to be subtracted from the number of selected events comes from the data, the correction matrices used to yield a result at the particle level are Monte Carlo-driven. It is thus essential to apply the proper data-driven corrections to the simulations, and to make sure of a proper agreement data/simulation before to build these matrices.

3.2 Event Reconstruction and Selection

3.2.1 Datasets and Simulation Samples

The data used in this work are those recorded in 2011 by the CMS detector from proton-proton collisions at 7 TeV at the LHC. Only data certified as good for physics are considered for this analysis, which corresponds to an integrated luminosity $\mathcal{L} = 5.0 \pm 0.2 \text{ fb}^{-1}$ [105]. More specifically will be used sets of data that have been selected by specific dilepton triggers which means some datasets contain events presenting two electrons and other datasets include events presenting two muons. The selection chain can be separated in a loose preliminary selection of the events of interest and a final tight selection in which the selection criteria, also called cuts, described in the following sections are applied. The initial requirement for keeping the event is based on the selection used in the measurement of the inclusive Z cross section [106], asking for the presence of two opposite charge and same flavour leptons (electrons or muons), reconstructed with a combined invariant mass compatible with the Z boson mass [2].

In this preselection step, the matching of leptons with the lowest p_T threshold unscaled double lepton trigger is performed. The lepton in the event is required to be close enough in distance and momentum to one of the two leptons that passed the trigger selection to be considered to be the same. Triggers used in this analysis are covering the different periods of data taking for the sample considered. Due to the steadily increase of the instantaneous luminosity delivered by the machine during the period of interest, the threshold for the muon trigger has been progressively increased from 6 up to 17 GeV. Therefore, the analysis uses a combination of the muon triggers listed in Table 3.1 where can also be found the unscaled electron triggers requested.

Table 3.1: *Triggers used in the analysis.*

Internal trigger name	Channel	Threshold GeV,GeV	Comments
<i>HLT_DoubleMu6</i>	dimuon	6,6	-
<i>HLT_DoubleMu7</i>	dimuon	7,7	-
<i>HLT_Mu13_Mu8</i>	dimuon	13,8	higher asymmetric threshold
<i>HLT_Mu17_Mu8</i>	dimuon	17,8	higher asymmetric threshold
<i>HLT_Ele17_CaloIdL_CaloIsoVL_Ele8_CaloIdL_CaloIsoVL</i>	dielectron	17,8	-
<i>HLT_Ele17_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL_Ele8_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL</i>	dielectron	17,8	tighter id for calorimeters id and iso requests for tracker

Datasets used for data are called dilepton skims: two electrons or two muons must be present in the event, with a transverse momentum p_T above 10 GeV,

without quality restriction on the leptons. The datasets are listed in Table 3.2 according to the different eras of data taking. Events are requested to pass selection cuts designed to remove the instrumental background coming e.g. from beam activity (beam scrapping, beam-gas events, etc.). Namely, the fraction of high purity tracks is required to be at least 20% in events with at least 10 tracks and presence of a selected primary vertex. The primary vertex is chosen as the vertex with the largest quadratic sum of the p_T of its constituent tracks, with at least four degrees of freedom, a z position within 24 cm from the nominal centre of the detector and a transverse distance from z -axis in the x - y plane, d_{xy} , smaller than 2 cm.

Table 3.2: *Data samples used in the analysis. The name indicates the skim and the reconstruction campaign date with the version.*

Sample name	Data taking period	Luminosity (fb^{-1}) delivered/recorded	Channel	Dataset name
EIA	Mar. - Sept. 2011	2.39 / 2.33	Electron	/DoubleElectron/Run2011A-08Nov2011-v1/AOD
EIB	Sept. - Oct. 2011	2.87 / 2.77	Electron	/DoubleElectron/Run2011B-19Nov2011-v1/AOD
MuA	Mar. - Sept. 2011	2.39 / 2.33	Muon	/DoubleMu/Run2011A-08Nov2011-v1/AOD
MuB	Sept. - Oct. 2011	2.87 / 2.77	Muon	/DoubleMu/Run2011B-19Nov2011-v1/AOD

The data will be compared with simulated samples for signal and background. The dominant background in this analysis is the inclusive Drell–Yan production in association with jets, named Z +jets or DY +jets. The production of a top-antitop quarks pair, named $t\bar{t}$ process, induces also a background representing about 5% of the total background for the selection Z +1b and 15% for the selection Z +2b. Minor background arises from the diboson ZZ process. Other processes such as QCD multijets (events with only production of one or several jets in the hard process), W +jets, single top, and diboson WW and WZ , producing a final state with misidentified leptons or b jets are found to give a negligible contribution.

In order to compare the data to the theoretical expectations, signal events and the expected backgrounds (Z +jets, $t\bar{t}$, and ZZ) are generated by Monte Carlo (MC) simulations. Samples of events with Z +jets or a $t\bar{t}$ pair are produced with MADGRAPH [60, 61] interfaced with PYTHIA 6 [57] with the Z2 tune [107, 108] for the parton showers, hadronisation, and MPIs. The CTEQ6L1 parton distribution functions [49, 50] are used. The matching is performed with the default k_T MLM scheme [68]. For Z +jets, events with a vector boson Z and up to four jets in the final state at leading order are generated using the matrix element calculation of MADGRAPH in the 5-flavour scheme. Further jets are

generated by the parton shower. The ZZ sample is simulated using PYTHIA 6 only, which also comes with its own ME generator.

Generated events are processed through the full GEANT4 [109] detector simulation and event reconstruction chain of the CMS experiment.

The Z+jets sample is also used to extract the signal efficiencies and for the comparison of kinematic distributions.

These simulated samples are normalised to the cross sections expected from theory. The cross section for the Z+jets sample (3048 pb) is normalised to match the NNLO prediction for inclusive Z production obtained with the FEWZ (Fully Exclusive W and Z Production) tool [110] and the CTEQ6m PDF set [49, 50]. NLO predictions obtained from MCFM [111] are used for the normalisation of the $t\bar{t}$ sample cross section (157.5 pb) and the ZZ sample one (5.9 pb) [112]. The simulated Z+jets sample is split into three sub-samples, according to the underlying production of b jets, c jets, or jets originating only from gluons or u,d,s quarks (hereafter called light jets), with no requirement on the p_T or η of the jets. These simulated jets are considered as coming from a b quark if a B hadron is found within a distance $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.5$ around the jet axis, and as coming from a c quark if a D hadron is found. The rest of the jets are classified as coming from light jets. These sub-samples are labelled Z+b, Z+c, and Z+l, respectively.

Table 3.3 summarises the simulation samples used in this analysis. In all cases, the samples correspond to an official CMS production ‘Fall11’ processed with the release 4.4.2 of the CMS software [113].

Table 3.3: *MC samples used in the analysis with their theoretical LO cross section and their theoretical NNLO or NLO cross section used to normalise the samples [114].*

MC sample	σ LO (pb)	σ (N)NLO (pb)	Events
Z+jets: /DYJetsToLL_TuneZ2_M-50_7TeV-madgraph-tauola	2475	3048 NNLO	36 264 432
$t\bar{t}$: /TTJets_TuneZ2_7TeV-madgraph-tauola	94.76	157.5 NLO	59 244 088
ZZ: /ZZ_TuneZ2_7TeV_pythia6_tauola	4.287	5.9 NLO	4 191 045

3.2.2 Leptons

The leptons in the event are required to originate from the primary vertex, which was defined in Section 3.2.1.

Both of the two electrons or muons are required to be reconstructed offline, for events that were selected by the HLT, with a transverse momentum above 20 GeV and a pseudorapidity $|\eta| < 2.4$ to be well inside the detector acceptance.

Electron reconstruction

Electrons are identified by combining tracker tracks and ECAL superclusters (SC) including the ECAL deposits from bremsstrahlung [115].

These SC are built in the following way: around the crystal with the largest energy deposit (seed crystal), a matrix of 5×5 crystals is considered and its energy content added to form a cluster. This matrix is expected to contain 97% of the energy of the incident electron. But before reaching the ECAL the electron can lose energy in the tracker due to bremsstrahlung. This energy is recovered by extending the search for energy in the ϕ direction of the seed crystal up to 17 crystals (0.3 rad); it is then added to the energy of the cluster and a SC is formed. The position of the SC is the one of the seed crystal shifted according to the position and weight of bremsstrahlung energy deposits. To ensure the matching between tracks and SC the differences in η and ϕ of the tracks position at the vertex and of the SC are required to be small. What is called ‘track position at the vertex’ is the track position in the inner layers extrapolated to the interaction vertex and then extrapolated to the calorimeter. Another criteria ensuring the quality of the matching is to require a small difference between the energy reconstructed in the supercluster and the momentum of the associated track, via the variable $|1/E - 1/p|$.

The electron is expected to shower in the crystals of the ECAL. To distinguish them from other showering particles, like a hadron giving a jet in a relatively largely spread shower, which would continue in the HCAL, a maximal value is given on both the transverse and the longitudinal shower shape. The transverse shower shape is accessed via the $\sigma_{in\eta}$ variable, measuring the width in η of the energy deposit in a 5×5 matrix around the seed crystal. The longitudinal shower shape can be estimated as the ratio of the energy deposited in the HCAL with respect to the energy deposited in the ECAL.

Electrons coming from conversion, i.e. interaction of a photon with a nucleus resulting in a dielectron e^+e^- pair, also have to be discarded. This is achieved by requiring a maximum distance of $200 \mu\text{m}$ of closest approach between the

track of the electron candidate and the primary vertex in the transverse plane. This distance is called the transverse impact parameter d_0 .

A vertex fit is also performed to determine if the track actually comes from the primary vertex, and there is a veto on candidate tracks presenting more than one missing hit in the inner most layers of the tracker (including the pixel detector).

To reduce the contamination by the pileup a cut on the longitudinal impact parameter to the primary vertex, d_Z is applied at 0.1 cm.

A fiducial region in the ECAL is defined which excludes electrons close to the transition between barrel and endcaps, and electrons in the first endcap trigger tower which lies in the shadow of cables and services exiting in that transition region. The fiducial region cut is applied on the position of the ECAL supercluster (SC) which is required to have $|\eta| < 2.5$ with the region $1.4442 < |\eta| < 1.566$ excluded.

Muon reconstruction

Muons are reconstructed through a Particle Flow (PF) algorithm [116, 117], that reconstructs all final state particles in the event by combining the information obtained from each of the CMS subdetectors. The combination of the information leads to an improved robustness of the momentum determination. The PF algorithm applies particular optimised selection criteria on muon candidates reconstructed with the Global and Tracker muon algorithms, which are two reconstruction approaches used in CMS. The Global muon one (outside-in) fits a global track by combining hits from a track in the inner tracker and hits from a segment in the muon chambers. For the Tracker muon approach (inside-out), any tracker track above a low momentum threshold, is extrapolated to the muon system and required to match one muon segment.

Thanks to the high efficiency of reconstructing tracks in the tracker and the very high efficiency of the muon system segments reconstruction, muons produced in the acceptance of the muon chambers are very often reconstructed as both Tracker and Global muons. The efficiency of muon reconstruction in one of the two algorithms is 99%.

If the muon is not isolated, or if its momentum is not compatible with the energy deposition in the calorimeters assigned to the candidate by the particle-flow event reconstruction, the PF algorithm will require an additional set of tight requirements in order to purify the collection from misidentified charged hadrons and other particles punching through the calorimetric system.

A more detailed description of the PF algorithm may be found in [116, 118, 119].

Non-prompt muons are rejected by asking the transverse impact parameter with respect to the primary vertex to be smaller than $200\ \mu\text{m}$.

Lepton isolation

An isolation variable is used to reject electrons (muons) that are embedded in jets, in order to select only the leptons coming from the Z boson. It is defined as the sum of the magnitude of the transverse momenta of the particles or tracks reconstructed in a cone around the lepton candidate, relative to the transverse momentum of the lepton. The cone being of a size $\Delta R = 0.3$ (0.4).

To reduce the effect from pileup which deteriorates the isolation of the leptons, charged particles not associated with the primary vertex are filtered and are not considered in forming the isolation variable. The contribution of neutral particles, for which this association is not straightforward, is corrected by subtracting the energy deposited in the isolation cone by charged particles not associated with the primary vertex, multiplied by a factor of 0.5. This factor corresponds approximately to the ratio of neutral to charged hadron production in the hadronisation process of pileup interactions [115, 120]. After this correction, the isolation variable is required to be less than 0.15 for electrons and 0.20 for muons.

The leptons are also expected to match the appropriate trigger of Table 3.1 according to the period of the dataset used. This means to check if the lepton is one of those which have triggered the selection and the reconstruction of the event. For this, the lepton candidate and the ‘triggering’ lepton should lie in a cone distance ΔR smaller than 0.3 and have a relative p_T difference smaller than 0.5.

Lepton pair selection

Opposite charges for the leptons are required when forming pairs and the dilepton invariant mass $M_{\ell\ell}$ is required to lie between 76 and 106 GeV. In the case of multiple lepton combinations, the lepton pair with the invariant mass the closest to the nominal Z-boson mass of 91.1876 GeV [2] is selected as the Z candidate. The mass window has been chosen with the goal to eliminate a large part of the $t\bar{t}$ background while keeping most of the Z+b-jets signal, as can be seen in Figure 3.1: from the mass range between 60 and 120 GeV considered, a final narrow [76–106] GeV mass window is selected.

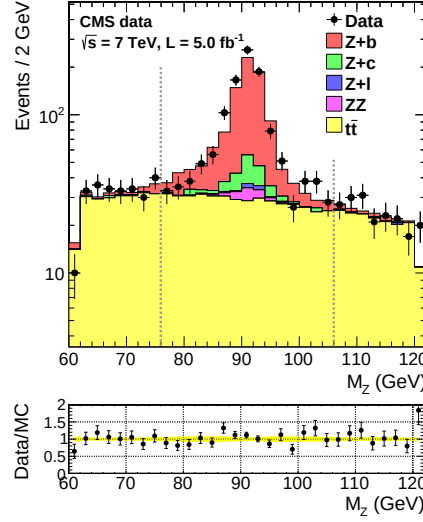


Figure 3.1: Wide mass window for the distribution of the Z mass. The yellow band in the lower plot represents the statistical uncertainty on the MC yield.

Selection cuts for electrons (muons) are summarised on Table 3.4 (3.5).

The efficiency of the dilepton selection is estimated in data and simulation using the *Tag and Probe* method [106], for events with two leptons and one jet satisfying the requirements detailed above for the leptons and in Section 3.2.3 for jets. The method will be discussed in detail in Section 3.2.6.2. For the data versus simulation comparisons, data/MC scale factors are estimated per lepton as a function of their p_T and η , and the simulation is then reweighted by this ratio.

The same idea applies for the pileup, for which data and MC estimations of the number of pileup interactions are derived and factorised as a ratio data/MC scale factor. Details will be discussed in Section 3.2.6.1.

The distributions of the main electron selection criteria are shown in Figure 3.2, and the ones for muons are shown in Figure 3.3. One can see that the simulation and data for these variables are in a general good agreement. In all these figures, the Monte Carlo samples are reweighted by the pileup and lepton scale factors, correcting for the differences in efficiencies.

Table 3.4: *Electron selection variables*

Type	Criterion	Variable	Cut	
Working Point Medium			Barrel $ \eta_{SC} \leq 1.479$	Endcap $1.479 < \eta_{SC} < 2.5$
	SC-track matching	$ \Delta\eta_{(\text{SC-track})} $	< 0.004	< 0.007
		$ \Delta\Phi_{(\text{SC-track})} $	< 0.06	< 0.03
		$\left \frac{1}{E} - \frac{1}{p}\right $	< 0.05	< 0.05
	Transverse shower shape	$\sigma_{\text{in}\eta}$	< 0.01	< 0.03
	Longitudinal shower shape	$\frac{E_{\text{HCAL}}}{E_{\text{ECAL}}}$	< 0.12	< 0.10
	Conversion rejection	$ d0 $ (cm)	< 0.02	< 0.02
		vertex fit probability	10^{-6}	10^{-6}
		missing hits	≤ 1	≤ 1
	Pileup rejection	$ d_Z $	< 0.1	< 0.1
Extra cuts	Cluster fiducial cuts	$ \eta_{SC} $	$1.4442 < \eta_{SC} < 1.566$	
	Isolation	$\frac{\sum_i p_T^i}{p_T}$ (in cone $\Delta R=0.3$)	< 0.15	
	Trigger matching	$\Delta R_{(\text{trig},e)}$ and $\frac{\Delta p_T(\text{trig},e)}{p_T}$	$\Delta R < 0.3$ and $\frac{\Delta p_T}{p_T} < 0.5$	
	Acceptance cuts	p_T	> 20 GeV	
		$ \eta $	< 2.4	

Table 3.5: *Muon selection variables*

Criterion	Variable	Cut
Reconstruction	Global and Tracker muon in PF algo.	
Tight requirements	Number tracker hits + Number pixel hits	> 10
	Number pixel hits	> 0
	Number muons hits	> 0
	Number of matched muon stations	≥ 2
	Number track layers hits	> 8
	Global muon track fit $ \chi^2 $	< 10
Non-prompt rejection	$ d0 $ (cm)	< 0.02
Isolation	$\frac{\sum_i p_T^i}{p_T}$ (in cone $\Delta R=0.4$)	< 0.20
Trigger matching	$\Delta R_{(\text{trig},\mu)}$ and $\frac{\Delta p_T(\text{trig},\mu)}{p_T}$	$\Delta R < 0.3$ and $\frac{\Delta p_T}{p_T} < 0.5$
Acceptance Cuts	p_T	> 20 GeV
	$ \eta $	< 2.4

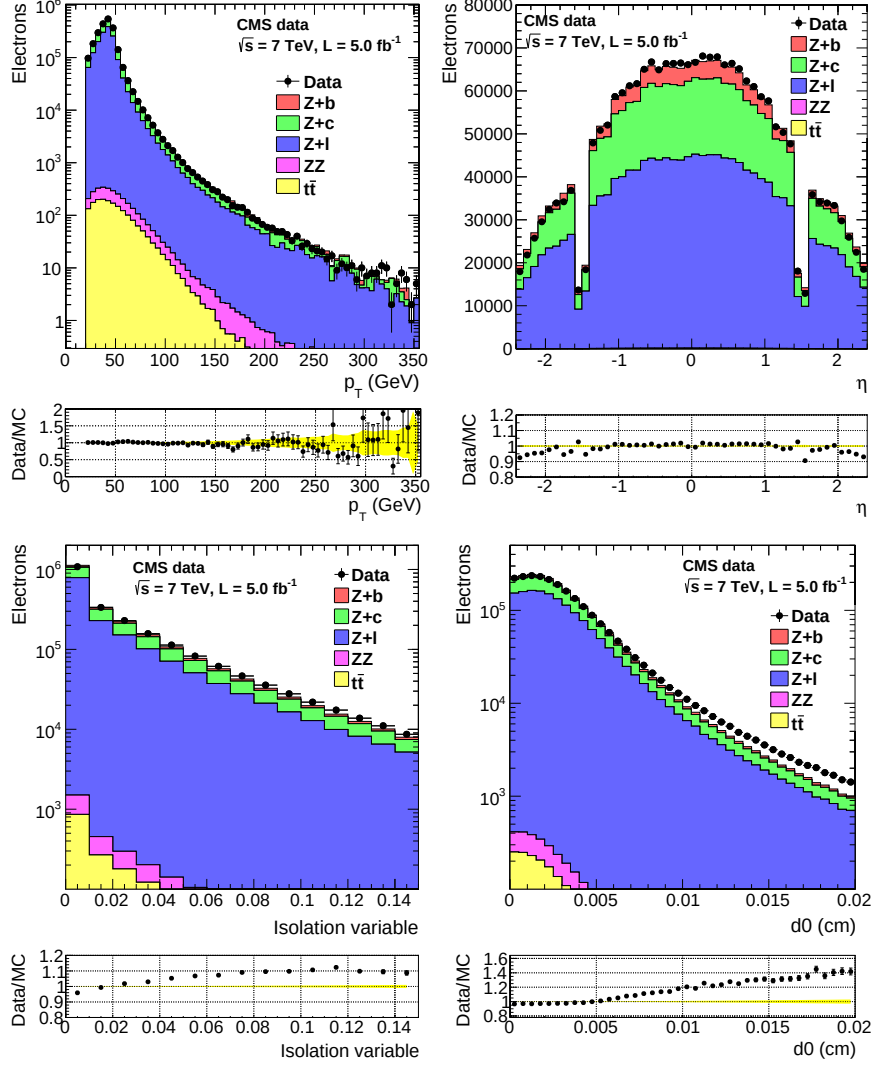


Figure 3.2: Distributions of variables used in the electron selection. The yellow bands in the lower plots represent the statistical uncertainty on the MC yield. A good agreement between data and simulation is found, except for some discrepancy in the tail of the isolation and of the d_0 variables. This can be explained by the presence of other particles or jets, falsely reconstructed as electron in the data, or by real electrons in events more affected by pileup and/or QCD activity.

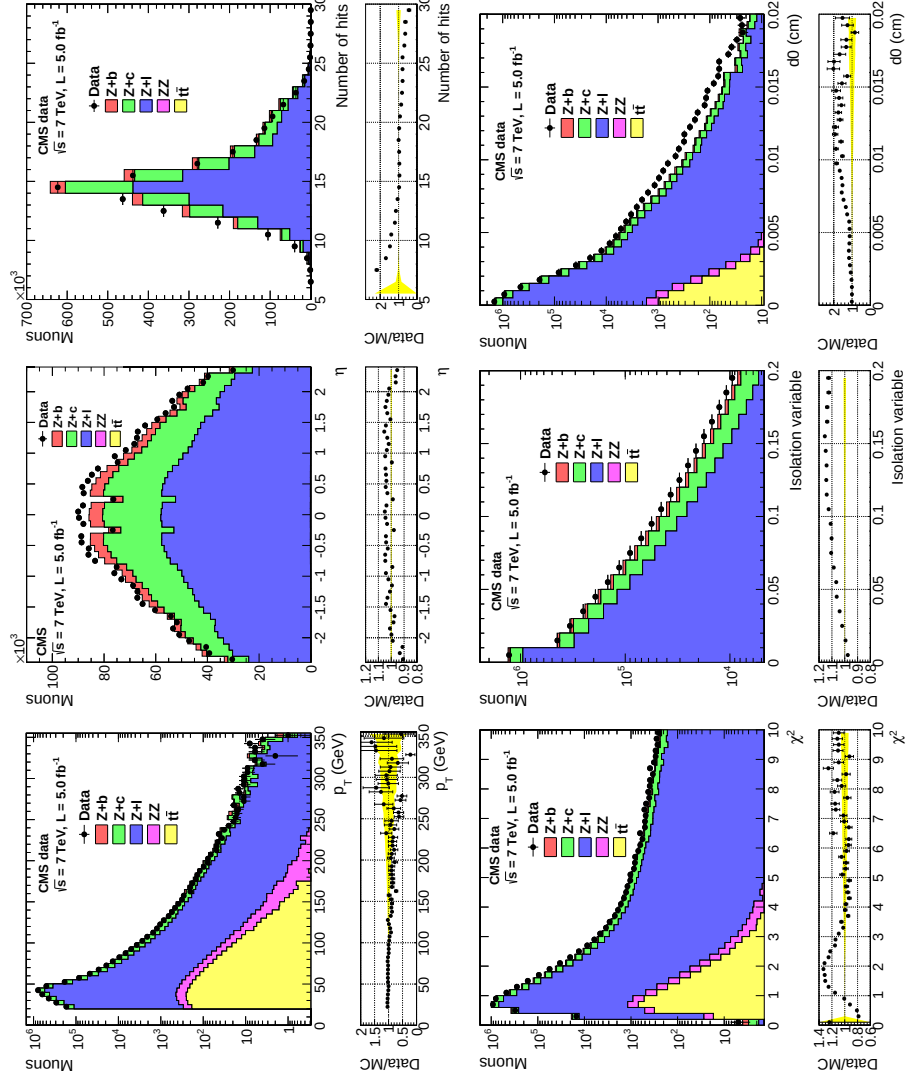


Figure 3.3: Distributions of variables used in the muon selection. The yellow bands in the lower plots represent the statistical uncertainty on the MC yield. A good agreement between data and simulation is found, except for some discrepancies in the central region of the η variable, and in the tails of the isolation and d_0 variables. This can be explained by the presence of other particles or jets, falsely reconstructed as muon in the data, or by real muons in events more affected by pileup and/or QCD activity.

3.2.3 Jets and b-tagging

The final selection requires at least two jets identified as coming from b quarks. Jets are reconstructed using the Particle Flow algorithm [118, 119]. Individual particles (leptons, photons, charged and neutral hadrons) are reconstructed by linking tracks, ECAL clusters, and HCAL clusters. Each particle is reconstructed with the optimal momentum or energy resolution by considering information from all sub-detectors. These particles are then clustered into jets using the anti- k_T jet clustering algorithm [121] with a distance parameter R of 0.5, as implemented in the FASTJET program [122, 123]. This algorithm regroups the objects into jets in an increasing order of distance between two objects or from one object to the beam. This distance being inversely proportional to the squared transverse momentum. Objects are combined in jets only if they are situated from each other at a distance ΔR smaller than the distance parameter. This algorithm drives the shape of the jet to be close to a cone of area $\pi \cdot R^2$. It does not depend on low- p_T objects, while still including them.

The jets are calibrated to ensure a uniform energy response in p_T and η . The Jet Energy Scale (JES) calibration technique implemented in CMS [124, 125] uses the fact that the true unbiased jet momentum can be related to the raw, uncorrected reconstructed jet p_T . A factorised approach can be used to correct the jets. First, an offset correction is performed to remove any contamination from pileup. This contribution to the jet transverse energy from pileup is estimated using the jet area method [126] on an event-by-event basis, and subtracted from the overall p_T response. Subsequently, a simulation driven correction is derived to flatten the JES as a function of p_T and η . Finally, in real collision events, the residual difference with the simulation is absorbed in a residual relative correction as a function of η , and in an absolute correction as a function of p_T . Both are measured using the p_T -balance of a back-to-back pair in γ +jet, Z+jet, and dijets events.

After correcting the jets with the JES corrections, the Jet Energy Resolution (JER) in data can be compared to simulation [124]. Once again a p_T -balance in dijets events can be used. A much worse JER was found in data, so on top of the JER of the simulated jets, a smearing of the resolution of 10% is performed to account for the difference.

The reconstructed jets are required to have $p_T > 25$ GeV and to be separated from each of the leptons from the Z boson by at least $\Delta R = 0.5$. This ensures that the leptons are not overlapping with the jet, whose radius can be considered to be approximatively of the size of the impact parameter used in the algorithm [121]. Furthermore, jets are required to have $|\eta| < 2.1$, which is

in principle smaller than the acceptance of $|\eta| < 2.5$ of the CMS tracker but can be understood since the cut is on the jet axis η value whereas the jet has a cone size. Loose identification criteria [124] are applied in order to reject jets coming from beam background, calorimeter noise, and isolated photons. These criteria are based on the requirements that the total energy is shared between more than one readout cell and not coming purely from deposits associated with neutral particles. The selection efficiency for real jets is close to 100% in both data and simulations. Jets originating from pileup in the Z+jets sample, and thereby contributing falsely to the cross section ratio, are suppressed by requiring the momentum of particle tracks from the jet, originating from the selected primary vertex, to be at least 10% of the total jet momentum. The remaining background caused by jets from pileup is $\sim 2\%$ in the Z+jets simulation sample.

Jets originating from b quarks are tagged by taking advantage of the long B hadron lifetime, which will be signed by a secondary vertex separated from the primary vertex, with the highest quadratic sum over the p_T of its constituent tracks, $\sum_{tracks} p_T^2$.

The Simple Secondary Vertex (SSV) b-tagging algorithm uses a discriminator which is a monotonic function of the three-dimensional flight distance significance from the reference primary vertex to the chosen secondary vertex. It takes values above one when a secondary vertex is found. The cut on the $|\eta|$ of the jet is then justified by the fact that the SSV algorithm requires the secondary vertex to be found within $\Delta R = 0.3$ with respect to the jet axis. A maximal value of $|\eta| = 2.1$ for the jet ensures to be still able to target the secondary vertex if present and to keep a good efficiency for this b-tagging algorithm.

To maximise the selection efficiency of the Z+b-jets process, especially when asking two jets, the high-efficiency version of the SSV is used (SSVHE), which considers all secondary vertices built from two tracks or more. The discriminant value to define b-tagged jets is chosen such that the rate of tagging a light-parton jet (mistagging rate ε_l) is below 1%, with a b-tag efficiency ε_b of $\sim 55\%$. Another version of the SSV with high-purity (SSVHP) exists and will be evoked later on. Characteristics of the two versions of the SSV algorithm are gathered in Table 3.6. Further details can be found in [127, 128]. The number of 2-tracks secondary vertices per jet is shown in Figure 3.4 (left), after the dilepton+jet selection. The distribution of the SSVHE discriminant is shown for the leading jet in Figure 3.4 (right).

Table 3.6: *Characteristics of the two versions of the Simple Secondary Vertex algorithm: High Efficiency and High Purity.*

Version	Requirements	Discriminant	Cut	ε_b (%)	ε_l (%)
SSVHE	1 secondary vertex with ≥ 2 tracks	$\log(1 + \frac{d_0}{\sigma_0})$	1.74	55	< 1
SSVHP	1 secondary vertex with ≥ 3 tracks		2.0	45	< 0.1

Even though the complete scale factors question will be discussed in detail in next Section 3.2.6, one can notice from Figure 3.4 (right) that the MC simulation tagging efficiency is higher than the one of the data. The efficiencies are measured in both data and MC as a function of p_T and η from inclusive jets events, and a weight is applied to the simulation for each b-tagged jet, depending on the generator level flavour (l, c or b), as detailed in Section 3.2.6.3. In all the following results, except if stated otherwise, reweighting from pileup, leptons and b-tagging will be applied to the simulation events, depending on the number of b jets requested at each selection step.

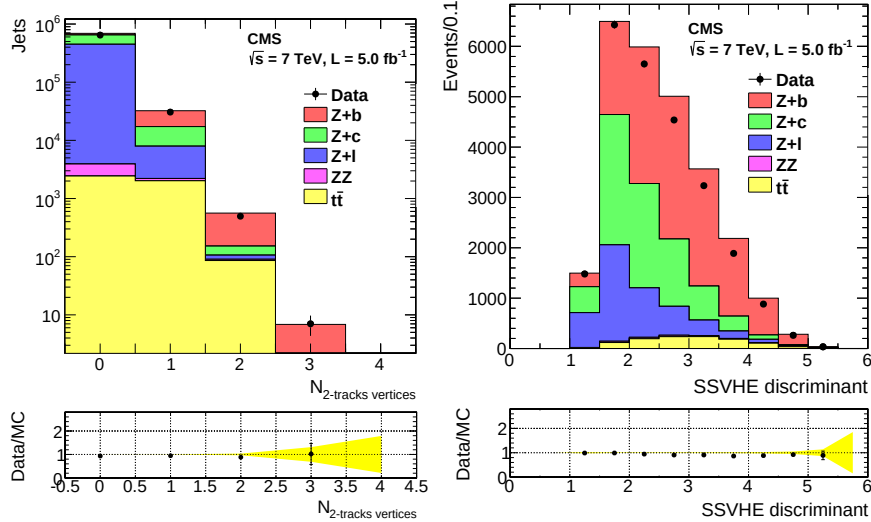


Figure 3.4: *Left: number of two-tracks secondary vertices per jet in dilepton + jet events. Right: SSVHE discriminant for the leading jet after the dilepton + jet selection. The yellow bands in the lower plots represent the statistical uncertainty on the MC yield.*

3.2.4 Missing Transverse Energy: MET

The missing transverse momentum is defined as the negative vectorial sum of the transverse momenta of all the visible particles in the detector, in the final state of one event. The missing transverse energy (MET) is the magnitude of this vector. It accounts for the energy of the neutral weakly interacting particles, such as neutrinos, which do not interact with any part of the detector and escape the collision without providing a direct response. Only the conservation of the vector momentum in the plane transverse to the beam direction allows to retrieve this energy. The conservation along the beam direction is not usable since it is impossible to determine which fraction of the initial energy has been involved in the collision and what is the energy loss along the beam pipe.

In Z plus b jets events final state, very little MET is expected and therefore this variable could be used to eliminate some background whose decay includes neutrinos. One example is the $t\bar{t}$ process, in which the two top quarks decay to two W bosons and two b quarks, the W bosons decaying subsequently each to a lepton and its associated neutrino.

In this analysis the particle flow MET (PF MET) is used, which is calculated using a complete particle flow technique [118]. This reconstruction uses all the detector information, leading to a complete list of identified and reconstructed PF objects, which can be either unclustered (leptons) or clustered (jets, without any correction). From this list the MET is calculated as the negative of the vectorial sum over all the PF objects transverse momentum.

The MET is corrected to account for the degradation of the reconstruction due to the pileup interactions. This is done by applying a charged hadrons subtraction, which removes the charged hadrons originating from the vertices of pileup interactions. Since the jets are used to calculate the PF MET, the JES corrections must be propagated, to account for the possible underestimation of its magnitude [129]. However jets are only properly corrected if they present a transverse momentum above 10 GeV. So jets with $p_T > 10$ GeV will be corrected for JES while jets with $p_T < 10$ GeV remain uncorrected.

Studies of the MET reconstruction in 2011 data [129] found an asymmetry in the azimuthal angle of the PF MET, while the distribution of true MET is independent of ϕ because of the rotational symmetry of the collisions around the beam axis. The cause for the shift has not yet been identified unambiguously. Most likely, the observed shift is due to several effects. A ϕ -asymmetry is also present in simulated events, albeit of different magnitude and pointing in a different direction.

It has been noticed, however, that the MET ϕ -asymmetry is due to a vectorial

shift of the E_x and E_y components, which increases approximately linearly with the number of reconstructed vertices. This correlation is utilised for a correction procedure called the ϕ correction.

A detailed description of the MET reconstruction goes beyond the scope of this manuscript, more details can be found in [118] and [129].

As seen in Figure 3.5 (top left) a cut on the MET at about 50 GeV would already remove a significant amount of $t\bar{t}$ background while keeping most of the signal. However, a more robust variable is used in this analysis: the significance of the missing transverse energy (MET) which assesses, on an event by event basis, the likelihood that an observed MET is consistent with a fluctuation from zero because of detector-related limitations, like finite measurement resolution. In the Gaussian case, the MET significance is equivalent to a two-dimensional χ^2 value and can be treated as such. The mathematical formalism to calculate this and the basic implementation in the CMS reconstruction software can be found in [129, 130]. The MET significance is based on the PF MET and uses jets and leptons reconstructed with the particle flow algorithm. The unclustered energy in each event, consisting of the remaining low- p_T objects extending down to zero energy, also enters in the MET significance calculation to provide basic information about the low- p_T hadronic sector.

The MET significance is more discriminating than the MET and at the same time leads to smaller systematic uncertainties. For the analysis final selection a cut below 10 u.s (units of significance) is applied on the MET significance which can be observed in Figure 3.5 (top right) after selection of a Z and 2 b-tagged jets. The effect of the MET significance cut on the MET can be observed by comparing Figure 3.5 (top left) before the cut, and Figure 3.5 (bottom) after the cut. One can see that this leads to a much smoother spectrum of the MET than if it would have been cut sharp at 50 GeV.

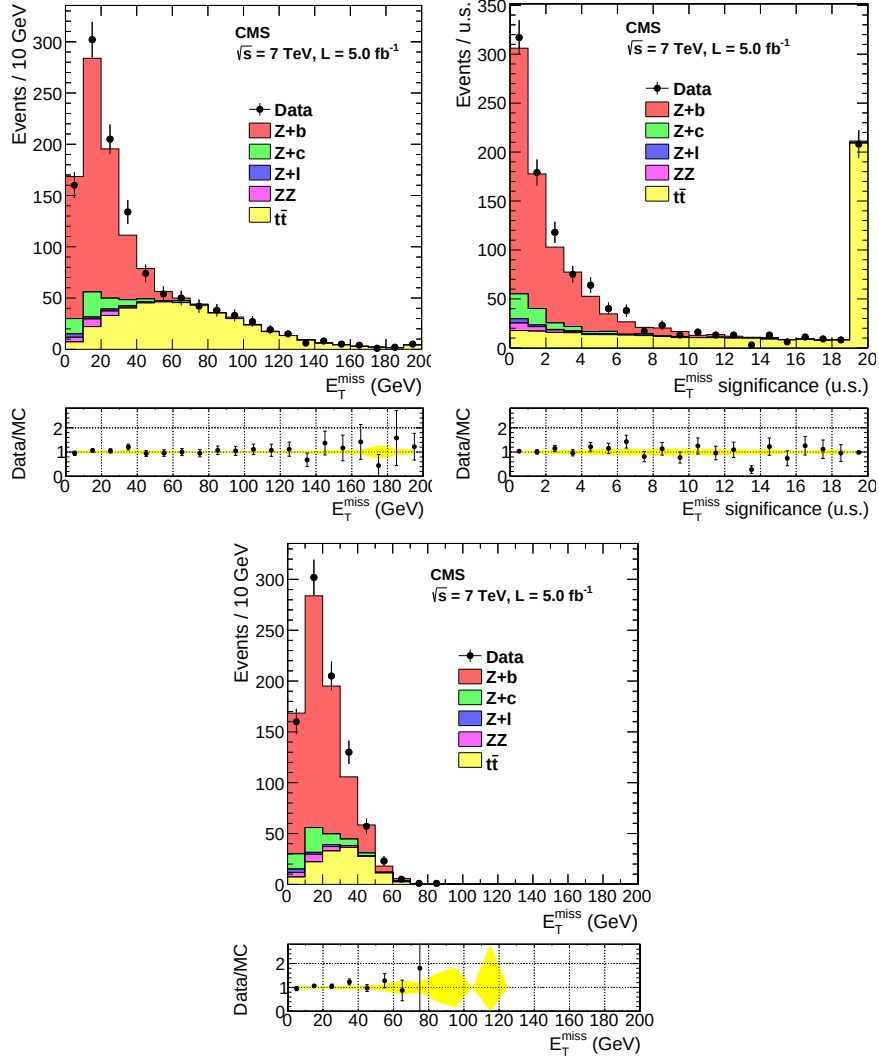


Figure 3.5: Top left: The missing transverse energy after dilepton + 2 b-tagged jets selection. Top right: The missing transverse energy significance after dilepton + 2 b-tagged jets selection. Bottom: The missing transverse energy after dilepton + 2 b-tagged jets selection and the cut on the MET significance. The yellow bands in the lower plots represent the statistical uncertainty on the simulation yield.

3.2.5 Summary of the Selection

The full list of selection criteria detailed previously, is given in Table 3.7.

Table 3.7: *Selection criteria used in the Z+b-jets cross section measurement.*

Object	Selection criterion
Leptons (PF muons, electrons)	$p_{\text{T}}^{\ell} > 20 \text{ GeV}$ $ \eta^{\ell} < 2.4$
Z(ll)	$76 < M_{\ell\ell} < 106 \text{ GeV}$
Jets (ak5 PF)	$p_{\text{T}}^j > 25 \text{ GeV}$ $ \eta^j < 2.1$
Jet-lepton separation	$\Delta R(\ell, j) > 0.5$
b-tagging	SSVHE: $disc. > 1.74$
PF MET (only for Z+2b)	MET significance < 10

3.2.6 Correction Factors

3.2.6.1 Pileup Reweighting

Simulation reweighting is done by assigning weights to a MC generated event such that the observed distribution of some variable in the simulation matches the observed distribution of the same variable in a data sample. The distribution has to be chosen for its robustness and can be either the observable quantity or a more fundamental variable.

In the case of the pileup distribution, the observable in the data is the number of reconstructed primary vertices. It is measured in a series of events for which the variation of instantaneous luminosity is small enough to be considered constant (a ‘LumiSection’ lasting 23 seconds). The average number of additional vertices corresponds more or less linearly to the number of additional interactions, modulo the vertex reconstruction efficiency which is approximately 70%. Therefore, the average number of reconstructed vertices divided by 0.7 is an estimate of the pileup present in one LumiSection. However, this estimate suffers from variations of the vertex reconstruction efficiency as a function of the pileup and as a function of time (e.g. because of misalignment).

The ‘true’ number of interactions in the data can however be retrieved. For this the measured instantaneous luminosity $\mathcal{L}$ for each LumiSection is used, obtained from the occupancy in the pixel detector as described in [131].

Multiplying $\mathcal{L}$ by the total inelastic cross section as in the Formula 3.1, one obtains the mean number of interactions for a bunch crossing.

$$\langle N \rangle = \mathcal{L} \times \sigma_{pp} \quad (3.1)$$

The total proton-proton cross section has been measured by CMS, ATLAS and TOTEM [132, 133, 134]. A central value of 68 mb was found for data at 7 TeV. This method relies on our knowledge of the instantaneous luminosity, of the total inelastic cross section and the forward physics relevant for the measurement of this cross section.

On the simulation side, the Monte Carlo samples used for the analysis have been simulated with a number of interactions per event to imitate the data conditions of pileup. This information is stored as a ‘true’ input, and its distribution is convoluted with a Poisson distribution at every bin, in order to create an observable quantity. In spite of a good simulation technique for the pileup, the simulation and the data distributions of the number of reconstructed primary vertices do not match. This is due to the fact that the simulation was generated

before the data taking, when the pileup distribution was not known. To correct this mismodelling one wants to reweight the simulation event-by-event.

In the case of pileup, the estimation of the ‘true’ distribution in data relies on a much more precise and robust method than the estimation of the observed distribution. In the simulation the true distribution is of course the genuine one.

Each event in the MC selection is therefore reweighed with a factor or weight coming from the number of interactions comparison in data and simulation. The distribution of the number of reconstructed primary vertices after pileup reweighting can be seen in Figure 3.6, where one can observe a good agreement between data and simulation, up to about 20 vertices.

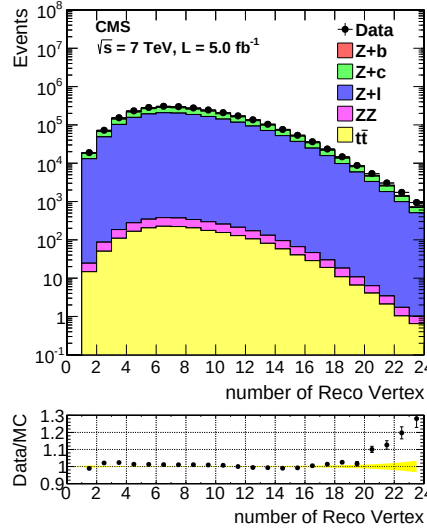


Figure 3.6: *Primary vertex multiplicity after pileup reweighting of simulated events compared to the data distribution.*

3.2.6.2 Lepton Scale Factors

Each step of the reconstruction of the leptons for the analysis is subject to an intrinsic efficiency, slightly different in Monte Carlo simulations and in data. To improve the simulation description of the data one wants to correct these simulation efficiencies by the corresponding data/MC ratio. The total efficiency measurement can be factorised in five sequential relative measurements. The

five steps are: tracking efficiency, reconstruction efficiency, identification efficiency, isolation efficiency and the online or total trigger efficiency. The total is thus given by the product:

$$\epsilon_{\text{lepton}} = \epsilon_{\text{Tracking}} \cdot \epsilon_{\text{RECO/Tracking}} \cdot \epsilon_{\text{ID/RECO}} \cdot \epsilon_{\text{ISO/ID}} \cdot \epsilon_{\text{Trigger/ISO}} \quad (3.2)$$

The tracking efficiency is considered to be close to 100% [100]. For the other lepton selection steps accurate measurements of the efficiencies from the data and from the simulation are needed separately. The problem then becomes to identify the desired particle types in data with a small, or easily subtracted, fake rate.

For this the so-called *Tag and Probe* method will be used, a generic tool using known resonances, like Z or J/ Ψ for measuring efficiencies [106, 135]. The principle is to select particles of the desired type from a data or simulation sample, and probe the efficiency of a particular selection criterion on those particles. The so-called ‘tag’ is an object passing a set of tight selection criteria that ensure that the fake rate is very small, for leptons it is smaller than 1%. In the same event the ‘probe’ is a generic object of the type of the particle studied, with loose selection criteria. The tag and the probe will be paired such that the invariant mass is consistent with a chosen resonance. The probe is defined according to the selection criterion that needs to be studied and the efficiency will be:

$$\epsilon_{\text{criterion}} = \frac{N_{\text{Probes Passing}}}{N_{\text{All Probes}}} . \quad (3.3)$$

The background is distinguished from the signal by simultaneously fitting the invariant mass spectra for the passing probes and the failing probes. The efficiency is then computed from the normalisation of the signal shapes in the two spectra and can be estimated in bins of p_T and η .

For the analysis, events with leptons resonating to a Z boson are selected because of the lepton momenta range, p_T greater than 20 GeV, where the kinematics are similar to the Z decay products. For lower p_T values a different resonance would have been used like J/ Ψ or Υ . A similar event topology to the one of the analysis is chosen to standardise the kinematics between the *Tag and Probe* selection and the analysis: two leptons and at least one jet are required in the event. The cuts on p_T and η for at least one jet and on the ΔR between the tag lepton and the jet would then be applied. In practice only in the muon case the jets have undergone those cuts, since for the electrons the finer binning in addition to the requirement of a jet in the event lead to a sensible decrease of the *Tag and Probe* pairs in each bin and therefore fits are less stable and can even not converge in some cases.

The data sample for the *Tag and Probe* study uses events from the single-lepton trigger dataset. The simulation sample for the signal is the inclusive MADGRAPH Drell–Yan process to leptons with QCD multijets as background. The events are reweighted for the pileup efficiency according to what has been described in Section 3.2.6.1.

The final efficiencies are computed by fitting the resonance peak with a Crystal Ball function and using an exponential function to fit the background. The Crystal Ball function is a probability density function used to model various lossy processes. It consists of a Gaussian convoluted with a power-law low-end tail, below a certain threshold. The function itself and its first derivative are continuous. In the estimation several parameters are left floating: normalisation of signal and background, efficiency of background and parameters controlling the shape of signal and background (sigma for Gaussian for example). An example of fit result and extraction of efficiency in the case of the muon can be seen in Figure 3.7.

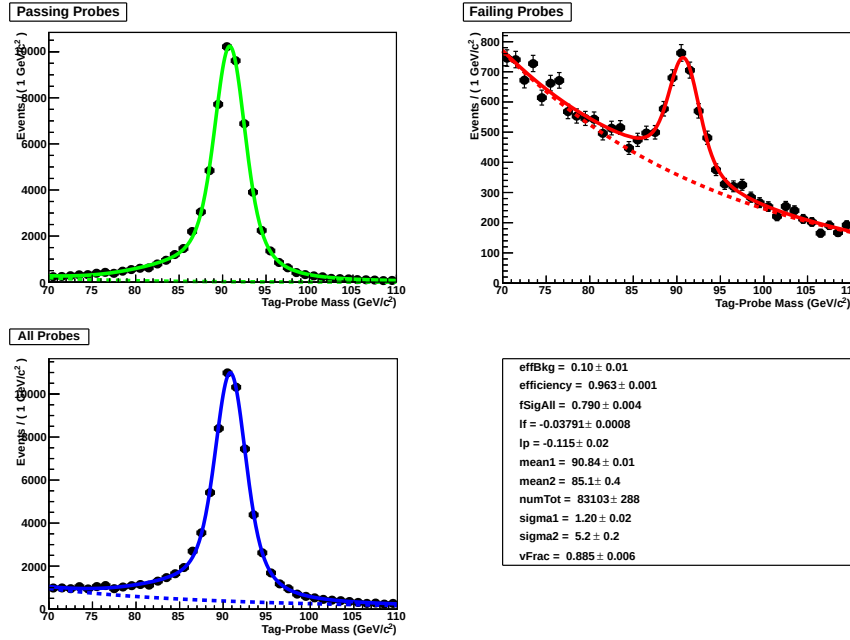


Figure 3.7: Example of efficiency extraction with the *Tag and Probe* method for the muon case. In the upper left plot the fit to the muon pairs passing the probe request is shown, while in the upper right plot, muon pairs failing the requirement are fitted, distinguishing between signal and background.

The uncertainty on the fitted efficiency is determined from the profile likelihood function. Since the parameters were left floating, the uncertainty includes naturally the contributions from the background subtraction procedure.

Eventually, the scale factors are calculated from the ratio of the efficiencies estimated on data and Monte Carlo simulation:

$$SF = \frac{\epsilon_{data}}{\epsilon_{MC}}. \quad (3.4)$$

The uncertainty on this ratio (σ_{SF}) is computed by propagating data and MC uncertainties derived from the efficiency calculation. The final value includes:

$$\sigma_{SF}^2 = \sigma_{stats+fit}^2 + \sigma_{syst}^2, \quad (3.5)$$

where $\sigma_{stats+fit}$ refers to the statistical plus fit uncertainty given by the fit result. The σ_{syst} is the intrinsic uncertainty from the *Tag and Probe* method. It is evaluated from the estimation of the efficiencies in each bin with two different methods: the *counting* and the *fitting* method. The fitting method as been described above. The counting method uses template shapes from simulation to fit the data and determine efficiencies, while MC efficiencies are extracted directly by counting the events for the passing and failing probe categories.

These systematic uncertainties have been calculated for muons in [136]. The values are conservatively taken to be 0.2% in the barrel and 0.4% in the endcaps, for muon isolation and identification. For electrons the uncertainty has been recomputed bin-by-bin as the difference between fitting and counting method as described in [106]. The results are summarised below. The complete tables of results can be found in Appendix B. When the uncertainties are specified they include the systematic contribution.

- **Scale factors for muons**

Reconstruction and ID. Efficiencies computed for the muon reconstruction and identification on data and simulation are evaluated in different η regions and in the p_T range [20–200] GeV. Tight cuts on the muon tag are requested, asking for a matching with a single muon trigger and also for a very high reconstruction quality following the criteria of a PF *tight* muon, detailed in Section 3.2.2.

Probes are all the muons reconstructed with any of the algorithms in the event. Reconstruction and identification criteria used for selecting the passing probes match also the ones of a PF *tight* muon from Section 3.2.2 with no further

requirements. The reconstruction and identification are entangled processes for which a common efficiency is attributed. The resulting scale factors between data and simulation range between 0.97 and 1.0 depending on the η range, with an uncertainty smaller than 1%.

Isolation. Efficiencies for muon isolation cuts have been computed requiring the passing probe to satisfy the criteria adopted for isolated leptons in this analysis as described in Section 3.2.2. Probes are now Particle Flow *tight* muons. Results for isolation scale factors are comprised between 0.99 and 1.0 depending on the η value, with uncertainties of 1% maximum.

- **Scales factors for electrons**

Reconstruction. Efficiencies computed for the electron reconstruction on data and simulation are evaluated in different η regions and different p_T ranges. Electrons are selected only in the ECAL acceptance, and therefore rejected in the gap for $1.4442 < |\eta| < 1.566$.

Tight cuts on the tag are requested, asking for a matching with an unrescaled single electron trigger and for very high quality criteria as described in Section 3.2.2. Probes are all the tracks originated from ECAL SC energy in the event and the reconstruction criteria only require them to be reconstructed as electrons. Results on data and simulation are reported in Appendix B for both fitting and counting methods. The scale factors are very close to 1.0 for all the ranges of p_T and η considered and the uncertainties never exceed 2%.

ID and Isolation. Efficiencies for isolation have been computed together with identification, by requiring the passing probe to satisfy most of the electron selection cuts of the analysis: the ‘medium working point’ type quality cuts, as detailed in Table 3.4. Results are found to range from 0.90 and 1.0, with an uncertainty varying from 0.1% to 10% for different p_T and η regions.

- **Trigger efficiencies for leptons**

The trigger efficiency is the probability that a lepton, selected with the analysis criteria, fires up one of the triggers imposed in the analysis. The tag is selected with tight requirements as previously, while the probe is kept unbiased. Then the probe is required to geometrically match with the offline trigger object within a cone of $\Delta R = 0.5$, and an efficiency is extracted. The matching inefficiency is of the order of one permil and therefore negligible. In order not to bias the *Tag and Probe* sample, the tag should match a different trigger than the one under study with the probe. Verifications have been performed to be

sure that a different threshold used in the trigger matched to the tag, whether it is a single or double trigger, does not affect the probe observables.

The trigger is not simulated in this study, therefore the efficiency measured on data is directly applied on all simulated samples.

For muons, the efficiencies of the four different triggers, covering the different data taking periods, have been evaluated. Two triggers being symmetric and two asymmetric in their thresholds. For the electrons two unprescaled asymmetric triggers efficiencies have been considered. The efficiency for the trigger paths has been computed for the single leg, requiring matching with trigger filters, and has been found to vary as a function of η of the lepton (also for muons). Therefore efficiencies for distinct regions have been evaluated. For muons, the efficiencies on the single leg range from 0.92 to 0.97 depending on the trigger considered and the range in p_T and η ; while for the electrons the efficiency is found to be very close to 1.0 within the entire ECAL fiducial region.

For combining the different legs in a per-event scale factor, the formula for a generic trigger having asymmetrical thresholds, labelled *high* and *low*, has been used:

$$\epsilon^{asym.trigger} = \epsilon_{lepton1}^{high} \cdot \epsilon_{lepton2}^{low} + \epsilon_{lepton1}^{low} \cdot \epsilon_{lepton2}^{high} - \epsilon_{lepton1}^{high} \cdot \epsilon_{lepton2}^{high} \quad (3.6)$$

This simply turns into the product of the single leg efficiency in the case of symmetric trigger. The final trigger efficiency is obtained by averaging over all triggers used, weighted by the corresponding integrated luminosity covered by each trigger. The agreement between data and reweighted MC simulation (both by pileup and lepton scale factors), after the selection of a Z in the event, can be seen in Figure 3.1.

Finally, a study on the dependence of the lepton scale factor from jet multiplicity in the event has been performed. By recomputing this scale factor in exclusive bin of multiplicity, the maximum relative difference is found to be less than 1% in the barrel region and up to 4% in the endcap region. This will enter in the systematic uncertainties for the lepton in the final computation of the results.

3.2.6.3 b-tagging Scale Factors

As already mentioned in Section 3.2.3 the b-tagging efficiencies differ between data and simulation, which calls for scale factors to correct for this difference. The weight assigned to an event will depend on the number of b's present and the way they are reconstructed. For one event several combinations are possible

depending on the number of real b's in the event at the generator level, the number of jets reconstructed in the event and the number of tags required.

If one considers an example with one generator level b, two reconstructed jets and the requirement of at least one b-tag in the final state: several combinations can happen. The probabilities for the different sources are:

- having the real b-jet tagged, and the light jet not tagged, i.e.
 $P = \varepsilon_b \times (1 - \varepsilon_l);$
- having the real b-jet mistagged and the light jet tagged, i.e.
 $P = (1 - \varepsilon_b) \times \varepsilon_l;$
- having the real b and the light jet both tagged, i.e.
 $P = \varepsilon_b \times \varepsilon_l.$

If the requirement would have been exclusive, so exactly one tag, the last probability would not have been considered. A generalised expression for one combination probability for the MC is given in Equation 3.7 with i being the flavour: b, c or l.

$$P_{MC} = \prod_{j=1}^{N_{tag}} \varepsilon_i(p_{Tj}, \eta_j) \times \prod_{k=1}^{N_{non-tag}} (1 - \varepsilon_i(p_{Tk}, \eta_k)) \quad (3.7)$$

So the total probability to observe the required final state in the MC is the sum over all the combinations of those probabilities. But each efficiency can be multiplied by the appropriate scale factor (SF) of each flavour and will give the data probability in Equation 3.8.

$$P_{data} = \prod_{j=1}^{N_{tag}} \varepsilon_i(p_{Tj}, \eta_j) \times SF_i(p_{Tj}, \eta_j) \times \prod_{k=1}^{N_{non-tag}} (1 - \varepsilon_i(p_{Tk}, \eta_k) \times SF_i(p_{Tk}, \eta_k)) \quad (3.8)$$

The per-combination weight is obtained for each acceptable combination by comparing the scaled and the uncorrected probability.

The total per-event weight is the fraction of the total efficiency for data and of the total efficiency for simulation; so the sum over all possible combinations

for data divided by the sum of all the possible combinations for simulation:

$$w_{\text{event}} = \frac{\sum_{\text{comb}} \left[\prod_{j=1}^{N_{\text{tag}}} \varepsilon_i(p_{Tj}, \eta_j) \times SF_i(p_{Tj}, \eta_j) \times \prod_{k=1}^{N_{\text{non-tag}}} (1 - \varepsilon_i(p_{Tk}, \eta_k) \times SF_i(p_{Tk}, \eta_k)) \right]}{\sum_{\text{comb}} \left[\prod_{j=1}^{N_{\text{tag}}} \varepsilon_i(p_{Tj}, \eta_j) \times \prod_{k=1}^{N_{\text{non-tag}}} (1 - \varepsilon_i(p_{Tk}, \eta_k)) \right]} \quad (3.9)$$

In the Formula 3.9 the values of the scale factor data/MC for b-tagging and light-tagging, as well as the light-tagging efficiency from simulation ε_l are taken from [128], for the b-tagging algorithm and version used in this analysis, namely the SSVHE. The c-tagging scale factor is assumed to be the same as the b-tagging scale factor. As ratios of efficiencies the scale factors do not depend on any selection in particular, and have been estimated by different methods which are extensions of the *Tag and Probe* technique detailed in Section 3.2.6.2. The b-tagging and c-tagging efficiencies are evaluated from the Drell-Yan MC simulation used in the analysis. The result versus the p_T of the leading tagged jet and in two different η regions can be seen in Figure 3.8. In the case of the b-tagging efficiencies they are compared to the results obtained from a different sample, QCD multijets, in [128].

Results for b-tagging and c-tagging efficiencies and scale factors are given in the Table 3.8, averaged on all the p_T bins. The values of $\varepsilon_l(p_{T,j}\eta_j)$ and $SF_l(p_{T,j}\eta_j)$ are given with a fine granularity allowed by the numerous simulated events available for this kind of light jet selection in the QCD sample in [128]. The maximal value for ε_l is of the order of 2% and the SF_l range between 0.89 and 0.97.

Table 3.8: *Simulation efficiencies (from Z+b/c simulated events) and scale factors for b and c jets, for the SSVHE tagger in two η regions of the jet.*

Tagger	SSVHE		
Bin	ε_b^{MC}	ε_c^{MC}	$SF_{b,c}$
$0 < \eta < 1.2$	0.571	0.167	0.95 ± 0.095
$1.2 < \eta < 2.4$	0.483	0.127	0.93 ± 0.093

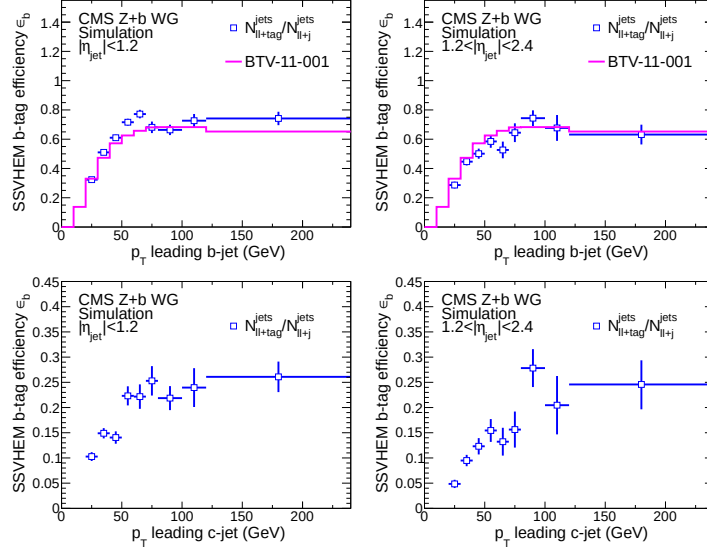


Figure 3.8: SSVHE MC efficiencies. Top: ϵ_b^{MC} as a function of p_T of the jet in the central $|\eta| < 1.2$ (left) and forward $1.2 < |\eta| < 2.4$ (right) regions. Bottom: ϵ_c^{MC} as a function of the p_T of the jet in the central $|\eta| < 1.2$ (left) and forward $1.2 < |\eta| < 2.4$ (right) regions. Markers denote the values computed using the Z+b/c MC sample, overlaid is the pink line corresponding to the ϵ_b^{MC} of the QCD sample used in [128].

3.2.7 Data and Simulation Yields Comparison

The complete table of yields for the different steps of the selection, and also for the high purity (HP) version of the SSV b-tagging algorithm can be found in Appendix A. The yields for the working points used for the cross section calculations, Z+1b and Z+2b selections are summarised in Table 3.9.

Table 3.9: Data yields in the selected samples and a comparison to the expectations based on MC simulations. The expected yields are estimated using the theoretical predictions for the cross sections. Uncertainties are statistical only.

Selection	Data	Total MC	Z+b	Z+c	Z+l	$t\bar{t}$	ZZ
$Z_{\mu\mu}+1b$	13 090	$12\,904 \pm 77$	6810 ± 58	3647 ± 41	1829 ± 29	549 ± 3	69 ± 1
$Z_{\mu\mu}+2b$	522	480 ± 13	350 ± 12	34 ± 4	5 ± 1	80 ± 1	11 ± 1
$Z_{ee}+1b$	9672	9924 ± 67	5218 ± 50	2844 ± 36	1364 ± 25	445 ± 3	53 ± 1
$Z_{ee}+2b$	362	357 ± 11	258 ± 10	27 ± 3	2 ± 1	62 ± 1	8 ± 1

3.2.8 Data and MC Comparison of Kinematic Variables

In Figure 3.9 several important variables distributions used in the analysis are shown for the Z+1b exclusive selection. The electron and the muon channels are combined. The small difference in luminosity for the two channels sample is taken into account in the combination. The resulting distribution will be said to belong to the combined channel. The data corresponding to 5.0 fb^{-1} are overlaid with the different MC simulations normalised to their theoretical cross sections and reweighted by the pileup, the leptons and the b-tagging scale factors as detailed above. The last bin in both data and MC distributions contains the overflow. A good agreement between data and simulation is found for the several distributions, namely: the transverse momentum and the mass of the Z boson (top, left and middle), the transverse momentum of the leading b-jet (bottom left), and the angular differences between the Z boson and the leading b jet in ΔR and $\Delta\phi$ (right, top and bottom). The disagreement between data and simulation observed in the η distribution of the leading b jet (bottom middle), is explained by the scheme used for the signal simulation from the Drell–Yan sample: produced with the 5F scheme for which the b’s are massless, when the b’s acquire their mass during the parton shower a shift in η takes place to fit the mass with the p_T . But around the central region, the value of η cannot be made smaller than zero, and simulated events are accumulated, creating an unphysical peak not present in the data.

For the Z+2b selection, the distributions of the interesting kinematic variables are shown and discussed in detail in Section 4.1. The agreement between data and simulation is overall good, confirming the confidence in the analysis selection.

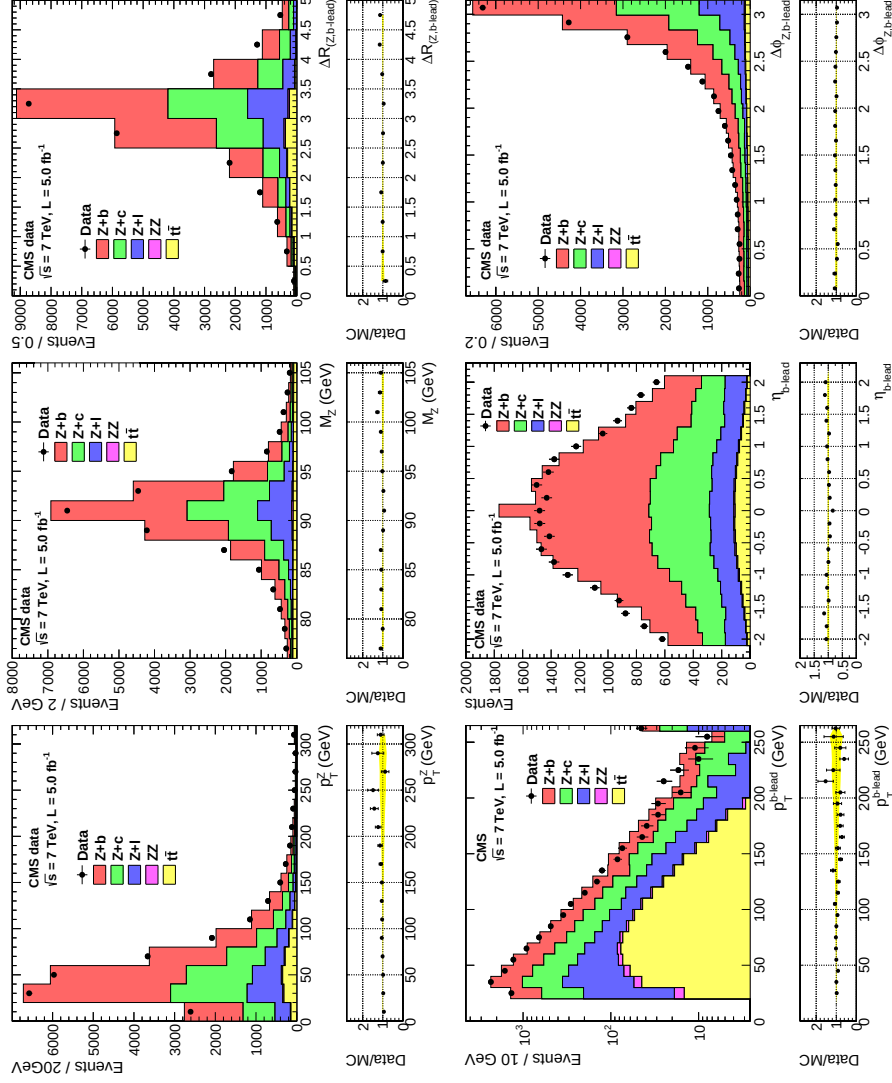


Figure 3.9: Data/MC comparison after the Z+1b selection in the combined channel. The yellow bands in the lower plots represent the statistical uncertainty on the MC yield.

3.3 Backgrounds

For the reconstruction of a Z in association with exactly one b-tagged jet (the Z+1b selection), the main background originates from Z+light-jets, whereas the $t\bar{t}$ background is relatively small (approximately 5%), and the diboson background is negligible.

For the Z+2b selection the contamination due to light jets is reduced, as proportional to the b-tagging efficiency squared, and hence the use of the b-tagging algorithm SSVHE on both jets suffices to suppress most of the background events originating from light jets. The signal simulation for the Z+2b selection is also contained in the so-called Z+b simulation sample from Drell–Yan plus jets and will be called in the text Z+b+b. On the contrary the $t\bar{t}$ and diboson backgrounds increase due to the presence of two real b jets in these processes. However, b jets from $t\bar{t}$ and ZZ production have different kinematics than b jets originating from Z+b+b production: in general the p_T spectrum of b jets originating from a heavy object is harder, the angular distribution is more central, and the kinematics are more correlated. Hence, the chance to reconstruct a second b jet in the case of $t\bar{t}$ and ZZ is larger than for Z+b+b production, since the b-tagging efficiency increases with the p_T (see Figure 3.8), and the contributions of these background sources increases. As already explained in Section 3.2.4, in order to suppress the dominating background from $t\bar{t}$, an extra selection criterion on the MET significance is employed. It makes the selection of a Z, two b-tagged jets with the SSVHE tagger and the application of a cut at 10 u.s on the MET significance the baseline for the Z+2b selection.

For both the Z+1b and Z+2b selections of the analysis a certain fraction of Z+lighter-jets (Z+l and Z+c) and of $t\bar{t}$ remains and has to be estimated, if possible without relying on MC simulations.

3.3.1 Background due to $t\bar{t}$ production

In the two selection working points, the background contribution from $t\bar{t}$ production is estimated by fitting Drell–Yan plus jets and $t\bar{t}$ templates to the dilepton invariant mass spectrum. These templates can be derived in three different ways:

- **From data:**

The DY template is taken from the dilepton invariant mass, $M_{\ell\ell}$ peak, after Z+jet selection, including a criterion on the MET: a value below 10 for the MET significance, to suppress contributions from $t\bar{t}$ in the DY

template. The $t\bar{t}$ template is built from an opposite flavour selection ($e\mu$) in a dilepton sample in which the $t\bar{t}$ contribution is enriched.

- **From simulation:**

For both DY and $t\bar{t}$, simulated events are selected with the same criteria as the data for that working point, and the template is parametrised as a binned probability density function (pdf). Since the resolution of the DY peak in MC is underestimated due to a misalignment in simulation, the DY peak is convolved with a Gaussian.

- **From ‘physics’:**

The DY signal is analytically modelled by a relativistic Breit–Wigner distribution which characterises the resonances in physics, convolved with a Gaussian which represents the smearing by the detector.

The $t\bar{t}$ background is modelled by a second order polynomial. This does not have a strong physics motivation but is analytically the best description of this background.

To perform the fit, if one considers all the possible combinations of these DY and $t\bar{t}$ templates, there could be up to nine methods. Out of the possible combinations, the following three representative options are used to extract the $t\bar{t}$ fraction and its uncertainty:

- **Data DY-MC $t\bar{t}$:** a data-derived pdf for DY and a simulation-derived $t\bar{t}$ template.
- **Physics:** a physics inspired pdf for DY and an analytic approach for $t\bar{t}$.
- **All MC:** simulation-based templates for both DY and $t\bar{t}$.

In order to improve the sensitivity to fit the background in the data, the mass window for the fit is widened by a factor two: $61 < M_{\ell\ell} < 121$ GeV. The result, from an extended likelihood fit, is subsequently interpolated to the signal region, $76 < M_{\ell\ell} < 106$ GeV, using the pdf shapes to estimate the expected fraction of signal and background events in the signal region.

The results of these three different fits for the muon channel are shown in Figure 3.10 for the $Z+2b$ selection.

In the extended fit the number of events is fitted. In the case of physics-inspired pdf, as well as when including a resolution function, all parameters of the fit are floating, in order to propagate the uncertainty due to the modelling to the estimate of the $t\bar{t}$ fraction. This means that uncertainties due to modelling of the polynomial distribution are automatically propagated to the estimate.

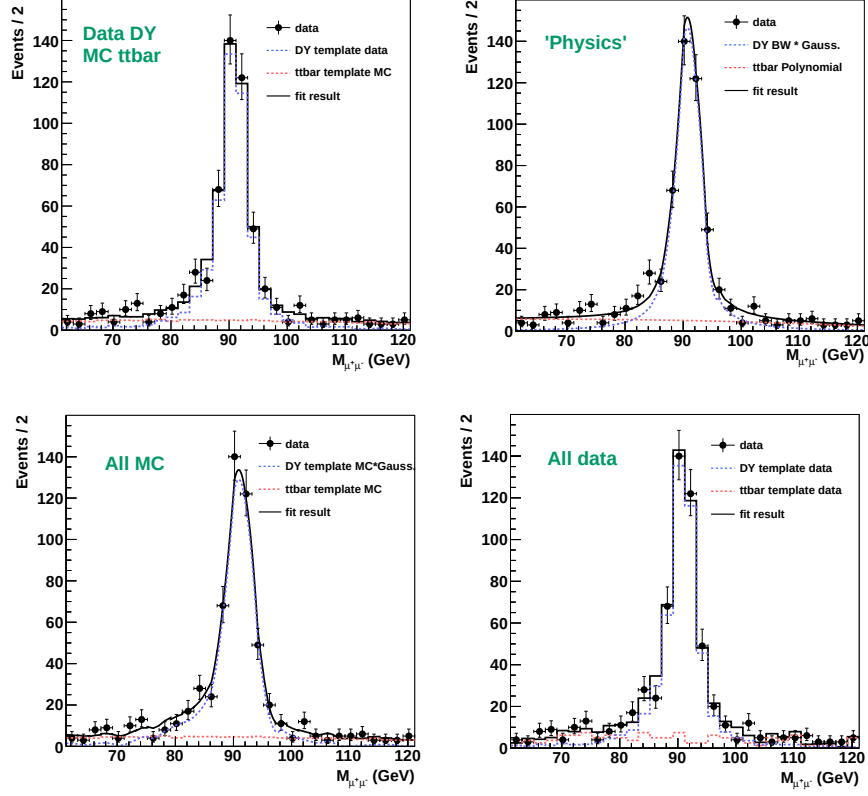


Figure 3.10: Fit of $t\bar{t}$ fraction to $M_{\ell\ell}$ distributions for the dimuon sample after $Z+2b$ selection, for the DY data-derived and $t\bar{t}$ MC-derived parametrisation (top left), for the physics-motivated parametrisation (top right) for the simulation-derived parametrisation (bottom left) and for the data-derived parametrisation (bottom right).

Another possible combination is the completely data-driven approach with both DY and $t\bar{t}$ templates extracted from data. However, the $e\mu$ data sample from which the $t\bar{t}$ template is derived has a small number of events and the fluctuations in the data sample are sizeable as visible on the bottom right plot of Figure 3.10. Beside, a different trigger is used, which introduces extra systematic uncertainties. Nevertheless, this combination, called the ‘All data’ parametrisation is used as a cross check.

Final plots for the distribution of the mass of the dimuon (dielectron) pair in the wide mass window, after selecting two leptons and two b-tagged jets, and the requirement on the MET significance variable, overlaid with the templates after a fit for the $t\bar{t}$ fraction with the ‘Data DY-MC $t\bar{t}$ ’ parametrisation, are shown in Figure 3.11 left (right).

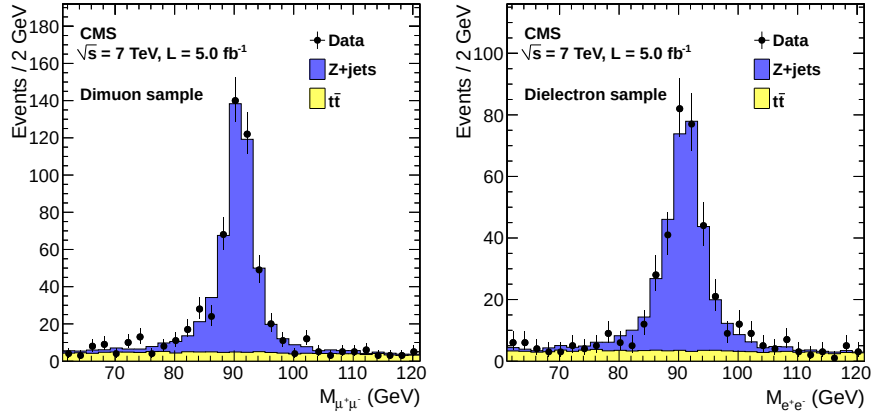


Figure 3.11: *Fit of the $t\bar{t}$ fraction to $M_{\ell\ell}$ distributions for the dimuon sample (left) and the dielectron sample (right) after the Z+2b selection, for the ‘Data DY-MC $t\bar{t}$ ’ parametrisation*

Interpolating the result to the analysis narrow mass window.

Finally, the fraction of $t\bar{t}$ is extracted by interpolation from the wide mass window to the signal region ($76 < M_{\ell\ell} < 106 \text{ GeV}$), and the result is given with its uncertainties, in Table 3.10 for all the considered combinations. The fully data-derived parametrisation cross check result in the muon channel, as well as the expectations from the MC yields from the Table 3.9 are also given and found compatible with the other combinations results. The comparison shows a slightly lower $t\bar{t}$ estimate in the Z+2b sample than what is expected from the MC yields. The estimated contamination due to $t\bar{t}$ in the electron sample is slightly larger than in the muon channel, which is in agreement with the simulation, although the difference is covered by the uncertainties. For the Z+2b selection the contamination is slightly lower, which is partly due to a larger DY yield, and partly due to a smaller $t\bar{t}$ yield.

Table 3.10: *Estimates of $t\bar{t}$ background, in the narrow mass window, using the 5.0fb^{-1} data sample. Results are estimated for the selection working points, per channel, for the different parametrisations. A comparison is made with the expectations from MC. The $t\bar{t}$ fraction in simulation is defined as $f_{t\bar{t}}^{MC} \equiv \frac{N_{t\bar{t}}}{N_{TotMC}}$, using the numbers of the data/MC yields comparison in Table 3.9.*

WP	$f_{t\bar{t}}^{dataDY}\%$	$f_{t\bar{t}}^{Phys}\%$	$f_{t\bar{t}}^{AllMC}\%$	$f_{t\bar{t}}^{e\mu}\%$	MC-exp %
Z($\mu\mu$)+1b	5.2 ± 0.3	6.1 ± 0.3	5.2 ± 0.3	5.2 ± 0.7	$549/13090 = 4.2$
Z(ee)+1b	5.0 ± 0.3	4.2 ± 0.5	5.5 ± 0.3	–	$445/9672 = 4.6$
Z($\mu\mu$)+2b	13.0 ± 1.8	13.9 ± 2.2	12.8 ± 1.8	13.0 ± 1.9	$80/480 = 16.7$
Z(ee)+2b	14.0 ± 2.3	14.2 ± 3.7	16.1 ± 2.4	–	$62/357 = 17.4$

The parametrisation ‘Data DY-MC $t\bar{t}$ ’, that combines data-driven template for DY and a simulation-based $t\bar{t}$ template, is chosen as the central value. The systematic uncertainty is derived from the combination of the results from ‘Data DY-MC $t\bar{t}$ ’, ‘Physics’ and ‘All MC’ parametrisations. The $t\bar{t}$ contamination, for the selection working points and in the default narrow mass window, is then summarised in Table 3.11.

Table 3.11: *Estimates of $t\bar{t}$ fraction, in the narrow analysis mass window $[76-106]\text{GeV}$, for the 5.0fb^{-1} data sample used in the analysis. The first uncertainty is statistical, and the second one is the systematic uncertainty related to the combination of the different results.*

WP	Muon Channel	Electron Channel
Z+1b:	$f_{t\bar{t}}^{1b}(\mu\mu) = (5.2 \pm 0.3 \pm 0.6)\%$	$f_{t\bar{t}}^{1b}(ee) = (5.0 \pm 0.3 \pm 0.6)\%$
Z+2b:	$f_{t\bar{t}}^{2b}(\mu\mu) = (13.0 \pm 1.8 \pm 0.6)\%$	$f_{t\bar{t}}^{2b}(ee) = (14.0 \pm 2.2 \pm 0.6)\%$

3.3.2 b-purity

After having reduced and estimated the background due to $t\bar{t}$ production, the remaining background due to lighter jets is estimated. For both the Z+1b cross section measurement [104], and the Z+2b measurement [103] the b-purity is estimated using the distributions of secondary vertex mass, $m(SV)$, from b-tagged jets, to extract the purity from the data. Templates are built from MC simulation by requiring the presence of a Z and a jet, and matching the flavour of the reconstructed jet to the required parton flavour.

Subsequently, after estimating the per-jet fractions $f_b^{1,2}$, this is transformed into a per-event purity $P_b^{1,2b}$, taking into account double b-tags in a single event. This is done for both the Z+1b selection as well as for the Z+2b selection.

3.3.2.1 Strategy: Per-jet Purity vs. Per-event Purity

Two definitions have to be taken into account in the following: the jet purity or jet fraction and the event purity.

The jet fraction f_i is the probability that a b-tagged jet is coming from a real i parton, where $i \in \{b, c, l\}$ with l being indifferently u, d, s or g . For Z+1b case the jet fraction is equal to the event purity: $P_b^{1b} = f_b$.

For Z+2b, the purity of true Z+b+b events differs from the fraction of correctly tagged b jets in a sample. The jet fraction f_{ij} is the probability that 2 b-tagged jets are coming from a real i parton and a real j parton, where indices refer to the true flavours of the two tagged jets: $i, j \in \{b, c, l\}$. The event purity P_b^{2b} is written as a function of the mistag fractions:

$$P_b^{2b} = 1 - f_{cc} - f_{ll} - f_{cb} - f_{bc} - f_{cl} - f_{lc} - f_{lb} - f_{bl}. \quad (3.10)$$

In other words: double mistags need to be included only once (as the mistagged jet is accompanied by another also mistagged jet), whereas single mistags need to be included twice, (as the mistagged jet is accompanied by a true b jet).

Subsequently, the fraction of events from singly mistagging Z+1/c+b needs to be migrated to the Z+1b signal estimation.

The size of the various contributions depends on the tagging efficiencies of lighter jets, as well as on the corresponding cross sections of the different Z+i+j processes. As mentioned in Section 3.2.6.3 and as can be seen in Figure 3.12 from [137], for the b-tagger used, the SSVHE, the efficiency of tagging a real c parton as a b jet, is about 10%, and for a real light parton it is about 1%.

Hence for the Z+2b selection, with two jets b-tagged with the SSVHE, the largest contribution of mistagged jets is expected to originate from mistagging two c's from Z+cc production, and from singly mistagging Z+lb.

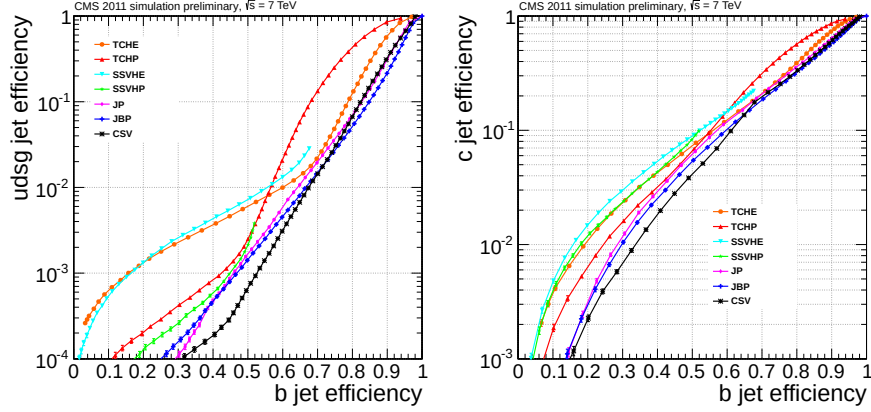


Figure 3.12: Performance curves obtained from simulation for different b-tagging algorithms described in [137]: light flavours (left) and c flavour (right) efficiencies as a function of the b efficiency.

3.3.2.2 Estimation of Jet Fractions with Templates Fit

The templates used to describe the different flavour contributions are extracted from QCD multijets simulations. Since the p_T spectrum of the inclusive jets differs from the one of jets in the selected Z+b-jets sample, and since the $m(SV)$ distributions change as a function of p_T , these templates are reweighed according to the p_T spectrum of the corresponding data sample.

Still due to the dependence of the $m(SV)$ distributions on the p_T of the jet, the templates are estimated separately for the leading and the subleading jets, since the p_T spectrum of the second jet is softer. This can also be observed in Figure 3.14: the $m(SV)$ templates of the leading b-tagged jet peak at higher masses than those corresponding to the subleading b-tagged jet. The templates are then created by selecting the jets according to the matching to a parton flavour: b, c or l. The fractions obtained separately for the leading and the subleading jet are then the per-jet fractions of each flavour f_i with $i \in \{b, c, l\}$.

The fits for the two different working points of the analysis, using the templates extracted from QCD MC sample, are shown in Figure 3.13 for Z+1b, and 3.14 for Z+2b. After the fit, the data shows an excellent agreement with the stack

of the simulations, over the full $m(\text{SV})$ range. One can observe, especially in Figure 3.13 which has a larger number of events:

- The peak related to the mass of the kaon, around 0.5 GeV, from the decay into two pions;
- The edge related to the D meson, around 2.0 GeV;
- The slope of the region pure in B meson decays, between 2 and 5 GeV.

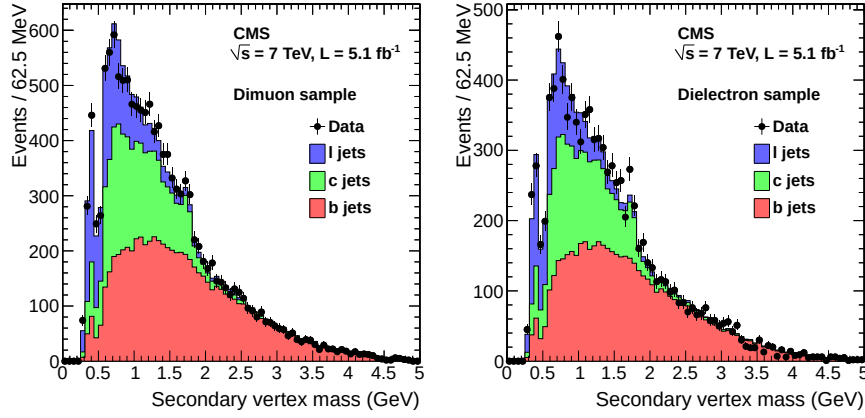


Figure 3.13: Secondary vertex mass distributions after the $Z+1b$ selection, for the muon (left) and electron (right) channels. Results are shown after a fit for the b fraction, overlaying the templates extracted from the QCD MC sample.

The results, for the jet-mistag fractions estimated from data, for the two selection working points, are summarised in Table 3.12. The results are acquired for an extended fit, which fits the number of events N_b, N_c, N_l together; as well as for a fit of the fractions, which fits separately the fractions f_b, f_c and the fractions f_c, f_l .

The b -jet signal estimates are shown in Table 3.13. As can be seen from the table, the purities for the sample of leading and the sample of subleading jets are consistent.

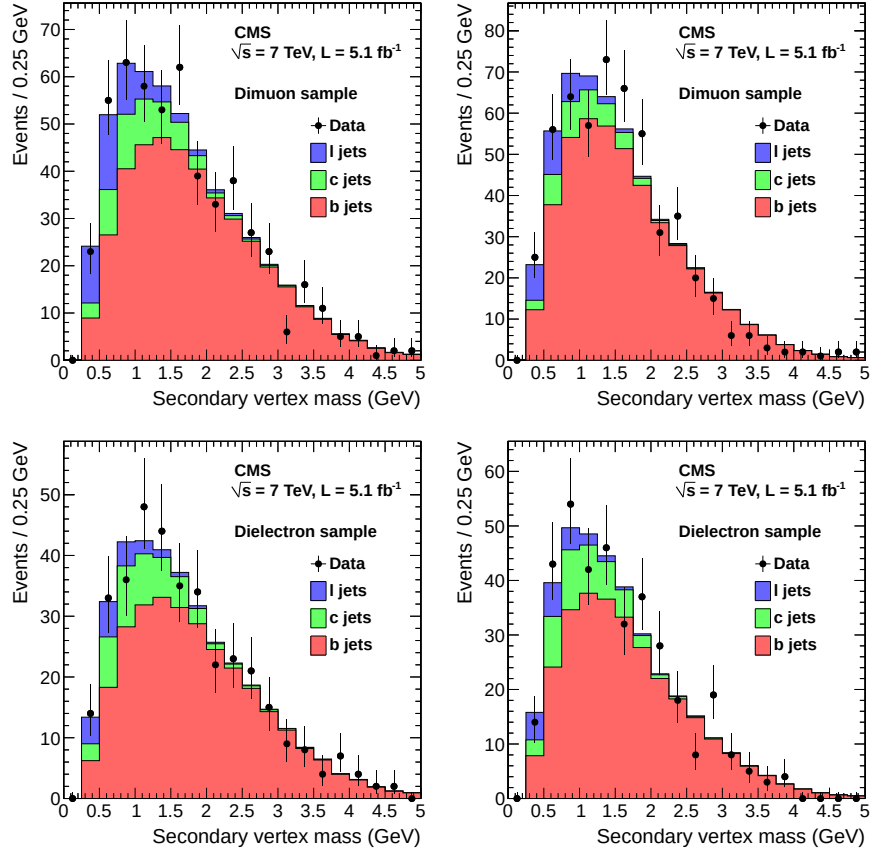


Figure 3.14: Secondary vertex mass distributions after the $Z+2b$ selection, for the leading (left) and subleading (right) b -tagged jets, for the muon (top) and electron (bottom) channel. Results are shown after a fit for the b fraction, overlaying the templates extracted from the QCD MC sample.

Table 3.12: *The per-jet fractions for the different working points after fits to the secondary vertex masses $m(SV_{1,2})$, for the 5.0fb^{-1} data sample.*

Selection	N_b	N_c	N_l	$f_c(\%)$	$f_l(\%)$
Z($\mu\mu$)+1b	7036 ± 157	3683 ± 214	2372 ± 130	28.1 ± 1.6	18.1 ± 1.0
Z(ee)+1b	5327 ± 136	2761 ± 184	1584 ± 110	28.5 ± 1.9	16.3 ± 1.1
Z($\mu\mu$)+2b-jet1	415 ± 34	54 ± 44	53 ± 25	10.2 ± 8.4	10.1 ± 4.9
Z(ee)+2b-jet1	295 ± 29	47 ± 36	20 ± 19	12.9 ± 9.8	5.4 ± 5.3
Z($\mu\mu$)+2b-jet2	451 ± 38	38 ± 44	33 ± 21	7.2 ± 8.4	6.4 ± 4.0
Z(ee)+2b-jet2	294 ± 30	48 ± 36	20 ± 19	13.3 ± 10.0	5.5 ± 5.3

Table 3.13: *The per-jet b-fractions for the different working points after fits to the secondary vertex masses $m(SV_{1,2})$.*

Selection	f_b^1 (data fit) %	f_b^2 (data fit) %
Z($\mu\mu$)+1b	53.7 ± 1.1	-
Z(ee)+1b	55.0 ± 1.3	-
Z($\mu\mu$)+2b	80.0 ± 5.9	85.7 ± 5.9
Z(ee)+2b	80.5 ± 6.2	83.6 ± 6.6

3.3.2.3 Expected b-jet Content and Comparison with Data

In order to know what is the expected content of truly b-flavoured jets and mistagged jets, the MC simulations used in the analysis are studied for both the Z+1b and the Z+2b selections. From the different samples a certain number of events are passing the selection working points, and the MC information allows to recover the flavour(s) of the b-tagged jet(s) in the event.

Z+1b

The expected jet fractions from the Drell–Yan MC sample, which has been divided according to the flavour of the parton present at generator level, for the Z+1b selection are summarised in Table 3.14. It is shown that 96% of the event passing the selection with one b-tagged jet are actually coming from the sample labelled ‘Zb’ (in which a b is produced at generator level).

The same flavour content study is performed for all the MC samples, this time with the Drell–Yan sample being considered unsplit, and the results are presented in Table 3.15. For the Z+1b selection 55.3% of the events are expected to have a b-tagged jet in the final state which is a truly b-flavoured jet at

generator level. This value is directly comparable to the b-tagging efficiency of 55% for the SSVHE algorithm.

Table 3.14: *Composition of the reconstructed MC samples for the Z+1b working point, ordered according to the generated flavour of b-tagged jets. The fraction is defined for each of the ‘Zi’ samples separately, with $i \in \{b, c, l\}$.*

Sample	f_{b1}	f_{c1}	f_{l1}
Zb	96%	2%	2%
Zc	-	82%	18%
Zl	-	-	100%

Table 3.15: *Number of jets tagged in simulation samples for the Z+1b working point, ordered according to true flavour, for the three samples. The fraction of jets is defined with respect to the total sum of the three MC samples.*

Z+1b selection	DY	$t\bar{t}$	ZZ	Total	Fraction
Z+b	6823.7	547.8	51.5	7423.0	55.3%
Z+c	3225.0	1.6	16.7	3243.3	24.2%
Z+l	2746.4	3.1	6.7	2756.2	20.5%

Z+2b

The expected jet fractions for the Z+2b selection in the Zb, Zc and Zl samples from the Drell–Yan MC sample are summarised in Table 3.16.

It is shown that 88% of the ‘Zb’ sample consists of correctly tagged Z+b+b. A sizable contribution of 12% originates from events in which at least one of the jets is mistagged, roughly equally distributed over Z+b+c and Z+b+l. Of the ‘Zc’ sample, about 2/3 of the events are mistagged Z+c+c, whereas 1/4 of the events are mistagged Z+c+l.

Table 3.16: *Composition of the MC samples for the Z+2b selection, ordered according to truly generated flavour of the b-tagged jets. The fraction is defined for each of the ‘Zi’ samples separately, with $i \in \{b, c, l\}$.*

Sample	f_{bb}	$f_{bc} + f_{cb}$	$f_{bl} + f_{lb}$	f_{cc}	$f_{cl} + f_{lc}$	f_{ll}
Zb	88%	5%	7%	< 1%	< 1%	< 1%
Zc	-	-	-	68%	26%	6%
Zl	-	-	-	-	-	100%

Jets fractions for all the analysis MC simulations are given in Table 3.17, from which it is expected that 82.1% of the events in the simulated Z+2b sample have two truly b-flavoured jets in the final state. One can also notice that both, the contributions from single mistags, and double mistags, are non-negligible: the expected fraction of single mistags is 9.0% and the fraction of double mistags is 8.9%. This gives an expected ratio of ‘double-mistags’/‘single-mistags’ ~ 1 .

It should be noted that the singly mistagged Z+b+c/l events are considered as Z+1b signal.

Table 3.17: *Number of jets tagged in simulations for the Z+2b selection, ordered according to true flavour, for the three MC samples. The fraction of jets is defined with respect to the sum of the three MC samples.*

Sample	DY	$t\bar{t}$	ZZ	Total	Fraction
Z+b+b	321.6	79.3	10.6	411.5	82.1%
Z+b+c	8.5	0.5	0.1	9.1	1.8%
Z+c+b	9.7	0.4	0.1	10.2	2.0%
Z+b+l	10.8	1.2	0.2	12.2	2.4%
Z+l+b	13.1	0.9	0.1	14.1	2.8%
Z+c+c	25.1	-	0.8	25.9	5.2%
Z+l+l	7.8	-	-	7.8	1.6%
Z+c+l	4.9	-	0.1	5.0	1.0%
Z+l+c	5.6	-	-	5.6	1.1%

Comparison between data and MC

A comparison of the results expected from simulation and from the data fits is given in Table 3.18 for the b-jet fraction, and in Table 3.19 for the mistag fractions. It has to be taken into account that the fit for data has been performed separately per channel, while the MC expectation has been extracted from the combined channel.

In Table 3.18 it is shown that the fraction of b jets estimated in the Z+1b sample is, within uncertainties, roughly in agreement with the expectation. The same holds for the Z+2b sample where in general the fraction of b jets is slightly lower than expected.

In the comparison Table 3.19 of the c/l fractions per jet, as found in data with the expectations from simulation, one can note that the expected fraction of mistags is the same for the samples of leading and subleading jets, despite the known increase of mistag probability at lower p_T .

Table 3.18: *The per-jet b fractions for the different selection working points after fits to the secondary vertex masses $m(SV_{1,2})$, compared to MC expectations in the combined channel for illustrative purpose.*

WP	f_b (data fit) %	f_b (MC) %
Z($\mu\mu$)+1b-jet	53.7 ± 1.1	55.3
Z(ee)+1b-jet	55.0 ± 1.3	
Z($\mu\mu$)+2b-jet1	80.0 ± 5.9	86.3
Z(ee)+2b-jet1	80.5 ± 6.2	
Z($\mu\mu$)+2b-jet2	85.7 ± 5.9	86.9
Z(ee)+2b-jet2	83.6 ± 6.6	

Table 3.19: *The per-jet mistag fractions for the different selection working points after fits to the secondary vertex masses $m(SV_{1,2})$, compared to MC expectations in the combined channel for illustrative purpose.*

WP	f_c (data fit) %	f_l (data fit) %	f_c (MC) %	f_l (MC) %
Z($\mu\mu$)+1b-jet1	28.1 ± 1.6	18.1 ± 1.0	24.2	20.5
Z(ee)+1b-jet1	28.5 ± 1.9	16.3 ± 1.1		
Z($\mu\mu$)+2b-jet1	9.3 ± 87.4	10.7 ± 4.0	8.2	5.5
Z(ee)+2b-jet1	14.5 ± 8.5	5.0 ± 4.5		
Z($\mu\mu$)+2b-jet2	7.1 ± 7.5	7.2 ± 3.4	8.1	5.0
Z(ee)+2b-jet2	11.0 ± 14.0	5.4 ± 8.9		

Within uncertainties, the observations in data are consistent with the expectations from simulation for the Z+2b sample. For the Z+1b sample the estimate differs slightly from the expectation, which will be considered when evaluating the systematic uncertainties of the purity.

Given the agreement between the b-, c-, and l-jet fractions estimated in data, and the expectations from simulations, one is confident to use MC simulation results for the estimate of the ratio of ‘single-mistags’/‘double-mistags’ in the Z+2b data sample.

3.3.2.4 Full Calculation of the Z+2b Purity

The fraction of light jets and c jets in every samples have been measured. The average mistag fractions from the leading and the subleading jets can be expressed with the following formulas:

$$\begin{aligned} f_l &\equiv f_{ll} + f_{lc} + f_{lb}, \\ f_c &\equiv f_{cc} + f_{cl} + f_{cb}. \end{aligned} \quad (3.11)$$

From data fits in Table 3.19, the following average mistag fractions are extracted:

$$\begin{aligned} \langle f_l^\mu \rangle &= 9.0\%, & \langle f_l^e \rangle &= 5.2\%, \\ \langle f_c^\mu \rangle &= 8.2\%, & \langle f_c^e \rangle &= 12.8\%. \end{aligned} \quad (3.12)$$

From the estimates from simulation in Table 3.17 the following ratios are extracted:

$$\begin{aligned} R_1 &= f_{lc}/f_{lb} = 0.40, \\ R_2 &= f_{cc}/f_{cb} = 1.37, \\ R_3 &= f_{ll}/(f_{lc} + f_{lb}) = 0.22. \end{aligned} \quad (3.13)$$

The use of these ratios is well motivated by the good agreement found between data and simulation in the previous section, and by the fact that reliable systematic uncertainty can be given for them related to the Zb,c,l production and to the tag efficiencies.

From the ratios in Equation 3.13 and taking the fractions f_l and f_c from the measurement in data in Equation 3.12, the system of Equation 3.11 can be solved.

In the default scenario, based on the measured fractions of c-/l-jets measured in data, the resulting purity for the muon channel is found to be:

$$\begin{aligned} P_{2b}(\mu\mu) = f_{bb}^\mu &= 1 - (f_{ll} + f_{cc}) - 2 \times (f_{lc} + f_{cb} + f_{lb}) \\ &= 100\% - (1.6\% + 3.5\%) - 2 \times (2.1\% + 2.6\% + 5.2\%) \\ &= 75.1\%. \end{aligned} \quad (3.14)$$

Similarly, for the electron channel, it is found:

$$P_{2b}(ee) = f_{bb}^e = 74.1\%. \quad (3.15)$$

The systematic uncertainties are estimated, varying both tagging efficiencies and production mechanisms according to their own uncertainties. Table 3.20 summarises the variables taken into account, the variation applied, and the

relative variation on the purity calculation in %. Two categories of uncertainties can be distinguished: the impacts of the hypothesis made from the simulations, which are small since they are propagated to the purity via the robust ratios from Equation 3.13; and the propagation of the uncertainties coming from the fit measurement of the jet fractions. The latter category dominates the systematic uncertainties on the purity. However, the uncertainties originating from variations of the different possible constituents of single/double mistags, in which c/l jets are involved, are small, as the effects are all highly correlated. The largest uncertainty is due to the uncertainty of the estimate of the b-fraction itself. All the systematics are summarised in Table 3.20. Nevertheless, it would not be realistic to vary the systematic uncertainties affecting f_l and f_c together up or down. A one sigma contour plot for those correlated quantities variations has been realised, and separately the effect of the systematics on the quantities affecting the value of the ratios in Equation 3.13 has been estimated. The final systematic uncertainty is taken as the difference between two extreme case scenarios.

Table 3.20: *Variables entering the calculation of the b-purity, the variation applied to the variables and their resulting effects on the b-purity as a systematic uncertainty percentage.*

Variable	Variation	Syst.
Mistag of c jets	+10%	0.4%
Mistag of l jets	+10%	0.1%
Mistag of c and l jets	+10%	0.5%
Tag efficiency of b jets	+5%	-0.3%
Tag efficiency of b jets	-5%	0.2%
Zbl production	$\pm 10\%$	0.4%
l fraction into c fraction	1σ	0.3%
c fraction into l fraction	1σ	-0.5%
b fraction	stat.unc = -5%	+9.7%
b fraction	stat.unc = +5%	-9.7%

3.3.2.5 Cross Checks

When fitting the SV mass separately for the leading and the subleading jets, reconstructed events in the Z+2b sample, in which zero, one, or two light-flavoured jets are mistagged can not be distinguished. In order to estimate the

contamination due to single/double mistags, a more optimised approach is to perform a two-dimensional fit of the SV of both the leading and the subleading b-tagged jet. However, given the limited number of events in both data and simulation, the sensitivity is limited, so this estimate is only used as a cross check. A MC simulation independent from the QCD sample used to construct the templates so far is chosen: the Drell–Yan sample. The results show a sensitivity to the Z+b+b contribution reasonable, and are in rough agreement with previous estimates within $\sim 10\%$.

In the future with more integrated luminosity (e.g. with the 2012 8 TeV data sample), the uncertainty on the b-purity might be decreased by using a tighter b-tagging algorithm (or version) in order to further suppress the l-jet contamination. With a smaller impurity in the sample the systematic uncertainty should decrease.

The efficiency of selection, even though decreased by the tagger, as can be seen in Figure 3.12, from $0.55^2 = 30\%$ (for two b jets HE and a light mistagging of 1%), to $0.45^2 = 20\%$ (for two b jet HP and a light mistagging of 0.1%), will be compensated by the luminosity four times greater than for the 7 TeV data sample.

The signal will become $(0.45/0.55)^2 \times 4 = 2.7$ times the one for the 7 TeV events, and the main background (light mistag) will be reduced by $(0.1/1) \times 4 = 0.4$ compared to 7 TeV. This results in:

$$\frac{S_{8TeV}}{\sqrt{B_{8TeV}}} = 4.2 \times \frac{S_{7TeV}}{\sqrt{B_{7TeV}}}.$$

Cuts on the secondary vertex mass could also be used like in [138] where the requirement on the hadrons of $m(SV) > 2 \text{ GeV}$ is used to get a purer Z+b+b sample.

With more collected events in data, the reliability on using information of both jet samples in a single fit increases. And one could build and fit combined variables like $m(SV)_1 * m(SV)_2$, $m(SV)_1 + m(SV)_2$. Another possibility could be to make 2 dimensions (2D), $m(SV)_1$, $m(SV)_2$, templates for bb, cc (+bl+bc, +cl+ll), distinguishing between single and double mistags. The $m(SV)_1$ and $m(SV)_2$ distributions can then be described by constructing a 2D pdf for the 2D distributions. With an even higher budget in integrated luminosity, one could think about using a 3D fit, simultaneously describing also the invariant mass of the two leptons from the Z.

3.3.3 Other Backgrounds: ZZ and ZH

As already mentioned the background due to ZZ production was found to be negligible for the Z+1b analysis. But due to kinematic reasons the probability to find a second b for the ZZ decay is larger than for the Z+b+b production. Indeed, the p_T spectrum of the b jets coming from a heavy object like a Z is harder and the b-tag efficiency increases with the p_T . Hence, the fraction of ZZ background is expected to increase in the Z+2b analysis.

The expected ZZ yield is estimated from simulation, using the cross section and the uncertainty from the CMS measurement [139] for the normalisation. The results are shown in Table 3.21. The systematic uncertainty is dominated by the uncertainty on the measurement of the cross section by CMS, which is approximately 30%.

Table 3.21: *Estimates of the ZZ background yields for the two selection working points and the two channels, with the statistical plus systematic uncertainties.*

WP	N_{ZZ}
Z($\mu\mu$)+1b	$86.2 \pm 0.8 \pm 29$
Z($\mu\mu$)+2b	$11.8 \pm 0.3 \pm 4$
Z(ee)+1b	$69.0 \pm 0.7 \pm 23$
Z(ee)+2b	$9.0 \pm 0.3 \pm 3$

The yield from a SM Higgs boson with a mass of 125 GeV [2, 140, 141, 142] that decays to two b jets and is produced in association with a Z boson, is expected to be approximately 20% of the ZZ contribution, i.e. 2.1 events in the Z($\mu\mu$)+2b-jets final state and 1.7 events in Z(ee)+2b-jets final state. The ZH contribution being smaller than the uncertainty on the ZZ background, this negligible background was not considered in this analysis.

3.3.4 Summary of the Backgrounds

A summary of the estimated backgrounds, and the number of estimated signal events is given in Table 3.22.

Table 3.22: *Summary of the estimated backgrounds and the estimated signal reconstructed yield for the 5.0fb^{-1} integrated luminosity data sample.*

WP	$N_{\text{reco}}^{Z+1/2b}$	$P_b^{1/2b}$	$f_{t\bar{t}}^{1/2b}$	$N_{ZZ}^{1/2b}$
Z($\mu\mu$)+1b	13090	$(53.7 \pm 1.1)\%$	$(5.2 \pm 0.7)\%$	73 ± 24
Z(ee)+1b	9672	$(55.0 \pm 1.3)\%$	$(5.0 \pm 0.7)\%$	56 ± 19
Z($\mu\mu$)+2b	522	$(75.1 \pm 6.4)\%$	$(13.0 \pm 1.9)\%$	12 ± 4
Z(ee)+2b	362	$(74.1 \pm 7.3)\%$	$(14.0 \pm 2.3)\%$	8 ± 3

To estimate the number of signal-only candidates, background events are subtracted from the reconstructed and tagged number of candidates:

$$N_{\text{sig}}^{Z+1/2b} = N_{\text{reco}}^{Z+1/2b} \times [P_b^{1/2b} - f_{t\bar{t}}^{1/2b}] - N_{ZZ}^{1/2b} \quad (3.16)$$

The categories for yields and backgrounds are still defined in the same way than previously: Z+1b corresponds to events with exactly one b-tagged jet in the final state, Z+2b to events with at least two b-tagged jets in the final state.

It has to be noted again that the Z+b+1/c contamination in the Z+2b final state should be considered as signal for the Z+1b final state.

This leads to the following more accurate expression of the signal only candidates:

$$\begin{aligned} N_{\text{sig}}^{Z+1b} &= N_{\text{reco}}^{Z+1b} \times (P_b^{Z+1b} - f_{t\bar{t}}^{Z+1b}) - N_{ZZ}^{Z+1b} + f_{1b}^{Z+2b} \times N_{\text{reco}}^{Z+2b}, \\ N_{\text{sig}}^{Z+2b} &= N_{\text{reco}}^{Z+2b} \times (P_b^{Z+2b} - f_{t\bar{t}}^{Z+2b}) - N_{ZZ}^{Z+2b}, \end{aligned} \quad (3.17)$$

with f_{1b}^{Z+2b} being the fraction of events in the Z+2b-jets sample for which one jet is mistagged, which is $16 \pm 5\%$, so a final correction at the percent level.

3.4 Efficiencies and Migrations

In order to extract a cross section to compare with theoretical predictions, the background corrected yields need to be corrected for the selection efficiency as well as for the detector effects. An estimation of the number of signal events at the particle level (also called generator level), from the data and in the acceptance can then be obtained. This number divided by the luminosity gives the cross section in the acceptance.

The selection efficiency is factorised in two parts: the lepton selection efficiency and the b-tagging efficiency. The efficiency from the MET significance cut is also taken into account as a separated correction. The correction for the detector resolution is equivalent to interpreting the results at the particle level from the reconstructed level.

Contrarily to the previous inclusive Z+b cross section measurement, where the respective proportions of events in the 1 and 2 or more b-jet(s) bins were not measured, the cross section of a Z in association with exactly one b jet and of a Z in association with two b jets or more depends on the migrations between the 1 and 2 b-jet(s) bins. Both, the b-tagging efficiency and the jet reconstruction can induce migrations between the two categories, since the number of b jets generated and reconstructed are generally different. To estimate these two cross sections the correction for the efficiencies and resolution has to be done as a function of the jet multiplicity. For this, a 2×2 matrix is used for each efficiency and leads to the Equation 3.18 where the matrices are applied in the order of the selection in the analysis. Off-diagonal terms in these matrices represent the mixing between the one and two b-jet(s) categories and yields, and permit to evaluate both cross sections simultaneously.

$$\begin{pmatrix} N_{Z+1b}^{gen} \\ N_{Z+2b}^{gen} \end{pmatrix} = \mathcal{A}_1^{-1} \times E_r^{-1} \times E_\ell^{-1} \times E_b^{-1} \times E_m^{-1} \times \begin{pmatrix} N_{sig}^{Z+1b} \\ N_{sig}^{Z+2b} \end{pmatrix} \quad (3.18)$$

E_ℓ and E_b deal with the efficiencies to reconstruct, respectively, leptons and b-tagging, E_m with the efficiency of the MET significance cut and E_r accounts for the efficiency for generator level events to be reconstructed as they were generated, dominated by the jet resolution. The matrix $\mathcal{A}_1$ is optional in this case, but could account for purely generator level corrections for the lepton acceptance cuts if the acceptance cuts at generator (gen) and reconstructed (reco) level are different.

The solution of Equation 3.18 is then directly used to determine the cross section:

$$\begin{pmatrix} \sigma_{Z+1b} \\ \sigma_{Z+2b} \end{pmatrix} = \frac{1}{\mathcal{L}} \begin{pmatrix} N_{Z+1b}^{gen} \\ N_{Z+2b}^{gen} \end{pmatrix}. \quad (3.19)$$

The efficiency matrices are derived from the DY MC sample introduced earlier (see Table 3.3). The reliance on this MC sample is motivated by the good agreement with the data especially for the transverse momentum of b jets and MET significance distributions, as already seen in all the figures in Section 3.2.8 and Figure 3.5 (bottom). Scale factors estimated in Section 3.2.6 are applied according to the step of selection: pileup reweighting, and possibly lepton scale factors, and possibly b-tagging scale factors. The calculations and estimations for each matrix are explained below.

3.4.1 Generator Level and Reconstructed Level Definitions

Since the numbers of events at generator level for different steps of the selection will be needed in the estimation of the efficiency factors, some precisions on this selection are added below:

- The leptons coming from a Z at the generator level are selected and constructed as the so-called ‘dressed leptons’. This consists in taking a lepton at the end of its decay chain in the parton shower (leptons of status 1 in PYTHIA), and adding to it the energy of the photons present in a cone of size $\Delta R = 0.1$ around the lepton.
- A real b at hadron level is defined by a generator level jet (called hereafter a genjet) matched with a B hadron in a cone of size $\Delta R = 0.5$.

For the estimation, the consistency of reconstructed leptons and jets with the generated ones have to be checked. This is done by requiring:

- Selected reconstructed leptons must be matched to the generator level leptons within a cone of size $\Delta R = 0.3$.
- Selected reconstructed jets must be matched to b-flavoured hadron genjets within a cone of size $\Delta R = 0.5$.

3.4.2 MET significance Efficiency

No migration in terms of b-jet bins is expected from the MET significance cut. So only the factors ϵ_m^{11} and ϵ_m^{22} need to be computed. The cut is only applied for the Z+2b selection which means that ϵ_m^{11} is equal to 1. The Table 3.23 below is summarising how the factors are related.

Table 3.23: *The different efficiency factors relating the MET significance selection criterion efficiency and b-jet multiplicities.*

		$N_{Z_{sel}+Xb}^{tag}$	
		1 b-tag	2 b-tag
$N_{Z_{sel}+Yb}^{tag} + \text{MET significance}$	1 b-tag	1	0
	2 b-tag	0	ϵ_ℓ^{22}

3.4.3 b-tagging Efficiency

The matrix E_b^{-1} is used to correct for the b-tagging efficiency, the number of reconstructed b-flavoured jets. It is applied to the yields of Z+1b-tag and Z+2b-tag events, hence already corrected for b-purity and $t\bar{t}$ contamination. So mistag does not enter in the efficiency terms. The b-tagging algorithm and the version applied in the analysis is used, namely the SSVHE. Each individual factor ϵ_b^{XY} corresponds to the efficiency for tagging Y jets, given X generated b-jets reconstructed as jets in the acceptance, as illustrated by Table 3.24.

Table 3.24: *The different efficiency factors relating the b-tagging efficiency and b-jet multiplicities.*

		$N_{Z_{sel}+Xb}^{reco}$		
		0 b (reco)	1 b jet	2 b jet
$N_{Z_{sel}+Yb}^{tag}$	0 b-tag	0	ϵ_b^{10}	ϵ_b^{20}
	1 b-tag	0	ϵ_b^{11}	ϵ_b^{21}
	2 b-tag	0	0	ϵ_b^{22}

The sum of efficiencies in each column, and hence for a given b-jet multiplicity, is necessarily constrained to be one,

$$\begin{aligned} \epsilon_b^{20} + \epsilon_b^{21} + \epsilon_b^{22} &= 1, \\ \epsilon_b^{10} + \epsilon_b^{11} &= 1. \end{aligned} \tag{3.20}$$

The resulting number of b-tagged events can be found for those efficiency factors through the systems of Equations 3.21 and 3.22.

$$\begin{aligned} N_{Z_{sel}+2b}^{tag} &= \epsilon_b^{22} N_{Z_{sel}+2b}^{reco} \\ N_{Z_{sel}+1b}^{tag} &= \epsilon_b^{11} N_{Z_{sel}+1b}^{reco} + \epsilon_b^{21} N_{Z_{sel}+2b}^{reco} \end{aligned} \quad (3.21)$$

$$\begin{aligned} N_{Z_{sel}+2b}^{reco} &= \frac{N_{Z_{sel}+2b}^{tag}}{\epsilon_b^{22}} \\ N_{Z_{sel}+1b}^{reco} &= \frac{N_{Z_{sel}+1b}^{tag}}{\epsilon_b^{11}} - \frac{\epsilon_b^{21}}{\epsilon_b^{11}\epsilon_b^{22}} N_{Z_{sel}+2b}^{tag} \end{aligned} \quad (3.22)$$

The equations result in the following system:

$$\begin{pmatrix} N_{Z_{sel}+1b}^{reco} \\ N_{Z_{sel}+2b}^{reco} \end{pmatrix} = \begin{pmatrix} \frac{1}{\epsilon_b^{11}} & -\frac{\epsilon_b^{21}}{\epsilon_b^{22}\epsilon_b^{11}} \\ 0 & \frac{1}{\epsilon_b^{22}} \end{pmatrix} \begin{pmatrix} N_{Z_{sel}+1b}^{tag} \\ N_{Z_{sel}+2b}^{tag} \end{pmatrix}, \quad (3.23)$$

from which the inverted matrix E_b^{-1} expression is taken.

3.4.4 Lepton Selection Efficiency

As no migration in b-jet bin is expected from the lepton selection, there are only two ϵ_ℓ factors to be estimated as shown in Table 3.25.

Table 3.25: *The different efficiency factors relating the lepton selection efficiency and b-jet multiplicities.*

		N_{Z+Xb}^{reco}	
		1 b (reco)	2 b (reco)
$N_{Z_{sel}+Yb}^{reco}$	1 b (reco)	ϵ_ℓ^{11}	0
	2 b (reco)	0	ϵ_ℓ^{22}

The factors ϵ_ℓ^{11} and ϵ_ℓ^{22} correct for inefficiencies in the dilepton selection. This includes contributions from the trigger, reconstruction, identification and isolation criteria. A separate factor is calculated for each b-jet multiplicity bin as, although jets are required to be separated from each lepton by at least $\Delta R = 0.5$, it is not guaranteed that the jet multiplicity does not introduce a bias in the lepton selection. The E_ℓ^{-1} matrix form is then:

$$\begin{pmatrix} N_{Z+1b}^{reco} \\ N_{Z+2b}^{reco} \end{pmatrix} = \begin{pmatrix} \frac{1}{\epsilon_\ell^{11}} & 0 \\ 0 & \frac{1}{\epsilon_\ell^{22}} \end{pmatrix} \begin{pmatrix} N_{Z_{sel}+1b}^{reco} \\ N_{Z_{sel}+2b}^{reco} \end{pmatrix}. \quad (3.24)$$

3.4.5 Reconstruction Efficiency

The ϵ_r are the efficiencies, measured in the MC signal sample, for Z+1b-jet and Z+2b-jet generator level events to be reconstructed as they were generated, in terms of number of b jets. Thus, ϵ_r^{XY} corresponds to the efficiency for obtaining a Z+Yb-jet event at the reconstructed level given a Z+Xb-jet event at the generator level. However, if the event does not pass the lepton selections at either reconstructed or generator level then X or Y is taken as zero, regardless of the b-jet multiplicity in the acceptance.

Equations for the generator level yields, $N_{Z+1b}^{gen-acc}$ and $N_{Z+2b}^{gen-acc}$ can be derived by considering the different factors in Table 3.26, which are subject to the constraints in equations 3.25 through 3.27. These constraints require the sum of efficiencies in each column of Table 3.26 (i.e. for each hadron-level b-jet multiplicity) to be one. One can notice that the factors ϵ_r^{0Y} are not null since the jet resolution can result in a generator level jet outside the acceptance being reconstructed in the acceptance, with the two objects (gen and reco) matched. They are however expected to be small.

Table 3.26: *The different efficiency factors relating the reconstruction efficiency and b-jet multiplicities.*

		N_{Z+Xb}^{gen}		
		No gen	1 b (gen-acc.)	2 b (gen-acc.)
N_{Z+Yb}^{reco}	No reco	ϵ_r^{00}	ϵ_r^{10}	ϵ_r^{20}
	1 b-tag	ϵ_r^{01}	ϵ_r^{11}	ϵ_r^{21}
	2 b-tag	ϵ_r^{01}	ϵ_r^{12}	ϵ_b^{22}

$$\epsilon_r^{00} + \epsilon_r^{01} + \epsilon_r^{02} = 1 \quad (3.25)$$

$$\epsilon_r^{10} + \epsilon_r^{11} + \epsilon_r^{12} = 1 \quad (3.26)$$

$$\epsilon_r^{20} + \epsilon_r^{21} + \epsilon_r^{22} = 1 \quad (3.27)$$

$$\frac{N_{NO}^{gen}}{N_{Z+1b}^{gen-acc} + N_{Z+2b}^{gen-acc}} = R \quad (3.28)$$

One can derive the expressions of the $N_{Z+1/2b}^{reco}$ as a function of the generator level yields in Equations 3.29 and 3.30.

$$N_{Z+1b}^{reco} = \epsilon_r^{01} N_{NO}^{gen} + \epsilon_r^{11} N_{Z+1b}^{gen-acc} + \epsilon_r^{21} N_{Z+2b}^{gen-acc} \quad (3.29)$$

$$N_{Z+2b}^{reco} = \epsilon_r^{02} N_{NO}^{gen} + \epsilon_r^{12} N_{Z+1b}^{gen-acc} + \epsilon_r^{22} N_{Z+2b}^{gen-acc} \quad (3.30)$$

Nevertheless, it is not possible to give a third equation for the number of non reconstructed events (or reconstructed without a b jet in the acceptance), as one cannot get this information from the data. But the assumption that the ratio of generator level events outside the acceptance to inside the acceptance in Equation 3.28 is correctly predicted by the simulation, allows the resolution of the system. The final matrix E_r^{-1} is given in Equation 3.31.

$$\begin{pmatrix} N_{Z+1b}^{gen-acc} \\ N_{Z+2b}^{gen-acc} \end{pmatrix} = \begin{pmatrix} \epsilon_r^{11} + R\epsilon_r^{01} & \epsilon_r^{21} + R\epsilon_r^{01} \\ \epsilon_r^{12} + R\epsilon_r^{02} & \epsilon_r^{22} + R\epsilon_r^{02} \end{pmatrix}^{-1} \begin{pmatrix} N_{Z+1b}^{reco} \\ N_{Z+2b}^{reco} \end{pmatrix} \quad (3.31)$$

3.4.6 Lepton Acceptance

This correction can be used if the lepton cuts at generator level for which one wants to give the final cross section are different from the generator level cuts used in the building of the correction matrices. Since only lepton cuts are involved, no migrations in terms of b-jet bins are expected. But the factors a_ℓ^{11} and a_ℓ^{22} are computed separately. The Table 3.27 below is summarising how the factors are related:

Table 3.27: *The different efficiency factors relating the number of events at the generator level cross section and the number of events in the analysis acceptance.*

		N_{Z+Xb}^{gen}	
		1 b-tag	2 b-tag
N_{Z+Yb}^{gen}	1 b-tag	a_ℓ^{11}	0
	2 b-tag	0	a_ℓ^{22}

In the present case the lepton cuts of the acceptance are the same than the ones the cross section will be given for. Namely $p_T > 20 \text{ GeV}$ and $|\eta| < 2.4$. Thus the matrix $\mathcal{A}_l$ is just equal to the unity matrix.

3.4.7 Final Matrix

Summarising the results of the previous sections the final expression of the correction matrices is given below in Equation 3.32:

$$\begin{pmatrix} N_{Z+1b}^{gen} \\ N_{Z+2b}^{gen} \end{pmatrix} = \begin{pmatrix} \frac{1}{a_\ell^{11}} & 0 \\ 0 & \frac{1}{a_\ell^{22}} \end{pmatrix} \begin{pmatrix} \epsilon_r^{11} + R\epsilon_r^{01} & \epsilon_r^{21} + R\epsilon_r^{01} \\ \epsilon_r^{12} + R\epsilon_r^{02} & \epsilon_r^{22} + R\epsilon_r^{02} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\epsilon_\ell^{11}} & 0 \\ 0 & \frac{1}{\epsilon_\ell^{22}} \end{pmatrix} \begin{pmatrix} \frac{1}{\epsilon_b^{11}} & \frac{-\epsilon_b^{21}}{\epsilon_b^{22}\epsilon_b^{11}} \\ 0 & \frac{1}{\epsilon_b^{22}} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\epsilon_m^{22}} \end{pmatrix} \begin{pmatrix} N_{Z_{sel}+1b}^{tag} \\ N_{Z_{sel}+2b}^{tag} \end{pmatrix} \quad (3.32)$$

- Muon channel:

$$\mathbf{E}_r^{-1} \times \mathbf{E}_\ell^{-1} \times \mathbf{E}_b^{-1} \times \mathbf{E}_m^{-1} = \begin{pmatrix} 1.1 & -0.5 \\ -0.02 & 1.4 \end{pmatrix} \begin{pmatrix} 1.36 & 0. \\ 0. & 1.37 \end{pmatrix} \begin{pmatrix} 2.1 & -3.8 \\ 0. & 3.7 \end{pmatrix} \begin{pmatrix} 1. & 0. \\ 0. & 1.03 \end{pmatrix} = \begin{pmatrix} 3.2 & -8.5 \\ -0.05 & 7.3 \end{pmatrix} \quad (3.33)$$

- Electron channel:

$$\mathbf{E}_r^{-1} \times \mathbf{E}_\ell^{-1} \times \mathbf{E}_b^{-1} \times \mathbf{E}_m^{-1} = \begin{pmatrix} 1.3 & -0.5 \\ -0.02 & 1.5 \end{pmatrix} \begin{pmatrix} 1.51 & 0. \\ 0. & 1.51 \end{pmatrix} \begin{pmatrix} 2.1 & -3.8 \\ 0. & 3.8 \end{pmatrix} \begin{pmatrix} 1. & 0. \\ 0. & 1.02 \end{pmatrix} = \begin{pmatrix} 4.1 & -10.8 \\ -0.06 & 9.1 \end{pmatrix} \quad (3.34)$$

The numerical values of the different matrices elements are determined from the Drell–Yan MC simulation as already mentioned above. Results per channel are given in Equations 3.33 and 3.34 below.

The numerical precision on each factor is in line with the associated systematic uncertainty.

Effect, from all those systematic uncertainties, on the cross section measurement will be treated and estimated in the following Section 3.5.

From the numerical values, one can notice that $\frac{1}{\epsilon_m^{22}}$ is about the same for muon and electron channel which means that the cut on the MET significance is not affecting one lepton flavour more than the other, and corresponds to the 2% of signal loss expected in the analysis from the yields in Table 3.9 (Section 3.2.7). The values of the efficiency factors for b-tagging are also similar in the electron and muon channel, which is expected from the use of the algorithm which should not be affected by the flavour of the leptons of the Z.

In the E_ℓ^{-1} matrices the values of the ϵ_ℓ^{11} and ϵ_ℓ^{22} differ between electron and muon channel, which shows a difference in efficiency in isolation and identification for electrons and muons.

But inside the same channel $\epsilon_\ell^{11} = \epsilon_\ell^{22}$, which confirms that the presence of one or two b jets is not influencing the efficiency of lepton identification and isolation, as expected since the requirement $\Delta R(\ell, j) > 0.5$ renders the lepton selection insensitive to the jet multiplicity.

Finally, in the resolution matrix it is noticed that the values for ϵ_r^{11} and for ϵ_r^{22} , for which no migrations due to jet energy happened, are still differing between electron and muon channel. This points to a difference between the muon and electron channel, for the efficiency from generator to reconstructed level. The reason for this is the non fiducial gap region in η in the electron selection; which implies a bigger correction in the correction matrix.

3.5 Systematic Uncertainties

The different sources of systematic uncertainties on the cross sections result are gathered below. Two different types have been considered:

- **Direct contribution to the background subtracted yields**, and its propagation to the final result via the Equation 3.17. The contributions in this category are the contributions from the systematic uncertainties on the b-purity, $t\bar{t}$ and ZZ contaminations and the luminosity.

- **Background from lighter parton jets:**

The main source of uncertainty for the determination of the background comes from b-jets mistagging, which affects the purity and the fraction of singly mistagged jets, and arises from the fit in the fractions of b,c,l jets for the $m(SV)$. The systematic uncertainty is estimated by varying the (mis)tagging efficiencies according to their uncertainties from the fit. No significant difference has been observed when using a different generator to create the templates used in the fit to the $m(SV)$ [128, 143].

- **Background from $t\bar{t}$:**

In the estimation of the fraction of $t\bar{t}$ background the main uncertainty is the one from the fit of the invariant mass $M_{\ell\ell}$. And additional uncertainty originates from the modelling of the signal and background template shapes. Different methods have been used to estimate this fraction as described in Section 3.3.1: with templates from simulation and data,

all simulation-based and empirical. The systematic uncertainty is computed from the difference in the fraction result obtained with the three methods.

- **The ZZ background:**

The uncertainty in the overall normalisation is taken from the CMS measurement [139]. Correlated sources of uncertainties (such as the luminosity) are ignored to avoid double counting.

- **Contributions via the correction matrices.**

Those are estimated by reweighting the reference MC sample used to estimate the efficiency factors in the correction matrices according to the systematic uncertainties. These are the contributions from theory, from the lepton and b-tagging selection, from the pileup, the JER, the JES and the MET uncertainties. Also grouped here are the other uncertainties from Equation 3.19, like the luminosity systematic uncertainty. Apart from the theory uncertainty, all these uncertainties have been estimated according to the recommendations from the respective Physics Object Group (POG) in CMS.

- The b-tagging efficiency and the mistagging fraction: These uncertainties affect the b-tagging efficiencies as described in Section 3.4.3.
The p_T -dependent uncertainties in the jet tagging efficiency, 3–8% for $p_T > 30$ GeV and 12% for $p_T < 30$ GeV, are propagated to the b-tagging data/simulation scale factors, as described in [128]. The uncertainty in the mistagging fraction, which enters the calculation of the event weight at second order, is found to have a negligible impact.
- Jet energy scale and resolution: The uncertainty in the JES is taken from [124] and amounts to 3–5% depending on the p_T and η of the jets. The uncertainty in the JER is taken to be 10%, after degrading the simulated resolution by 10% to match the one measured in the data. Both affect the matrix E_r^{-1} .
- Effect from pileup: The total inelastic cross section used to reproduce the pileup in data from the instantaneous luminosity is varied by $\pm 5\%$, thereby affecting the pileup distribution in the simulated samples and covering the uncertainties due to pileup modelling. It is then propagated to the estimation of the correction matrices where it affects mainly the lepton efficiency factors through the lepton isolation requirements.
- Dilepton selection efficiencies: The systematic uncertainty of the scale factor per lepton, which is applied to simulated events to compensate for

data/simulation differences, is obtained with the *Tag and Probe* method, and is less than 0.4% for muons and 1.0% for electrons.

- Requirement on MET: The requirement on the MET significance removes $\sim 2\%$ of the Z+2b signal contribution, which is evaluated from simulation. The systematic uncertainty is estimated by varying each component entering the MET calculation within its uncertainty. This includes contributions from JES and JER as discussed above, unclustered energy (10%), τ leptons (3%), electrons and photons (0.6–1.5%), and muons (0.2%) [129]. The effect on the cross section measurement was however found negligible.
- Simulated event samples size: While the size of the MC samples suffice for the Z+1b selection sample, they lead to uncertainties of several percent in all the correction factors involving the Z+2b selection sample.
- Theory: The measured cross section is compared to different theoretical expectations as will be seen in the final cross section in Table 3.30. The effect of uncertainties in the renormalisation and factorisation scales on those theoretical cross sections is estimated using MCFM [77]. The impact of scale variations on the p_T of the b jets is then used in the correction to generator level procedure to estimate the effect on the cross section. Similarly, the p_T of the dilepton system is varied according to the difference observed between data and simulation to estimate the impact on the factor of the correction matrices. Furthermore, the effect due to the Multi Parton Interactions on the acceptance of Z+b-jets events is studied by artificially reducing their contribution by a factor 2. This is done by applying a veto on the azimuthal angle $\Delta\Phi_{Z,bb}$, which has been shown to be a discriminant observable for MPIs [144]. The effect of this requirement on the cross sections has been found to be less than 0.5%. Together, this leads to an uncertainty of at most 3% in the cross sections.
- Luminosity: The uncertainty of the integrated luminosity recorded by CMS is 2.2% in the 2011 data set [105].
- Vertex association: For the estimate of the cross section ratio $\sigma_{Z+b}/\sigma_{Z+j}$, an additional uncertainty arises from the contribution of jets not associated with the primary vertex from which the Z boson arises. After the requirement on the momentum fraction of tracks originating from the primary vertex, the background due to pileup is estimated from simulation to be 2.2%. The efficiency of the requirement is estimated from the distribution of this observable in data before applying the requirement. The corresponding systematic uncertainty is evaluated by comparing the

distributions of this observable in data and simulation. It is assumed that the difference observed for the variable used for the vertex association is entirely due to events originating from pileup. This assumption results in a systematic uncertainty of 18% in the pileup contamination, fully correlated between the electron and muon channel.

All the systematic uncertainties are summarised in Table 3.28, where they are presented separately for the muon and the electron channel and for the selection working points: Z+1b and Z+2b. They are also classified according to the correlation between channels.

Table 3.28: *Fractional uncertainties on the measured cross sections, grouped according to the correlation between the channels.*

	$\mu\mu$ (%)		ee (%)	
	Z+1b	Z+2b	Z+1b	Z+2b
Uncorrelated				
b-purity	3.0	12.7	3.3	15.1
$t\bar{t}$	1.7	3.8	1.7	4.8
Dilepton sel.	1.0	1.0	2.0	2.0
MC samples size	0.9	4.2	1.2	5.1
Correlated				
b-tag eff SFs	3.6	9.0	3.6	9.0
JES	2.0	3.6	2.0	3.6
Theory	1.8	3.0	1.8	3.0
Luminosity	2.3	2.3	2.3	2.3
ZZ	0.4	1.2	0.5	1.4
Pileup	0.3	0.3	0.3	0.3
JER	0.2	0.3	0.2	0.3
Mistag	0.0	0.1	0.0	0.1
Tot. stat. unc.	0.9	4.5	1.0	5.4
Tot. syst. unc.	6.3	17.4	6.7	19.8

3.6 Cross Sections Results

Data

The cross sections are estimated per b-jet multiplicity bin and for each lepton flavour separately. The results are summarised in Table 3.29.

Table 3.29: *Cross sections at the particle level for the production of a Z boson with exactly one b jet, with at least two b jets, and with at least one b jet. In the last row the ratio for the production of a Z boson with at least one b jet with respect to the production of a Z boson in association with at least one jet of any flavour is given. The first uncertainty is statistical, and the second systematic.*

Cross section	$\mu\mu$	ee
σ_{Z+1b} (pb)	$3.52 \pm 0.03 \pm 0.22$	$3.51 \pm 0.04 \pm 0.23$
σ_{Z+2b} (pb)	$0.38 \pm 0.02 \pm 0.07$	$0.32 \pm 0.02 \pm 0.06$
σ_{Z+b} (pb)	$3.91 \pm 0.04 \pm 0.23$	$3.84 \pm 0.04 \pm 0.24$
$\sigma_{Z+b/Z+j}$ (%)	$5.23 \pm 0.04 \pm 0.24$	$5.08 \pm 0.05 \pm 0.24$

Using the best linear unbiased estimator (BLUE) method [145], results for the $\mu\mu$ and ee channels are found to be consistent with a χ^2 probability of 42% for the Z+1b and 78% for the Z+2b cases. They are therefore combined in a single measurement using the optimal set of coefficients that minimises the total uncertainty in the combined result, taking into account statistical and systematic uncertainties and their correlations. The results are summarised in Table 3.30 and are then compared with various predictions.

A similar measurement has been performed by another group of the CMS collaboration [138], but with a final state presenting two B hadrons instead of b jets in association with the Z boson. Because of this difference in the final state definition, but also in the acceptance, the measurements cannot be directly compared. Nevertheless, it can be said that the results values are coherent as reported in Table 3.31.

More recently the ATLAS collaboration has released results from measurements analogous to the ones in the present thesis [146]. Even if the processes considered are the same, namely the production of a Z boson with at least one b jet and with at least two b jets, the acceptance is different and does not allow a direct comparison. However, the definitions are close and if one looks at the cross sections given in Table 3.31 they are compatible within the uncertainties or within less than 1σ with the present analysis ones.

Theoretical expectations

The expectations from MADGRAPH, in both the 5-flavour and the 4-flavour schemes which were described in Section 1.2.2.2, are estimated using the LO estimations and a global K factor of value 1.23 to correct the inclusive Drell–Yan cross section for NNLO effects [110].

For the current MG5F prediction, up to four partons have been produced at the matrix element level in association with the Z boson, with the b quarks assumed massless and included in the proton PDF. The PDF set used is CTEQ6L1 [49], and the jet matching is performed using the standard k_T -MLM scheme at a matching scale $Q_{match} = 20$ GeV [68].

The MG4F prediction which considers massive b quarks in the ME calculation with the mass set to $m_b = 4.7$ GeV, was derived with two additional light partons produced in the ME in association with the Z+bb final state. The jet matching scheme is also the k_T -MLM with $Q_{match} = 30$ GeV.

The NLO expectations from aMC@NLO are also estimated using both 5F and 4F calculations with the MSTW2008 NLO PDF set [51]. This generator adds to the LO diagrams the ones with one extra hard radiation, real or virtual.

The events simulated with MADGRAPH and aMC@NLO are interfaced with the PYTHIA parton shower simulation, which will be responsible for all the extra jets in the events.

The NLO prediction from MCFM is at the parton level. The MCFM calculations are estimated with the CTEQ6mE PDF, and the renormalisation and factorisation scales μ_F and μ_R are set to the invariant mass of the dilepton.

Uncertainties in the theoretical predictions are estimated by varying the renormalisation and factorisation scales up and down by a factor two. For the MADGRAPH 5F prediction, the uncertainties are smaller than for the other predictions: this is because the scales are varied in a correlated way for MG5F, while for the other predictions the scales are varied in an uncorrelated manner, which leads to a larger uncertainty. The uncertainties in the 4F predictions amount to 15–20%, as expected [147]. Uncertainties due to the choice of PDF (using MSTW2008 [51], CTEQ6 [49], and CT10 [148] sets), jet matching scale (up to a factor two), and mass of the b quark (varied between 4.4 and 5.0 GeV), have all been found to be smaller and negligible compared to the scale uncertainties.

In Table 3.30 results are gathered for the five theoretical predictions detailed above and compared to the measurements from data. In addition to the two selection working points, the exclusive Z+1b and the inclusive Z+2b, the cross section results are also given for the inclusive Z+b-jets and for the ratio (Z+b)/(Z+jets).

Table 3.30: Cross sections for the production of a Z boson with exactly one b jet, with at least two b jets, with at least one b jet, and the ratio with respect to at least one jet of any flavour, showing the statistical and systematic uncertainties. The expectations from MADGRAPH, MCFM and aMC@NLO include uncertainties due to scale variations.

Cross section	Measured	MADGRAPH (5F)	aMC@NLO (5F)	MCFM (parton level)	MADGRAPH (4F)	aMC@NLO (4F)
σ_{Z+1b} (pb)	$3.52 \pm 0.02 \pm 0.20$	3.66 ± 0.22	$3.70^{+0.23}_{-0.26}$	$3.03^{+0.30}_{-0.36}$	$3.11^{+0.47}_{-0.81}$	$2.36^{+0.47}_{-0.37}$
σ_{Z+2b} (pb)	$0.36 \pm 0.01 \pm 0.07$	0.37 ± 0.07	$0.29^{+0.04}_{-0.04}$	$0.29^{+0.04}_{-0.04}$	$0.38^{+0.06}_{-0.10}$	$0.35^{+0.08}_{-0.06}$
σ_{Z+b} (pb)	$3.88 \pm 0.02 \pm 0.22$	4.03 ± 0.24	$3.99^{+0.25}_{-0.29}$	$3.23^{+0.34}_{-0.40}$	$3.49^{+0.52}_{-0.91}$	$2.71^{+0.52}_{-0.41}$
$\sigma_{Z+b}/Z+j$ (%)	$5.15 \pm 0.03 \pm 0.25$	5.35 ± 0.11	$5.38^{+0.34}_{-0.39}$	$4.75^{+0.24}_{-0.27}$	$4.63^{+0.69}_{-1.21}$	$3.65^{+0.70}_{-0.55}$

Table 3.31: Cross sections at the particle level for the production of a Z boson with at least two b jets, and with at least one b jet, for events with $Z \rightarrow \ell\ell$ (where $\ell\ell = e^+e^-$ or $\mu^+\mu^-$). Three analysis results using 7 TeV data samples are presented: the one of the present thesis (Zbb CMS), the CMS analysis with B hadrons in the final state [138] (ZBB CMS) and the ATLAS analysis [146] (Zbb ATLAS). When two are displayed, the first uncertainty is statistical, and the second systematic.

Analysis	Zbb (CMS)	ZBB (CMS)	Zbb (ATLAS)
Acceptance	$p_T^{\ell} > 20 \text{ GeV}, \eta^{\ell} < 2.4, 76 < M_{\ell\ell} < 106 \text{ GeV}, p_T^{b, \text{jet}} > 25 \text{ GeV and } \eta < 2.1, \Delta R(\ell, j) > 0.5$	$p_T^{\ell} > 20 \text{ GeV}, \eta^{\ell} < 2.4, 81 < M_{\ell\ell} < 101 \text{ GeV}, p_T^{B, \text{hadron}} > 15 \text{ GeV and } \eta < 2, \Delta R(\ell, h) > 0.5$	$p_T^{\ell} > 20 \text{ GeV}, \eta^{\ell} < 2.5, 76 < M_{\ell\ell} < 106 \text{ GeV}, p_T^b > 20 \text{ GeV and } \eta < 2.4, \Delta R(\ell, j) > 0.5$
σ_{Z+2b} (pb)	$0.36 \pm 0.01 \pm 0.07$	$p_T^{\ell} > 0 \text{ GeV: } 0.71 \pm 0.08$ $p_T^{\ell} > 40 \text{ GeV: } 0.44 \pm 0.05$	$0.52 \pm 0.02 \pm 0.07$
σ_{Z+1b} (pb)	$3.88 \pm 0.02 \pm 0.22$	-	$4.82 \pm 0.06 \pm 0.37$

Comparisons

The measured cross sections are consistent, within uncertainties, with the expectations in the 5F scheme from both MADGRAPH and aMC@NLO. Compared to the predictions from MADGRAPH and aMC@NLO in the 5F scheme, the predictions from MCFM are approximately 20% lower.

The predictions by MADGRAPH and aMC@NLO from calculations in the 4F scheme, compared to the predictions in the 5F scheme, show a reduction of the Z+1b production rate when the other b jet in the final state is produced outside of the acceptance.

A difference of approximately two standard deviations is observed when comparing to the parton-level prediction from MCFM for the Z+b-jets cross section. Since the correction factor from parton level to hadron level is smaller than one [104], this difference is not explained by hadronisation effects. The difference remains when measuring the cross section ratio, which excludes an explanation based on experimental systematic effects that are shared between the Z+jets and the Z+b-jets final states, such as luminosity, and the reconstruction of jets and leptons. These results indicate that the difference observed with MCFM is specific to the modelling of the Z+b-jets final state.

The largest discrepancy is observed when comparing the measured Z+1b cross section with the predictions in the 4F scheme. In particular, the prediction from aMC@NLO in the 4F scheme shows a discrepancy of more than two standard deviations compared to the measurement.

All these comparisons are summarised in Figure 3.15 for both the exclusive one b jet (left) and the inclusive two b jets (right) final states.

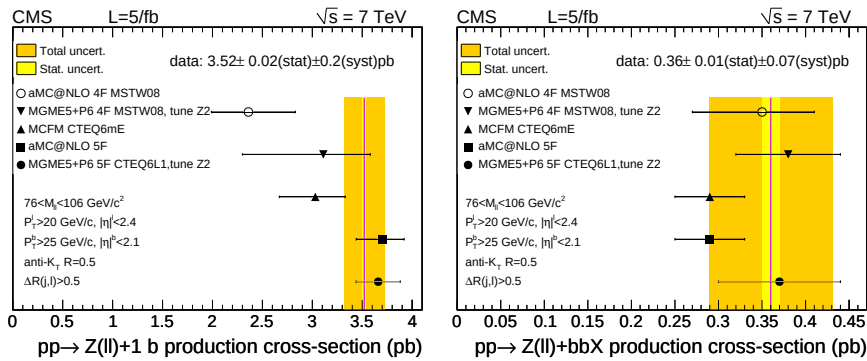


Figure 3.15: Measured cross section for the exclusive Z+1b (left) and the inclusive Z+2b (right) final states compared with different simulation predictions.

Comments on the previous $Z + b$ -jets cross section measurement

As mentioned in the beginning of the chapter, another measurement has been published in 2013 with 2.2 pb^{-1} of data, but superseded by the second one presented in detail before. The final cross section for Z plus at least one b jet of the first paper ($5.84 \pm 0.08(\text{stat.}) \pm 0.72(\text{syst.})_{0.55}^{+0.25}(\text{th.}) \text{ pb}$) is compatible with the second paper result within the uncertainties.

The second measurement was performed with more than double the number of data events, which allowed to separate the b jet multiplicities. But the main improvement lies in the definition of the signal at generator level. For the first measurement this latter was defined at the parton level, i.e. asking for a b quark in the genjet. This definition is biased and unpractical from a phenomenological point of view. In the second measurement the signal definition is done at the hadron level, i.e. that the event must contain a genjet matched to a B hadron. This is also exactly the level of definition of the cross section measurement.

Some other improvements can be mentioned:

- **Acceptance cuts of the leptons:** the p_T and η cuts are chosen such that the kinematic acceptance is the same for both lepton flavours, and maximal (given the limitations of the detector): $p_T > 20 \text{ GeV}$, $|\eta| < 2.4$. This means that the kinematic comparisons are the same for both lepton flavours, and the acceptance term can be ignored in the estimate of the combined cross section.
- **Mass window of the Z boson:** the mass window has been narrowed from $[60\text{--}120] \text{ GeV}$ to $[76\text{--}106] \text{ GeV}$ in order to remove a larger fraction of $t\bar{t}$ background.
- **Veto on MET:** after the first measurement an intermediate result was presented with a cut on the MET [149]. In the second measurement, for the events selection, instead of the MET itself, the MET significance was used to suppress background originating from $t\bar{t}$. This is more discriminating and at the same time leads to smaller systematic uncertainties.
- **Purity estimate:** previously the per-jet fraction contribution to the purity, through the ratio $Z+2b/Z+1b$, was entirely driven by simulation. In the second measurement the fractions are all evaluated from data.

The tagger used in the first measurement was the SSVHP 3.6, which was ensuring a smaller mistagging from the light jets fraction than the SSVHE (0.1% versus 1%). The switch to SSVHE in the second measurement is driven by its higher efficiency, necessary for the $Z+2b$ selection to get a higher number of events selected.

Chapter 4

Kinematic Observables and Unfolding

4.1 Interesting Kinematic Observables

The comparison between the data and simulation can take on several aspects and goals. The first is to check that there is a control on the analysis selection by checking some control variables like e.g. the momentum of the reconstructed particles. A general agreement is expected and tensions can then be related to a special aspect of the modelling in the simulation. To have access to the expression of the underlying events, the angular distributions between particles are examined and the modelling of these events in the simulation can then be confirmed or questioned. Another goal for the comparison is to check for eventual signs of new physics; a significant discrepancy in the kinematic variables may not be explained by a wrong modelling of the simulation anymore, but could come from a new theory aspect not considered in the standard model and thus not present in the SM simulation. Furthermore, spectra of the invariant mass of the particles or systems of particles are particularly sensitive to unknown resonances and will then be controlled.

Hereafter, the distributions are all presented for the Z+2b selection in the combined channel, meaning a Z boson (decaying to e^+e^- or $\mu^+\mu^-$ pairs) and at least 2 b-tagged jets from the SSVHE algorithm and a cut on the MET significance. The plots comply with the following characteristics: the data

corresponding to 5.0 fb^{-1} are overlaid with the different Monte Carlo simulations normalised to the data luminosity using the theoretical cross sections, and reweighted by the pileup, the lepton scale factors and the b-tagging scale factors as detailed in Section 3.2.6. The last bin in both data and MC distributions contains the overflow. The hatched uncertainty band on the simulation has been obtained by varying the jet energy scale and the b-tagging scale factors to obtain a systematic band which was then combined in quadrature with the statistical uncertainty. The lower plot is the data/MC ratio which shows the separate contributions to the total uncertainty: the band represents the statistical uncertainty on the simulated yield and the lines indicate the uncertainties due to the JES (dashed) and the b-tagging scale factors (solid). All the plots are published in [103] or as additional public material in [150].

The control variables are for example the momentum distributions of the leading and the subleading b jets, shown in Figure 4.1. An overall good agreement is found within the uncertainties, statistical for data and statistical and systematic for the simulation.

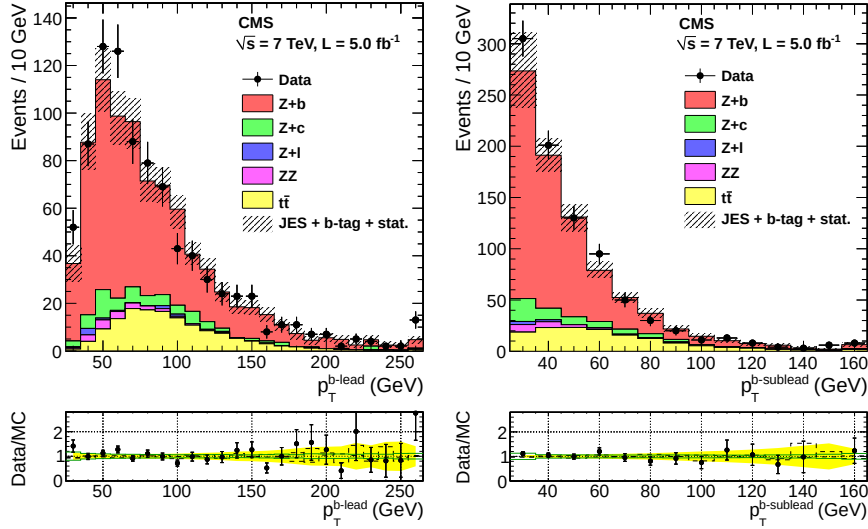


Figure 4.1: The p_T of the leading (left) and subleading (bottom right) b jet.

Kinematic observables of interest in the Z+2b-jets final state, both for the analysis validation and for searches, are the transverse momentum of the dilepton (p_T^Z) and the dijets (p_T^{bb}). More than two jets are b-tagged in less than 2% of the Z+2b events, and in this case the two highest- p_T jets are considered.

The spectrum of the dijets momentum shown in Figure 4.2 (left) shows a good agreement between data and Monte Carlo.

The momentum of the dilepton system shown in Figure 4.2 (right), has a limited agreement between data and Monte Carlo. An overall excess of events can be observed for p_T^Z above 80 GeV, and in particular in the region around 100 GeV. The spectrum is harder in the data than in the predictions.

This trend is consistent with the earlier CMS publication [104], where a similar discrepancy is observed for the variable in the Z+b-jets final state. The effect of the disagreement on the estimate of the cross sections has been studied and found negligible.

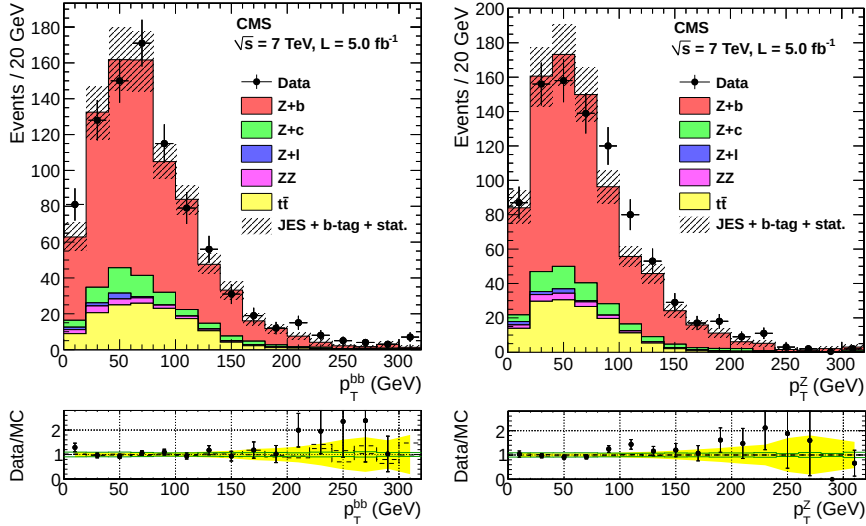


Figure 4.2: The p_T of the dijets bb pair (left) and of the Z boson (right).

The angle between the dilepton and the dijets $\Delta\phi_{Z,bb}$, and the difference between the two b jets in the $(\eta-\phi)$ plane $\Delta R_{b,b}$, are interesting angular distributions of the Z+2b final state. The distribution of $\Delta\phi_{Z,bb}$, shown in Figure 4.3 (left), shows agreement with the predictions as well, both in the collinear and back-to-back regions. This is especially relevant with respect to contributions from Multi Parton Interactions, which are expected to have less correlated kinematics than those from the Z+2b-jets process, and will, therefore, give a uniform distribution in $\Delta\phi_{Z,bb}$. The $\Delta R_{b,b}$ distribution in Figure 4.3 (right) on the contrary shows some tensions between data and simulation at low and high values. In another analysis on the same data sample [138], using a final state presenting a Z and this time two B hadrons instead of two b jets, this

tension was also observed in the ΔR distribution between the B hadrons. It is interesting to notice that two different analyses with different approaches and kinematic cuts, give the same result for the bb system, in a phase space region interesting for modelling in simulations and in particular concerning the gluon splitting.

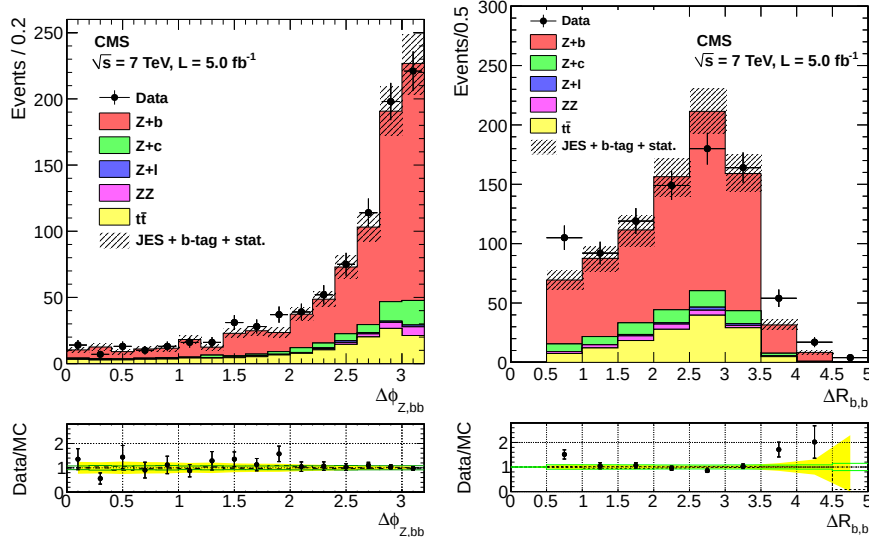


Figure 4.3: Azimuthal angular difference between the Z and the bb system (left) and the difference ΔR between the two b jets (right).

Two of the observables in the Z+2b final state, relevant to searches for undiscovered processes, are the invariant mass of the b-jets pair M_{bb} , and of the Z boson plus dijets system M_{Zbb} . The observable M_{bb} is, for example, used in the study of the Higgs boson produced in association with a Z boson and decaying to two b jets, in the $Z(\ell\bar{\ell})H(b\bar{b})$ final state [151, 152].

The observable M_{Zbb} could show a resonance according to some beyond the standard model theories like the 2 Higgs Doublet Models mentioned in Section 1.2.2.3. Such a resonance could come, for example, from the decay of a heavy pseudo-scalar particle: $A \rightarrow Z(\ell\bar{\ell})H(b\bar{b})$, where H is also a new particle decaying into a b-quark pair.

The two invariant mass distributions are shown in Figure 4.4, and no trace of obvious disagreement between measured and expected distributions can be observed. The highest discrepancy being of one standard deviation for M_{Zbb} around 450 GeV.

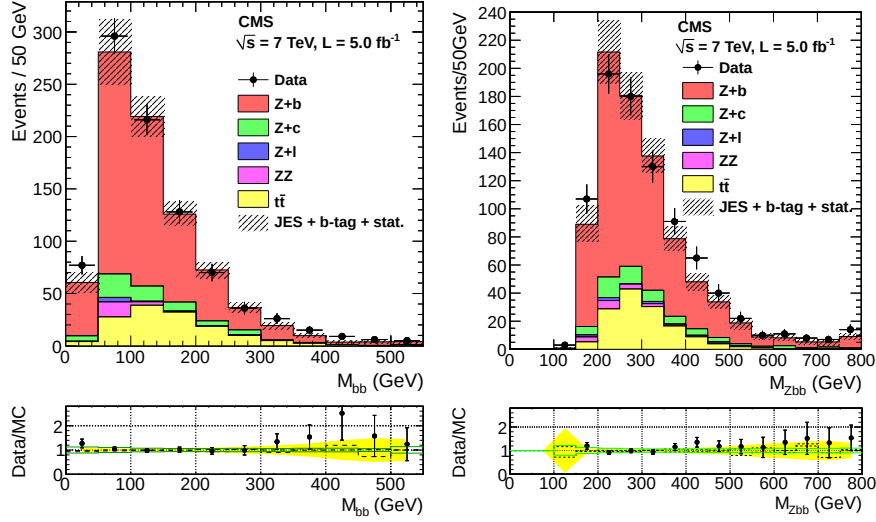


Figure 4.4: The invariant mass of the b -jets pair (left) and of the Z boson plus b -jets pair system (right).

It is interesting to mention that a bin-by-bin reweighting of the predictions according to the observed discrepancy in the p_T^Z observable has been performed, which improves also the agreement in other observables where differences are observed like M_{Zbb} . This behaviour drives us to think that the fluctuations are not only statistical, but have a physics meaning.

Just like for the calculation of the absolute cross section, for which one could not use directly the number of events measured as pure signal, the kinematic distributions still contain the background contribution and are affected by all the inefficiencies of reconstruction and selection of the objects and the resolution of the detector.

When the comparison is done at the reconstructed level this implies that for each generator a complete production and reconstruction has to be done for the analysis. This is highly time consuming and can introduce errors and systematic uncertainties.

The comparison is given here with only one type of Monte Carlo simulation, so especially for the signal one looks at only one type of modelling of the Z + b -jets process. Here it comes from MADGRAPH in the 5-flavour scheme. The ideal situation would be to be able to compare the distributions with all the available types of generator with the goal to see if the tensions can be explained

and fixed by the modelling, or if a better description is emerging from a certain simulation. For example a harder spectrum for the p_T^Z observable might be predicted in 4-flavour calculations with massive b quarks at next-to-leading order, which might explain the observed disagreement.

The solution would then be to do just like for the cross section measurement, and to estimate the distributions at the generator level from the reconstructed level.

In this way a differential cross section per bin of the considered variable distribution is obtained, and can be compared at generator level to the available generators distributions in the acceptance.

This would allow anyone, even non-CMS collaboration member, to perform further studies with any other generator.

The process to go from a reconstructed distribution from data to the generator level, while correcting for the detector effect, is called ‘unfolding’ and can be performed by means of different methods.

4.2 Unfolding

4.2.1 Unfolding Methods

As in any experiment, the measured observables of the analysis differ from the corresponding true physical quantities because of the physical process and the detectors. One is able to build up variables which have a physical meaning alike the true genuine quantities, but with respect to those ones they will be deformed due to the selection and reconstruction of the objects implied, to the background noise and to the limited performance of the detector. The true distribution is then related to the measured distribution by migrations, distortions and transformations. The inversion of this measured distribution requires the knowledge of all the distortion effects on the true quantity. This can be acquired by the intermediate of a Monte Carlo simulation.

In the case of particle detection the true quantity is the particle at generator level, while the measured quantity originates from the detected particles in the detector. The detection depends on the effects of the detector such as resolution and efficiencies, but also on the contaminations by other processes mimicking the presence and nature of the particle.

The most naive method to get back to the true distribution would be to perform a bin-by-bin correction, meaning to evaluate the ratio, in a certain bin, of the number of events reconstructed over the number of true events. This efficiency

estimated from simulation can then be applied to the same bin for the data distribution in order to evaluate the number of true events in this bin. However, this method is completely discarding the migrations between different bins, and, moreover, neglects the fact that the smearing can be larger than the bin size. Hence, this cannot be considered as a reliable unfolding method. A more correct way to solve the problem of migrations between different bins is to build from the simulation true and measured distributions a matrix of size: $N_{bins} \times N_{bins}$, where N_{bins} is the number of bins identical in the true and measured distribution of the variable. This matrix can then be inverted and applied on the measured data distribution to get the true one. However, this method rises the problem of the singular matrix since there is no reason for the matrix to be invertible. Moreover, even if the matrix can be inverted this method cannot deal with large statistical fluctuations: for example in the case of a particle momentum distribution the build up matrix is expected to be mostly diagonal with migrations whose values are getting smaller as one drifts away from the diagonal. The direct inverse of those small fluctuations would then be huge values which do not have any physical meaning. The problem is said to be ill-posed, hence, unstable against small variations in the initial system. To overcome the problem, different techniques have been set up based on different approaches. The most commonly used in particle physics are the Iterative Bayes [153, 154], and the Singular Value Decomposition [155, 156] methods. They are presented below and they will be tested for the variables to unfold in this analysis.

4.2.1.1 Bayesian Unfolding

As expected, the procedure of Bayesian Unfolding is based on an iterative use of the Bayes theorem, which is enunciated below as a reminder:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (4.1)$$

The unfolding method [153, 154] can be explained by the means of causes C and effects E . The causes correspond to the true values before smearing and the effects correspond to the measured values after smearing by detector effects.

Each cause can result in a different effect but for one given effect, meaning one bin of the measured variable, the original cause is not known. However, the probability for an effect E_j to happen knowing the prior cause C_i , $P(E_j|C_i)$, can be estimated via the knowledge of the smearing. This knowledge is acquired by Monte Carlo simulation and will account for all the migrations, efficiency and resolution effects. The indices i and j can be linked to the bin number in

the true or cause distribution and in the measured or effect distribution, and $P(E_j|C_i)$ is in fact the element R_{ij} of the so-called smearing or response matrix R of size $N_E \times N_C$.

The goal is then to estimate the probability of a cause C_i to produce the effect E_j , which can be estimated via the Bayes theorem as:

$$P(C_i|E_j) = \frac{P(E_j|C_i)P_0(C_i)}{\sum_{l=1}^{N_C} P(E_j|C_l)P_0(C_l)}, \quad (4.2)$$

where $P_0(C_i)$ is the initial probability or prior and $\sum_{l=1}^{N_C} P_0(C_l) = 1$, and N_C is the total number of causes, which means the total number of bins in the true definition.

If one observes $n(E_j)$ events in the j^{th} bin of the measured variable (with a total of N_E bins) the expected number of events in the cause bin i can be evaluated by:

$$\hat{n}(C_i) = \frac{1}{\epsilon_i} \sum_{j=1}^{N_E} n(E_j) P(C_i|E_j) \quad \text{with} \quad \epsilon_i \neq 0, \quad (4.3)$$

where $\epsilon_i = \sum_{j=1}^{N_E} P(E_j|C_i)$ is the efficiency of detecting the cause C_i in any of the possible effects. It can be also formulated as the probability that the cause i has an effect. If $\epsilon_i = 0$ then $\hat{n}(C_i)$ will be set to zero since the experiment is not sensitive to the cause C_i .

The estimated number of events in each cause bin i can be expressed as a function of the response matrix bin $R_{ij} = P(E_j|C_i)$:

$$\hat{n}(C_i) = \frac{1}{\sum_{j=1}^{N_E} P(E_j|C_i)} \sum_{j=1}^{N_E} n(E_j) \frac{P(E_j|C_i)P_0(C_i)}{\sum_{l=1}^{N_C} P(E_j|C_l)P_0(C_l)}. \quad (4.4)$$

In the Formula 4.4 one can question the fact that the probabilities are depending on a prior knowledge of the cause. In reality the Bayes formula has the power to increase the knowledge of $P(C_i)$ with the increasing number of observations.

An iterative process works as follow:

- 1) Choose the prior distribution $\mathbf{P}_0(C)$ from the best knowledge of the process under study, which will be here the true generated distribution from simulation. The initial expected number of events $n_0(C_i) = P_0(C_i) \times \sum_{k=1}^{N_C} n_0(C_k)$.
- 2) Calculate $\hat{\mathbf{n}}(C)$ for each bin i from Equation 4.4, with the knowledge of $P_0(C_i)$ from the generated distribution, $n(E_j)$ from the measured distribution and $P(E_j|C_i)$ from the response matrix.
- 3) Calculate $\hat{P}(C_i) = \frac{\hat{n}(C_i)}{\sum_{k=1}^{N_C} \hat{n}(C_k)}$.
- 4) Make a χ^2 comparison between $\hat{\mathbf{n}}(C)$ and $\mathbf{n}_0(C)$.
- 5) Replace $\mathbf{P}_0(C)$ by $\hat{\mathbf{P}}(C)$, and $\mathbf{n}_0(C)$ by $\hat{\mathbf{n}}(C)$. If after the second iteration the value of χ^2 is small enough stop the iteration, otherwise go to step 2.

The regularisation parameter specifies the number of iterations, starting with the training sample truth (iterations=0). One should choose a small integer greater than 0. Since only a few iterations are needed, a reasonable performance can usually be obtained without fine-tuning the parameter. Empirically, one learns that in most of cases only two or three steps are necessary to recover the true spectrum with accuracy [157]. Most of the time four iterations is chosen as regularisation parameter.

Estimation of the uncertainties

After the iterative process the unfolded distribution $\hat{\mathbf{n}}(C)$ has been obtained. However, the physical Poisson distribution is $\mathbf{n}(E)$ and not $\hat{\mathbf{n}}(C)$ so the uncertainties cannot be taken simply as the square root of these numbers. Besides, since the number of events of one effect bin $n(E_j)$ is shared between several different causes, the uncertainties on $n(C_i)$ are correlated. The Equation 4.4 can be rewritten as:

$$\hat{n}(C_i) = \sum_{j=1}^{n_E} M_{ij} n(E_j), \quad (4.5)$$

where

$$M_{ij} = \frac{P(E_j|C_i)P_0(C_i)}{\sum_{l=1}^{N_C} P(E_j|C_l)P_0(C_l) \sum_{l=1}^{N_E} P(E_j|C_i)}. \quad (4.6)$$

M_{ij} can be seen as the elements of the unfolding matrix M that allows to go from the measured to the generated distribution, which is obviously not the mathematical inverse of the response matrix R .

One wants to consider the different sources of uncertainties contributing to the covariance matrix of $\hat{n}(C_i)$: V . The complete estimation of those uncertainties are detailed in [153], with one omission which was added in the reference paper of the RooUnfold package [158], as well as in the package itself [159], which will be discussed in Section 4.2.2. The covariance matrix is calculated by error propagation. In [153], M_{ij} is considered to be independent of $n(E_j)$ which is in fact only true for the first iteration. In the following iteration $n_0(C_i)$ is replaced by the new estimation $\hat{n}(C_i)$ which depends on $n(E_j)$. To take this dependence into account one should compute the error propagation matrix, which can be then used to obtain the covariance matrix on the unfolded distribution (see [158] for details):

$$\sum_{i,j=1}^{n_E} \frac{\partial \hat{n}(C_k)}{\partial n(E_i)} V(n(E_i), n(E_j)) \frac{\partial \hat{n}(C_l)}{\partial n(E_j)}, \quad (4.7)$$

where $V(n(E_i), n(E_j))$ is the covariance matrix of the measurements.

4.2.1.2 Singular Value Decomposition

Another idea is to provide a regularisation method re-formulated in terms of the Singular Value Decomposition (SVD) of the response matrix $\hat{A}$. This regularisation method allows the suppression of spurious, quickly oscillating components of the solution, and leaves only the statistically significant terms.

Here below will be followed the discussion detailed in [155].

A SVD of a real $m \times n$ matrix A is its factorisation of the form:

$$A = USV^T, \quad (4.8)$$

where U is an $m \times m$ orthogonal matrix, V is an $n \times n$ orthogonal matrix, while S is an $m \times n$ diagonal matrix with non-negative diagonal elements:

$$\begin{aligned} UU^T &= U^T U = I, & VV^T &= V^T V = I, \\ S_{ij} &= 0 \text{ for } i \neq j, & S_{ij} &\equiv s_i \geq 0. \end{aligned} \quad (4.9)$$

The quantities s_i are called *singular values* of the matrix A , and columns of U and V are called the left and the right *singular vectors*, respectively. If A itself is orthogonal all $s_i = 1$, while if A is degenerated at least one $s_i = 0$. The s_i can be ordered in decreasing order by swapping simultaneously corresponding columns of U and V .

In our case if one has an initial high statistics MC sample one can create the matrix simulating the detector response. In other words one has access to the probability for an event generated in the true bin j of the *truth* distribution x^{ini} , to be found in the bin i of the measured distribution b^{ini} :

$$\sum_j \hat{A}_{ij} x_j^{ini} = b_i^{ini}. \quad (4.10)$$

The goal is to determine the *truth* distribution x from a real data measured distribution b given the probability matrix $\hat{A}_{ij}$ built from simulation:

$$\hat{A}x = b. \quad (4.11)$$

The problem clearly enters the *ill-defined* category since there are inevitable statistical errors on the measured distribution and because $\hat{A}$ is most likely to be degenerate, so any direct and ‘exact’ solution of the problem will give unphysical spurious oscillations.

With SVD the linear system of equations $\hat{A}x = b$ can be diagonalised by introducing the rotated vectors z and d , such that the solution corresponds to the following ensemble:

$$\begin{aligned} USV^T x = b &\Rightarrow z \equiv V^T x, \quad d \equiv U^T b, \\ s_i z_i = d_i &\Rightarrow z_i = \frac{d_i}{s_i} \Rightarrow x = Vz. \end{aligned} \quad (4.12)$$

However, at this stage the system is still not solvable directly due to the remaining errors in b but also due to the fact that some singular value s_i maybe be very small or null which might magnify the contributions of fluctuations or poorly known coefficients. In order to make the equations and their unknowns more uniform the system has to be rescaled in a way that does not suppress the significant information while the non-significant information is not enhanced. A suggestion on how to rescale can be given by looking at the initial problem $\hat{A}x = b$ from another point of view and seeing that it represents the solution of the following least square problem:

$$\sum_i \left(\sum_j \hat{A}_{ij} x_j - b_i \right)^2 = \min, \quad (4.13)$$

which is appropriate only if the variations in error of b are suppressed to give all equations the same significance. This ends up in the following weighted square problem:

$$\sum_i \left(\frac{\sum_j \hat{A}_{ij} x_j - b_i}{\Delta b_i} \right)^2 = \min, \quad (4.14)$$

where Δb_i is the error in b_i . This can also be expressed in a more general way like:

$$(\hat{A}x - b)^T B^{-1} (\hat{A}x - b) = \min, \quad (4.15)$$

where B is the covariance matrix of the measured vector b .

Rescaling of the unknowns

If the data and the MC are highly fluctuating functions one might want to estimate x relatively to x^{ini} used to build the matrix $\hat{A}$. This can be done by considering a new unknown vector $w_j = x_i/x_j^{ini}$ and by multiplying each columns of A by x_j^{ini} . This is changing the matrix from probability to ‘number of events’ matrix which is undoubtedly more useful since the bins with higher statistics and hence more significance are enhanced. After this the system becomes:

$$\sum_j A_{ij} w_j = b_i. \quad (4.16)$$

Rescaling of the equations

The idea of the algorithm is to make all the equations ‘equal’ by checking that the error on the r.h.s. (right-hand side) is always ± 1 , which means that the covariance matrix B of the r.h.s. must be made equal to the unit matrix I . After dividing each equation by the corresponding error Δb_i one obtains a balanced system. After this rescaling the system $Aw = b$ has turned into:

$$\sum_j \tilde{A}_{ij} w_j = \tilde{b}_i. \quad (4.17)$$

Regularisation

The rescaled system is the following:

$$\begin{aligned} \tilde{A}w = \tilde{b} &\Rightarrow \tilde{A} = USV^T \Rightarrow USV^T w = \tilde{b} \Rightarrow z \equiv V^T w, \quad d \equiv U^T \tilde{b}, \\ s_i z_i = d_i &\Rightarrow z_i = \frac{d_i}{s_i} \Rightarrow w = Vz. \end{aligned} \quad (4.18)$$

By construction all the d_i are independent and present errors of ± 1 . If $s_i \neq 0$ then $\{z_i\}$ forms the exact solution of the initial system, but the exact solution of Equation 4.17 still gives rapidly oscillating distributions, with smaller amplitude but still not usable.

If the general weighted least square system from Equation 4.15 is considered on the new rescaled system, the later is equivalent to the minimisation of the residual χ^2 :

$$\chi^2 \equiv (\tilde{A}w - \tilde{b})^T (\tilde{A}w - \tilde{b}) = \min. \quad (4.19)$$

The spurious oscillating components can be suppressed by adding an *a priori* requirement that the regularised solution is smooth, which technically is achieved by introducing into the χ^2 minimisation a *regularisation* or *stabilisation* term [160, 161]:

$$\chi^2 \equiv (\tilde{A}w - \tilde{b})^T (\tilde{A}w - \tilde{b}) + \tau \cdot (Cw)^T Cw = \min, \quad (4.20)$$

where C is a matrix which defines the *a priori* condition on the solution, and τ is the regularisation parameter determining the relative weight of this condition. C should be chosen from general considerations, and a crucial one is that the solution histogram w should be smooth with small bin-bin variations. Therefore Cw is chosen to be the second finite difference of w , or in other words the ‘curvature’ of the discrete distribution w_j which is defined as the sum of the squares of its second derivative. This leads to the expression:

$$(Cw)^T Cw = \sum_i [(w_{i+1} - w_i) - (w_i - w_{i-1})]^2. \quad (4.21)$$

The new minimisation problem leads to the following over-constrained system:

$$\begin{bmatrix} \tilde{A} \\ \sqrt{\tau} \cdot C \end{bmatrix} w = \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix}. \quad (4.22)$$

One could call the SVD for each value of τ , but the more efficient *damped least squares* method is available [161]. For $\tau = 0$ the problem is non regularised and Equation 4.22 is equivalent to Equation 4.17 and can be solved with SVD. First, the regularisation term is made proportional to the unit matrix I :

$$\begin{bmatrix} \tilde{A}C^{-1} \\ \sqrt{\tau} \cdot I \end{bmatrix} Cw = \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix}. \quad (4.23)$$

The matrix C is singular and needs to be regularised before to be inverted. One of the methods is to add a small diagonal component, large enough to make the inversion possible, but small enough to not change the condition of minimum curvature (see [155] for details). The product of matrices $\tilde{A}C^{-1}$ can be decomposed with SVD as:

$$\tilde{A}C^{-1} = USV^T, \quad (4.24)$$

with U and V orthogonal and S diagonal with s_i . Then, as performed in Equation 4.18, $\tilde{b}$ and Cw are rotated and the system looks like:

$$s_i z_i = d_i \quad z \equiv V^T Cw, \quad d \equiv U^T \tilde{b}. \quad (4.25)$$

Since the covariance matrix of $\tilde{b}$ was made equal to the unit matrix, it will also be the case for the covariance matrix of d . The exact solution of the non-regularised system is then:

$$z_i^{(0)} = \frac{d_i}{s_i}, \quad w^{(0)} = C^{-1} V z^{(0)}. \quad (4.26)$$

Using the detailed procedure of [161], which in summary proves that introducing a non-zero τ is equivalent to changing d_i by a regularised distribution of the form:

$$d_i^{(\tau)} = d_i \frac{s_i^2}{s_i^2 + \tau}, \quad (4.27)$$

the solution becomes:

$$z_i^{(\tau)} = \frac{d_i}{s_i} \cdot \frac{s_i^2}{s_i^2 + \tau}, \quad w^{(\tau)} = C^{-1} V z^{(\tau)}. \quad (4.28)$$

One can see how τ is regularising the singularities due to small s_i 's: for large, significant $s_i \gg \tau$ the suppression factor $s_i^2/(s_i^2 + \tau)$ is close to 1; but for small s_i it works as a low-pass filter. And so for a smooth way to eliminate the wildly oscillating contributions (values of i with small s_i and non-significant d_i) one should choose $\tau = s_k^2$, where k is the index of the last significant d . In practice what one is actually asked to provide is the k index and not the value of τ . This information can be obtained by plotting the d_i distribution or better the $\log |d_i|$ versus i . On the plot two patterns should be visible: for small i , d_i should be statistically significant and fall gradually (most likely exponentially) towards a Gaussian distributed random value, which will in many cases appear first like a plateau before to fluctuate. So the critical value $i = k$ is usually clearly visible on the plot of $\log |d_i|$ as the value of i where the behaviour of d_i changes from exponentially falling to a constant. The advantage of the k index over the value of τ is that it is an integer comprised between 0 and the number of bins of the considered distribution, which makes it easier to discuss.

The solution for the regularised version of an example of initial system $\hat{A}x = b$ is taken from [155] and shown in Figure 4.5.

Error propagation

The covariance matrix of the exact ($\tau = 0$) solution was diagonal. The covariance matrix of the regularised $z_i^{(\tau)}$ solution of Equation 4.28 is still diagonal, because the errors corresponding to insignificant d_i are suppressed.

$$Z_{ik}^{(\tau)} = \frac{s_i^4}{(s_i^2 + \tau)^2} \delta_{ik} \quad (4.29)$$

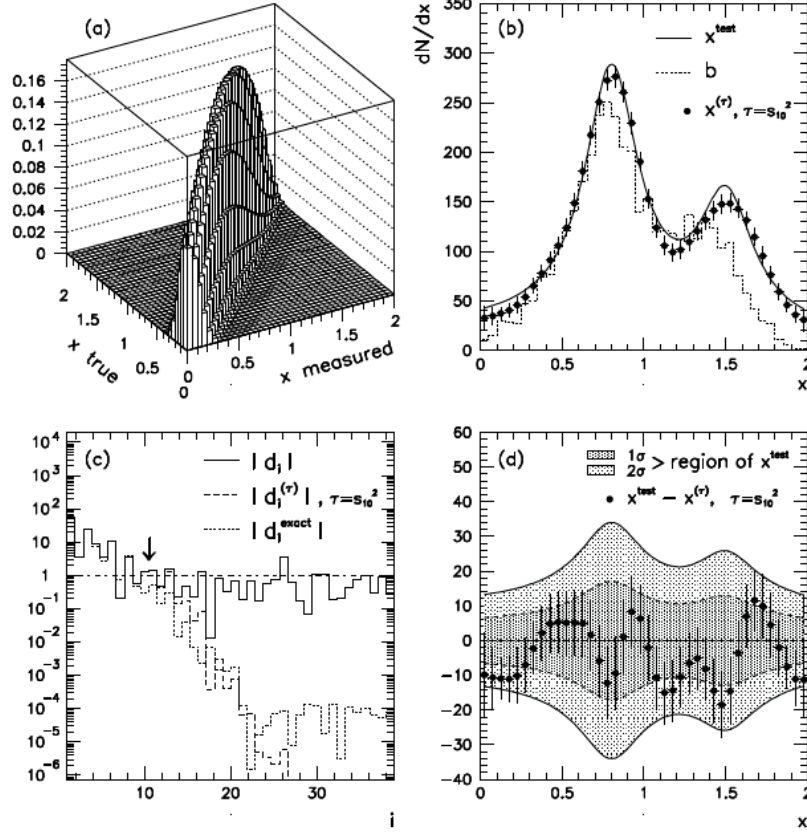


Figure 4.5: Academic example from [155] of the unfolding of a simulated counting experiment. **(a)** The response matrix $\hat{A}$. **(b)** The true distribution x^{test} compared to the measured histogram b and the unfolded distribution $x^{(\tau)}$ for $\tau = s_{10}^2$. **(c)** The absolute values of d_i compared to the regularised d_i (dashed line) and to the one unaffected by the statistical fluctuations (dotted line). The arrow indicates the boundary between significant and non-significant equations. **(d)** The deviation of the unfolded distribution from the true one, compared to 1σ and 2σ error bands.

These errors can be propagated back to the covariance matrices of w and x :

$$W^{(\tau)} = C^{-1} V Z^\tau V^T (C^T)^{-1}, \quad (4.30)$$

$$X_{ik}^{(\tau)} = x_i^{ini} W_{ik}^{(\tau)} x_k^{ini}.$$

For any non-zero τ the covariance matrices in Equation 4.30 are not diagonal, and the larger the value of τ , the larger the bin-to-bin correlations in errors. The value of τ (or in the RooUnfold package the value of k), the unfolded vector x and its covariance matrix X form the solution of the unfolding problem.

4.2.2 Performing the Unfolding

The RooUnfold package [159] was designed to provide a framework for the different unfolding algorithms using existing ROOT [162] classes. The implemented interfaces are for the Iterative Bayes and the Singular Value Decomposition (SVD), presented above, but also for TUnfold [163, 164], bin-by-bin correction factors, and unregularised matrix inversion methods. The two latter methods are strongly discouraged by statistics experts, one for the strong biases from the Monte Carlo model on which the method relies and the second for the risks to enhance statistical fluctuation and to have large bin-bin correlations. The Figure 4.6, taken from [158], shows how the classes are working together. The user has to fill his ‘training’ samples from the MC simulation, build the response matrix, and finally provide the data distribution to be unfolded.

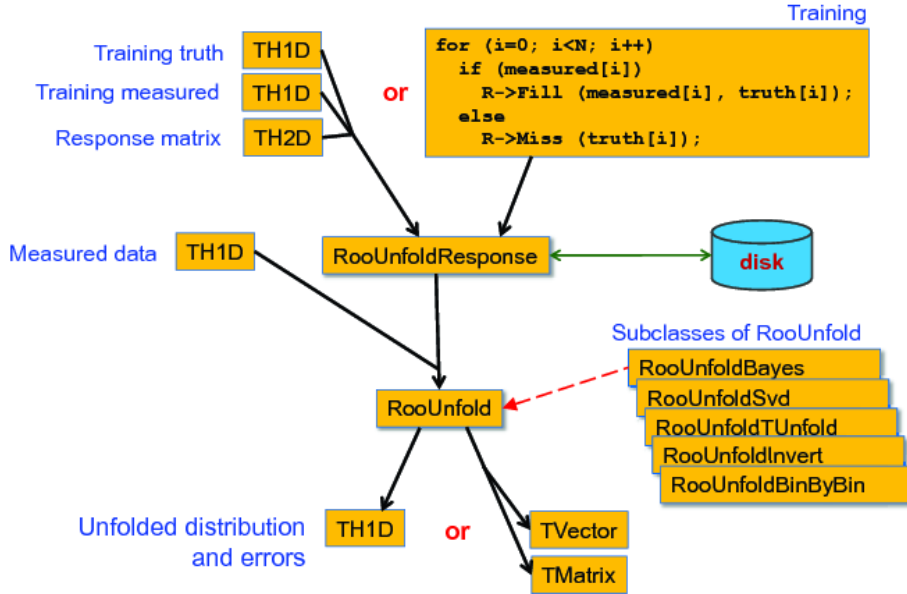


Figure 4.6: The RooUnfold classes from [158]

The complete process of the unfolding will be detailed in the next section, taking the example of the transverse momentum of the Z boson distribution in the electron channel, p_T^Z . The same procedure will then be identically applied to all the other eight interesting variables shown in Section 4.1 and for the three channels (electron, muon and combined). The choice to show the separate electron (e^+e^-) and muon ($\mu^+\mu^-$) channels in addition to the combined one is driven by the fact that some differences in the shape of the distribution from the

two channels can appear. With the relatively small amount of events selected in the analysis, the gain in events in the combined channel could not always compensate for the loss of information in the combination. One can have a feeling of this by looking at the distribution of the p_T^Z for the three channels in Figure 4.7.

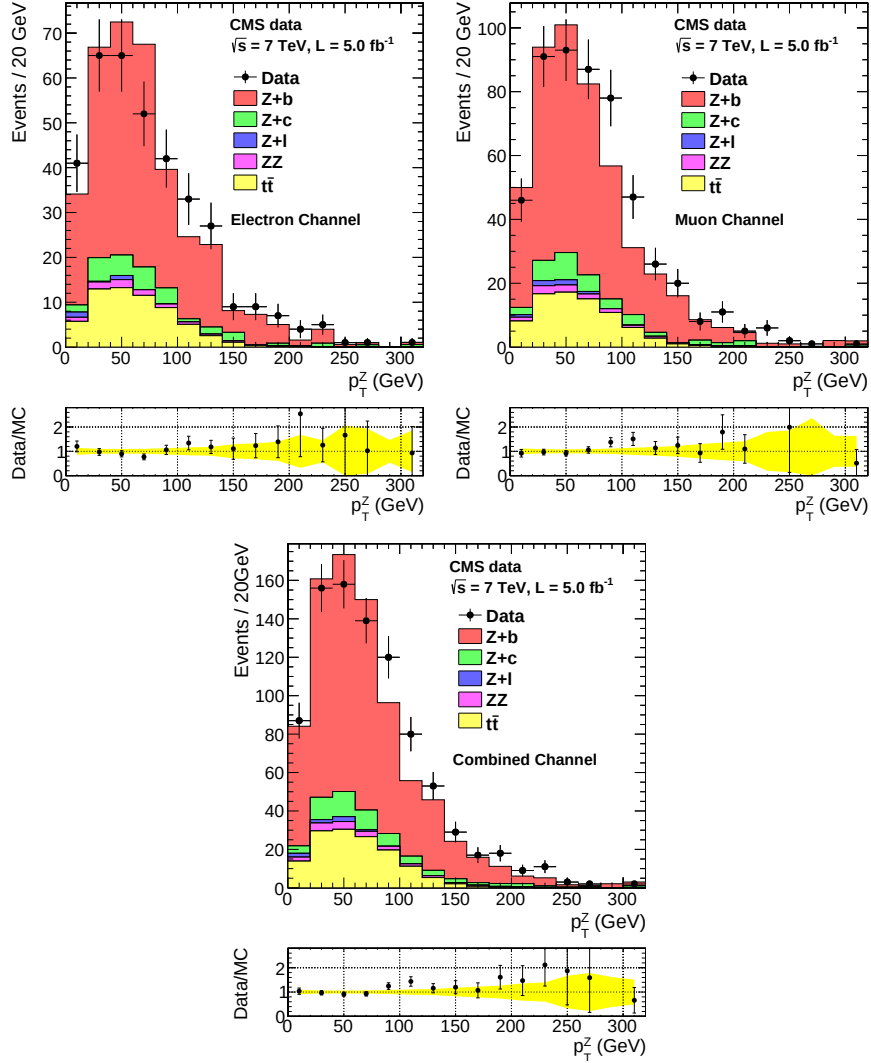


Figure 4.7: The p_T of the Z boson in the electron (top left), muon (top right) and combined channel (bottom).

4.2.2.1 Construction of the Response Matrix

The response matrix is constructed from three inputs:

- **The generator level distribution**, which is defining the generator level and acceptance for the unfolding. In this study one defines as ‘generator level Zbb events’ events with a Z boson build up from two dressed leptons in the acceptance cuts, and two genjets matched with two B hadrons in the acceptance cuts. The acceptance cuts being the ones of the analysis, summarised in Section 3.2.5.

The dressed leptons are constructed by selecting the leptons at the end of the decay chain in the PYTHIA parton shower [57], when they have radiated their photons (leptons of status 1 in PYTHIA). They are subsequently ‘dressed’ with the surrounding photons in a cone of size $\Delta R = 0.1$, which corresponds to the detector resolution. The dressed leptons are paired to form a Z candidate and are asked to be within the acceptance cuts. The selected Z will be the one with the mass the closest to the nominal mass of the Z boson (91.1876 GeV [2]). The genjets are required to match a B hadron, meaning to stand in the vicinity of a B hadron within a cone of $\Delta R = 0.5$. They are then required to have a $p_T > 25$ GeV, $|\eta| < 2.1$ and not to overlap with the selected leptons within a cone of $\Delta R = 0.5$. These are the cuts of the analysis presented in Chapter 3.2.

- **The reconstructed level distribution**, which is constituted of all the events passing the cuts of the analysis for the reconstructed objects as described in Section 3.2 and summarised in Table 3.7. This means asking for the presence of a Z candidate from reconstructed leptons within the acceptance and of two reconstructed b-tagged jets also in the acceptance. In this selection a distinction can be made between events for which the reconstructed jets are matched to the genjets or not. The latter are so-called ‘fakes’ in the RooUnfold jargon that will be adopted in this chapter.

- **The dimension-two matrix of the generated distribution versus the measured one** which is properly speaking the response matrix. The later has the dimensions of the number of bins of the two distributions (which ought to be the same). The measured distribution is this time only a subsample of the generator level distribution for which the events after passing the generator level selection, are also passing the cuts of the analysis for the reconstructed objects. It corresponds to the reconstructed distribution minus the fake distribution.

The Drell–Yan + jets sample used here for the construction of the response matrix is the same that was used for the construction of the samples Zb, Zc and Zl of the analysis (see Table 3.3), as well as for the estimation of the efficiencies and migrations in the calculation of the cross section in Section 3.4. This

generation of the simulated events was done with the MADGRAPH generator, set up for a 5-flavour scheme (MG5F).

The Figure 4.8 (left) presents this response matrix for the dilepton transverse momentum distribution in the electron channel. The associated ‘generated’ and ‘reconstructed’ distributions are also shown in Figure 4.8 (right).

Initially, the ‘fake’ distribution is supposed to be constituted of events passing the reconstructed level cuts and not the generator level cuts. However, in the present study this contribution is already taken into account in the background subtraction as detailed in the next Section 4.2.2.2. The only events which still have to be considered by the unfolding, and that will enter the ‘fake’ distribution by the intermediate of the difference between the events selected in the response matrix and in the reconstructed distribution, are the ones for which the matching is not done between reconstructed and generated jet objects. This contribution is however relatively small as one can see in Figure 4.8 (right).

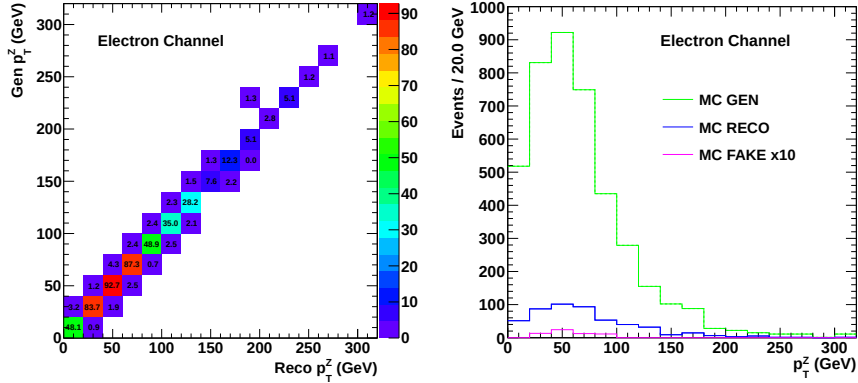


Figure 4.8: The response matrix of the p_T of the Z distribution (left) in the electron channel, and the associated ‘generated’, ‘reconstructed’ and ‘fake’ distributions (right).

One can notice that, compared to the definition of the generator level in the ‘unfolding’ of the cross section, Section 3.4.1, the definition of a generator level Z has been slightly changed, in the sense that all the leptons are here considered to build a Z candidate, and not anymore only the ones for which their resulting from a Z decay can be checked. This has been chosen in the scope of the comparison of the data with *any* generator, which do not necessarily keep the ancestor informations of the leptons in their event record. This is sensible, since quantum-mechanically the candidate could come from a Z boson or a photon

which interfere with each other. Moreover just like in reconstructed data, the leptons can now be non-prompt, coming from conversion for example.

Tests have been performed with the same Z candidate selection than for the cross section unfolding, and the results for the cross sections are coherent, within the uncertainties, for all channels (electron, muon and combined).

4.2.2.2 Background Subtraction

The data to be unfolded should contain only measured ‘Zbb’ events. The remaining contribution from the backgrounds, $t\bar{t}$, ZZ and Z+light-jets or Z+c-jets, in the data has thus to be subtracted. For convenience the contributions coming from the DY+jets sample will be redefined as ‘Zbb’ and ‘Zxx’ samples, with:

- **‘Zbb’**: events presenting a generator level Z boson build up from two dressed leptons in the acceptance cuts, two genjets matched with two B hadrons in the acceptance cuts.
This definition is exactly the same than the one for the generator level distribution used to build the response matrix.
- **‘Zxx’**: the orthogonal sample from ‘Zbb’, meaning all the events of the DY + jets sample which did not pass the ‘Zbb’ selection.
This includes Z+bc, Z+bl, Z+lc, Z+ll, Z+cc, and the Z+bb outside of the acceptance cuts defined in the ‘Zbb’ sample.

Then those generator level events will have to pass the reconstructed level cuts of the analysis as described in Section 3.2 and summarised in Table 3.7. Figure 4.9 shows the data compared to the simulation for the transverse momentum of the Z distribution with this new definition of the contributions from the DY + jets sample. The total number of simulated events is the same at the permil level than for the repartition between ‘Zb’, ‘Zc’ and ‘Zl’ in the analysis, as was shown in Figure 4.2.

The Zxx events are thus the ‘fake’ events described in the previous section. They do not have to be taken into account in the unfolding since they will be removed from the data to be unfolded.

The remaining events from backgrounds Zxx, $t\bar{t}$ and ZZ must be removed from the data. For this the background simulation shapes will be used, but they will be rescaled by the estimations obtained from data of the b purity and of the fraction of $t\bar{t}$, performed in Section 3.3. The background coming from ZZ does

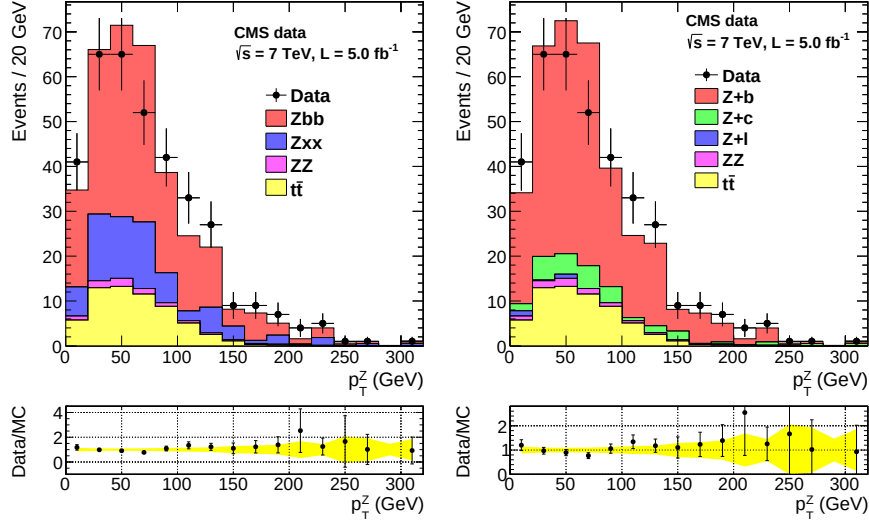


Figure 4.9: Comparison of data and simulation for the p_T of the Z distribution in the electron channel. The $DY + \text{jets}$ sample being divided at generator level according to the new definitions (left). The division according to the analysis definition being shown for comparison (right).

not need to be rescaled as it was already normalised to the cross section from the CMS measurement [139].

One could also think about doing a data driven shape estimation of the backgrounds, for example for $t\bar{t}$ by doing the selection on a dilepton sample of opposite flavours $e^\pm\mu^\mp$, but the available amount of data at 7 TeV is too low for such an exercise.

The computations of the scale factors of the Drell–Yan background SF_{DY} , and of the $t\bar{t}$ background $SF_{t\bar{t}}$, are shown in the Formulas 4.31.

$$\begin{aligned}
 SF_{DY} &= \frac{N_{data}^{tot} \times (1 - P_b^{2b})}{N_{Zxx}^{tot}} \\
 SF_{t\bar{t}} &= \frac{N_{data}^{tot} \times f_{t\bar{t}}^{2b}}{N_{t\bar{t}}^{tot}}
 \end{aligned} \tag{4.31}$$

The value of those scale factors are presented per channel in Table 4.1, using the values of the purity and the $t\bar{t}$ fraction obtained previously by channel (see Table 3.22). The systematic uncertainties related to the b purity and $t\bar{t}$ fraction, are propagated to the background subtracted data, bin-by-bin, by addition in quadrature with the statistical uncertainties. They will then be propagated

with the statistical uncertainties by the unfolding to the unfolded data distribution. Since they are intrinsically bounded to the data to be unfolded, for simplicity of representation and notation the total uncertainty on the unfolded data will be called only statistical. The Figure 4.10 shows the process of the background subtraction for the transverse momentum of the Z boson, in the electron channel.

Table 4.1: *Estimation of the background rescaling scale factors for the 5.0fb^{-1} luminosity data sample.*

Channel	Electron	Muon
P_b^{2b}	$(74.1 \pm 7.3)\%$	$(75.1 \pm 6.4)\%$
$f_{t\bar{t}}^{2b}$	$(14.0 \pm 2.3)\%$	$(13.0 \pm 1.9)\%$
SF_{DY}	1.28 ± 0.29	1.28 ± 0.28
$SF_{t\bar{t}}$	0.82 ± 0.16	0.85 ± 0.18
SF_{ZZ}	1.0	1.0

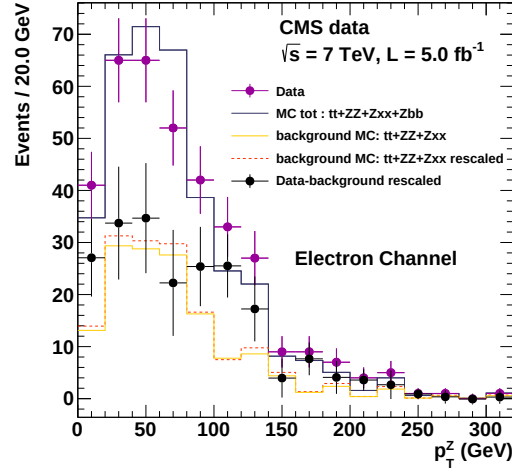


Figure 4.10: *Background subtraction on the data distribution from the 5.0fb^{-1} luminosity data sample. Initial background and total (signal plus background) MC distributions are shown, as well as the rescaled background distribution. Initial and subtracted data distributions are shown with their error bars (comprising the systematic uncertainties for the background-subtracted data).*

4.2.2.3 Choice of Method and Regularisation Parameter

Within the available methods of unfolding, focus is made on the Singular Value Decomposition and the Bayesian unfolding, which are today the state of the art of unfolding methods [165]. As a background subtraction is performed that might render the data to be unfolded non Poissonian anymore [166], the Bayesian method is only used as a cross check, and the SVD becomes the main method.

Once the response matrix is specified, the unfolding is performed based on the chosen algorithm. One needs to pass as an argument the so-called ‘regularisation parameter’. In the case of SVD the optimal parameter is the integer k in the expression of $\tau = s_k^2$, where k is the index of the last significant term of the vector d related to the measured distribution, and being defined in Equation 4.27. As already mentioned the information is obtained by plotting $\log |d_i|$ versus i . This is done in Figure 4.11 and one can see that the value $i = k$ where the behaviour of d_i changes from exponentially falling to a gaussian distributed constant can be evaluated to be $k = 9$.

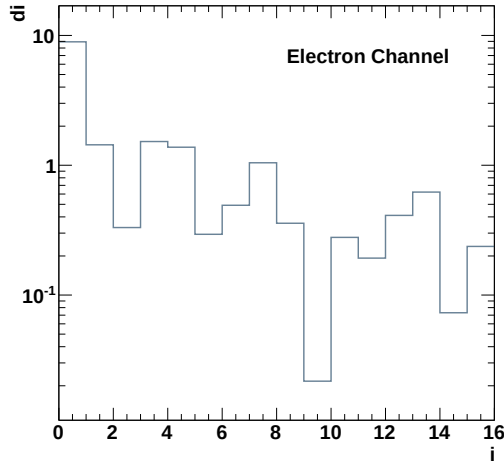


Figure 4.11: The decomposition of the vector d versus the parameter i in logarithmic scale, corresponding to the total number of bins of the concerned distribution (here p_T of the Z boson). For $i = k = 9$ the behaviour of vector d changes and this value is chosen as regularisation parameter to perform the SVD unfolding.

The SVD unfolding with $k = 9$ is applied first on the Monte Carlo reconstructed distribution, which by construction should give an unfolded distribution identical to the generated one. This closure test performed for the p_T of the Z distribution, is shown in Figure 4.12 (left) where the red unfolded distribution is perfectly in agreement with the generated one. The error bars displayed only for the unfolded distribution are statistical and were propagated in the unfolding.

The Bayesian unfolding is also performed on simulations first. The regularisation parameter is this time the number of iterations. In this case the unfolding is very stable in front of the number of iterations and 4 iterations are sufficient. One can use a higher number but as shown on the Figure 4.12 (right) the unfolded result will not be affected.

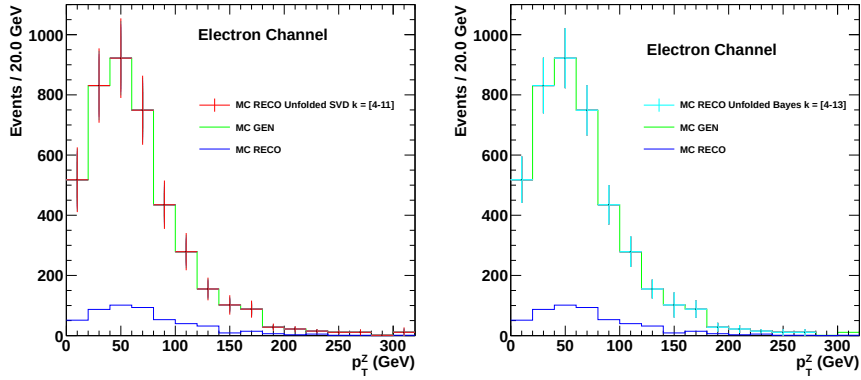


Figure 4.12: Closure test showing the reconstructed distribution in blue, the generator level distribution in green, and the unfolded reconstructed distribution with its statistical uncertainties in red (left) or light blue (right) matching the generated distribution. Different regularisation parameters have been used (see legend) and give the same result, so the distributions are superimposed for both, the case of the SVD method (left) and the Bayesian method (right).

The data before and after unfolding are shown in Figure 4.13 (left) together with the generated and the reconstructed distribution. The statistical plus systematic uncertainties on the background subtracted data distribution are propagated by the unfolding and shown in red. In the cross section measurement analysis the systematic uncertainties are dominated by b-tagging and JES, as seen in Section 3.5. The effect on the unfolded distribution has been estimated by realising the unfolding with the different b-tagging scale factors and JES

variations. For each bin the greatest difference with the central value in the unfolded data is taken as systematic uncertainty from the b-tagging on one side, and the JES on the other side. The two uncertainties are added together in quadrature and with the statistical one, and are displayed as the violet part of the error bar on Figure 4.13 (top, left and right). The systematic uncertainty coming from the unfolding method is calculated by varying the k parameter between values that still give acceptable uncertainties on the unfolded distribution (but always greater than 2 to avoid a too important simulation dependence). It is taken for each bin as the greatest difference between the distribution unfolded with the optimal $k = 9$ parameter and the distribution unfolded with another possible k parameter. This uncertainty is added in quadrature with the previous one (statistical plus b-tagging and JES), and shown as the green part of the error bar on Figure 4.13 (top, left and right). The values of these ‘extreme’ k parameters are given in the legend and the corresponding unfolded distributions are displayed with a dashed line on the right-hand side plot. The unfolding is also performed with the Bayesian algorithm and 4 iterations with the resulting shape given with the blue line distribution. The ratio ‘data/MC’ is the ratio between the unfolded data distributions with SVD and $k = 9$, and the generator level MC distribution. It is the first comparison of the data at generator level, with the MG5F simulation. One can see that with respect to the same comparison at reconstructed level in Figure 4.9, the same trend remains between data and simulation.

The covariance matrix of the different bins of the unfolded distribution is an integral part of the unfolding result and is shown in Figure 4.13 (bottom left) in logarithmic scale. As expected, the correlation is the most important between same bins and then in between adjacent bins, so the matrix is almost diagonal. In order to have a more precise evaluation of the correlations, the correlation matrix is computed from the covariance matrix, the result, also mainly diagonal, is shown in Figure 4.13 (bottom right).

An extra attention has been brought to the next to last bin of the distributions, presenting a negative value for data before unfolding. This is due to absence of data in this bin contrarily to the MC background estimation. Before background subtraction an uncertainty of 1 is attributed to the 0 data value. In the corresponding bin diagonal of the response matrix, one can see in Figure 4.8 that there is also an absence of simulated events. For the data distribution, this bin presents before unfolding a relative uncertainty greater than its two neighbouring ones. But after unfolding, while the unfolded central value becomes positive, the uncertainty seems to be under-estimated with respect to the evolution encountered by the neighbouring bins uncertainties. This behaviour is better illustrated in Figure 4.14 which is a zoom on the last four bins of the p_T

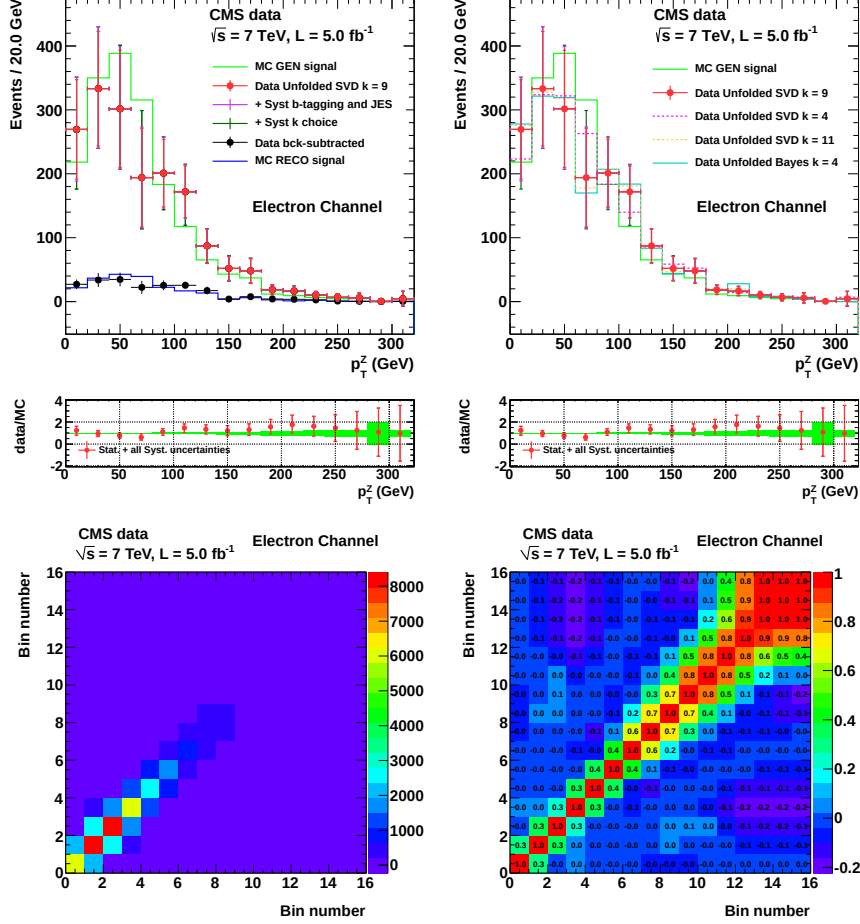


Figure 4.13: The result of the unfolding with SVD of the data distribution from the 5.0fb^{-1} luminosity data sample of p_T^Z in the electron channel. The distributions of data before unfolding (blue line), and after unfolding with the optimal (red line, top left), and boundaries (pink and yellow lines, top right) regularisation parameters are shown. Statistical uncertainties are shown as the red part of the error bars. Systematic uncertainties from b-tagging and JES are added in quadrature and shown as the violet part of the error bars. Systematic uncertainties on the method are added in quadrature as well and shown as the green part of the error bars. The correspondent covariance matrix (bottom left) from the unfolding with optimal regularisation parameter $k = 9$ is displayed. The correlation matrix (bottom right) is computed from the covariance one.

of the Z distribution shown in Figure 4.13. It is most likely due to the absence of information in the response matrix, which makes it not suitable for the unfolding uncertainties propagation treatment. What the unfolding seems to do is to evaluate the uncertainty with respect to the neighbouring uncertainties and so setting it to an averaged value, as one can see on the ratio plot. The uncertainty is then smaller than the one of the last bin and greater than the one on the next-to-next-to last bin, while it would have been expected to be the biggest of the three. The choice has been made not to modify that error bar by hand but to keep in mind in further studies that it is under-estimated. Another approach could be to neglect the small background in this bin. Then, in a background free region, the observation of 0 event excludes a signal of more than 3 events with a confidence level of 95%. Indeed a Poisson distribution with a mean of 3 has a 5% probability to find a zero value. One can then set the uncertainty on the data to 3. But the corresponding empty bin in the response matrix will still prevent a correct uncertainty treatment during the unfolding.

The same will be valid for another quantity shown in Appendix C: the M_{Zbb} in the muon channel, where the response matrix is empty for the next to last diagonal bin. One can also mention in Appendix C the p_T^Z in the muon and in the combined channel for the last two bins; where even if the response matrix is not empty for that diagonal bin, the data present negative values and the uncertainties are underestimated. The other variables do not present any problematic bin of this type.

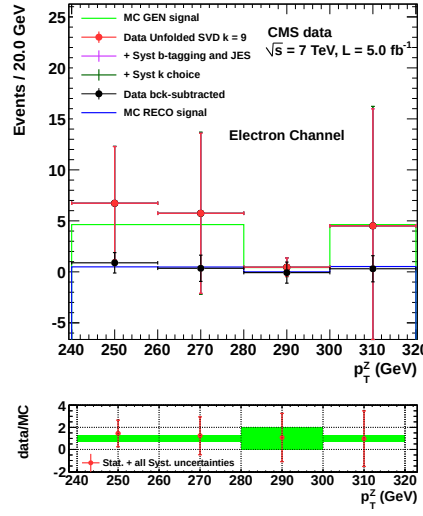


Figure 4.14: The result of the unfolding with SVD of the data distribution from the 5.0fb^{-1} luminosity data sample of p_T^Z in the electron channel, zoomed on the last four bins. The distributions of data before unfolding (black), and after unfolding (red) with the optimal regularisation parameter $k=9$. Statistical uncertainties before unfolding are shown in black. Statistical uncertainties after unfolding are shown as the red part of the error bars. Systematic uncertainties from b -tagging and JES are added in quadrature and shown as the violet part of the error bars. Systematic uncertainties on the method are added in quadrature as well and shown as the green part of the error bars.

Generator independence of the unfolding

In order to prove that the unfolding is not dependent on the type of generator used to build the response matrix, a test is performed with distributions coming from the MADGRAPH 4-flavour generator (MG4F). In Figure 4.15 (left), one can see the reconstructed distribution from MG4F being unfolded using the response matrix build from MG5F generated events. The resulting unfolded distribution is compatible within the statistical uncertainties with the generated MG4F distribution. The method used is the SVD with the optimal k parameter of 9. The MC sample used for the MG4F distributions is a private production only suitable for Z+2b and not for Z+1b, with a NNLO cross section estimated to be with a K factor of 1.23:

$$\sigma_{NLO} \times K = 44.5 \times 1.23 = 54.74 \text{ pb}^{-1}, \quad K = \sigma_{NNLO}/\sigma_{NLO}.$$

The same procedure is performed on the reconstructed distribution from MG5F, this time unfolded using the response matrix build from MG4F produced events, in Figure 4.15 (right).

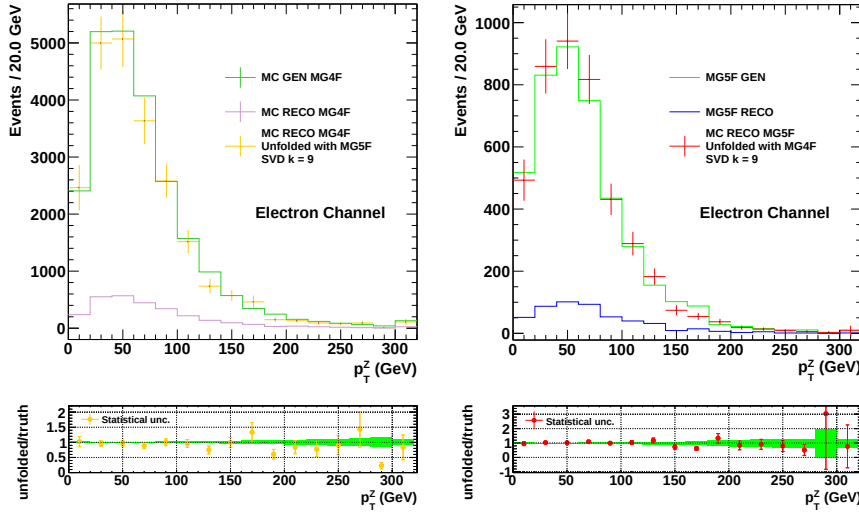


Figure 4.15: Tests: with MG4F reconstructed distribution of p_T^Z being unfolded by SVD ($k = 9$) using a response matrix build from MG5F events (left); with MG5F reconstructed distribution of p_T^Z being unfolded by SVD ($k = 9$) using a response matrix build from MG4F events (right). The unfolded distribution is matching the generator level one within the statistical uncertainties.

The problematic next to last bin is showing again the same behaviour than for the unfolding on the data, this time for the unfolding on MG4F using the MG5F response matrix, visible on the left plot of Figure 4.15. This means the uncertainty is underestimated due to the absence of information in the response matrix. This is visible as a false discrepancy in the ratio plot since the uncertainty on the ratio in this bin should be much larger. On the contrary in the unfolding on MG5F using a MG4F matrix with information from simulation in all bins, the uncertainty is correctly propagated as one can see on the right plot of Figure 4.15.

To go further one can compare the data unfolded from one side with the response matrix build from MG5F, and from the other side with the response matrix build from MG4F. The method used in both cases is again the SVD unfolding with the optimal k parameter of 9. The results are compared in Figure 4.16, and are found to agree within one standard deviation (1σ), within their systematic uncertainties which are partially correlated as coming from the same method. The statistical uncertainties are not displayed since they are propagated from the background-subtracted data distribution which is the same in both cases, and thus highly correlated after unfolding with MG4F and MG5F.

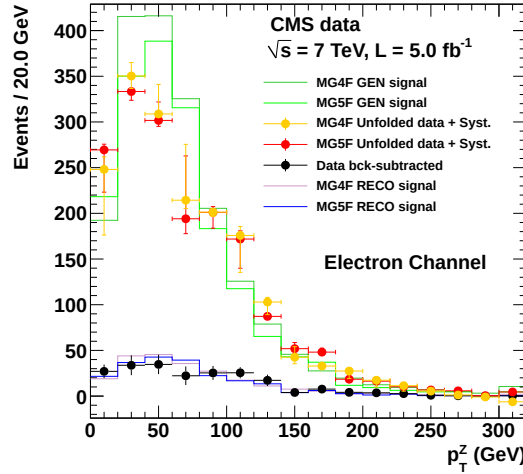


Figure 4.16: Unfolding of the data from the 5.0 fb^{-1} luminosity data sample, with SVD ($k = 9$) method, using the response matrix build from MG4F generated events (in orange) and the response matrix build from MG5F generated events (in red). Uncertainties are systematic only (JES, b -tag and choice of k).

Level of comparison and sensitivity

When doing the unfolding on the reconstructed data, the errors are magnified by the procedure which propagates also the uncertainties on the response matrix used by the unfolding. Looking at the resulting very large uncertainties one can wonder if the comparison between data and simulation is as sensitive at the generator level than at the reconstructed level. The comparisons for both levels are given in Figure 4.17, with the simulations coming from the MG5F (left) and MG4F (right) productions seen in the previous sections.

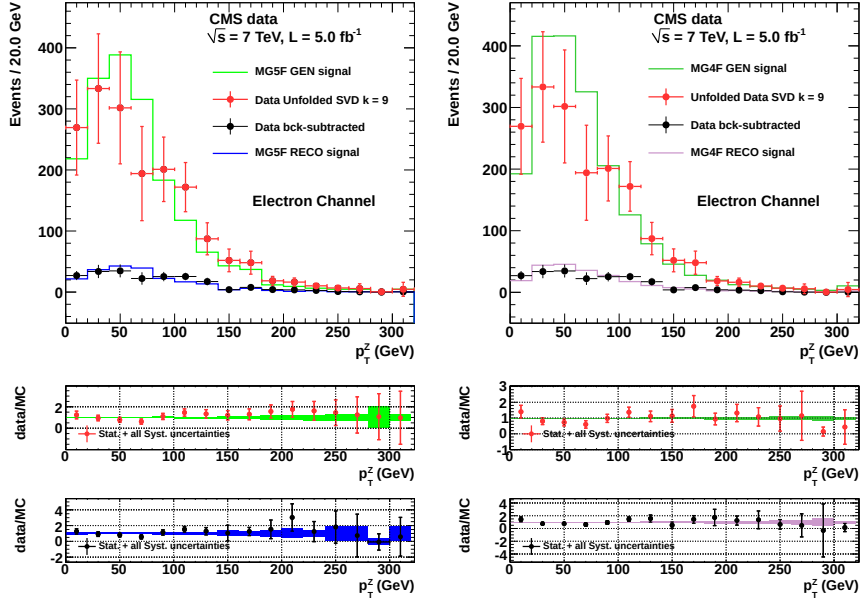


Figure 4.17: Unfolding of the data from the 5.0 fb^{-1} luminosity data sample, with SVD ($k = 9$) method, using the response matrix build from MG5F generated events (in red). Data uncertainties are statistical and systematic coming from the background subtraction. MC uncertainties, shown as the colour bands in the ratio plots, are statistical only. Comparisons are given with MG5F (left) and MG4F (right) simulations before and after unfolding.

To probe the sensitivity of the comparison a Kolmogorov test is performed between the data and simulation distributions before and after unfolding. The results of the test are summarized in Table 4.2. All four resulting probabilities show a compatibility of the tested distributions, and no difference can be established between MG4F and MG5F. The probability of the test before and after unfolding is of the same order, and this in both 5F and 4F simulation

cases. There is no loss of sensitivity in spite of the large uncertainties resulting from the unfolding process. This reinforces the merits of the procedure.

Table 4.2: *Kolmogorov test results for the comparison of data and simulations distributions, from Figure 4.17, at the reconstructed and generator levels. Since the probabilities are not $\ll 1$, the distributions are considered compatible.*

		Simulation	
		MG5F	MG4F
Comparison level	Reconstructed	0.46	0.74
	Generator	0.59	0.77

• Results for muon and combined channels / Results for all the kinematic variables

The unfolding approach is exactly the same in the muon channel and in the combined channel. The optimal k parameter has to be re-evaluated since it must be tuned for each distribution. The results in the muon channel show the same type of agreement and behaviour than in the electron channel, which makes the channels be called compatible. Those channels are then combined meaning the unfolding is entirely performed again with a selection of Z coming from a pair electrons or muons indistinctly.

Full results for the three channels and all the eight variables mentioned before, are presented in Appendix C. The same method will be applied for all the kinematic variables with the response matrix build from MG5F events. In the appendix one can find the response matrix, the logarithmic d_i distribution, which suggests the regularisation parameter to choose, and the data distribution unfolded with the SVD method (and cross checked with Bayesian method), as well as its associated correlation matrix, calculated from the covariance matrix.

4.3 Generators Comparisons

4.3.1 RIVET

With the data unfolded to the generator level, comparisons with more simulations are now possible. Those simulations should of course still undergo the same selection, meaning the set up of a framework containing the same object definitions and acceptance cuts than what is done for the generator level selection in the unfolding process.

The RIVET (Robust Independent Validation of Experiment and Theory) project [167] is such a framework for the selection of simulated events for different Monte Carlo generators. It consists of a set of programs and libraries which produce simulated distributions to be directly compared to real data coming from a high-energy physics experiment, unfolded at generator level. The idea is that the user, theorist, phenomenologist or experimentalist, can compare all the available generators with the data, within any analysis encoded in RIVET.

For the current analysis, a RIVET code has been put in place to reproduce the selection performed at the generator level in the unfolding. It should be integrated in the RIVET tool in the near future. The cuts already mentioned in Section 4.2.2.1 are summarised below on Table 4.3.

Table 4.3: *List of the acceptance cuts for the generator level selection of the simulated events.*

Objects	Cuts
genleptons status 1	dressed with photons in a cone of $\Delta R = 0.1$
dressed leptons	$76 < M_{\ell\ell} < 106 \text{ GeV}$, $p_T^\ell > 20 \text{ GeV}$, $ \eta^\ell < 2.4$
Z candidate	$\min (M_{Z_{\text{candidate}}} - 91.1876 \text{ GeV})$
B hadrons	matched genjet within a cone of $\Delta R = 0.5$
2 genjets matched to B hadrons	$p_T^j > 20 \text{ GeV}$, $ \eta^j < 2.1$, $\Delta R(\ell, j) > 0.5$

As a cross check of this implementation, a comparison is performed for the event selection from the DY + jets sample simulated with the MADGRAPH 5F scheme, that has been used in the unfolding process. One expects the same selection of events from RIVET on one side and from the unfolding generated selection on the other side.

With events generated with MADGRAPH, one has to run directly on the lhe (les Houches events) format files [168]. The hadronisation via PYTHIA is performed on the fly and the resulting events are directly passed through the cuts of the RIVET analysis. In the case of the unfolding the selection was made from the lhe of the corresponding CMS official Monte Carlo production.

Since the hadronisation is randomly seeded for each production it is impossible to perform a synchronisation event by event on both side. However a comparison has been performed on an intermediate skim of the CMS central production which was already hadronised. This synchronisation, performed on about a quarter of the total production gives an identical selection of ‘Zbb’ events from both sides at the permil level.

4.3.2 Comparison at Generator Level

The obtained unfolded data and the RIVET settings of the corresponding generator level cuts, give an unique opportunity to compare with experimental data the different types of Monte Carlo generators mentioned in the discussion Section 1.2.

The main differences in the philosophy of event simulation for b-initiated processes are laying in the choice of a 4-flavour or 5-flavour scheme, but also in the possibility of a leading order or next-to-leading order production. All these studies can be performed with the new version of MADGRAPH [62], where one can choose the scheme and the ‘QCD order’ of the production. All the generated events are passed after parton shower through the RIVET cuts detailed in the Section 4.3.1, offering a coherent comparison with the data unfolded at generator level. All simulations and data distributions in the coming figures are normalised to 1 and to the bin width.

A summary of all the Monte Carlo productions, based on the MADGRAPH generator, used in this study is given in Table 4.4. The table details the cross section (after the parton shower), the type and version of parton shower used, the tuning(s) with respect to the default settings at the matrix element level production step, and the tune used for the settings of the parton shower (described in [169] and whose settings are summarised on [170]).

Table 4.4: *List of the different MADGRAPH based productions used for comparison with the data.*

Production name	Cross section (pb)	Parton shower	Matching scheme	Tuning ME	Tune PYTHIA
MG4F LO Pythia6	18.37	PYTHIA 6	k_T -MLM	-	Z2
MG4F LO Pythia8	18.21	PYTHIA 8	k_T -MLM	-	4C
MG4F LO Pythia6 $\alpha_S(m_T^b)$	30.67	PYTHIA 6	k_T -MLM	reweight. $\alpha_S(m_T^b)$	Z2
MG4F LO Pythia8 $\alpha_S(m_T^b)$	30.19	PYTHIA 8	k_T -MLM	reweight. $\alpha_S(m_T^b)$	4C
MG5F LO Pythia6	2497	PYTHIA 6	k_T -MLM	-	Z2
MG5F LO Pythia8	2475	PYTHIA 8	k_T -MLM	-	4C
MG4F aMC@NLO Pythia6	51.55	PYTHIA 6	- -	$\mu_R = \mu_F = \frac{\sum_i m_{T,i}^2}{2}$ $\mu_S = [0.025 \times Q' - 0.25 \times Q']$	D6T
MG4F aMC@NLO Pythia8	51.53	PYTHIA 8	- -	$\mu_R = \mu_F = \frac{\sum_i m_{T,i}^2}{2}$ $\mu_S = [0.025 \times Q' - 0.25 \times Q']$	CUETP8M1

4.3.3 Comparison with Leading Order Generators

As a first comparison one will look at LO generators and the differences between a 4-flavour and a 5-flavour scheme. The 5F production was taken from a CMS official production of Drell–Yan plus jets (from 0 to 4 jets) at 7 TeV. The 4F scheme is a private production of Drell–Yan plus 2 b jets, plus 2 extra jets at 7 TeV. This production has been done, once with the default setting for α_S reweighting which is performed with only the masses of the 4 lightest quarks, and once asking to include the mass of the b quark in this reweighting. Both versions (6.4 [57] and 8.1 [58]) of the PYTHIA parton shower have been used on the 5F production and on the two 4F productions. In both versions the matching scheme is the default k_T -MLM matching [68] discussed in Section 1.2.1.2. The matching scale so-called Q_{match} (or ‘ $qCut$ ’ in the PYTHIA configuration), determining the energy scale for which the description will be handled by ME or PS, depends on the flavour of the scheme: 30 GeV for the 4F and 20 GeV for the 5F. The effect of the variation of $qCut$ in the PYTHIA 8 version for the p_T of the Z boson in the combined channel, can be observed at the end of Appendix D.

A compatibility of the PYTHIA versions is expected and is observed in Figure 4.18, which displays the unfolded data/MC comparison for the p_T of the Z boson in the electron channel. In this figure a comparison between the two different MG4F LO productions is also shown (left), which do not exhibit any clear difference or tendency. For simplicity, on the following plots only the MG4F LO production with the default α_S reweighting will be displayed. The uncertainties represented for the data are statistical plus systematic, resulting from the unfolding procedure; the ones on the simulations are statistical only. This is also the case in the ratio plot: the yellow band contains the total uncertainty on the data and the transparent coloured band the statistical uncertainty on the simulation.

The 4F/5F differences can be looked at in Figure 4.19, where both schemes simulations at LO (default setting production for 4F) are superimposed with the unfolded data. The left plot uses the PYTHIA 6 version of the parton shower, and the right one the PYTHIA 8 version.

For both, Figure 4.18 and Figure 4.19, no discrepancy with any of the simulations is observed, but the uncertainty on the unfolded data is large and makes the comparison not very sensitive, even with a rather small uncertainty on the MC simulations. Nothing can be concluded in the region of tension at reconstructed level, around 100 GeV, visible in Figure 4.7. However, one can clearly see the trends of the ratio plots being almost identical for the same flavour and different PYTHIA versions or reweighting at ME level. Compatibility is also visible between the two different flavour schemes with a tendency for the MG5F LO ratio to be lower than the MG4F LO ratio above 80 GeV.

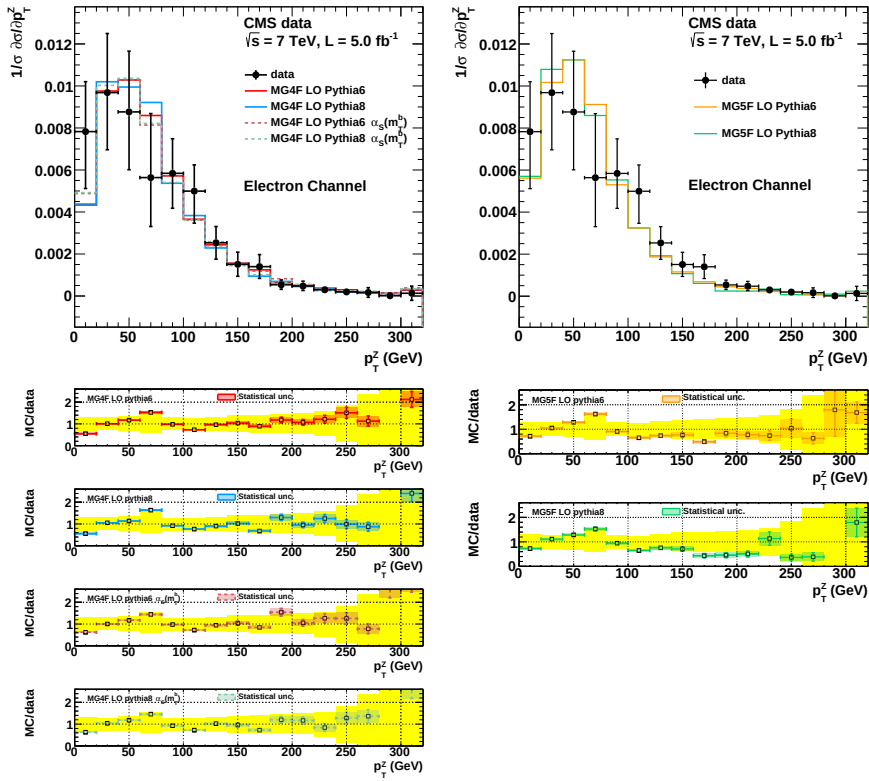


Figure 4.18: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the Z boson in the electron channel.

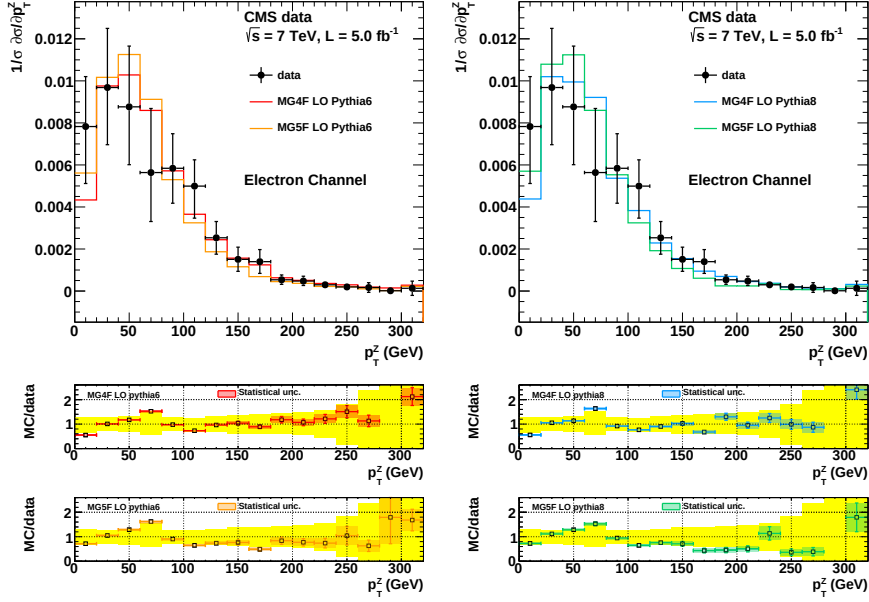


Figure 4.19: Comparison of unfolded data with simulated events from MG4F and MG5F LO productions, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the Z boson in the electron channel.

4.3.4 Comparison with Next-to-Leading Order Generator

A comparison is performed as well for the same 4F scheme but looking at the difference between a LO generation and a NLO generation [62]. The production of 5F scheme at NLO is still a recent development that could not be included in this thesis. In Figure 4.20, the unfolded data are superimposed with events from MG4F LO production (already seen previously in Figure 4.18 and in Figure 4.19) and with events from MG4F aMC@NLO production. On the left plot the hadronisation is performed with PYTHIA 6 and on the right plot with PYTHIA 8.

Two independent MG4F aMC@NLO samples were produced, specially dedicated to be hadronised either with PYTHIA 6 or PYTHIA 8 (which was not the case for the production at LO where both versions of the parton shower can be used for the same parton level production). There is no afterwards matching performed, since an equivalent result arise from the NLO corrections at the parton level generation step.

However, the two MG4F aMC@NLO productions share the same modified settings with respect to the default MADGRAPH ones. First, the scales of renormalisation and factorisation of the process, μ_R and μ_F , are taken as the half of the sum of the transverse mass of the b's in the process; while the default setting in MADGRAPH is half of the sum of all the final state particles of the process. In the current case that would be the Z and the b's. This choice of lowering the scales is driven by the discussion of Section 1.2.2.2, where one learned that for a process with a scale Q much higher than the mass of the quark, like the mass of a weak Z boson, apparition of logarithms of the type $\log \frac{Q^2}{m_b^2}$ might spoil the convergence of the calculation at fixed order. The second tuning performed concerns the shower scale interval μ_S , which by default is set to $[0.1 \times Q' - 1 \times Q']$ in MADGRAPH (Q' being defined by the aMC@NLO matching formalism and computed from the partonic centre-of-mass energy) and which has been lowered to $[0.025 \times Q' - 0.25 \times Q']$. Again, this choice is driven by the wish to lower the scale closer to the mass of the b quark. These settings define the central value of the MG4F aMC@NLO productions for each bin of the distributions in Figure 4.20.

Uncertainties on data are still both systematic and statistical, and statistical only for MG4F LO simulation. For MG4F aMC@NLO the uncertainties coming from theory are added in quadrature with the statistical one. Those theoretical uncertainties are estimated as follows: on one side the scales of renormalisation and factorisation of the process, μ_R and μ_F , are varied independently by 0.5 and by 2 times their central value. On the other side the shower scale interval μ_S , is varied by 0.5 times its default value and by 2 times its default value, without having μ_R and μ_F changed from the central value. The envelop of the highest variation from the central value for, on one side μ_R and μ_F , and on the other side μ_S , is taken as the theory uncertainty for each bin of the distribution. The same uncertainty types are displayed on the ratio plots.

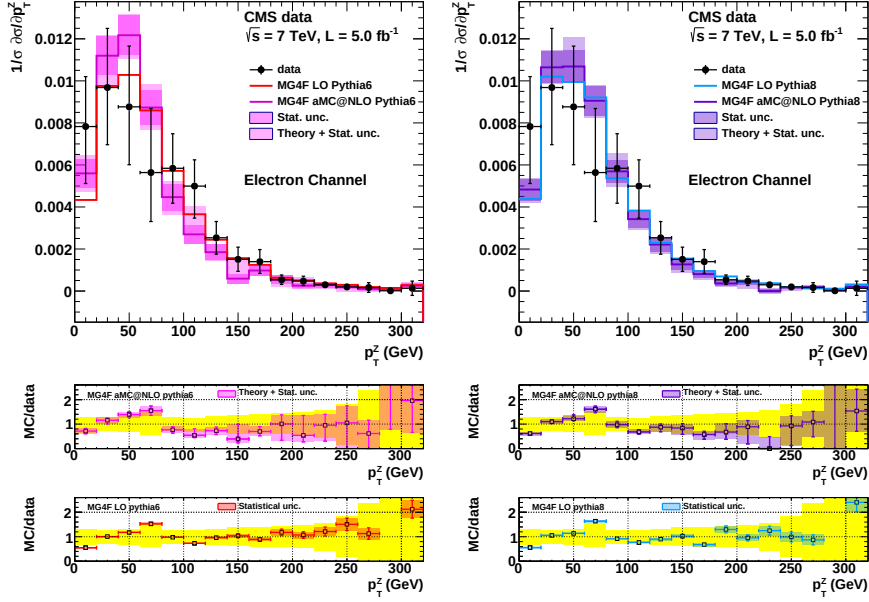


Figure 4.20: Comparison of unfolded data with events from the MG4F LO and MG4F aMC@NLO generator productions, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the Z boson in the electron channel.

In Figure 4.21 only the MG4F aMC@NLO distributions (hadronised with PYTHIA 6 and PYTHIA 8) are shown and compared to the data.

In the three plots of Figure 4.20 and 4.21, no discrepancy is observed for the simulations with respect to the data. The trend of the simulations at NLO is compatible with the ones at LO, even though the fluctuations are higher in the tail of the distribution for the NLO simulations, but covered by larger uncertainties. The versions of PYTHIA are again compatible.

The change for the central value of the scales μ_R and μ_F of the process with respect to the default one ($\frac{\sum_i m_{T_i}^{b_i}}{2}$ instead of $H_T = \frac{\sum_j m_{T_j}^j}{2}$) was mentioned earlier. The effect of this modification can be observed in Figure 4.22, where the productions MG4F aMC@NLO with the modified setting, are compared with the productions MG4F aMC@NLO with the default setting. The latter present a cross section of about 50 pb, for both PYTHIA 6 and PYTHIA 8, and are using the same tunes than the corresponding other MG4F aMC@NLO. A very good agreement between the two different scales MG4F aMC@NLO samples

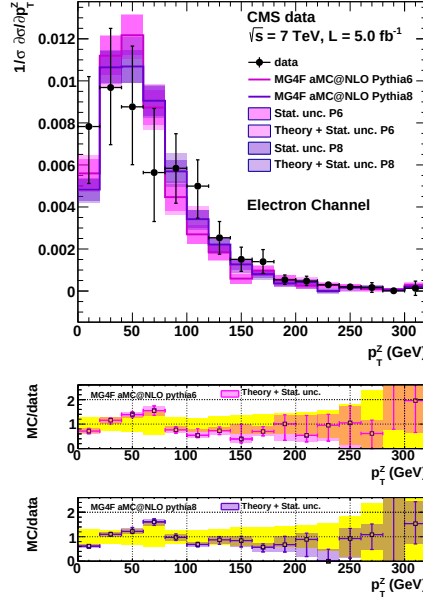


Figure 4.21: Comparison of unfolded data with events from MG4F aMC@NLO generator production and hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the Z boson in the electron channel.

is observed, with respect to both the trend and the size of the uncertainties. This can be explained by the cut on the p_T of the b jets (required to be greater than 25 GeV) which will avoid cases where the $\frac{\sum_i m_{T_i}^{b_i}}{2}$ scale is in a non physical region and not only ‘shifted’ with respect to the H_T scale.

In all the presented results for this specific p_T distribution of the Z boson, one recalls that the next to last bin presents a data error most likely under estimated after unfolding, as seen in Section 4.2.2.3. This explains that the Monte Carlo simulations ratio plots with an uncertainty only statistical are off the data by about 1σ in this bin. The NLO ratio plots present larger uncertainties due to the addition of the theoretical contribution, which cover the effect. The choice has been made to discard this bin and to adapt the scale to see the trend in all the other bins.

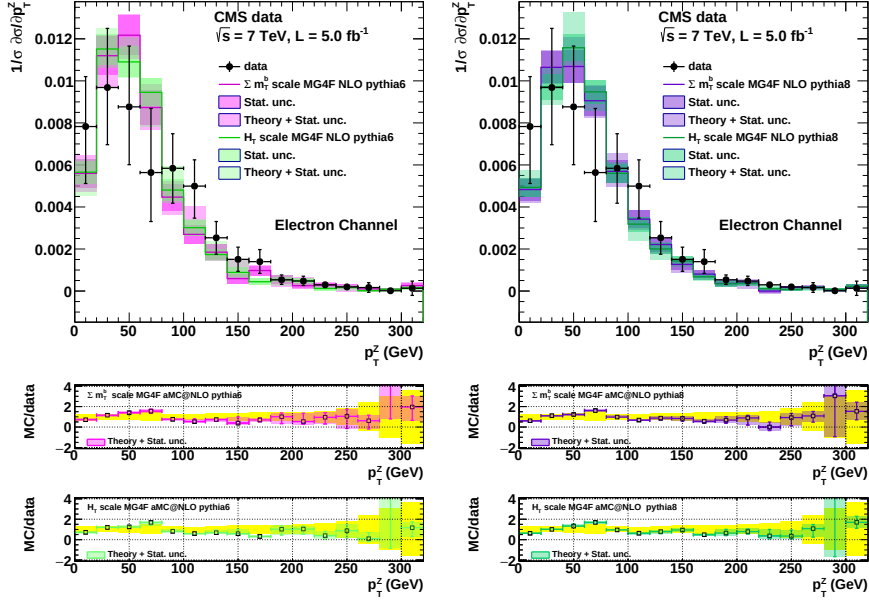


Figure 4.22: Comparison of unfolded data with events from MG4F aMC@NLO generator production with the default and modified μ_F and μ_R scales settings, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the Z boson in the electron channel.

4.3.5 Results

Following the same procedure and conventions, exhaustive results for the three channels (electron, muon and combined) and for all the eight variables of interest can be found in Appendix D.

However, hereafter the main results for all the variables in the combined channel are presented, meaning:

- the comparison with unfolded data of MG4F LO and MG5F LO events hadronised with PYTHIA 6
- the comparison with unfolded data of MG4F LO and MG4F aMC@NLO events hadronised with PYTHIA 8

4.3.5.1 Transverse momentum of the Z boson

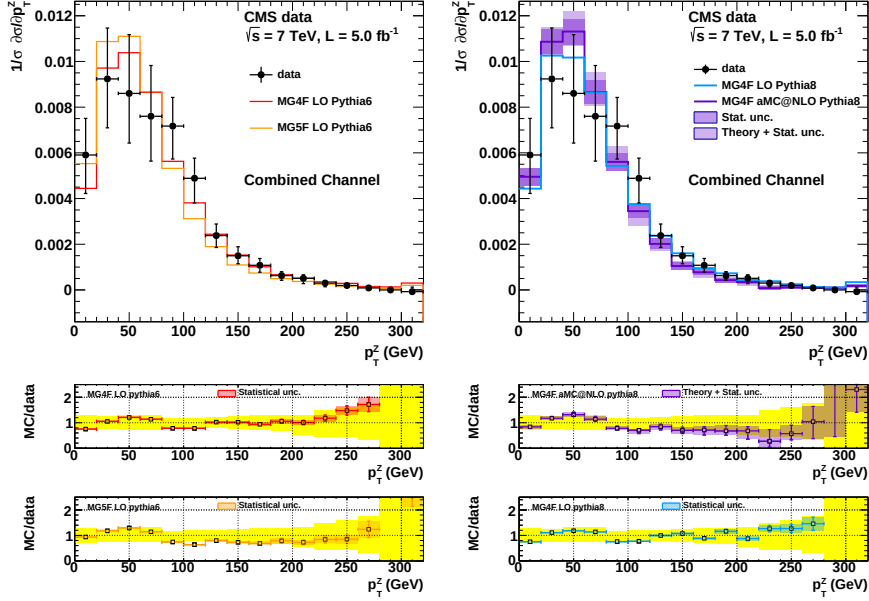


Figure 4.23: Distribution of the p_T of the Z boson in the combined channel: comparison of the unfolded data with events generated with MG4F LO and MG5F LO, hadronised with PYTHIA 6 (left); with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 8 (right).

No discrepancy can be observed in Figure 4.23 for the distribution of the p_T of the Z boson in the combined channel, between the unfolded data and any of the MC simulations. In this distribution, like for the p_T^Z in the electron channel, the next to last bin suffers from an underestimated uncertainty on the unfolded data. The same choice of scale to see the trend of all the other bins points has been made, and results in not seeing the ratio for the next to last bin, which is off by about one standard deviation, and for the last bin where it matches the data within the large uncertainties. The trends can be considered similar for MG5F LO and MG4F aMC@NLO, while MG4F LO presents a ratio plot with higher values than the other simulations for p_T above 120 GeV.

4.3.5.2 Transverse momentum of the leading b jet

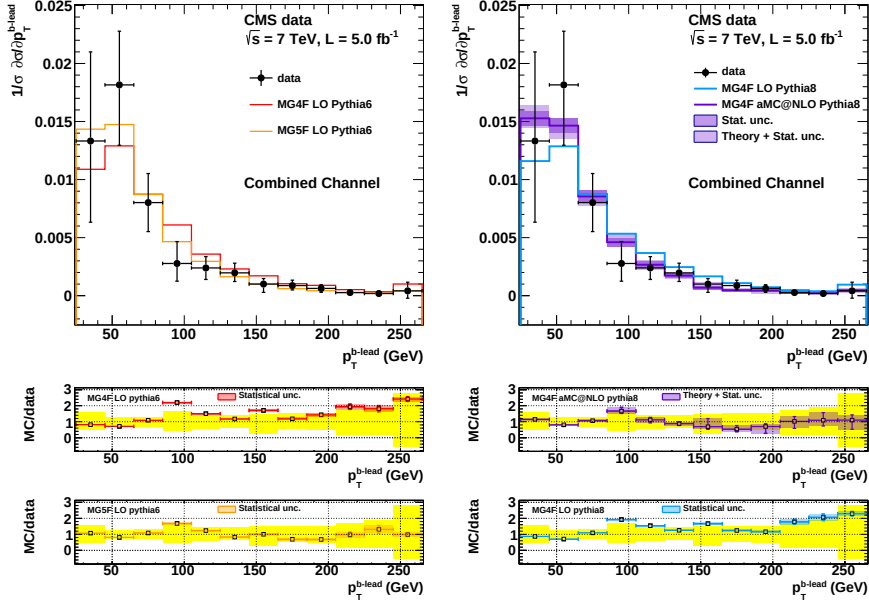


Figure 4.24: Distribution of the p_T of the leading b jet in the combined channel: comparison of the unfolded data with events generated with MG4F LO and MG5F LO, hadronised with PYTHIA 6 (left); with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 8 (right).

For the distribution of the p_T of the leading b jet in the combined channel in Figure 4.24, again partially due to the large uncertainty on the data, no significant discrepancy can be observed. The trends are similar for MG5F LO and MG4F aMC@NLO: a rather flat spectrum with a upward fluctuation around 95 GeV. While MG4F LO presents a ratio plot with higher fluctuations and values above 1 from 130 GeV onwards increasing along the tail.

4.3.5.3 Transverse momentum of the subleading b jet

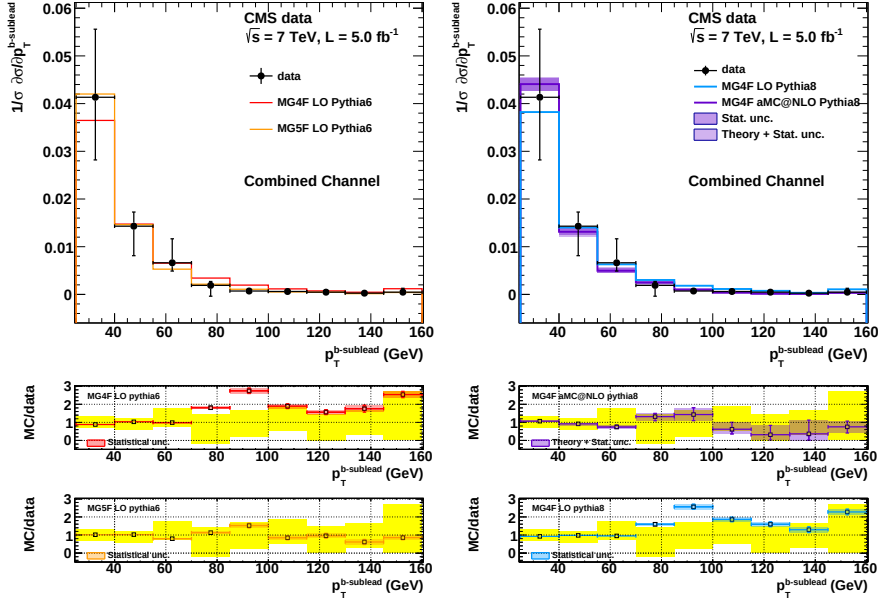


Figure 4.25: Distribution of the p_T of the subleading b jet in the combined channel: comparison of the unfolded data with events generated with MG4F LO and MG5F LO, hadronised with PYTHIA 6 (left); with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 8 (right).

The comparison for the p_T of the subleading b jet in the combined channel in Figure 4.25 draws the same conclusions than for the leading b jet p_T : MG4F has a more fluctuating spectrum than the other simulations, with values above 1 in the tail, and presents a discrepancy of almost 2σ in the bin from 85 to 100 GeV. Discrepancy that disappears in the MG4F aMC@NLO or MG5F LO, displaying the same rather flat trend.

4.3.5.4 Transverse momentum of the $b\bar{b}$ system

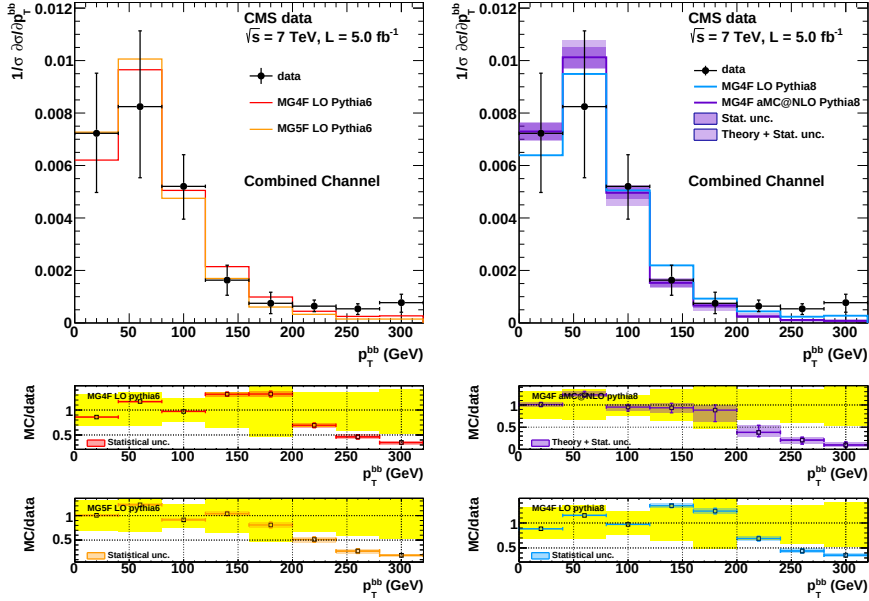


Figure 4.26: Distribution of the p_T of the b -jets pair in the combined channel: comparison of the unfolded data with events generated with MG4F LO and MG5F LO, hadronised with PYTHIA 6 (left); with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 8 (right).

For the p_T of the b -jets pair in the combined channel distributions shown in Figure 4.26, the MC simulations are off the data by up to almost 2σ in the last three bins, which are also the least statistically significant. These discrepancies are more important for MG5 LO and MG4F aMC@NLO productions, which present again a more similar trend than with MG4F LO. If one sums up these three bins, the discrepancy for MG4F aMC@NLO reaches 2.5σ , while for MG4F LO it accounts for 2σ . The plot with this variable binning can be found in Appendix D.

4.3.5.5 Invariant mass of the bb system

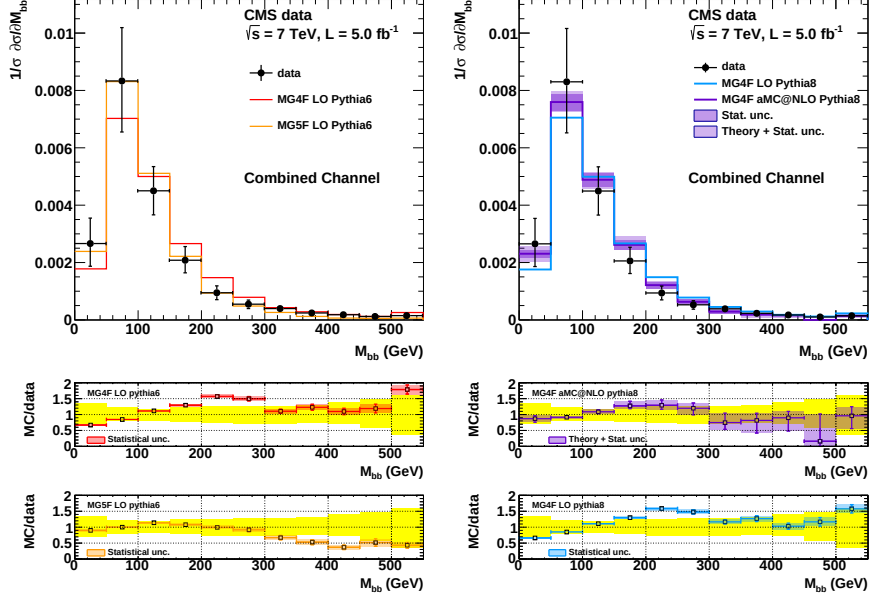


Figure 4.27: Distribution of invariant mass of the b -jets pair in the combined channel: comparison of the unfolded data with events generated with MG4F LO and MG5F LO, hadronised with PYTHIA 6 (left); with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 8 (right).

The comparisons of the simulation with unfolded data of the invariant mass of the b -jets pair, displayed in Figure 4.27, present a wavelike trend for the different ratio plots. The difference between the 4F and 5F scheme is visible and gives clear advantage to the MG5F up to 320 GeV. The data results look like favouring the MG4F aMC@NLO simulations on the overall distribution.

4.3.5.6 Invariant mass of the Zbb system

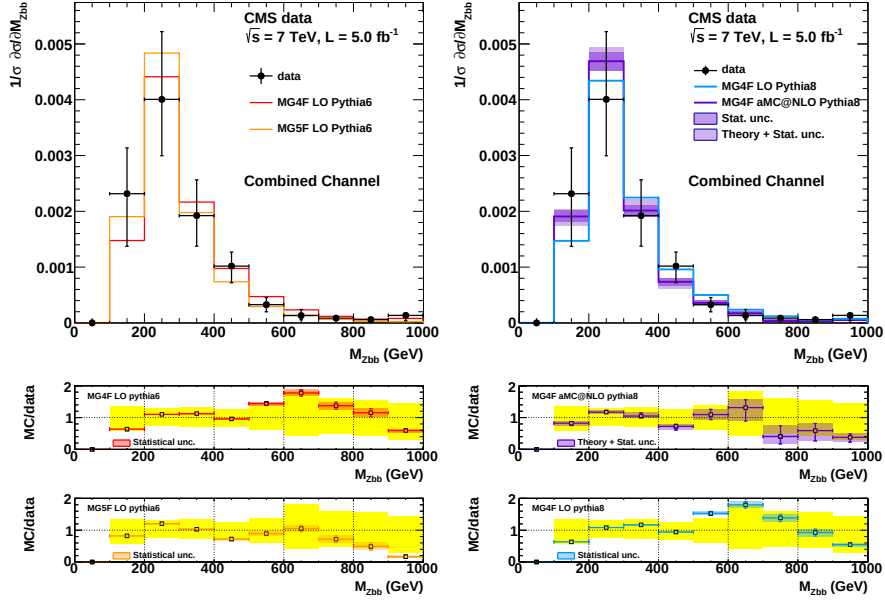


Figure 4.28: Distribution of invariant mass of Zbb system in the combined channel: comparison of the unfolded data with events generated with MG4F LO and MG5F LO, hadronised with PYTHIA 6 (left); with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 8 (right).

For the invariant mass of Zbb system in the combined channel shown in Figure 4.28, the simulations and the unfolded data are in agreement within the uncertainties, in spite of a rather fluctuating ratio spectrum for all considered simulations. For the first four bins (with a significant number of events), the MG5F LO trend is comparable to the MG4F aMC@NLO. The tail being then more fluctuating.

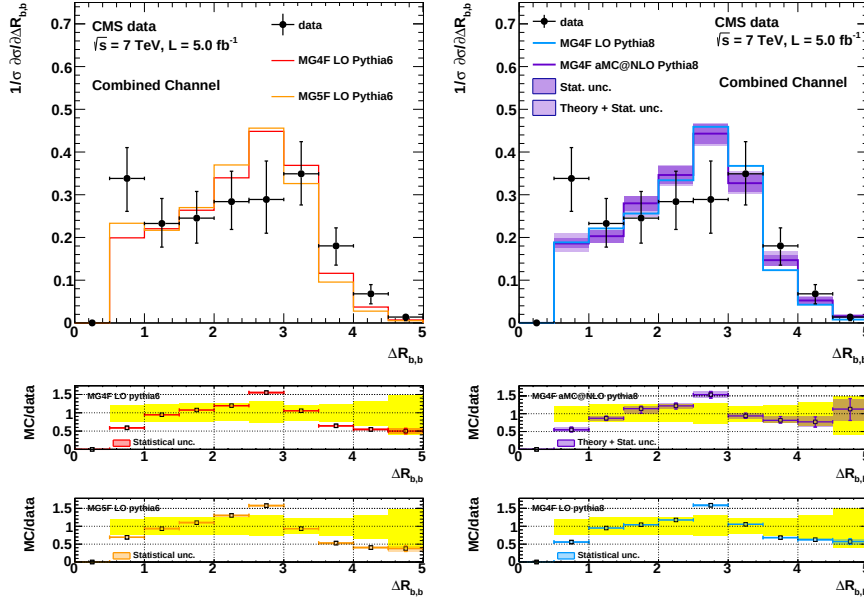
4.3.5.7 Difference ΔR between the two b jets

Figure 4.29: Distribution of the angular difference ΔR between the two b jets in the combined channel: comparison of the unfolded data with events generated with MG4F LO and MG5F LO, hadronised with PYTHIA 6 (left); with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 8 (right).

For the distribution of the angular difference ΔR between the two b jets in the combined channel, it is interesting to notice that some of the tensions mentioned at the reconstructed level, persist at the unfolded generator level, without becoming however, significant. The regions of tensions in Figure 4.3 were:

- the lower bin, for which in Figure 4.29 all ratio plots present a small discrepancy (smaller than 2σ).
- the two next to last bins, for which only the MG5F LO simulation in Figure 4.29 presents a small discrepancy with data at generator level.

One can also notice a small discrepancy for the central bin [2.25–3] for all simulation ratio plots. This upward discrepancy in the ratio could also be compensated by the downward one at the lower bin and come from a statistical fluctuation.

4.3.5.8 Azimuthal angular difference between the Z and the bb system

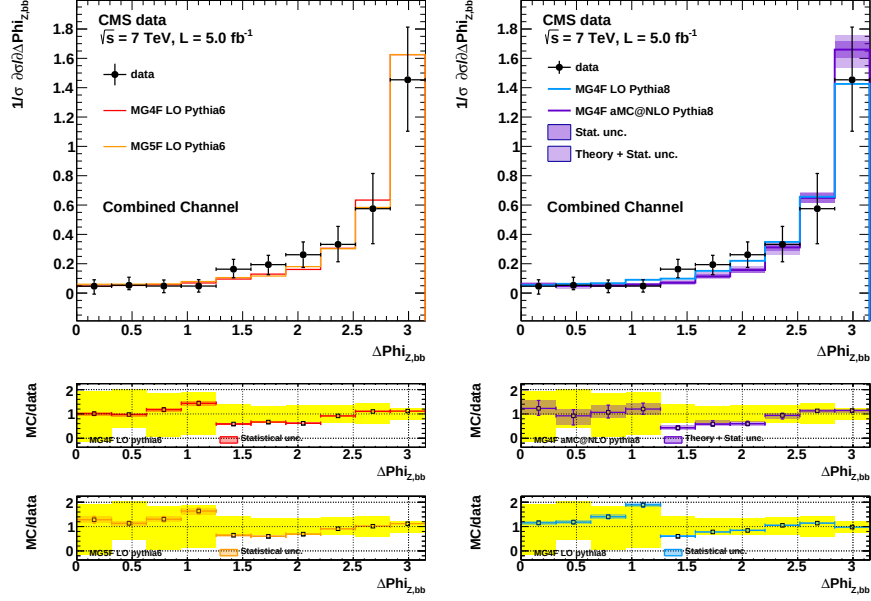


Figure 4.30: Distribution of the azimuthal angular difference between the Z and the bb system in the combined channel: comparison of the unfolded data with events generated with MG4F LO and MG5F LO, hadronised with PYTHIA 6 (left); with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 8 (right).

For the azimuthal angular difference between the Z and the bb system in the combined channel, the distributions at generator level both for data and simulations are presented in Figure 4.30. The simulations MG5 LO and MG4F @NLO show a similar trend for their respective ratio plots. The MG4F simulation ratio plot spectrum agrees with the other simulations as well in the most statistically significant bins.

4.4 Summary

In this chapter the unfolding to generator level of interesting kinematic variables and their comparisons to different MADGRAPH-based simulations were described.

These variables are taken from the cross section measurement at 7 TeV discussed in Chapter 3 and are interesting either for the control of the analysis, the modelling scheme in the Monte Carlo simulation, or eventual signs of new physics.

The unfolding of the data was performed with the Singular Value Decomposition method and both statistical and systematic uncertainties are included in the final result. The unfolded data were then compared to different Monte Carlo simulations, each representing a different type of modelling the process. For this purpose, a RIVET analysis has been set up and used as a framework to select the simulated events in the generator level acceptance cuts.

In spite of the lack of data in the Zbb selection dataset at 7 TeV centre-of-mass energy, the results present a first comparison of different MADGRAPH based simulations for a b-initiated process. They give a feeling on the different questions for the 4F/5F comparison and the LO/NLO comparison with a relatively clear trend of the MC/data ratio spectrum.

The study could be completed with the implementation of the comparison with the recent MG5F aMC@NLO PYTHIA 8 simulation.

Conclusions

Thanks to the high instantaneous and integrated luminosity of the LHC, it is becoming possible to measure rare processes with precision. This is the case of the associated production of a Z boson with b jets. The process can be treated in two different ways in the calculation of its cross section and in the simulation of corresponding events. The first, called 5-flavour includes the PDF of the b quark in the description of the proton. Calculations can be performed in a relatively simple way considering the b's massless, but the interface with the parton shower can be complicated when simulating events. The second scheme, called 4-flavour does not account for the b quark in the composition of the proton and consider the b's as massive with a full kinematic description at LO, and to even higher accuracy at NLO. The divergences that could appear at fixed order calculations are believed to be small at hadron colliders.

The first measurement of the Z boson with b jets production cross section has been realised, using the 5.0 fb^{-1} of data collected by the CMS detector in 2011, for LHC p-p collisions at a centre-of-mass energy of $\sqrt{s} = 7 \text{ TeV}$.

For events with $Z \rightarrow \ell\ell$ (where $\ell\ell = e^+e^-$ or $\mu^+\mu^-$), $p_T^\ell > 20 \text{ GeV}$, $|\eta^\ell| < 2.4$, and the invariant mass $76 < M_{\ell\ell} < 106 \text{ GeV}$, at least one b jet at the hadron level with $p_T > 25 \text{ GeV}$ and $|\eta| < 2.1$, and with a separation between the leptons and the jets of $\Delta R > 0.5$: the cross section at the generator level is $\sigma_{Z+b} = 3.88 \pm 0.02 \text{ (stat.)} \pm 0.22 \text{ (syst.) pb}$.

At the reconstructed level, the four main backgrounds, namely Drell–Yan plus light jets, Drell–Yan plus c jets, $t\bar{t}$ and ZZ, derived from simulations, have been reweighted to fit the data for the pileup, lepton selection and b-tagging. Their contributions have been estimated from simulation and when possible from data. The number of background-free events has then been deduced

for two selections: the exclusive production of a Z boson and exactly one b jet (Z+1b), and the inclusive production of a Z boson and at least 2 b jets (Z+2b). To correct for the detector effects, matrices taking into account efficiencies of selection, resolution and migrations between b jets multiplicity bins, have been built from simulation. The inverted matrices allow to access the ‘true’ number of signal events and thus the cross sections at the generator level. The cross sections evaluated separately in the muon channel and in the electron channel are found compatible and are combined in a single measurement. Finally, the cross sections of the Z+1b selection and the Z+2b selection can be added to yield the cross section of Z plus at least one b jet. In addition, still for the same acceptance, the ratio (Z+b/Z+jets) has been measured to be 5.15 ± 0.03 (stat.) ± 0.25 (syst.) %.

The measured cross sections and the measured ratio have been compared with the theoretical predictions coming from both schemes, 5F and 4F, at leading order with MADGRAPH, and at next-to-leading order with aMC@NLO. The MADGRAPH predictions are corrected for NNLO effects. An additional prediction comes from MCFM at the parton level. The comparisons are summarised in Figure 4.31 for the Z+1b (left) and the Z+2b (right) selection. The 5F scheme predictions at LO and NLO are in agreement with the measurement for Z+1b and Z+2b cross sections. The MCFM prediction shows a one standard deviation difference to the measurement, specific to its modelling of the Z+1b final state. The highest disagreement yields for two standard deviations in the case of the 4F scheme NLO predictions for the Z+1b cross section. The agreements and discrepancies propagate to the total Z+b-jets cross section.

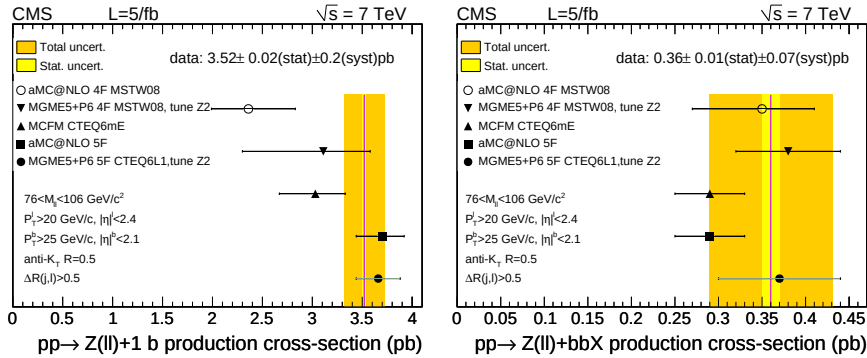


Figure 4.31: Measured cross section for the exclusive Z+1b (left) and the inclusive Z+2b (right) final states compared with different simulation predictions.

Several kinematic variables of the Z+2b selection are interesting for the understanding of the process, the way to model it, and for the search of new physics. The approach adopted in this thesis is to compare these variables to different predictions, but not after reconstruction in the detector as they appear, but as they were generated. One wants to account for the deformations induced by the detector as well as the inefficiencies of selection. The result will then be experiment-agnostic and available to the scientific community to be compared to any generator level simulation.

The unfolding from reconstructed to generator level was performed on eight kinematic variables of the Z+2b selection, using the Single Value Decomposition method and the Bayesian method as a cross check. Both rely on a response matrix built from simulation, accounting for efficiencies and migrations. The final result is generator independent, given for a choice of regularisation parameter inherent to the method, and largely dominated by statistical uncertainties. To easily test several generators, the selection at generator level for the affected variables was implemented using the RIVET framework.

Finally, comparisons have been performed with the simulations from the matrix element generators MADGRAPH, 4F and 5F and aMC@NLO 4F, interfaced with the two versions of the parton shower program PYTHIA. The low statistical significance of the results does not allow to draw sharp conclusions, with the theoretical predictions overall in agreement with the data within the uncertainties. Nevertheless, it is possible to see tendencies in the kinematic spectra, attesting for differences in 4F versus 5F and LO versus NLO.

A selection of results has been given, and the exhaustive production can be found in Appendix D, in the spirit to create a database for further comparisons of a particular variable in a particular channel at unfolded level.

This thesis work took place in one of the most exciting time for particle physics, with the availability of data produced at an unprecedented energy, in an unprecedented quantity, at the LHC. It started shortly after the first collisions at a centre-of-mass energy of 7 TeV, I could set up the analysis on the first data, witnessed the announcement of the discovery of the Higgs boson, and the thesis is ending now, at the dawn of the LHC restart.

As part of a small analysis group, I was able to participate to all parts of the analysis, from the setting up of the first measurement to the final results of the second one. I took care of all aspects of events selection, and I was more specially involved in the pileup reweighting and the measurements of the leptons efficiencies with the *Tag and Probe* method. I participated to the calculation of the corrections matrices and the implementation of one of them. I estimated most of the systematics and elaborated the kinematic variables plots included in the publications. The unfolding part of my thesis was an entirely personal

work. From the choice of the definitions, the technicalities, the implementation in RIVET, to the comparison with the different simulations from the lhe files. This work introduced me to the expertise of a relatively recent aspect of an analysis, still in development.

The associated production of a Z boson and b jets is still a process of interest in the future data of the LHC. With the foreseen integrated luminosity in 2015, the comparisons with the unfolded data will no longer be statistically limited and the observed tendencies could become certitudes about the modelling schemes. The quest for a 5σ confirmation in the vector boson associated SM Higgs production, $VH(b\bar{b})$, is still topical, and the search for new physics in the $Z(\ell\ell)b\bar{b}$ topology shows promising signs [171], putting the Z plus b jets process in the first rows of SM processes necessary to be understood.

Appendix A

Yields

Data yields for the cross section measurement for the 5.0 fb^{-1} luminosity sample compared to the predictions from the simulation samples are gathered in Table A.1.

The complete selection steps are presented, in both the electron and the muon channels. For the b-jet tagging the two versions of the simple secondary vertex are considered: high efficiency (HE) chosen in the analysis, and high purity (HP), as well as their combination in the case of two b jets.

All the steps are here inclusive, so the selection working point Z+2b is the step ‘Z($\ell\ell$)+HEHE+MET sig’ and the selection working point Z+1b is obtained from the yields of ‘Z($\ell\ell$)+HE’ step minus the yields of ‘Z($\ell\ell$)+HEHE’ step.

Table A.1: The data yields for the 2011 5.0 fb⁻¹ sample for the different selection working points, and a comparison with the expectations from different sources based on simulation. Yields for the Zb/Zc/Zl samples are based on the production mechanism (and not on the flavour of the reconstructed jets). Also note that all working points here are inclusive, i.e. Z+HE is defined as the working point containing **at least** one b-tagged jet.

WP	Data	$t\bar{t}$	ZZ	Z+1	Z+c	Z+b	Total MC
Z($\mu\mu$)	1606095 $\pm$ 1267.3	1116.3 $\pm$ 3.9	762.7 $\pm$ 2.3	1081702.9 $\pm$ 677.6	405300.0 $\pm$ 414.8	87298.7 $\pm$ 192.5	1576180.5 $\pm$ 817.5
Z($\mu\mu$)+jet	309215 $\pm$ 556.1	1090.0 $\pm$ 3.8	554.4 $\pm$ 2.0	209001.4 $\pm$ 297.8	92464.9 $\pm$ 198.1	24818.4 $\pm$ 102.6	327929.0 $\pm$ 372.2
Z($\mu\mu$)+HE	13784 $\pm$ 117.4	790.4 $\pm$ 3.3	80.4 $\pm$ 0.8	1833.8 $\pm$ 27.9	3682.0 $\pm$ 39.5	7167.2 $\pm$ 55.2	13553.8 $\pm$ 73.4
Z($\mu\mu$)+HP	6567 $\pm$ 81.0	605.9 $\pm$ 2.9	48.3 $\pm$ 0.6	117.8 $\pm$ 7.1	1097.6 $\pm$ 21.6	4680.7 $\pm$ 44.6	6550.3 $\pm$ 50.1
Z($\mu\mu$)+HE+MET sig	13016 $\pm$ 114.1	243.7 $\pm$ 1.8	77.8 $\pm$ 0.7	1827.8 $\pm$ 27.9	3647.9 $\pm$ 39.3	7059.6 $\pm$ 54.7	12856.8 $\pm$ 73.0
Z($\mu\mu$)+HP+MET sig	6012 $\pm$ 77.5	188.1 $\pm$ 1.6	47.0 $\pm$ 0.6	117.3 $\pm$ 7.1	1088.0 $\pm$ 21.5	4618.8 $\pm$ 44.3	6059.2 $\pm$ 49.7
Z($\mu\mu$)+HEHE	694 $\pm$ 26.3	241.0 $\pm$ 1.8	11.2 $\pm$ 0.3	4.9 $\pm$ 1.4	34.9 $\pm$ 3.8	357.3 $\pm$ 12.3	649.4 $\pm$ 13.1
Z($\mu\mu$)+HEHP	587 $\pm$ 24.2	219.8 $\pm$ 1.7	9.9 $\pm$ 0.3	0.9 $\pm$ 0.6	13.4 $\pm$ 2.4	304.6 $\pm$ 11.4	548.5 $\pm$ 11.8
Z($\mu\mu$)+HPHP	298 $\pm$ 17.3	120.2 $\pm$ 1.3	4.9 $\pm$ 0.2	0.5 $\pm$ 0.5	2.3 $\pm$ 1.0	145.8 $\pm$ 7.9	273.6 $\pm$ 8.0
Z($\mu\mu$)+HEHE+MET sig	522 $\pm$ 22.8	78.8 $\pm$ 1.0	10.9 $\pm$ 0.3	4.9 $\pm$ 1.4	34.3 $\pm$ 3.8	350.4 $\pm$ 12.2	479.3 $\pm$ 12.9
Z($\mu\mu$)+HEHP+MET sig	423 $\pm$ 20.6	71.7 $\pm$ 1.0	9.6 $\pm$ 0.3	0.9 $\pm$ 0.6	12.9 $\pm$ 2.3	298.7 $\pm$ 11.3	393.8 $\pm$ 11.6
Z($\mu\mu$)+HPHP+MET sig	208 $\pm$ 14.4	39.1 $\pm$ 0.7	4.8 $\pm$ 0.2	0.5 $\pm$ 0.5	2.3 $\pm$ 1.0	143.5 $\pm$ 7.8	190.2 $\pm$ 7.9
Z(ee)	1120064 $\pm$ 1058.3	899.9 $\pm$ 3.5	578.7 $\pm$ 2.0	767146.6 $\pm$ 567.2	291989.6 $\pm$ 349.9	63705.7 $\pm$ 163.5	1124320.5 $\pm$ 686.3
Z(ee)+jet	224675 $\pm$ 474.0	879.2 $\pm$ 3.4	423.5 $\pm$ 1.7	152369.0 $\pm$ 252.8	68176.9 $\pm$ 169.1	18661.1 $\pm$ 88.5	240509.7 $\pm$ 316.8
Z(ee)+HE	10162 $\pm$ 100.8	634.6 $\pm$ 2.9	61.3 $\pm$ 0.7	1366.1 $\pm$ 23.9	2870.7 $\pm$ 34.7	5481.5 $\pm$ 47.9	10414.2 $\pm$ 63.9
Z(ee)+HP	4894 $\pm$ 70.0	484.8 $\pm$ 2.5	37.2 $\pm$ 0.5	93.7 $\pm$ 6.3	869.7 $\pm$ 19.1	3580.7 $\pm$ 38.8	5066.0 $\pm$ 43.7
Z(ee)+HE+MET sig	9533 $\pm$ 97.6	195.4 $\pm$ 1.6	59.3 $\pm$ 0.6	1358.7 $\pm$ 23.9	2848.7 $\pm$ 34.6	5397.2 $\pm$ 47.6	9859.3 $\pm$ 63.5
Z(ee)+HP+MET sig	4460 $\pm$ 66.8	151.1 $\pm$ 1.4	36.2 $\pm$ 0.5	93.7 $\pm$ 6.3	862.2 $\pm$ 19.0	3527.7 $\pm$ 38.5	4671.0 $\pm$ 43.4
Z(ee)+HEHE	490 $\pm$ 22.1	190.0 $\pm$ 1.6	8.1 $\pm$ 0.2	2.3 $\pm$ 1.0	26.9 $\pm$ 3.4	263.2 $\pm$ 10.5	490.4 $\pm$ 11.2
Z(ee)+HEHP	417 $\pm$ 20.4	173.1 $\pm$ 1.5	7.1 $\pm$ 0.2	0.5 $\pm$ 0.5	11.8 $\pm$ 2.2	226.5 $\pm$ 9.7	418.9 $\pm$ 10.1
Z(ee)+HPHP	195 $\pm$ 14.0	94.8 $\pm$ 1.1	3.4 $\pm$ 0.2	0.0 $\pm$ 0.0	3.2 $\pm$ 1.2	108.2 $\pm$ 6.7	209.6 $\pm$ 6.9
Z(ee)+HEHE+MET sig	362 $\pm$ 19.0	61.5 $\pm$ 0.9	7.9 $\pm$ 0.2	2.3 $\pm$ 1.0	26.5 $\pm$ 3.3	258.3 $\pm$ 10.4	356.4 $\pm$ 11.0
Z(ee)+HEHP+MET sig	297 $\pm$ 17.2	56.5 $\pm$ 0.9	6.9 $\pm$ 0.2	0.5 $\pm$ 0.5	11.8 $\pm$ 2.2	222.5 $\pm$ 9.7	298.1 $\pm$ 10.0
Z(ee)+HPHP+MET sig	131 $\pm$ 11.4	30.8 $\pm$ 0.6	3.3 $\pm$ 0.2	0.0 $\pm$ 0.0	3.2 $\pm$ 1.2	105.6 $\pm$ 6.7	142.9 $\pm$ 6.8

Appendix B

Lepton Scale Factors

In this appendix complete results can be found for the lepton scale factors, calculated from the ratio of the efficiencies estimated on data and Monte Carlo simulation in the leptons selection, with the *Tag and Probe* method. The method is explained in detail in Section 3.2.6.2. For all the results, the data are from the 2011 5 fb^{-1} sample and the simulated samples have been reweighted according to the pileup distribution observed in data as explained on Section 3.2.6.1. In the tables σ_{stat} refers to the statistical plus fit uncertainty given by the *Tag and Probe* method result. The σ_{syst} is the systematic uncertainty, intrinsic to the *Tag and Probe* method.

◦ **Results for muons**
Reconstruction and ID

Table B.1: *Muon reco/ID scale factor values from 2011 data and simulation efficiencies using the Tag and Probe method and computed for the analysis. Values from standard CMS estimations are given for comparison.*

$SF \pm \sigma_{stat} \pm \sigma_{syst}$ (analysis)			
p_T range	$ \eta < 0.9$	$0.9 < \eta < 2.1$	$ \eta > 2.1$
20–200 GeV	$0.995 \pm 0.001 \pm 0.002$	$0.972 \pm 0.001 \pm 0.002$	$0.978 \pm 0.004 \pm 0.004$
$SF \pm \sigma_{stat} \pm \sigma_{syst}$ (standard)			
p_T range	$ \eta < 0.9$	$0.9 < \eta < 2.1$	$ \eta > 2.1$
20–200 GeV	$0.996 \pm 0.0004 \pm 0.002$	$0.977 \pm 0.0005 \pm 0.002$	$0.983 \pm 0.0005 \pm 0.004$

Isolation

Table B.2: *Muon isolation scale factor values from 2011 data and simulation efficiencies using the Tag and Probe method and computed for the analysis. Values from standard CMS estimations are given for comparison.*

$SF \pm \sigma_{stat} \pm \sigma_{syst}$ (analysis)			
p_T range	$ \eta < 0.9$	$0.9 < \eta < 2.1$	$ \eta > 2.1$
20–200 GeV	$0.993 \pm 0.0008 \pm 0.002$	$0.997 \pm 0.0079 \pm 0.002$	$1.010 \pm 0.0027 \pm 0.004$
$SF \pm \sigma_{stat} \pm \sigma_{syst}$ (standard)			
p_T range	$ \eta < 0.9$	$0.9 < \eta < 2.1$	$ \eta > 2.1$
20–200 GeV	$0.9964 \pm 0.00034 \pm 0.002$	$0.999 \pm 0.00032 \pm 0.002$	$1.010 \pm 0.001 \pm 0.004$

Trigger

Table B.3: *HLT_DoubleMu7* single leg trigger efficiency values from 2011 data using the Tag and Probe method.

$\epsilon_{DoubleMu7leg}$	
η	$p_T > 20 \text{ GeV}$
$ \eta < 1.2$	$0.971 \pm 0.0018 \pm 0.0005$
$1.2 < \eta < 2.4$	$0.948 \pm 0.0031 \pm 0.0005$

Table B.4: *HLT_Mu13Mu8* single leg trigger efficiency values from 2011 data using the Tag and Probe method.

$\epsilon_{Mu13leg}$			
η	20–30 GeV	30–50 GeV	> 50 GeV
$ \eta < 1.2$	$0.967 \pm 0.0015 \pm 0.0005$	$0.967 \pm 0.0007 \pm 0.0005$	$0.968 \pm 0.00096 \pm 0.0005$
$1.2 < \eta < 2.4$	$0.924 \pm 0.0025 \pm 0.0005$	$0.927 \pm 0.00013 \pm 0.0005$	$0.934 \pm 0.0019 \pm 0.0005$
ϵ_{Mu8leg}			
η	20–30 GeV	30–50 GeV	> 50 GeV
$ \eta < 1.2$	$0.968 \pm 0.0015 \pm 0.0005$	$0.967 \pm 0.0006 \pm 0.0005$	$0.968 \pm 0.00096 \pm 0.0005$
$1.2 < \eta < 2.4$	$0.935 \pm 0.0023 \pm 0.0005$	$0.933 \pm 0.00012 \pm 0.0005$	$0.939 \pm 0.0019 \pm 0.0005$

Table B.5: *HLT_Mu17Mu8* single leg trigger efficiency values from 2011 data using the Tag and Probe method.

$\epsilon_{Mu17leg}$			
η	20–30 GeV	30–50 GeV	> 50 GeV
$ \eta < 1.2$	$0.960 \pm 0.0023 \pm 0.0005$	$0.965 \pm 0.001 \pm 0.0005$	$0.966 \pm 0.0014 \pm 0.0005$
$1.2 < \eta < 2.4$	$0.907 \pm 0.0036 \pm 0.0005$	$0.919 \pm 0.0002 \pm 0.0005$	$0.923 \pm 0.0033 \pm 0.0005$
ϵ_{Mu8leg}			
η	20–30 GeV	30–50 GeV	> 50 GeV
$ \eta < 1.2$	$0.962 \pm 0.0038 \pm 0.0005$	$0.967 \pm 0.0009 \pm 0.0005$	$0.967 \pm 0.00015 \pm 0.0005$
$1.2 < \eta < 2.4$	$0.919 \pm 0.0037 \pm 0.0005$	$0.929 \pm 0.00019 \pm 0.0005$	$0.930 \pm 0.0032 \pm 0.0005$

◦ **Results for electrons**
Reconstruction

Table B.6: *Electron reconstruction SF values using the Tag and Probe method. Evaluation of the efficiencies is done by counting (upper table) and fitting (lower table). The central values adopted in the analysis are the ones from the counting method.*

$SF \pm \sigma_{stat}$ (Counting method)				
$\eta \mid p_T$	20–30 GeV	30–40 GeV	40–50 GeV	50–200 GeV
[0–0.8]	$0.998 \pm 0.011 - 0.011$	$0.994 \pm 0.033 - 0.0017$	$0.997 \pm 0.0002 - 0.0002$	$0.996 \pm 0.0022 - 0.0022$
[0.8–1.44]	$1.01 \pm 0.0093 - 0.0093$	$0.995 \pm 0.0023 - 0.0021$	$0.999 \pm 0.0012 - 0.0013$	$0.996 \pm 0.0027 - 0.0028$
[1.44–1.55]	$1.01 \pm 0.059 - 0.021$	$0.994 \pm 0.0056 - 0.0058$	$0.988 \pm 0.0047 - 0.0047$	$1.01 \pm 0.0018 - 0.002$
[1.55–2.0]	$0.996 \pm 0.0084 - 0.0085$	$0.993 \pm 0.0033 - 0.0029$	$1.00 \pm 0.00043 - 0.00043$	$1.00 \pm 0.0037 - 0.0007$
[2.0–2.4]	$0.999 \pm 0.0086 - 0.0077$	$0.993 \pm 0.0037 - 0.0038$	$0.993 \pm 0.049 - 0.0029$	$0.996 \pm 0.005 - 0.0052$
$SF \pm \sigma_{stat}$ (Fitting method)				
$\eta \mid p_T$	20–30 GeV	30–40 GeV	40–50 GeV	50–200 GeV
[0–0.8]	$0.993 \pm 0.004 - 0.004$	$0.995 \pm 0.0013 - 0.0013$	$0.998 \pm 0.0017 - 0.00087$	$0.997 \pm 0.0017 - 0.0017$
[0.8–1.44]	$0.995 \pm 0.004 - 0.004$	$0.996 \pm 0.0015 - 0.0015$	$0.999 \pm 0.0019 - 0.0011$	$0.991 \pm 0.013 - 0.00056$
[1.44–1.55]	$1.01 \pm 0.013 - 0.013$	$0.992 \pm 0.0052 - 0.0054$	$0.991 \pm 0.0048 - 0.0052$	$1.000 \pm 0.0082 - 0.0094$
[1.55–2.0]	$0.993 \pm 0.0055 - 0.0056$	$0.994 \pm 0.0026 - 0.0026$	$1.000 \pm 0.0032 - 0.0018$	$1.000 \pm 0.0032 - 0.0033$
[2.0–2.4]	$0.994 \pm 0.0058 - 0.006$	$0.993 \pm 0.0032 - 0.0033$	$0.993 \pm 0.0028 - 0.0026$	$0.995 \pm 0.0044 - 0.0046$

ID and Isolation

Table B.7: *Electron id-isolation SF values using the Tag and Probe method. Evaluation of the efficiencies is done by counting (upper table) and fitting (lower table). The central values adopted in the analysis are the ones from the fitting method while the counting method is used to calculate the uncertainties.*

$SF \pm \sigma_{stat}$ (Counting method)				
$\eta \mid p_T$	20–30 GeV	30–40 GeV	40–50 GeV	50–200 GeV
[0–0.8]	$0.983 \pm 0.0091 - 0.0075$	$0.993 \pm 0.16 - 0.0032$	$0.993 \pm 0.0011 - 0.0011$	$0.993 \pm 0.12 - 0.0057$
[0.8–1.44]	$0.957 \pm 0.01 - 0.01$	$0.982 \pm 0.004 - 0.004$	$0.99 \pm 0.0017 - 0.0038$	$0.99 \pm 0.0022 - 0.0022$
[1.44–1.55]	$0.894 \pm 0.036 - 0.038$	$0.969 \pm 0.54 - 0.019$	$0.981 \pm 0.025 - 0.025$	$0.943 \pm 0.011 - 0.011$
[1.55–2.0]	$1.02 \pm 0.017 - 0.017$	$0.987 \pm 0.0074 - 0.0075$	$0.994 \pm 0.0027 - 0.0027$	$1.00 \pm 0.0044 - 0.0045$
[2.0–2.4]	$1.01 \pm 0.015 - 0.015$	$1.01 \pm 0.0042 - 0.0042$	$1.01 \pm 0.0036 - 0.0036$	$1.03 \pm 0.0063 - 0.0064$
$SF \pm \sigma_{stat}$ (Fitting method)				
$\eta \mid p_T$	20–30 GeV	30–40 GeV	40–50 GeV	50–200 GeV
[0–0.8]	$0.988 \pm 0.0064 - 0.0064$	$0.996 \pm 0.0029 - 0.0029$	$0.993 \pm 0.11 - 0.0022$	$0.994 \pm 0.0041 - 0.0041$
[0.8–1.44]	$0.968 \pm 0.0088 - 0.0088$	$0.983 \pm 0.004 - 0.004$	$0.991 \pm 0.0028 - 0.0028$	$0.99 \pm 0.0083 - 0.0046$
[1.44–1.55]	$0.913 \pm 0.043 - 0.033$	$0.975 \pm 0.019 - 0.019$	$0.978 \pm 0.015 - 0.016$	$0.92 \pm 0.027 - 0.028$
[1.55–2.0]	$1.01 \pm 0.015 - 0.015$	$0.992 \pm 0.0074 - 0.0074$	$0.996 \pm 0.26 - 0.0055$	$0.984 \pm 0.049 - 0.005$
[2.0–2.4]	$1.02 \pm 0.015 - 0.015$	$1.01 \pm 0.0087 - 0.0088$	$0.998 \pm 0.027 - 0.0038$	$1.03 \pm 0.013 - 0.013$

Trigger

Table B.8: *HLT_Ele17Ele8 single leg trigger efficiency values from 2011 data using the Tag and Probe method.*

$\epsilon_{Ele17leg}$					
$\eta \mid p_T$	20-30 GeV	30-40 GeV	40-50 GeV	50-200 GeV	
[0-0.8]	$0.984 \pm 0.0018 - 0.0019$	$0.994 \pm 0.00022 - 0.0012$	$0.995 \pm 0.0046 - 0.00041$	$0.998 \pm 0.00041 - 0.00047$	
[0.8-1.44]	$0.986 \pm 0.002 - 0.0021$	$0.995 \pm 0.00013 - 0.0003$	$0.997 \pm 0.0005 - 0.00039$	$0.997 \pm 0.00065 - 0.00075$	
[1.55-2.0]	$0.988 \pm 0.0033 - 0.0037$	$0.994 \pm 0.0011 - 0.0012$	$0.995 \pm 0.0009 - 0.00067$	$0.998 \pm 0.00073 - 0.00094$	
[2.0-2.4]	$0.98 \pm 0.0035 - 0.0038$	$0.995 \pm 0.001 - 0.0011$	$0.997 \pm 0.0014 - 0.00047$	$0.998 \pm 0.00092 - 0.0013$	
$\epsilon_{Ele8leg}$					
$\eta \mid p_T$	20-30 GeV	30-40 GeV	40-50 GeV	50-200 GeV	
[0-0.8]	$0.994 \pm 0.0011 - 0.0012$	$0.997 \pm 0.00037 - 0.0004$	$0.998 \pm 0.00028 - 0.00029$	$0.999 \pm 0.00028 - 0.00035$	
[0.8-1.44]	$0.998 \pm 0.00092 - 0.0011$	$0.998 \pm 0.00031 - 0.00036$	$0.999 \pm 0.00024 - 0.00023$	$0.999 \pm 0.00038 - 0.00048$	
[1.55-2.0]	$0.997 \pm 0.0014 - 0.0018$	$0.999 \pm 0.00045 - 0.00055$	$0.998 \pm 0.00066 - 0.00032$	$0.999 \pm 0.00045 - 0.00065$	
[2.0-2.4]	$0.997 \pm 0.0013 - 0.0017$	$0.999 \pm 0.00044 - 0.00059$	$1 \pm 1.6e-11 - 0.00011$	$0.999 \pm 0.00038 - 0.00074$	

Appendix C

Unfolding Results for all the Kinematic Observables

Result of the unfolding with SVD of the data distribution from the 5.0 fb^{-1} luminosity data sample. Statistical uncertainty in red contains the systematics contribution from background subtraction, but since it was propagated with the unfolding and intrinsically bounded to the data to be unfolded, for simplicity in the caption this will be called the statistical uncertainty.

- Transverse momentum of the Z boson
- p_T^Z Electron Channel

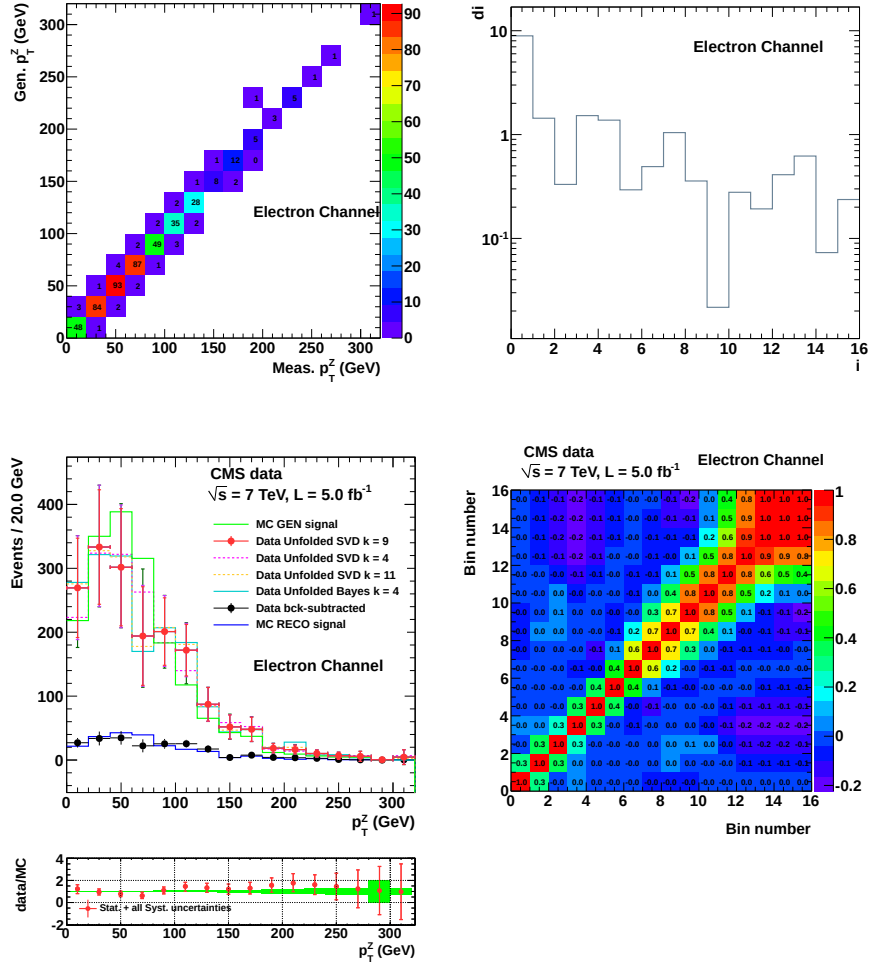


Figure C.1: Response matrix of the p_T^Z distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b-tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• p_T^Z Muon Channel

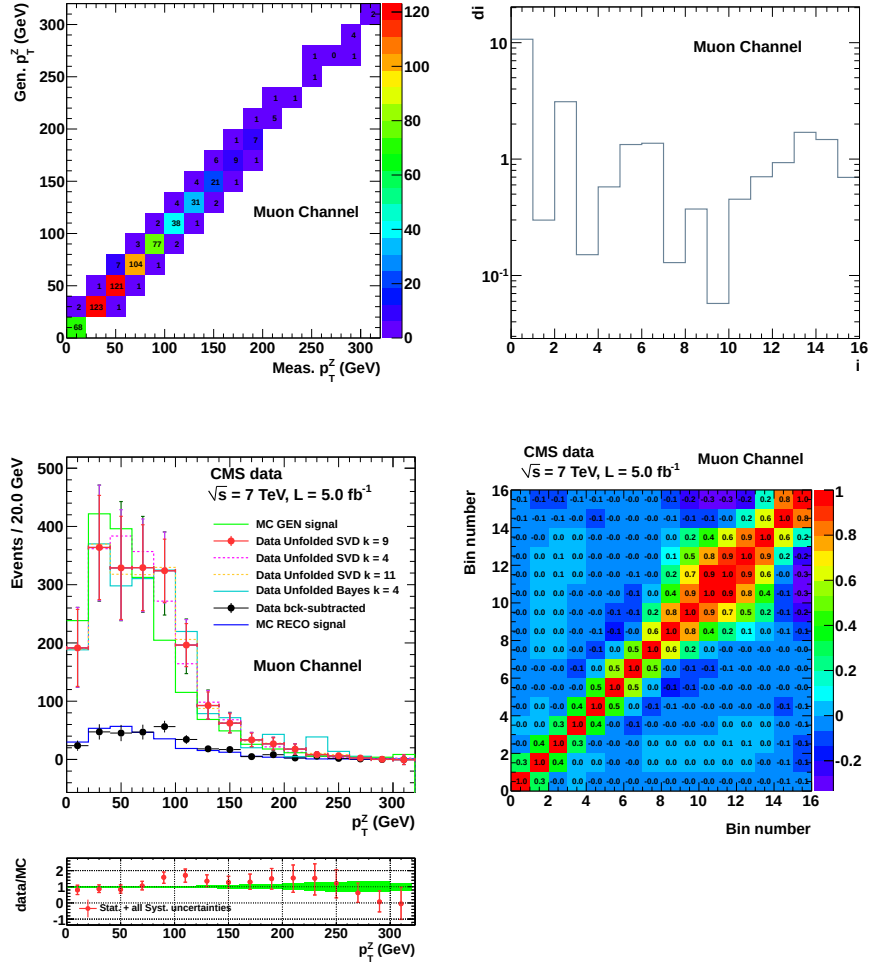


Figure C.2: Response matrix of the p_T^Z distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b-tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• p_T^Z Combined Channel

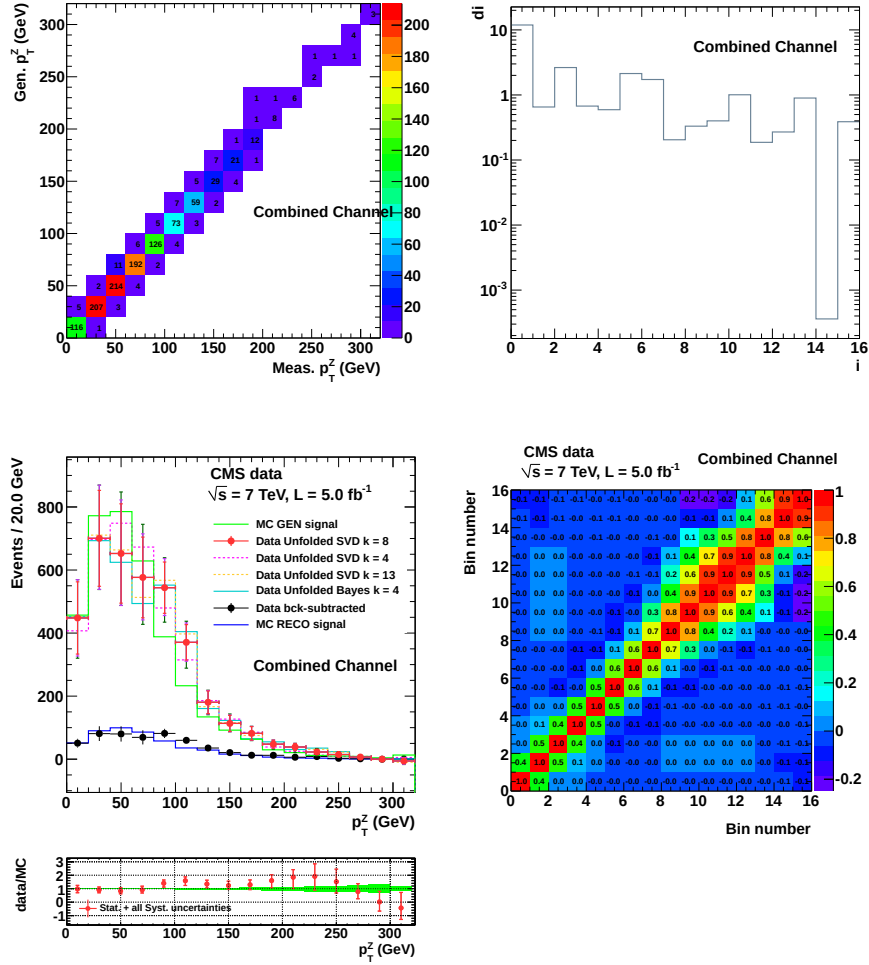


Figure C.3: Response matrix of the p_T^Z distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b-tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

- Transverse momentum of the leading b jet
- $p_T^{b\text{-leading}}$ Electron Channel

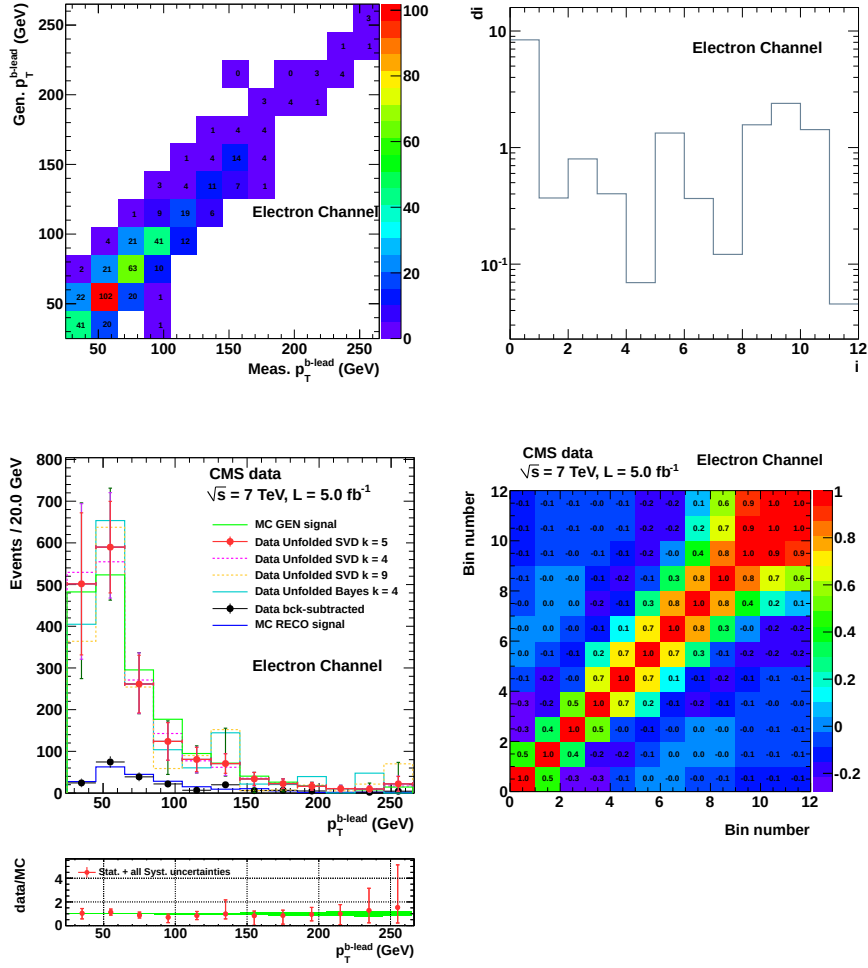


Figure C.4: Response matrix of the p_T leading b jet distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b-tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• $p_T^{\text{b-leading}}$ Muon Channel

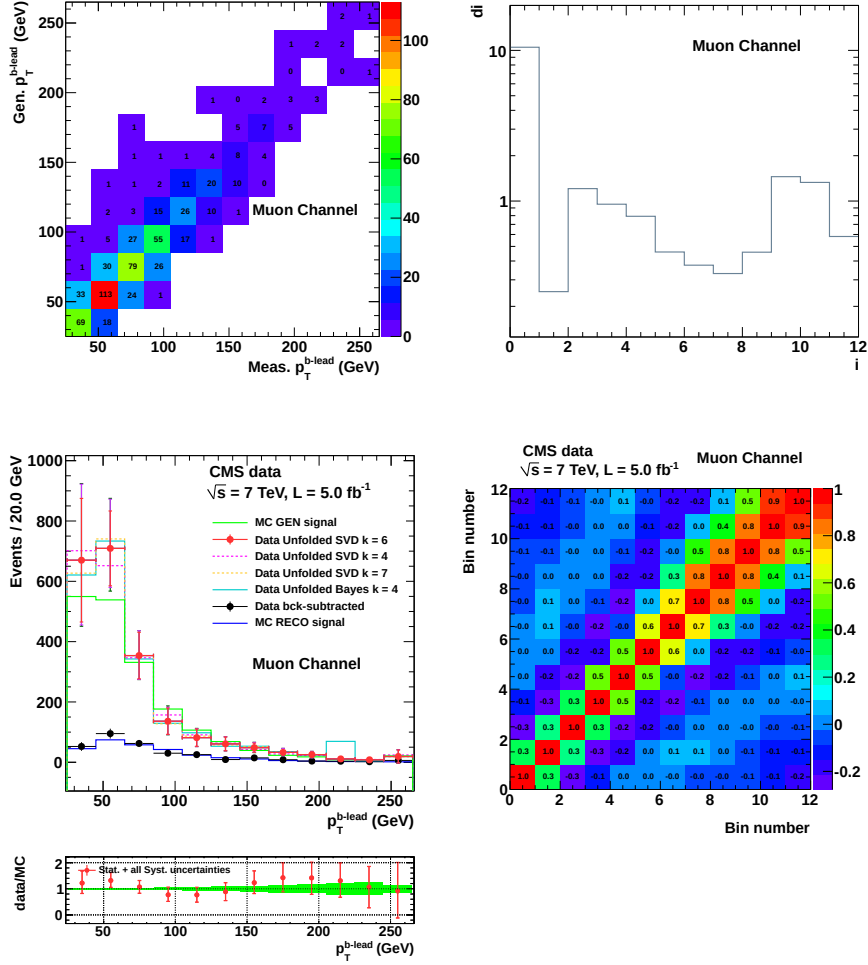


Figure C.5: Response matrix of the p_T leading b jet distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• $p_T^{\text{b-leading}}$ Combined Channel

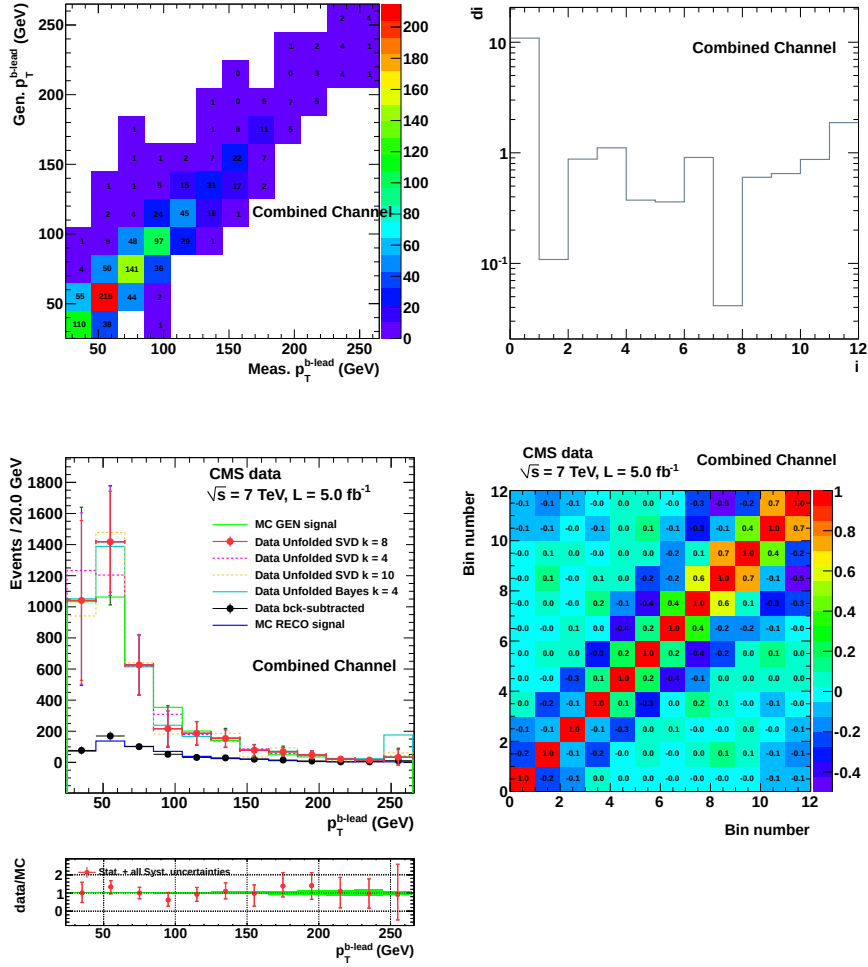


Figure C.6: Response matrix of the p_T leading b jet distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

- Transverse momentum of the subleading b jet
- $p_T^{b\text{-subleading}}$ Electron Channel

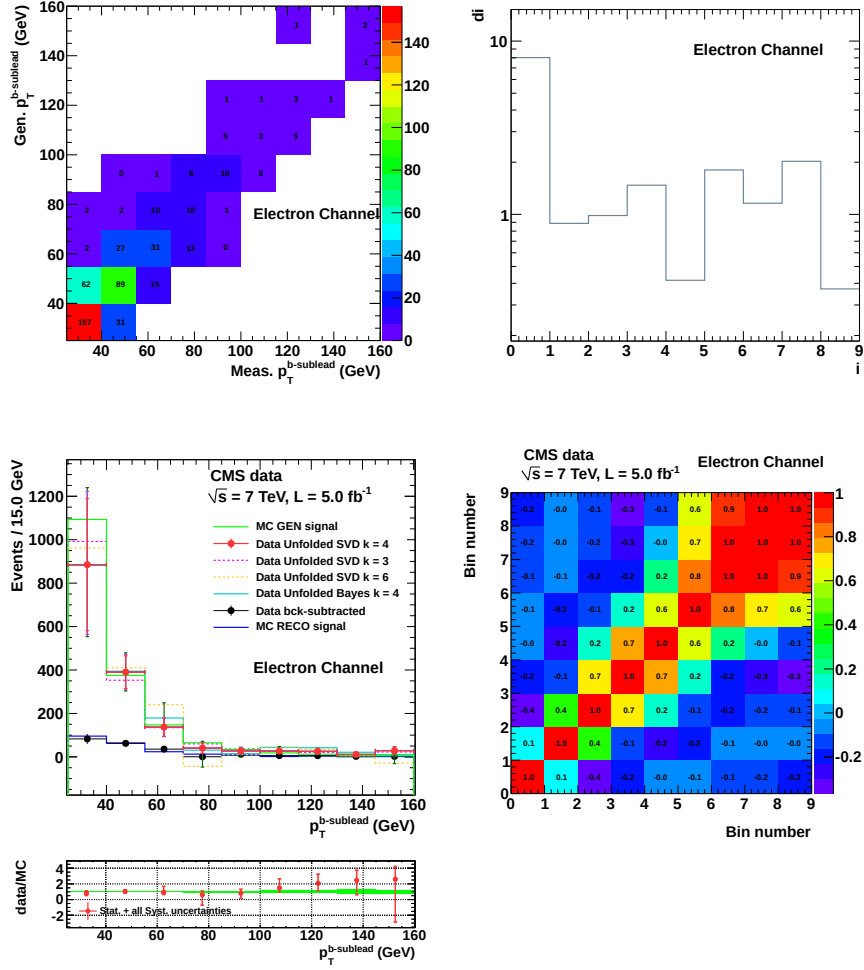


Figure C.7: Response matrix of the p_T subleading b jet distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• $p_T^{b\text{-subleading}}$ Muon Channel

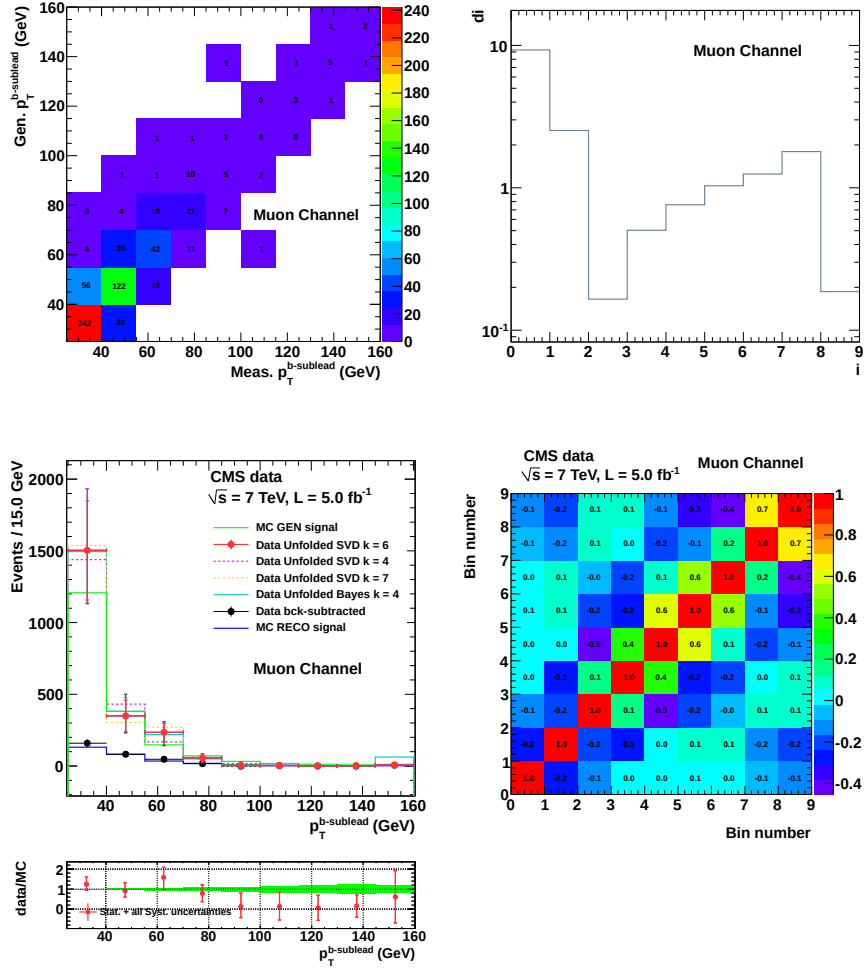


Figure C.8: Response matrix of the p_T subleading b jet distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• $p_T^{b\text{-subleading}}$ Combined Channel

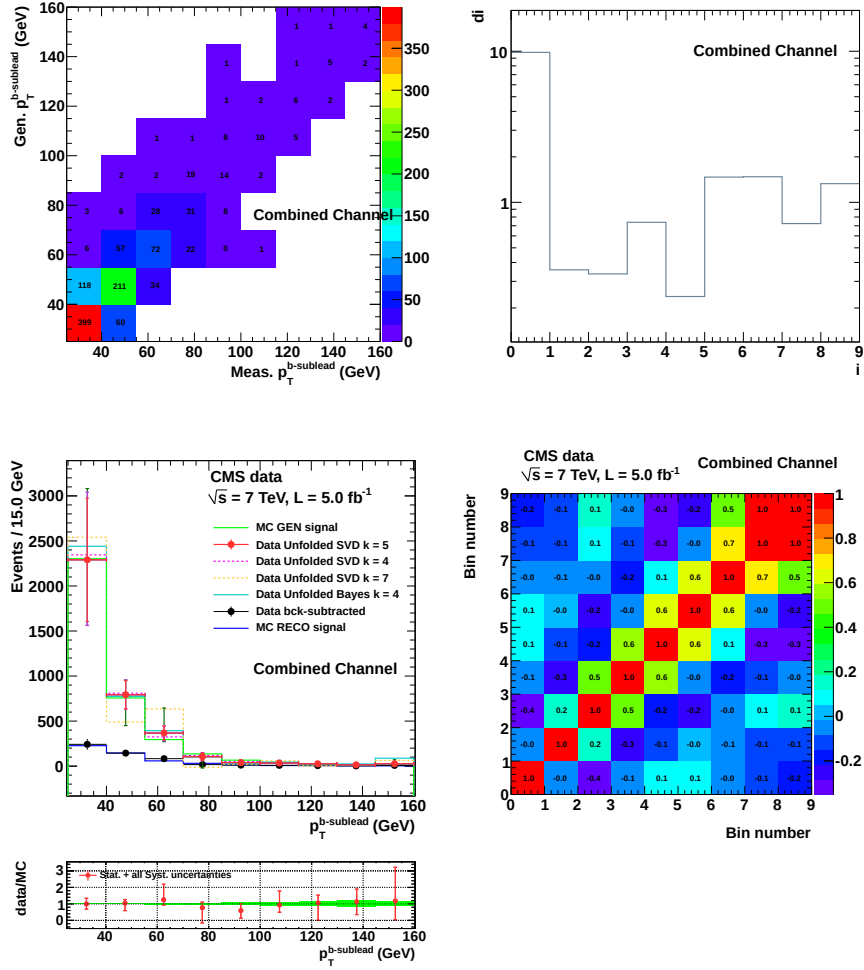


Figure C.9: Response matrix of the p_T subleading b jet distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

- Transverse momentum of the $b\bar{b}$ system
- $p_T^{b\bar{b}}$ Electron Channel

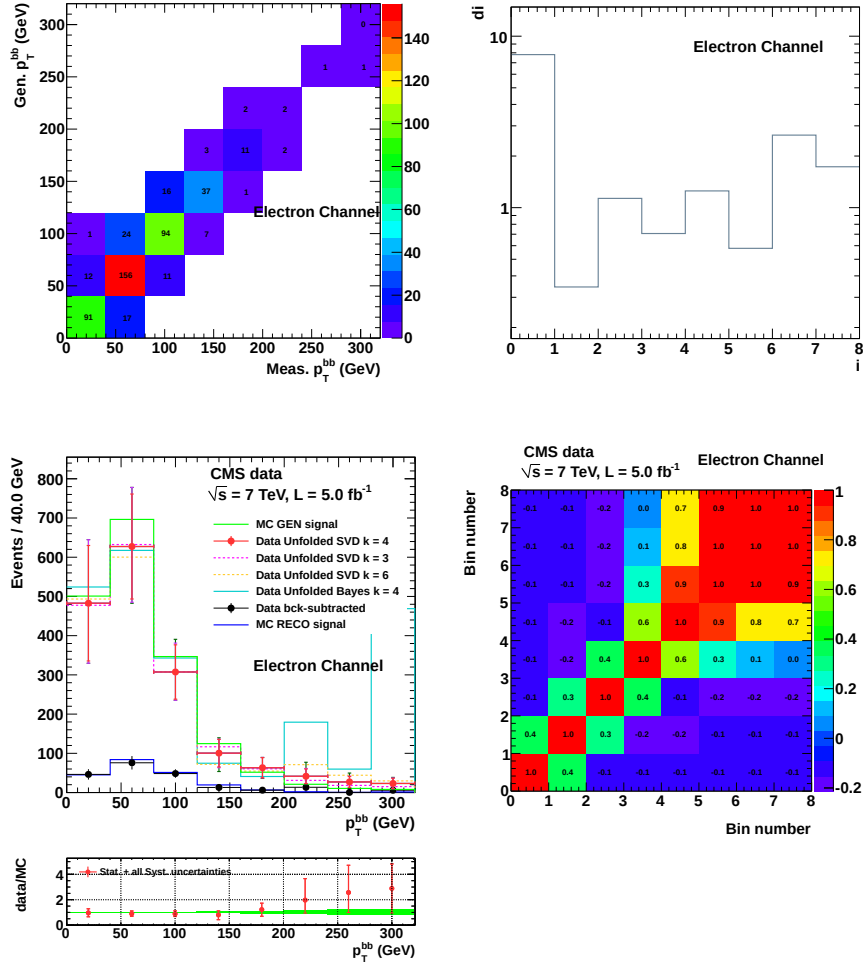


Figure C.10: Response matrix of the $p_T^{b\bar{b}}$ pair distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• p_T^{bb} Muon Channel

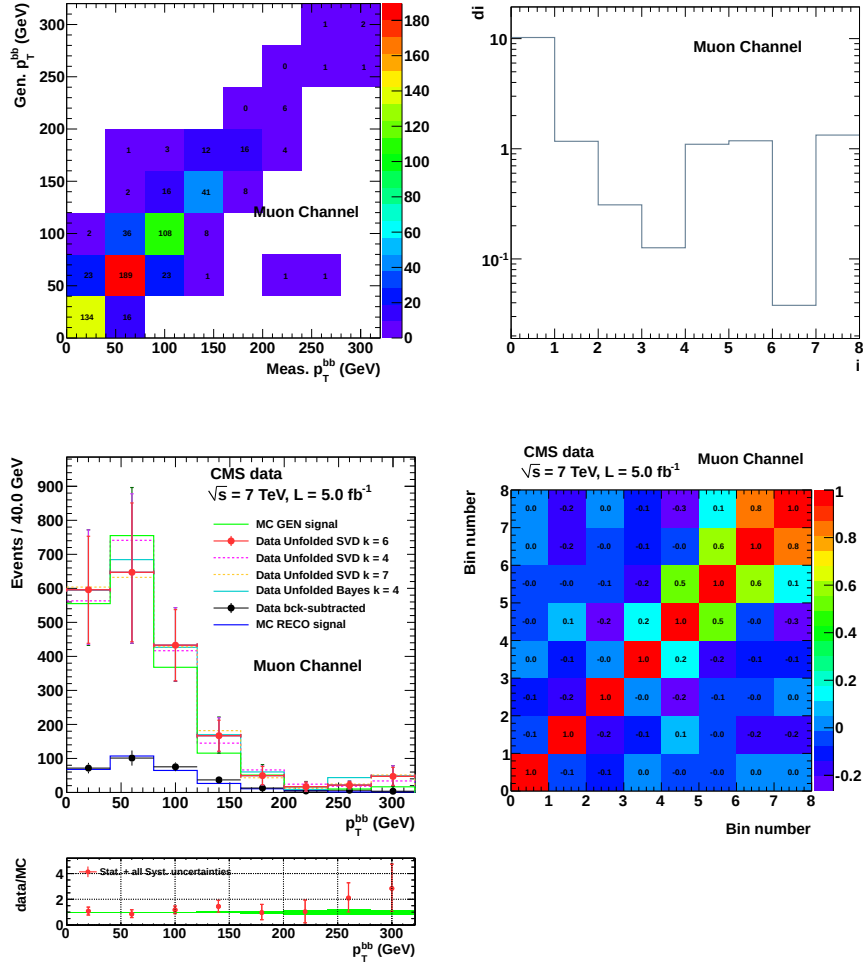


Figure C.11: Response matrix of the p_T^{bb} pair distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• p_T^{bb} Combined Channel

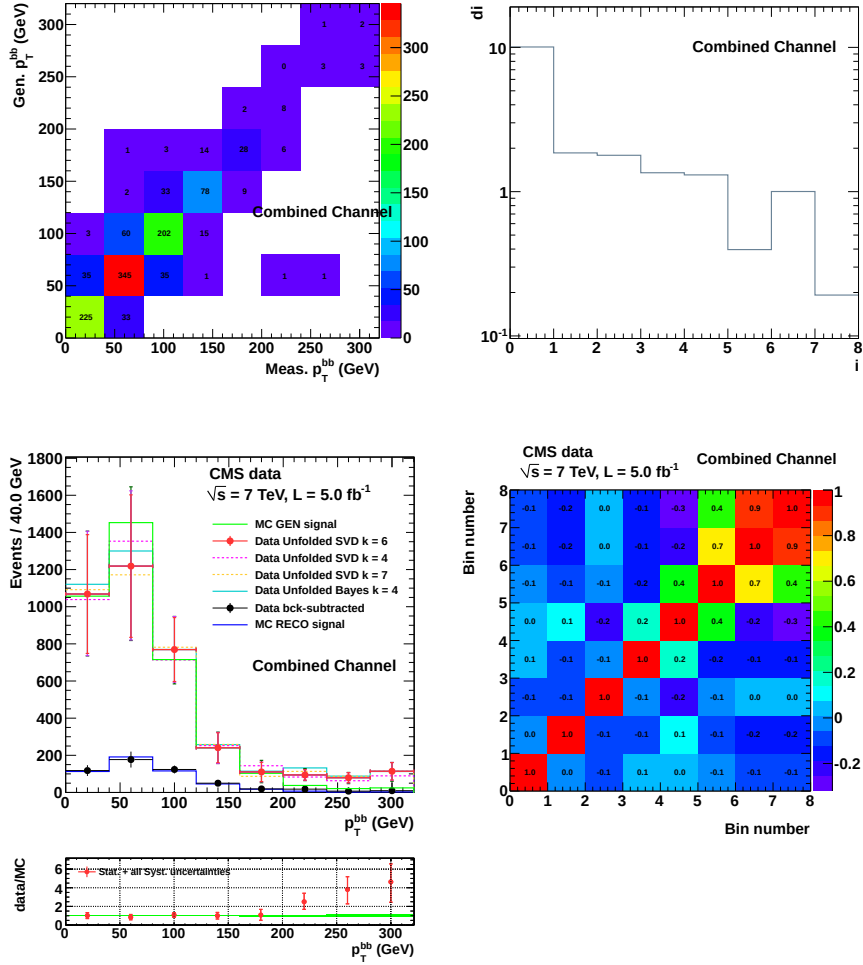


Figure C.12: Response matrix of the p_T^{bb} pair distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

- Invariant mass of the $b\bar{b}$ system
- $M_{b\bar{b}}$ Electron Channel

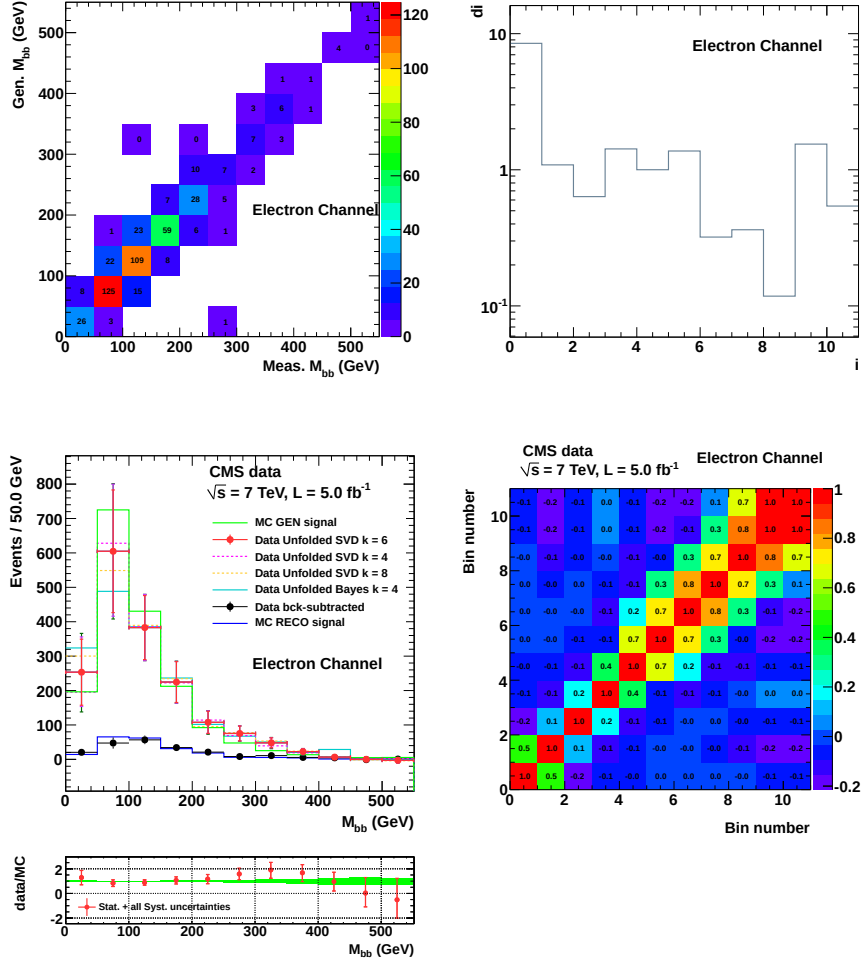


Figure C.13: Response matrix of the $M_{b\bar{b}}$ distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• M_{bb} Muon Channel

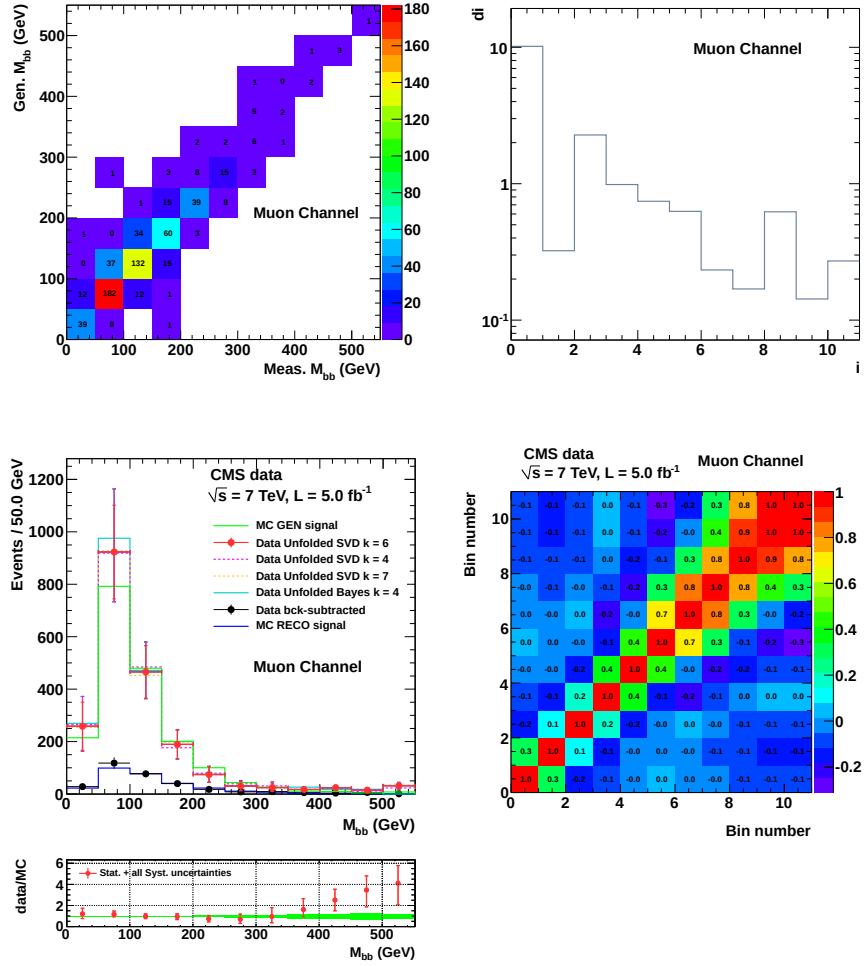


Figure C.14: Response matrix of the M_{bb} distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b-tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• M_{bb} Combined Channel

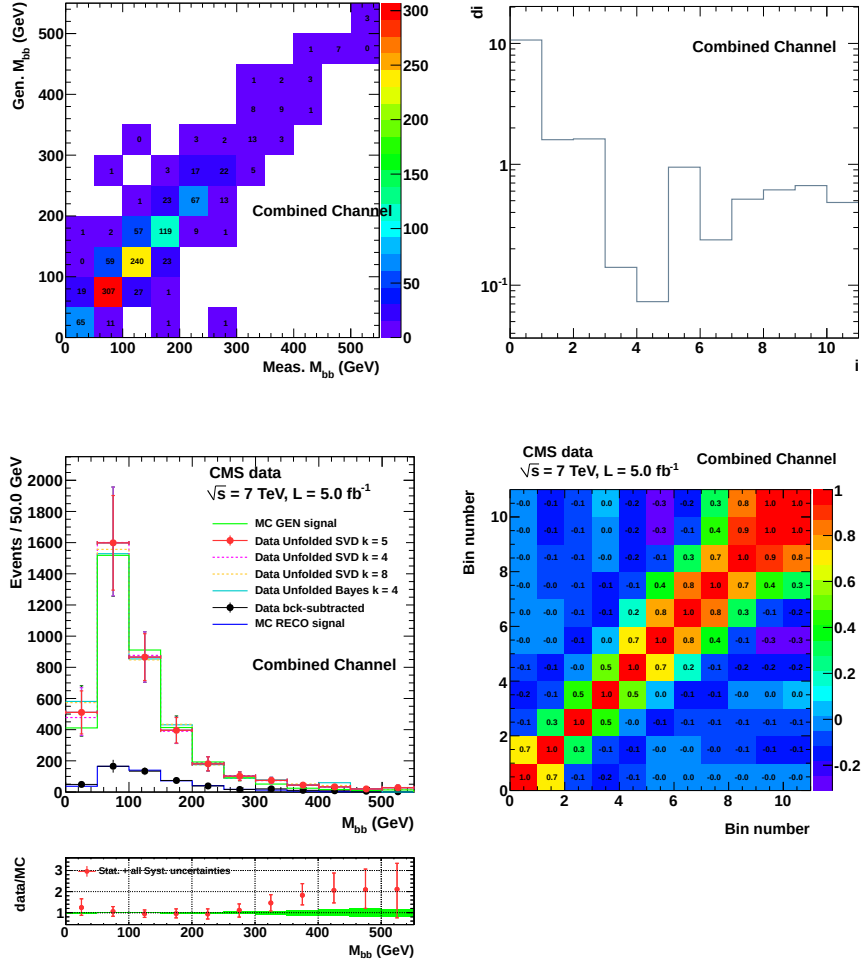


Figure C.15: Response matrix of the M_{bb} distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b-tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

- Invariant mass of the $Zb\bar{b}$ system
- $M_{Zb\bar{b}}$ Electron Channel

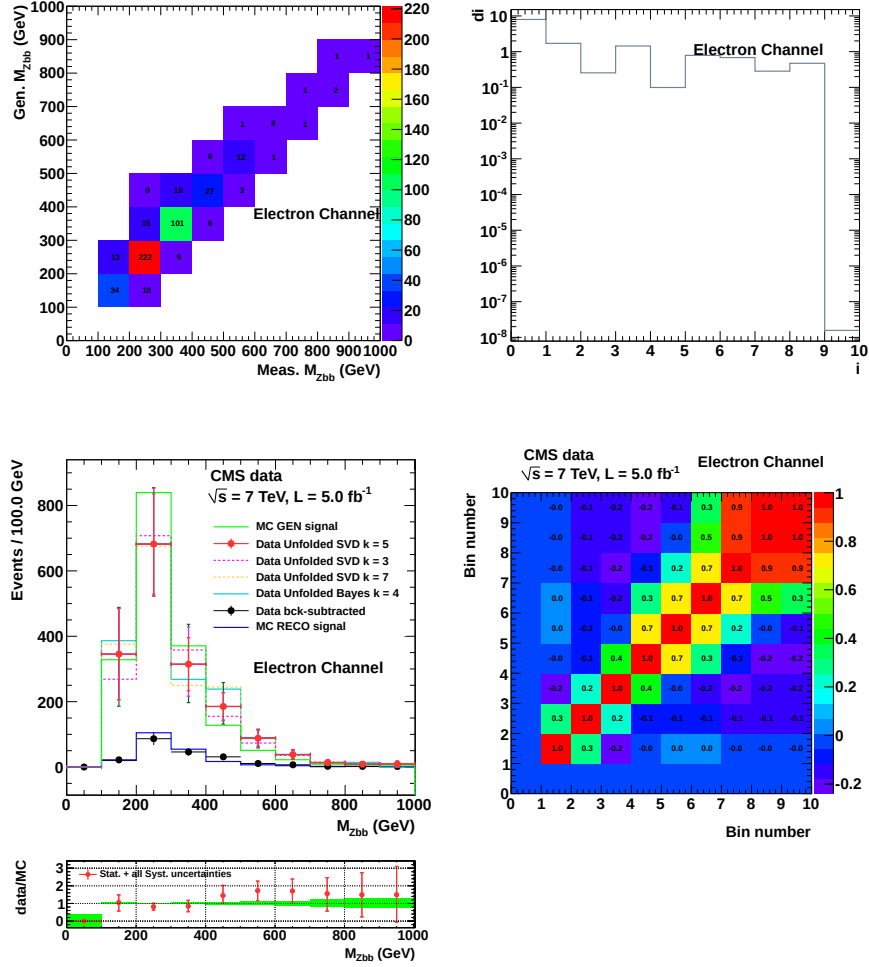


Figure C.16: Response matrix of the $M_{Zb\bar{b}}$ distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• M_{Zbb} Muon Channel

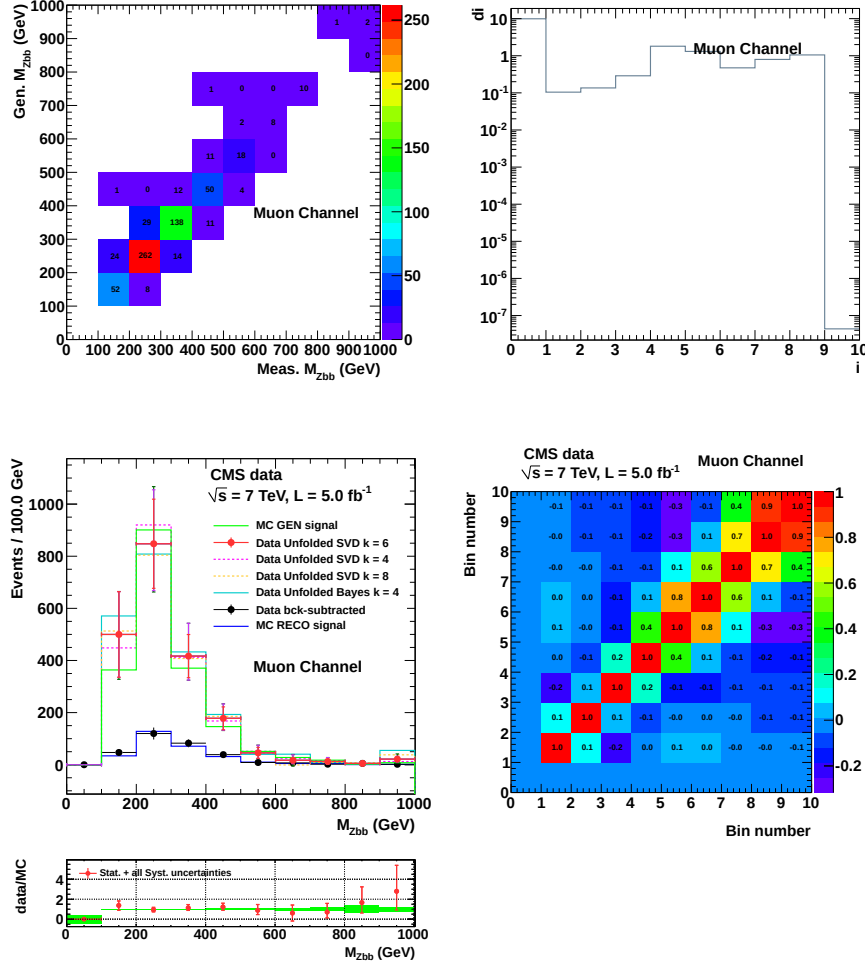


Figure C.17: Response matrix of the M_{Zbb} distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b-tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• M_{Zbb} Combined Channel

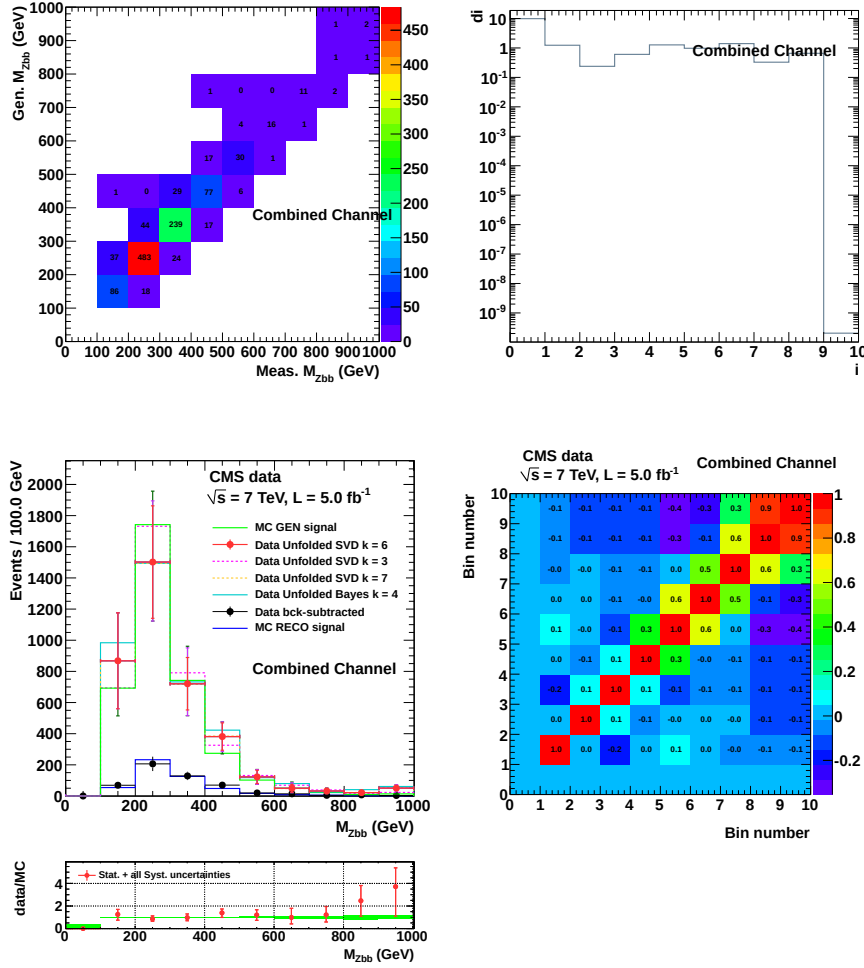


Figure C.18: Response matrix of the M_{Zbb} distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b-tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

- Difference ΔR between the two b jets
- $\Delta R_{b,b}$ Electron Channel

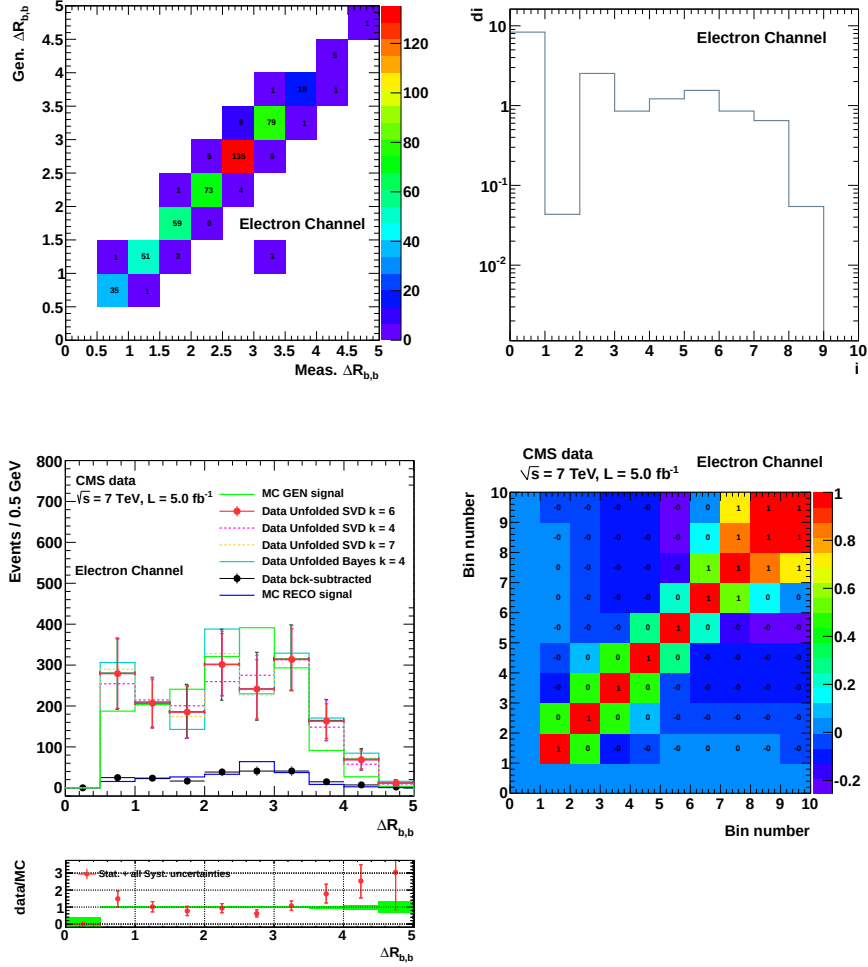


Figure C.19: Response matrix of the $\Delta R_{b,b}$ distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• $\Delta R_{b,b}$ Muon Channel

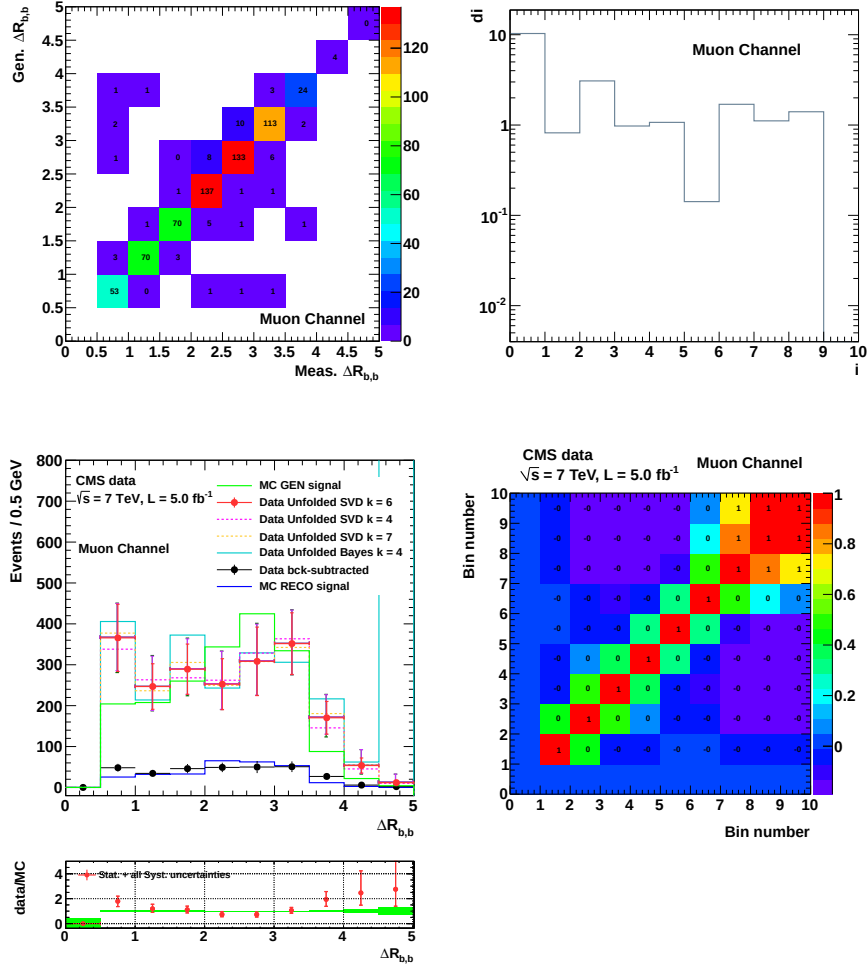


Figure C.20: Response matrix of the $\Delta R_{b,b}$ distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• $\Delta R_{b,b}$ Combined Channel

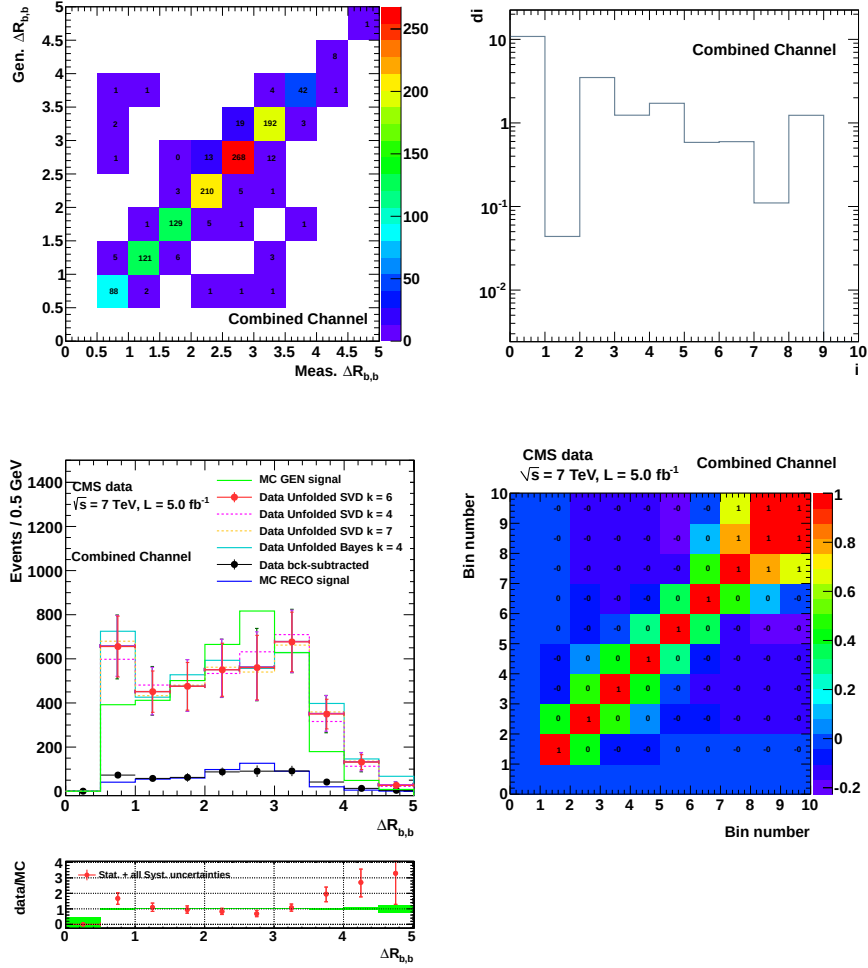


Figure C.21: Response matrix of the $\Delta R_{b,b}$ distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

- Azimuthal angular difference between the Z and the $b\bar{b}$ system
- $\Delta\Phi_{Z,b\bar{b}}$ Electron Channel

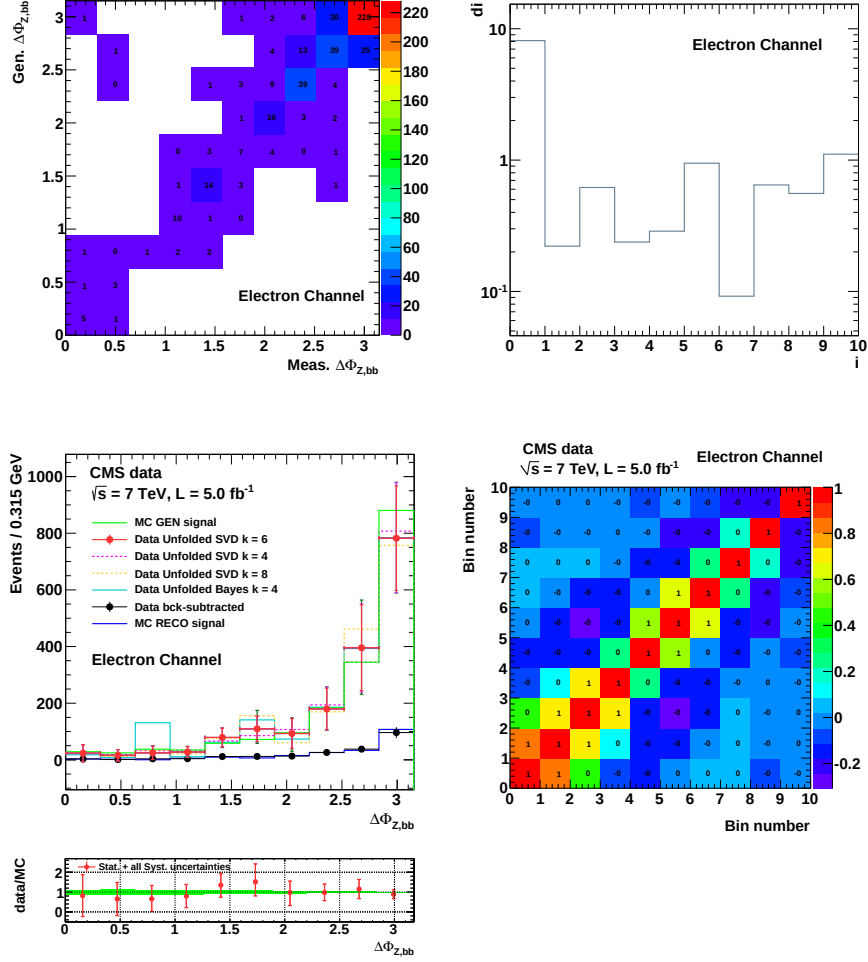


Figure C.22: Response matrix of the $\Delta\Phi_{Z,b\bar{b}}$ distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• $\Delta\Phi_{Z,bb}$ Muon Channel

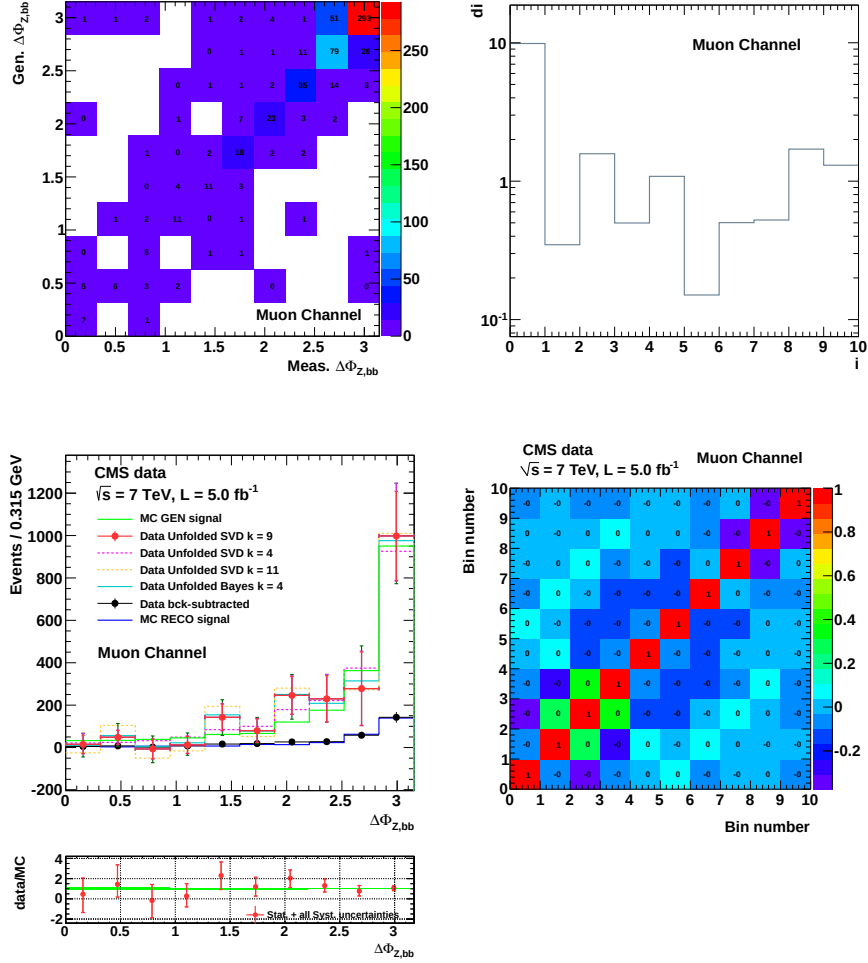


Figure C.23: Response matrix of the $\Delta\Phi_{Z,bb}$ distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

• $\Delta\Phi_{Z,bb}$ Combined Channel

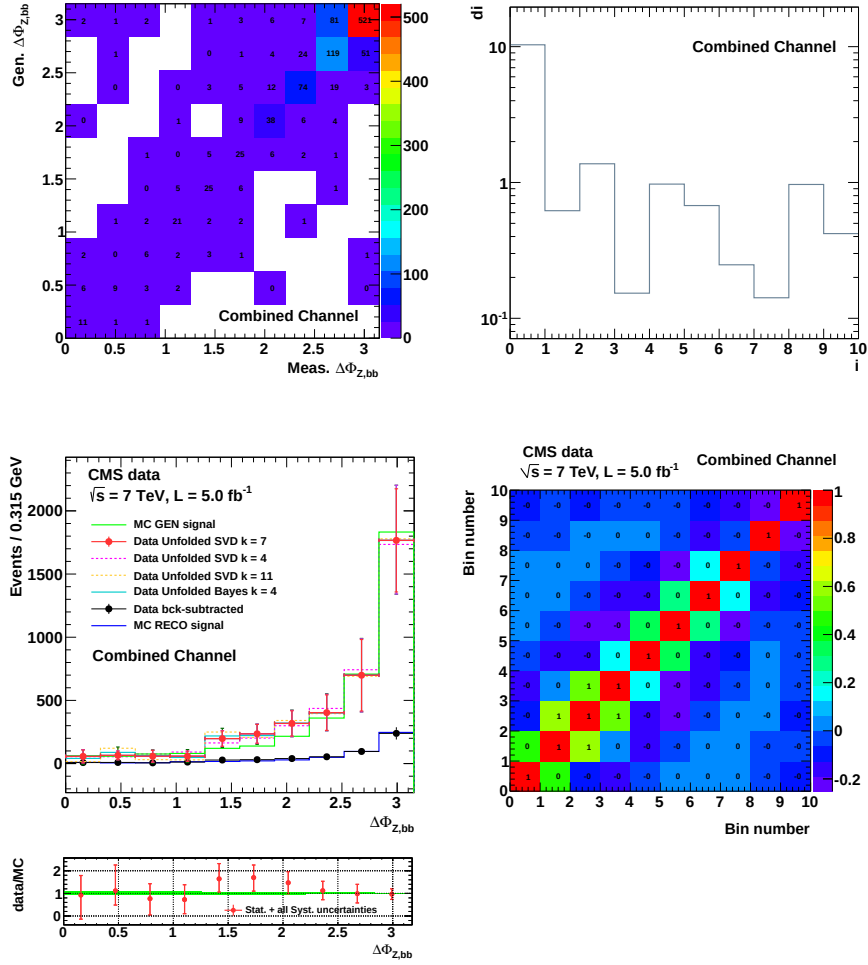


Figure C.24: Response matrix of the $\Delta\Phi_{Z,bb}$ distribution (top left). Decomposition of the vector d versus the parameter i (top right). Distributions of data before unfolding, and after unfolding with the optimal, and boundary SVD regularisation parameters (bottom left). Statistical uncertainty shown as the red part of the error bars. Systematic uncertainty from the method (from b -tag and JES) added in quadrature and shown as the green (violet) part of the error bars. Correspondent correlation matrix for the optimal unfolding (bottom right).

Appendix D

Generator Comparisons for all the Kinematic Observables

Results of the unfolding with SVD of the data distribution from the 5.0 fb^{-1} luminosity data sample are compared with the different MADGRAPH events productions listed in the Table 4.4, for the electron, muon and combined channels.

- Transverse momentum of the Z boson
- p_T^Z Electron Channel

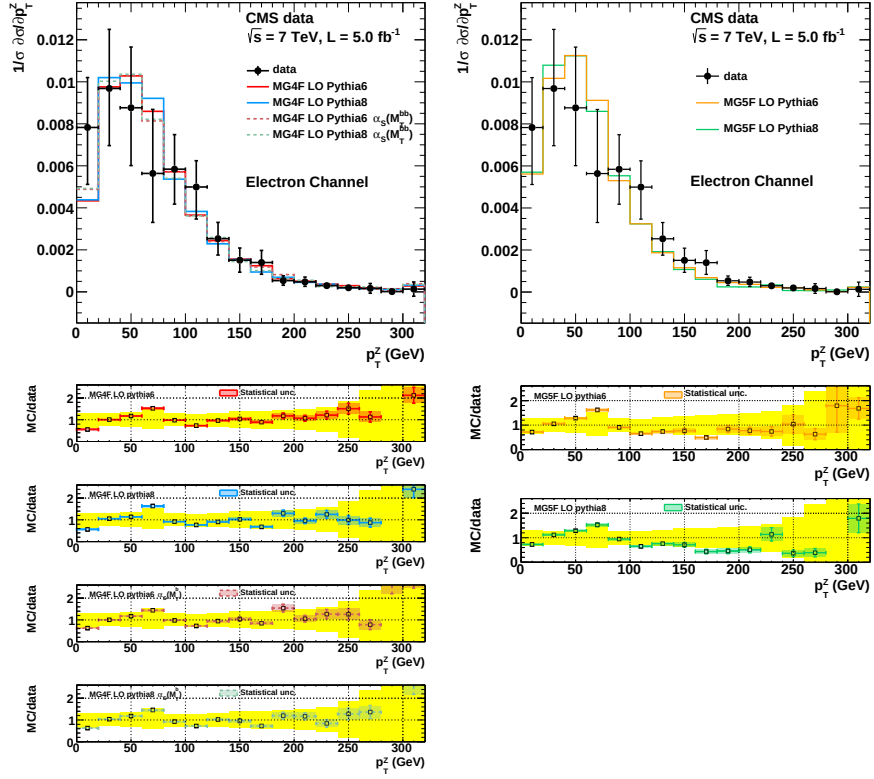


Figure D.1: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the Z boson in the electron channel.

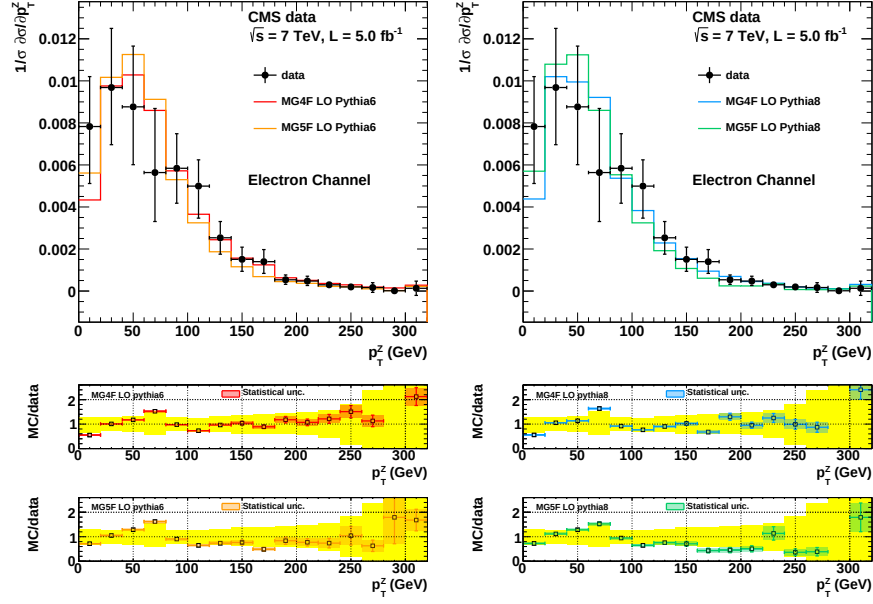


Figure D.2: Comparison of unfolded data with events generated with MG4F and MG5F generators, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the Z boson in the electron channel.

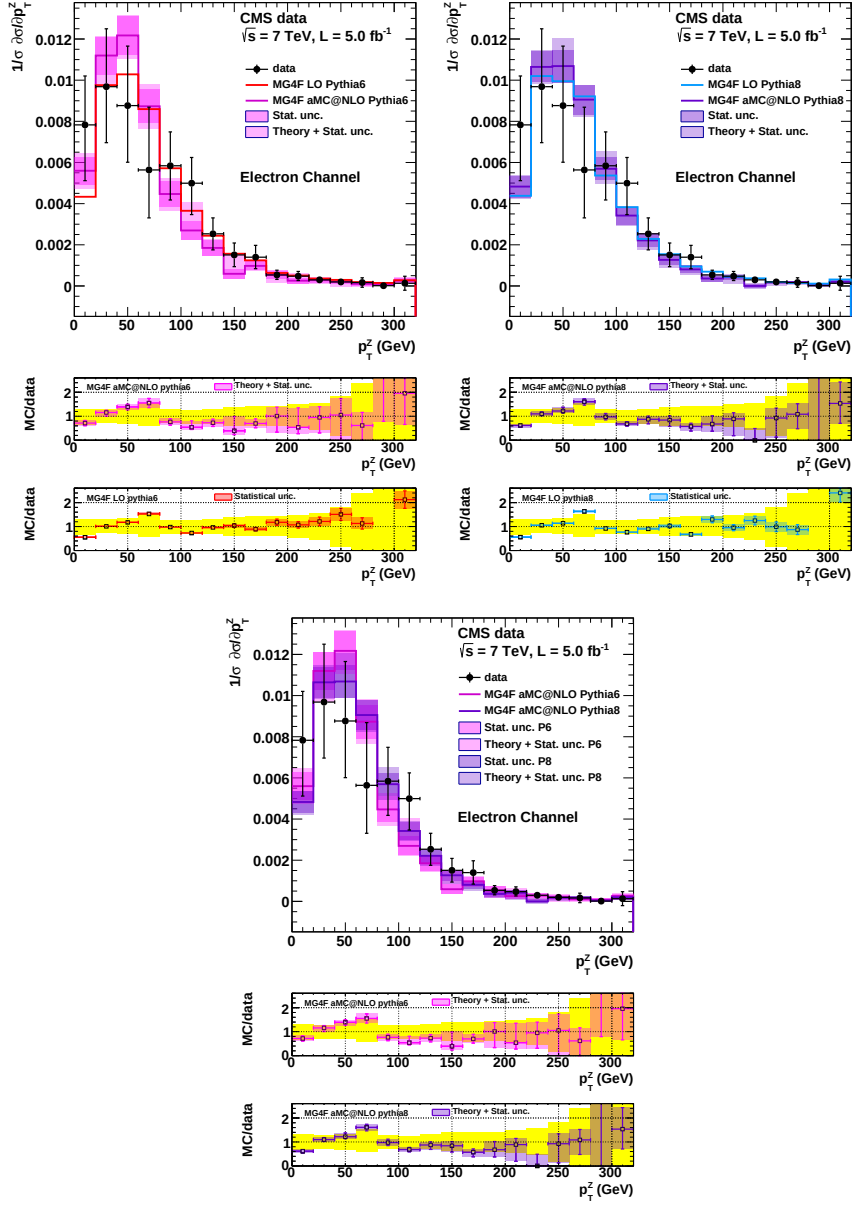


Figure D.3: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the p_T of the Z boson in the electron channel.

• p_T^Z Muon Channel

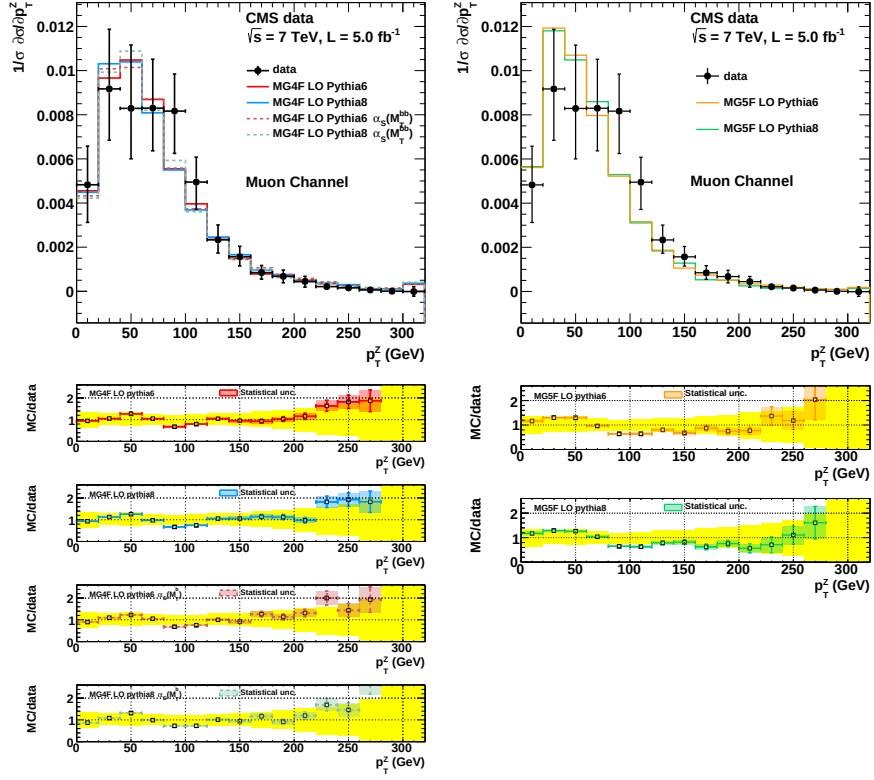


Figure D.4: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the Z boson in the muon channel.

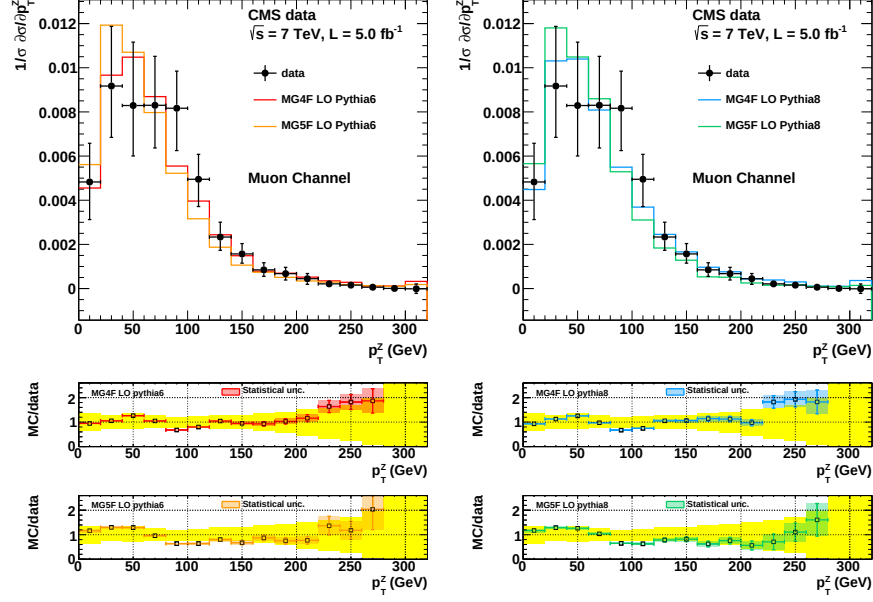


Figure D.5: Comparison of unfolded data with events generated with MG4F and MG5F generators, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the Z boson in the muon channel.

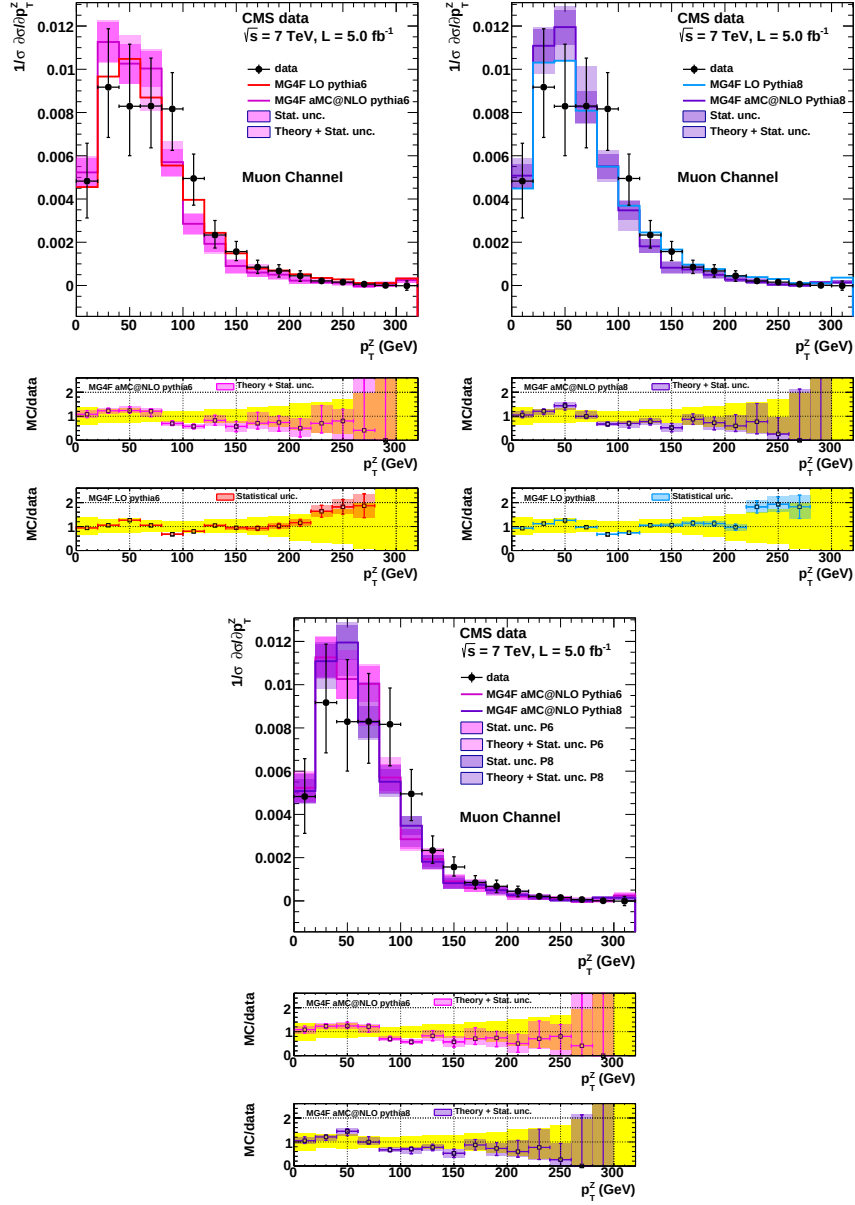


Figure D.6: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the p_T of the Z boson in the muon channel.

• p_T^Z Combined Channel

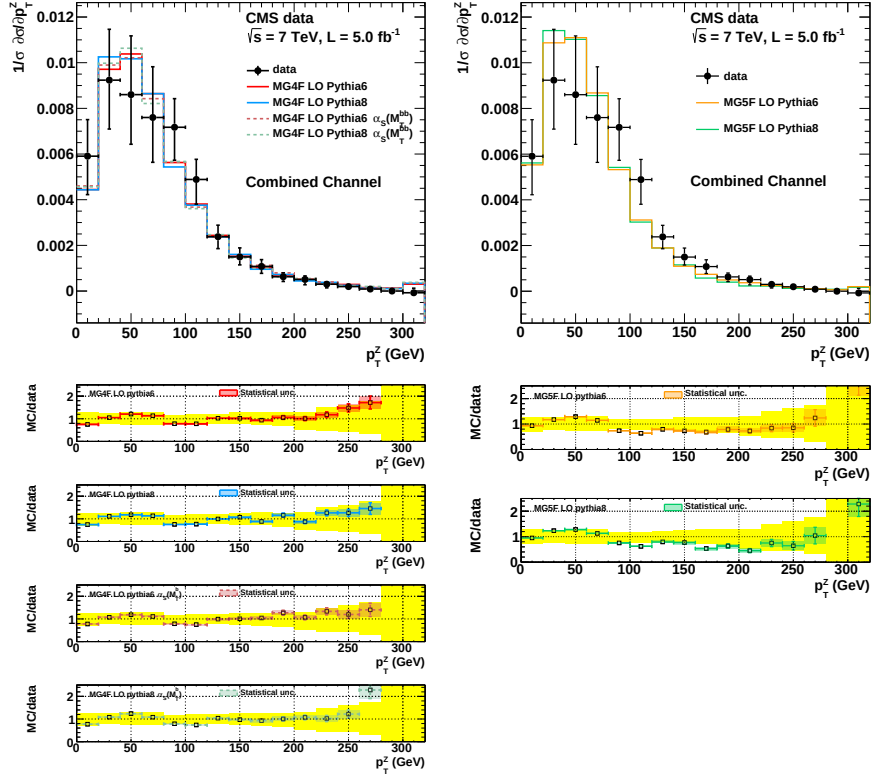


Figure D.7: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the Z boson in the combined channel.

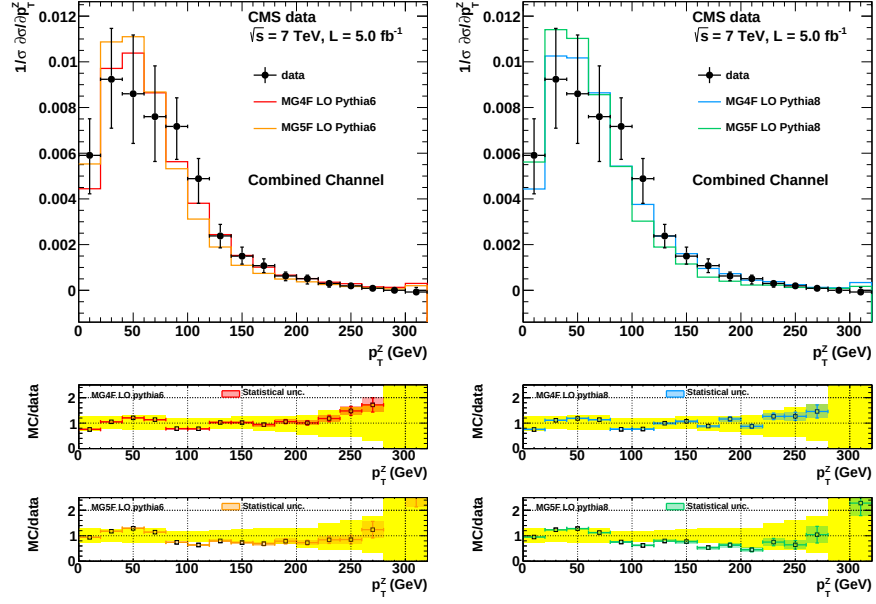


Figure D.8: Comparison of unfolded data with events generated with MG4F and MG5F generators, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the Z boson in the combined channel.

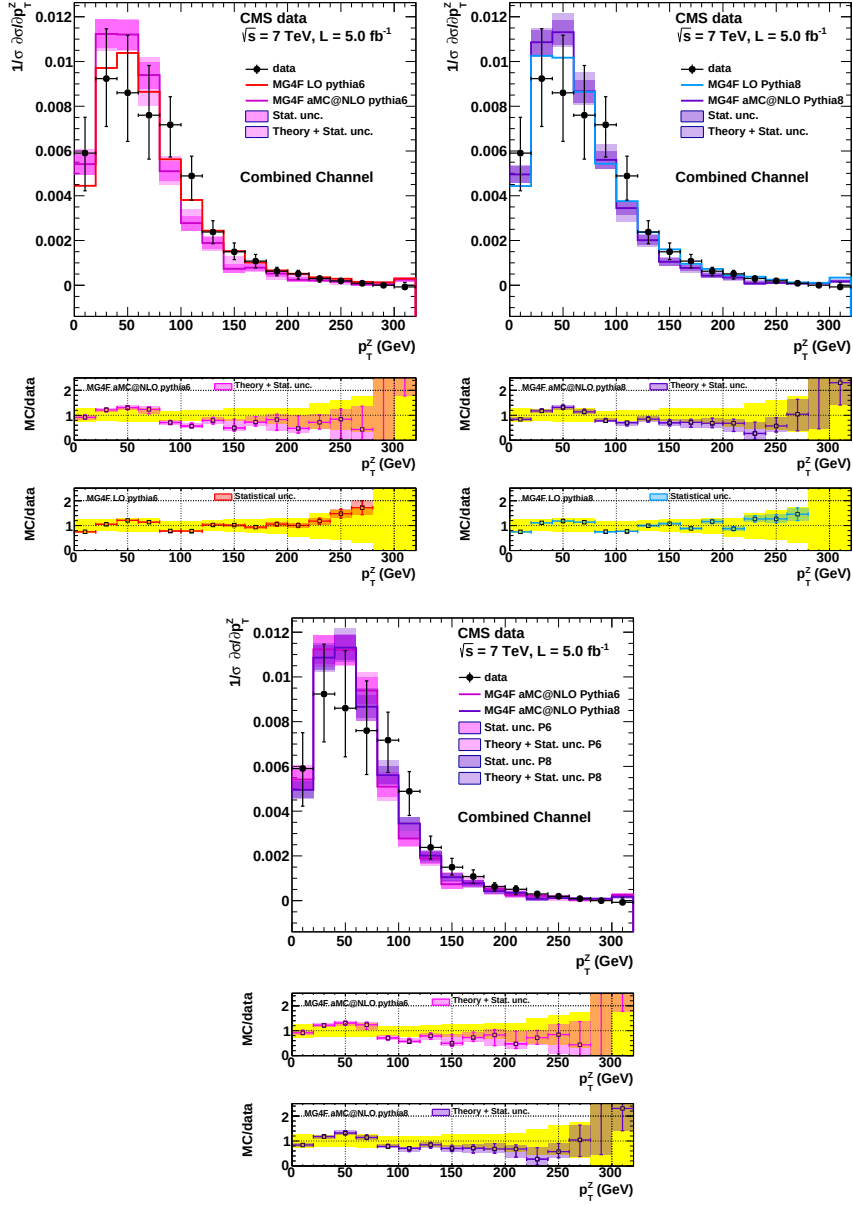


Figure D.9: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the p_T of the Z boson in the combined channel.

- Transverse momentum of the leading b jet
- $p_T^{b\text{-leading}}$ Electron Channel

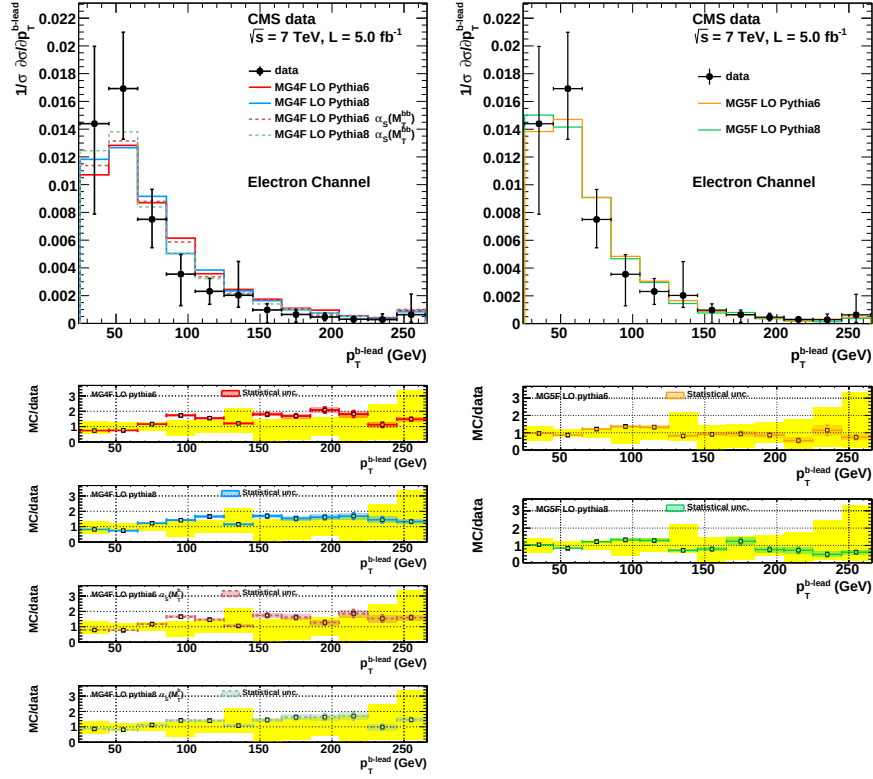


Figure D.10: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the leading b jet in the electron channel.

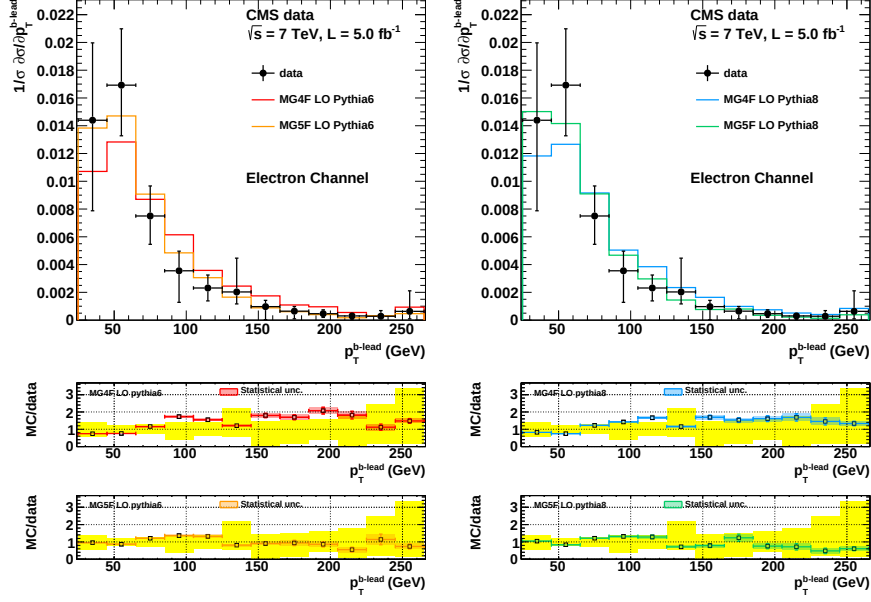


Figure D.11: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the leading b jet in the electron channel.

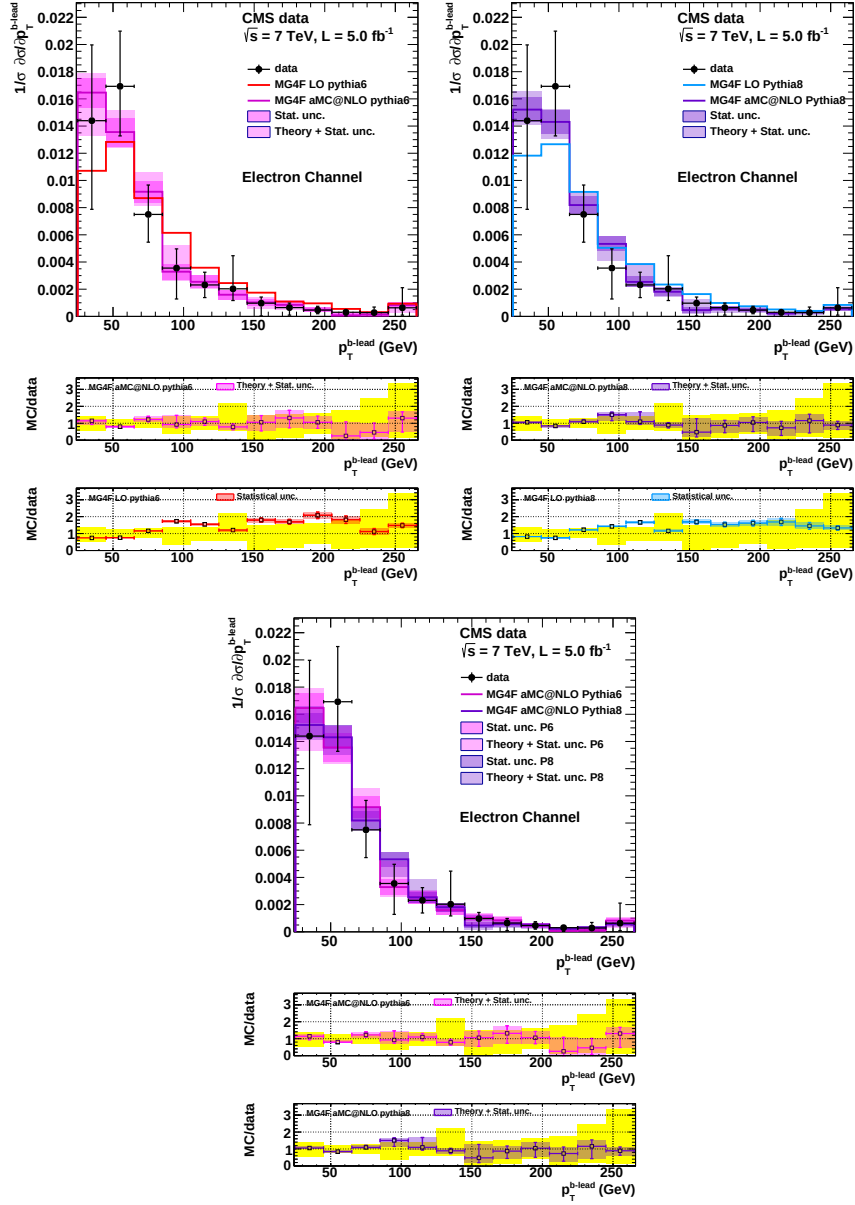


Figure D.12: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the p_T of the lead, b jet in the electron channel.

• $p_T^{b\text{-leading}}$ Muon Channel

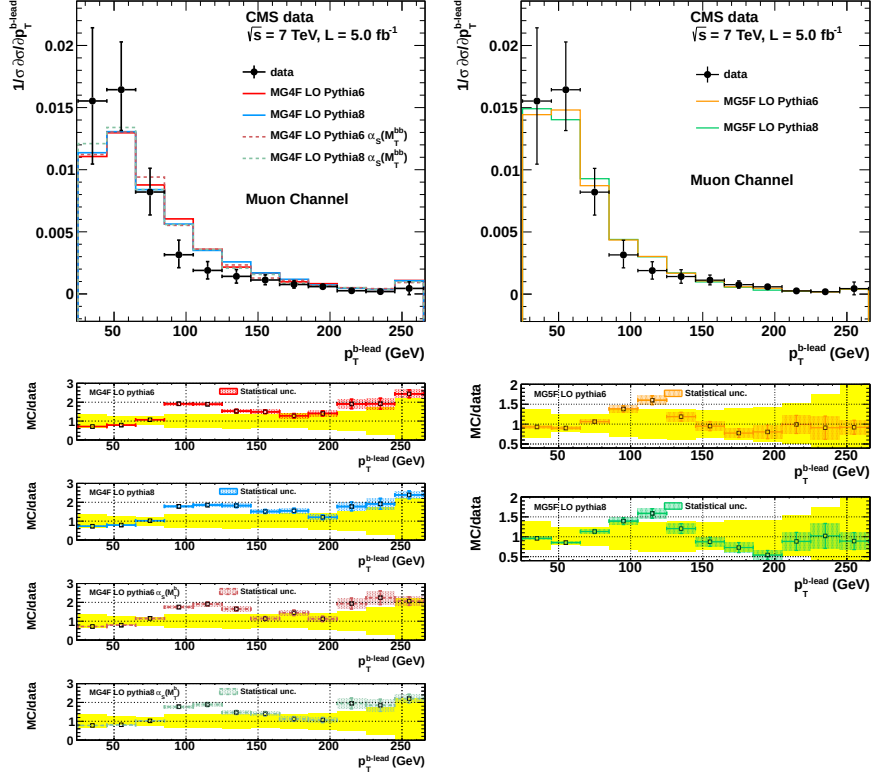


Figure D.13: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the leading b jet in the muon channel.

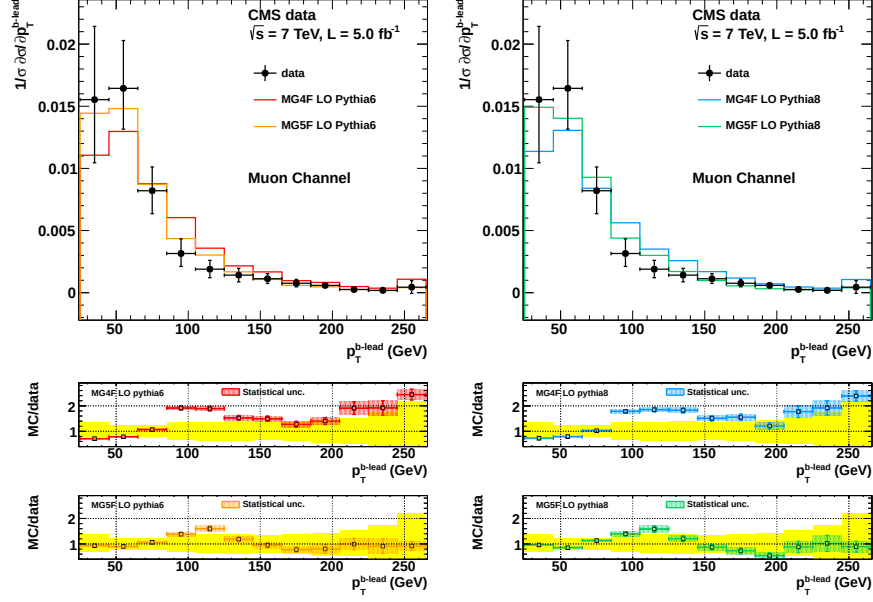


Figure D.14: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the leading b jet in the muon channel.

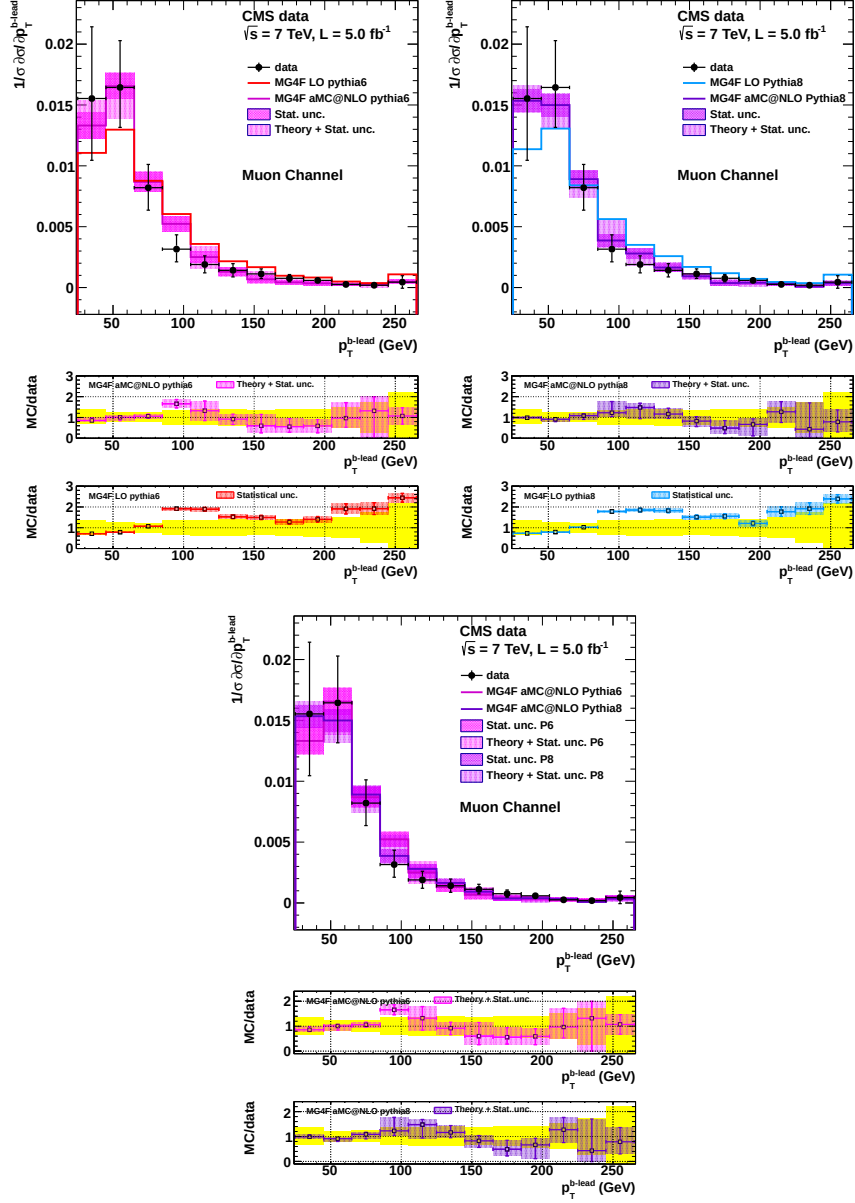


Figure D.15: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the p_T of the leading b jet in the muon channel.

• $p_T^{b\text{-leading}}$ Combined Channel

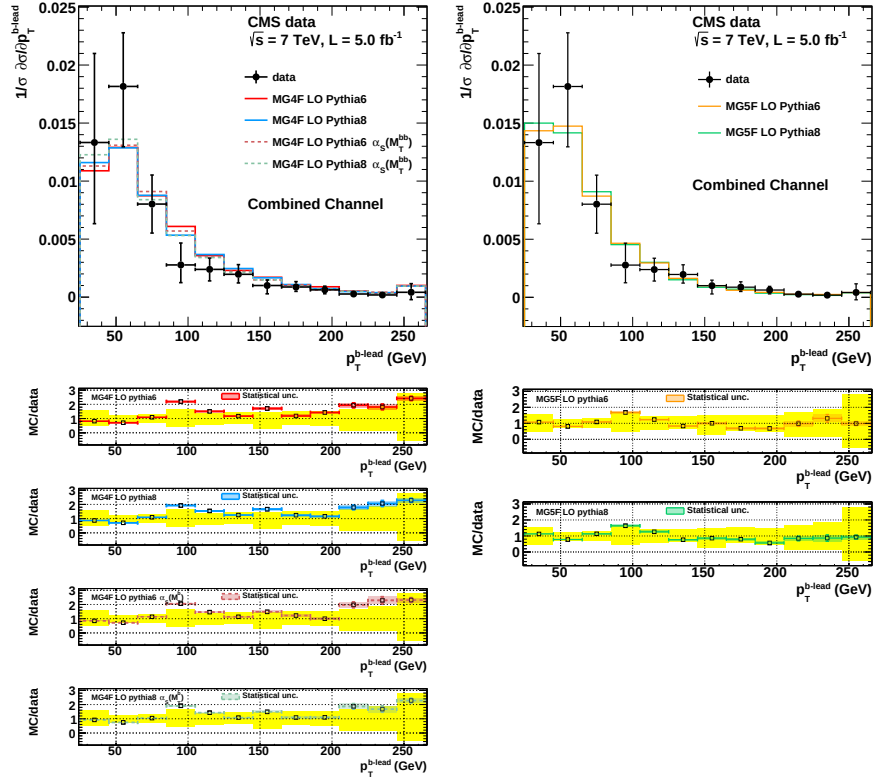


Figure D.16: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the leading b jet in the combined channel.

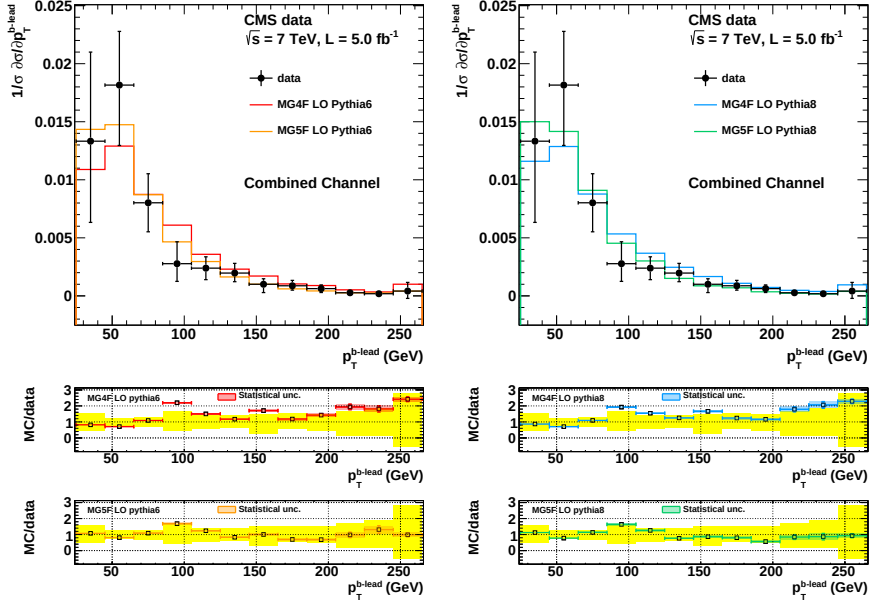


Figure D.17: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the leading b jet in the combined channel.

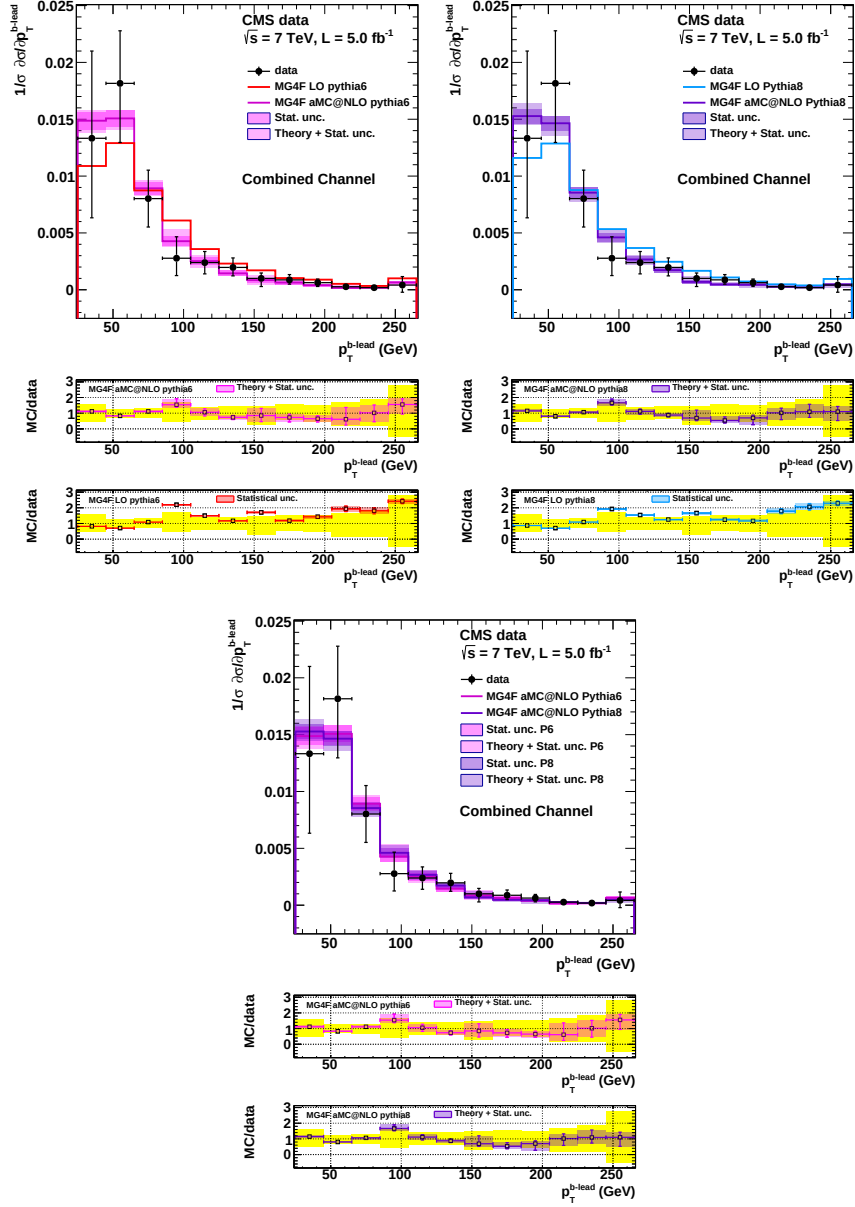


Figure D.18: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of p_T of the lead b jet in the combined channel.

- Transverse momentum of the subleading b jet
- $p_T^{\text{b-subleading}}$ Electron Channel

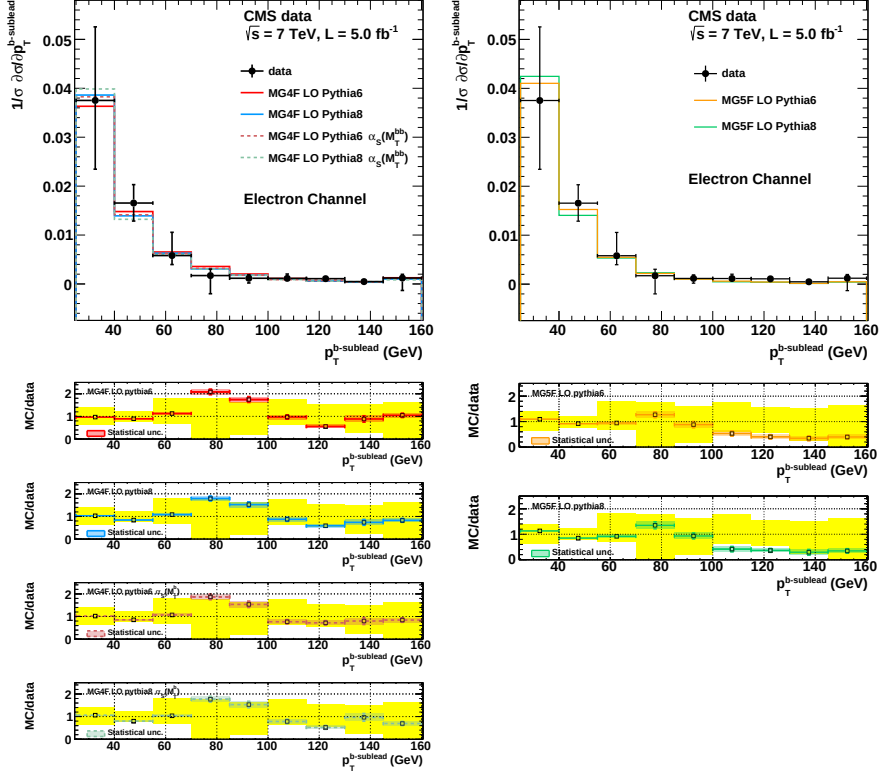


Figure D.19: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the subleading b jet in the electron channel.

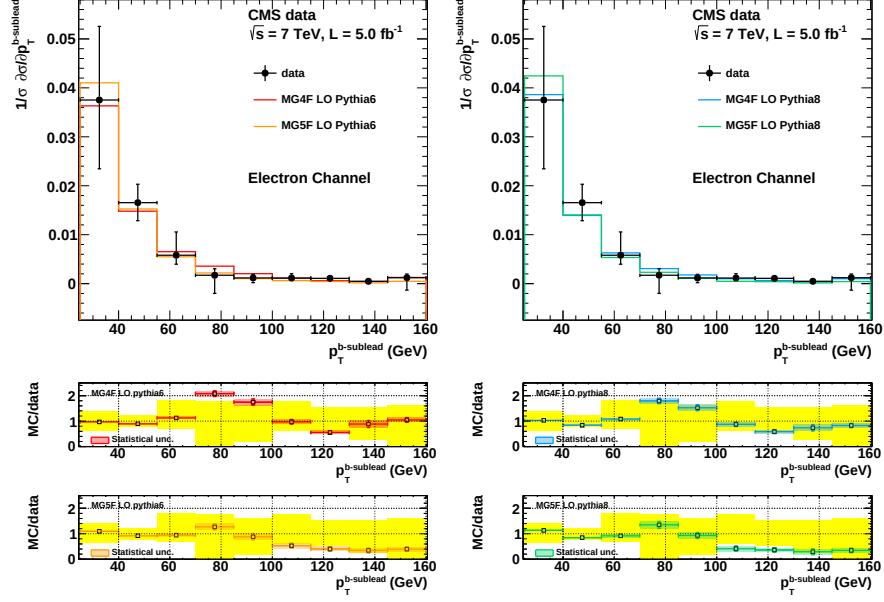


Figure D.20: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the subleading b jet in the electron channel.

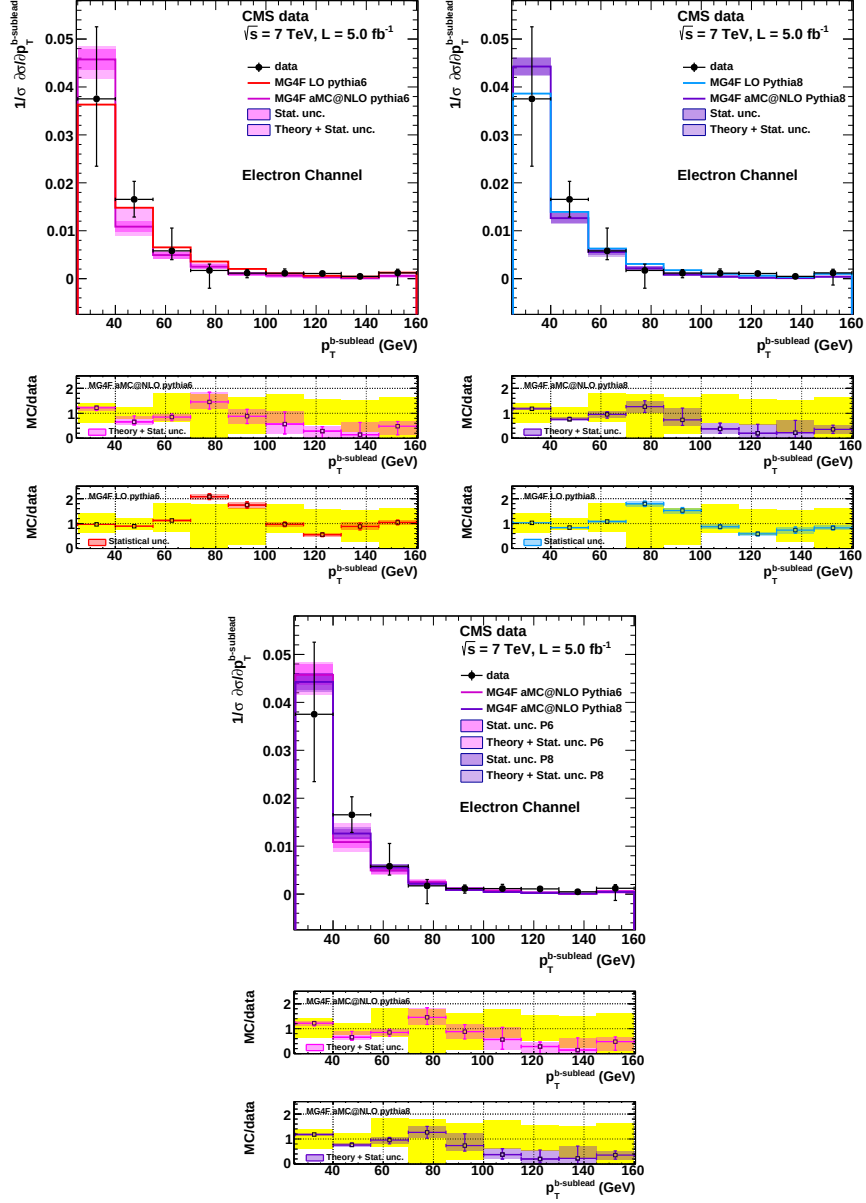


Figure D.21: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the subleading b jet in the electron channel.

• $p_T^{b\text{-subleading}}$ Muon Channel

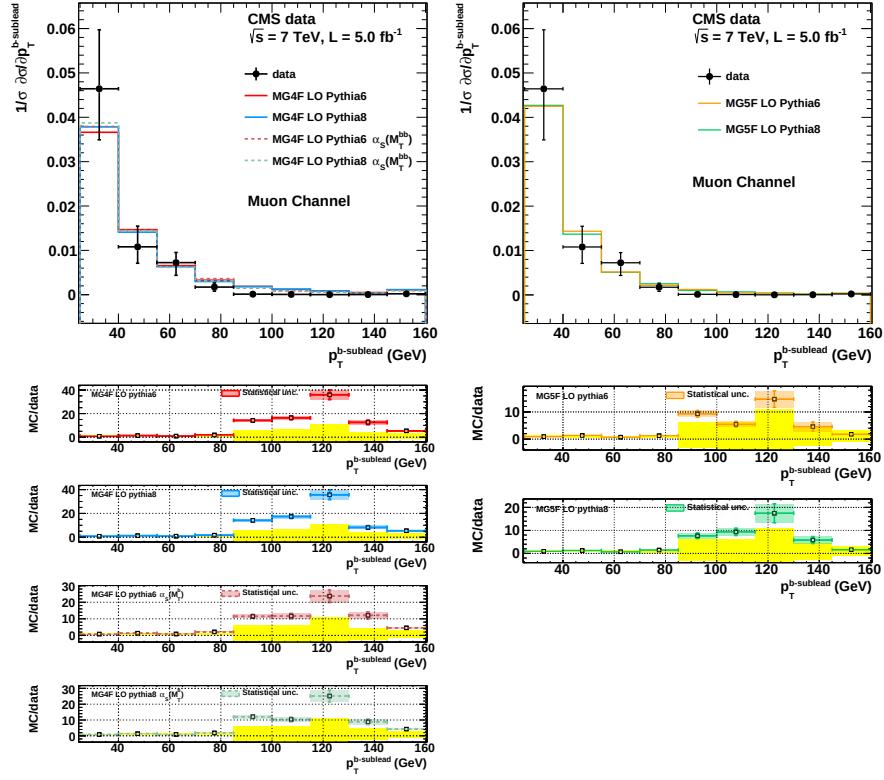


Figure D.22: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the subleading b jet in the muon channel.

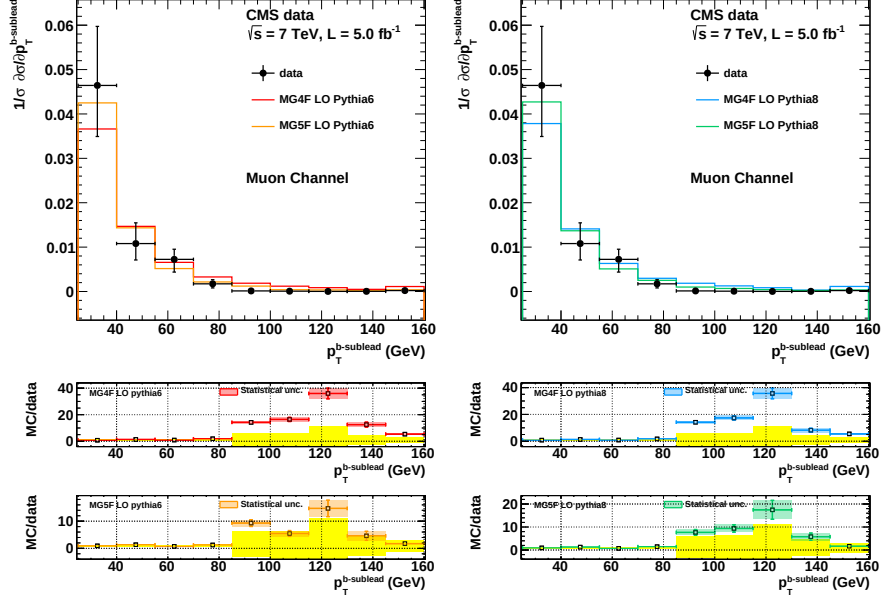


Figure D.23: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the subleading b jet in the muon channel.

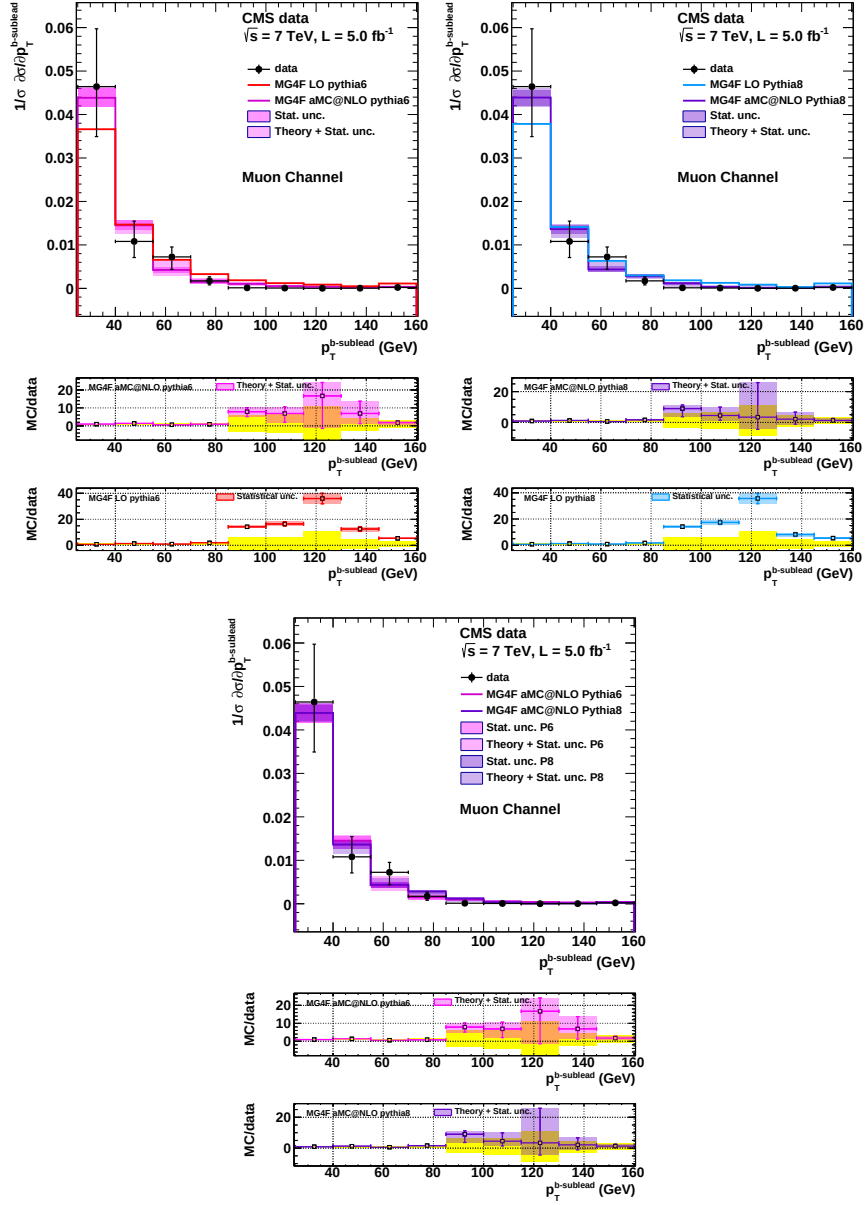


Figure D.24: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the subleading b jet in the muon channel.

• $p_T^{\text{b-subleading}}$ Combined Channel

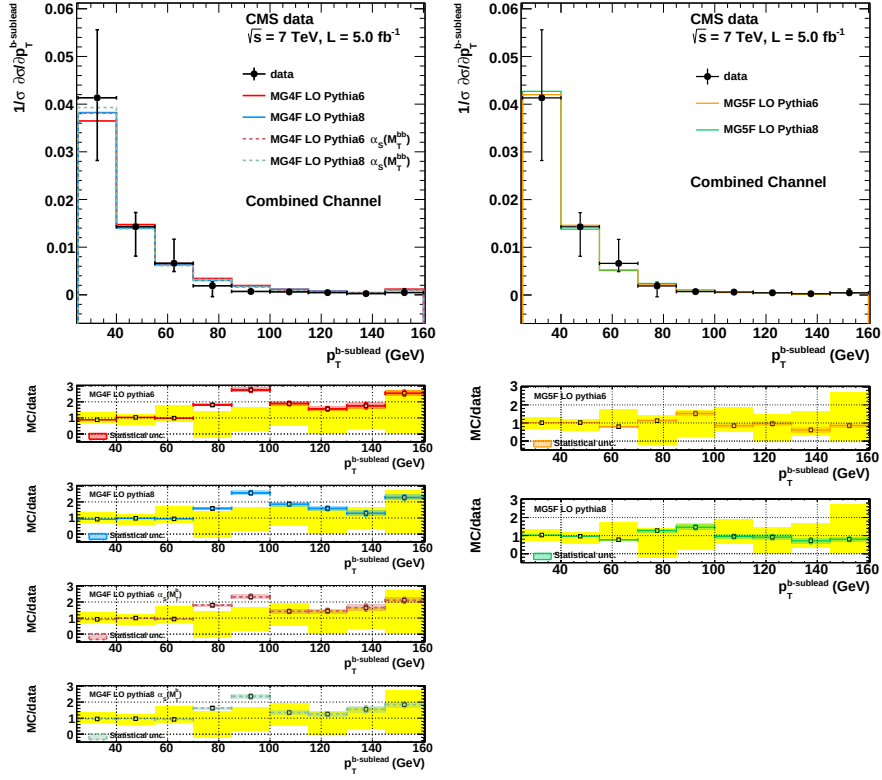


Figure D.25: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T of the subleading b jet in the combined channel.

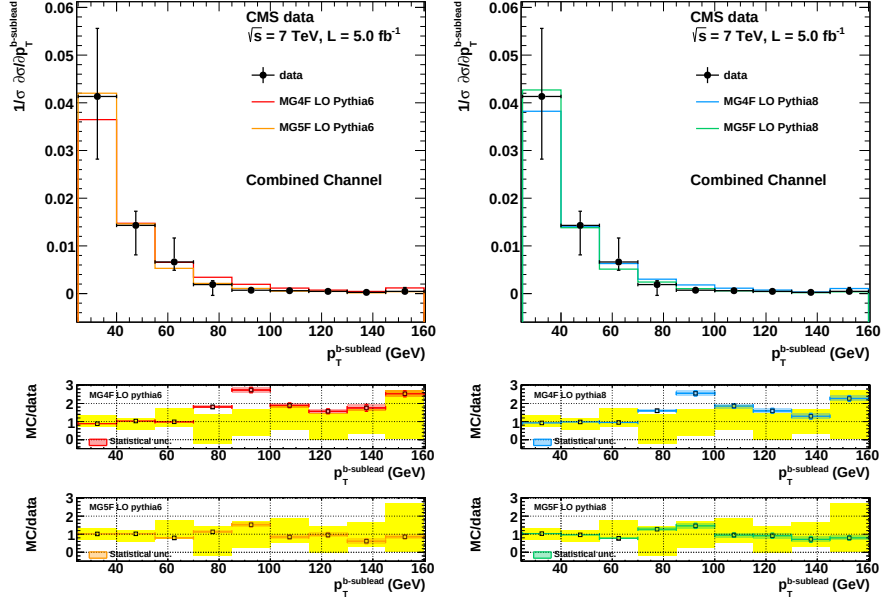


Figure D.26: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T of the subleading b jet in the combined channel.

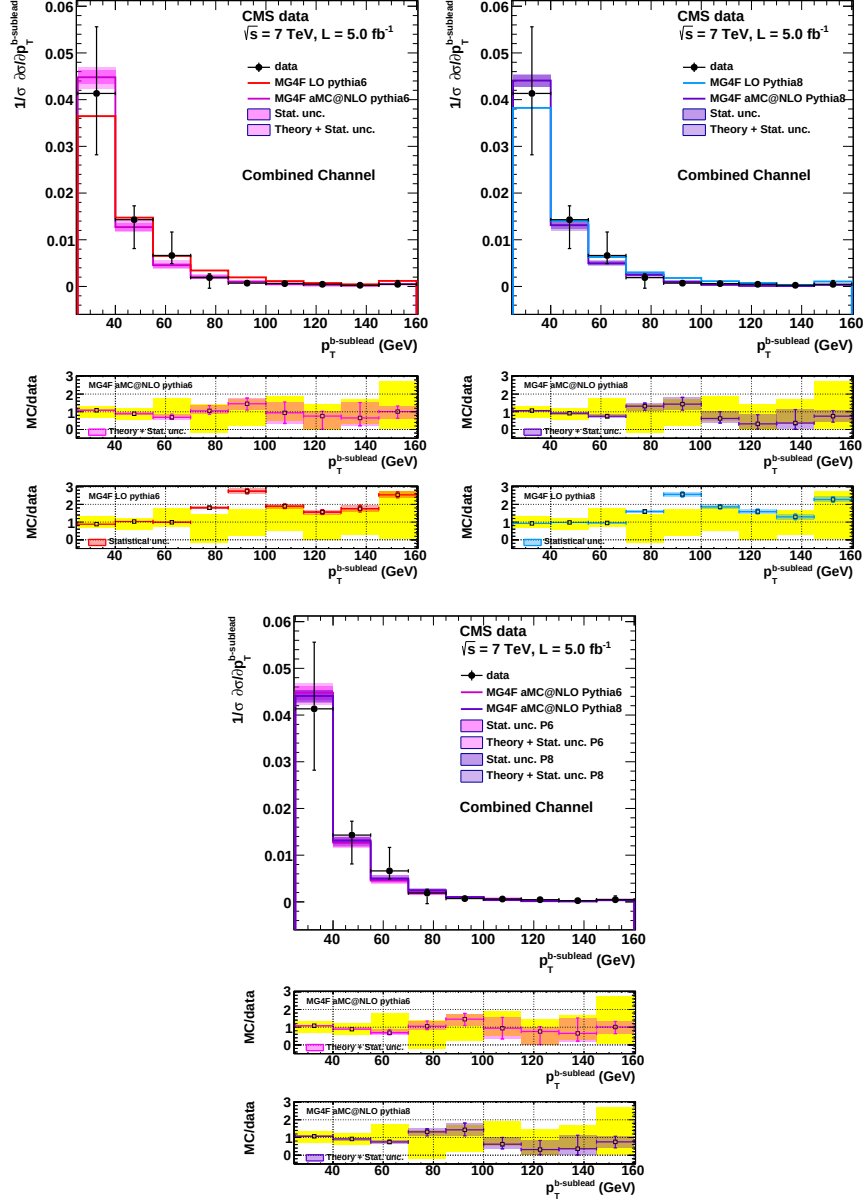


Figure D.27: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the $p_T^{b\text{-subleading}}$ in the combined channel.

- Transverse momentum of the $b\bar{b}$ system
- $p_T^{b\bar{b}}$ Electron Channel

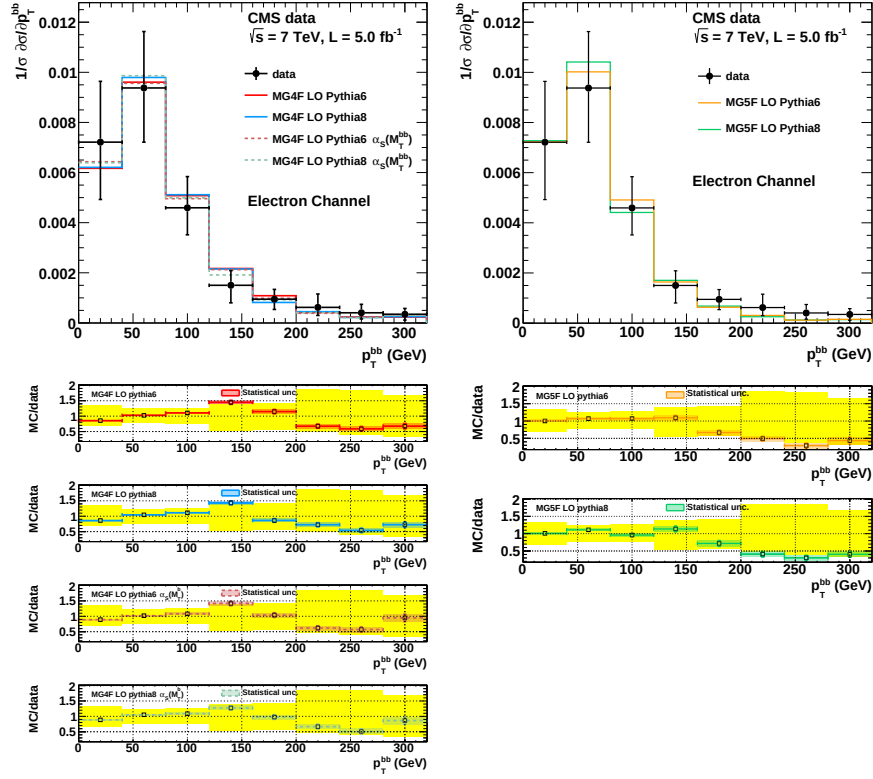


Figure D.28: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T $b\bar{b}$ pair in the electron channel.

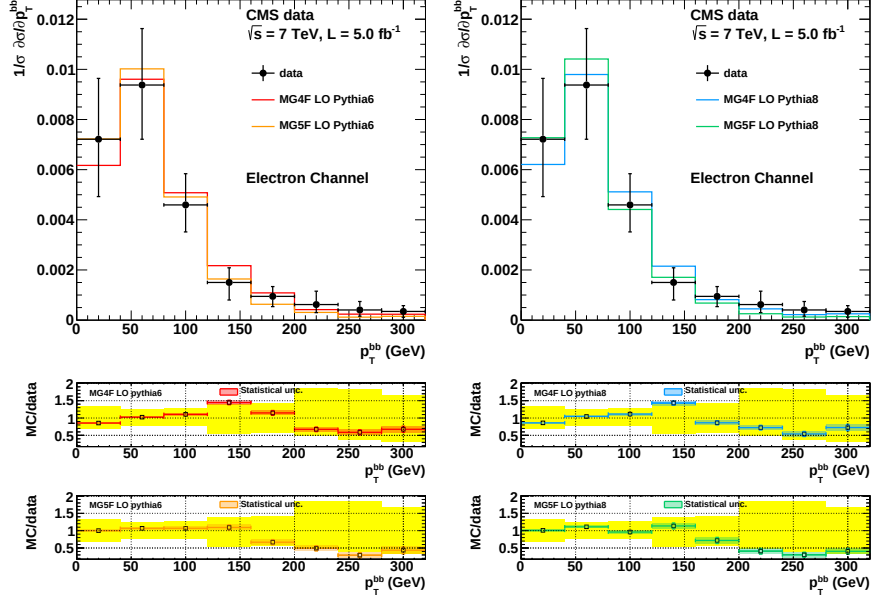


Figure D.29: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T bb pair in the electron channel.

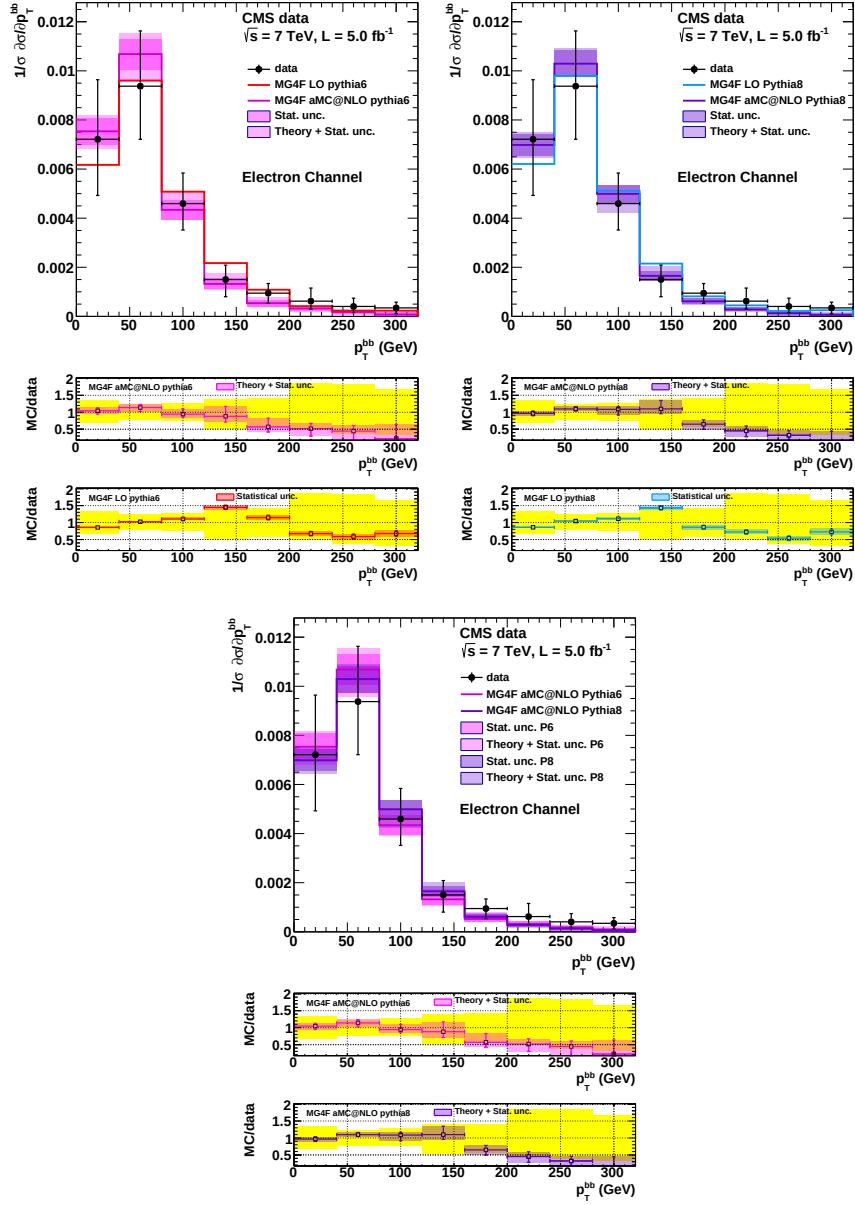


Figure D.30: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the p_T^{bb} pair in the electron channel.

• p_T^{bb} Muon Channel

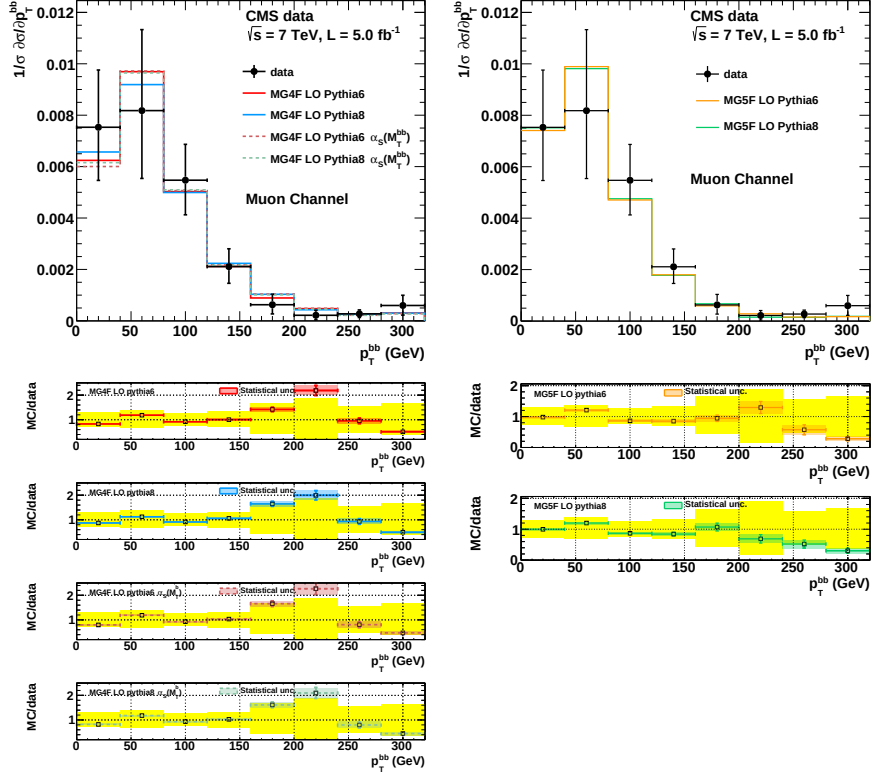


Figure D.31: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T^{bb} pair in the muon channel.

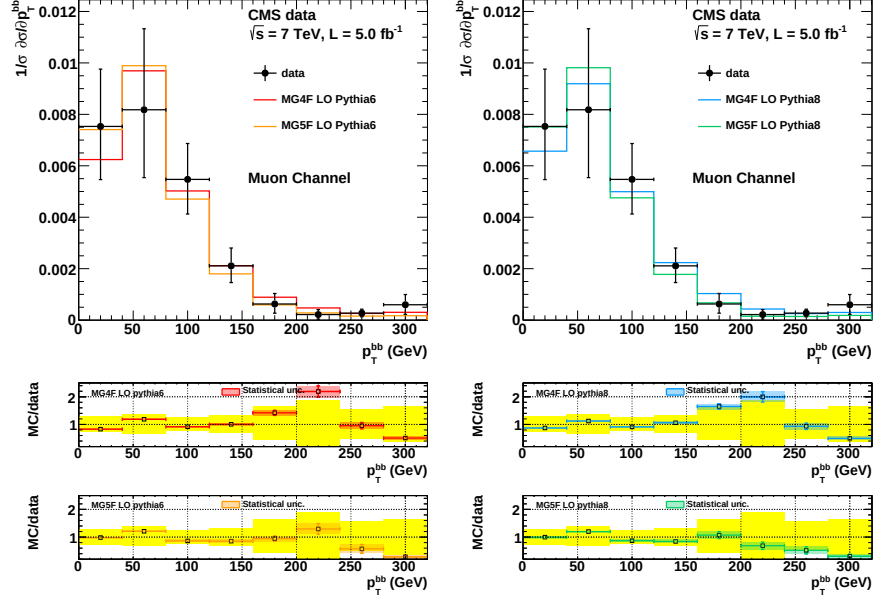


Figure D.32: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T bb pair in the muon channel.

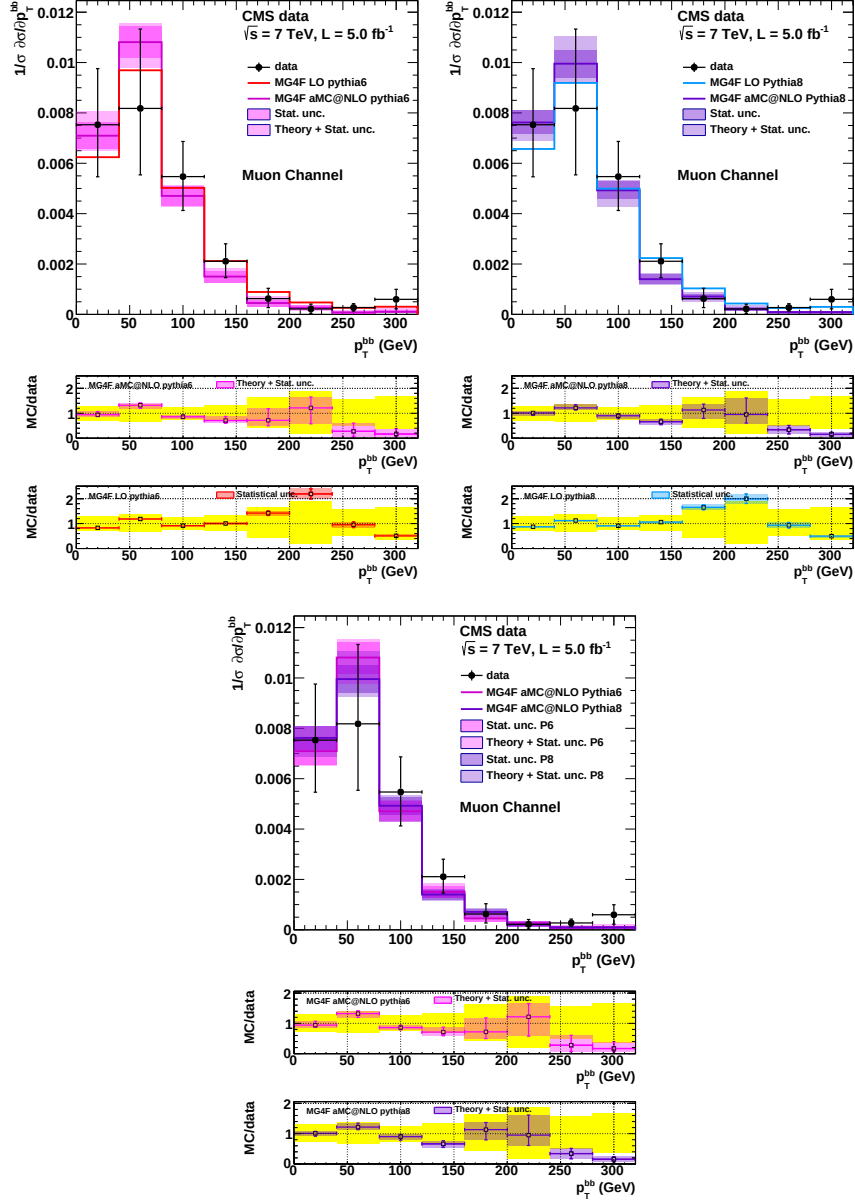


Figure D.33: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the p_T^{bb} pair in the muon channel.

• p_T^{bb} Combined Channel

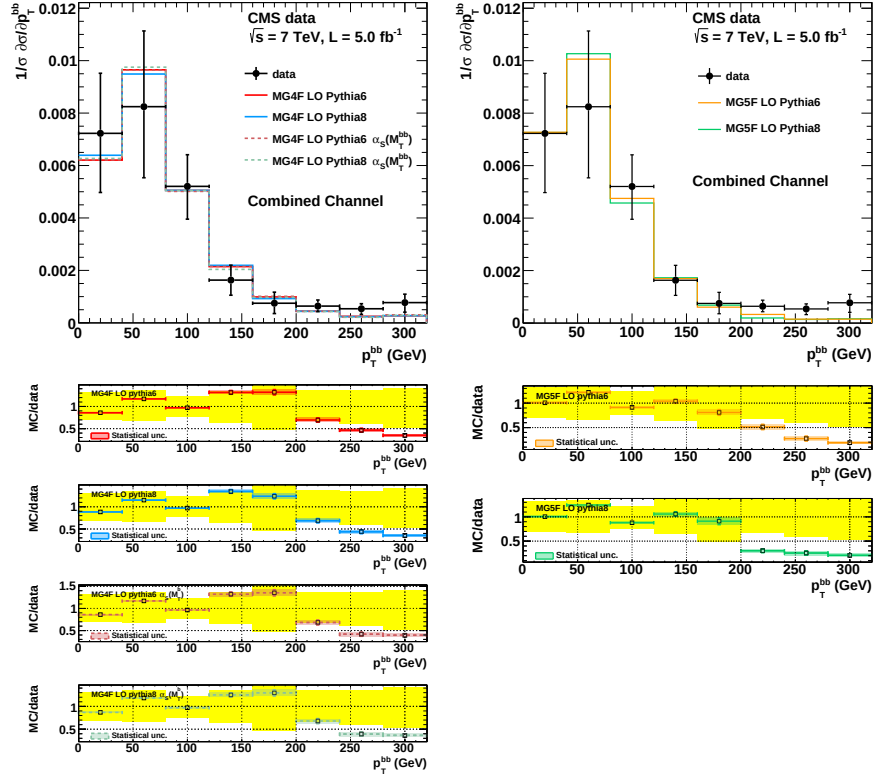


Figure D.34: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the p_T^{bb} pair in the combined channel.

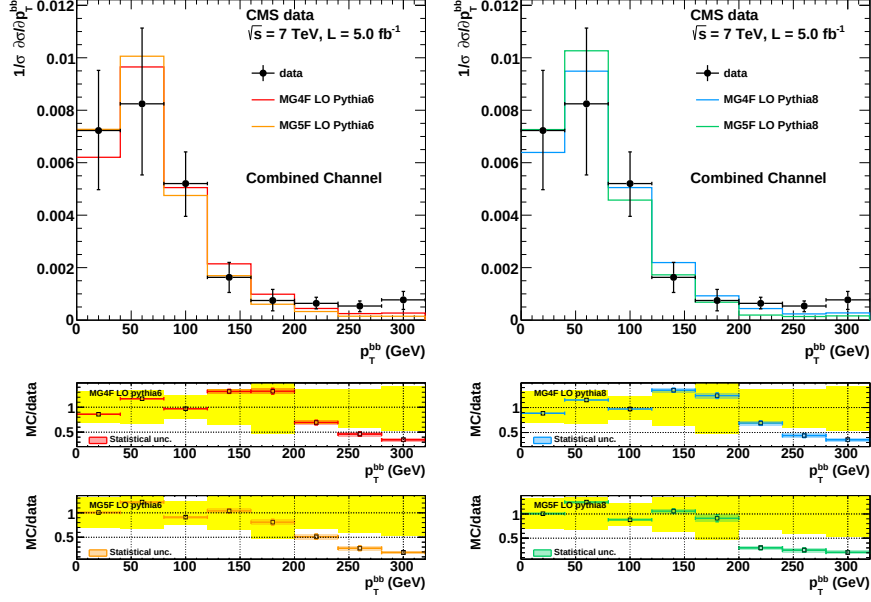


Figure D.35: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the p_T^{bb} pair in the combined channel.

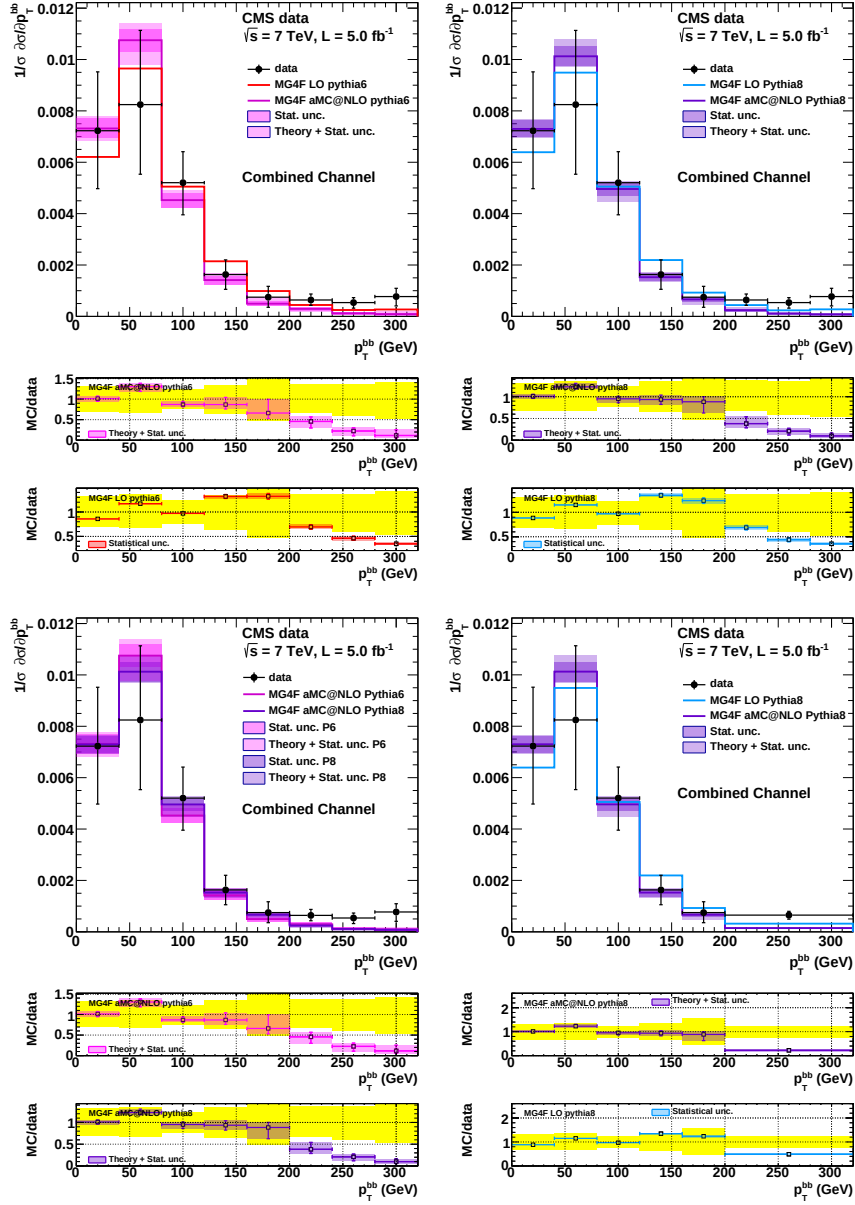


Figure D.36: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (top left) and PYTHIA 8 (right, top and bottom); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (bottom left), for the p_T^{bb} distribution in the combined channel.

- Invariant mass of the $b\bar{b}$ system
- $M_{b\bar{b}}$ Electron Channel

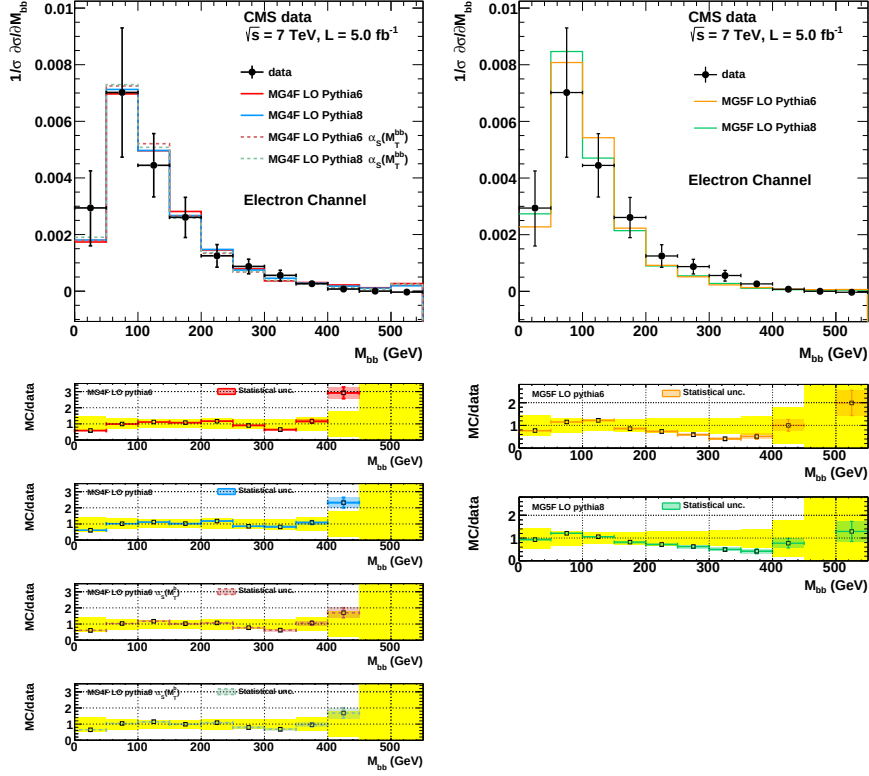


Figure D.37: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the $M_{b\bar{b}}$ in the electron channel.

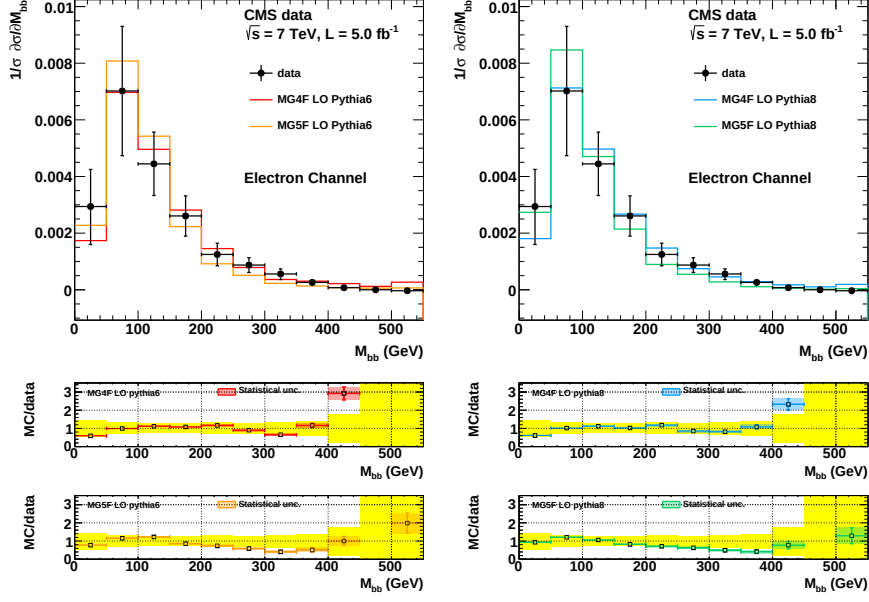


Figure D.38: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the M_{bb} in the electron channel.

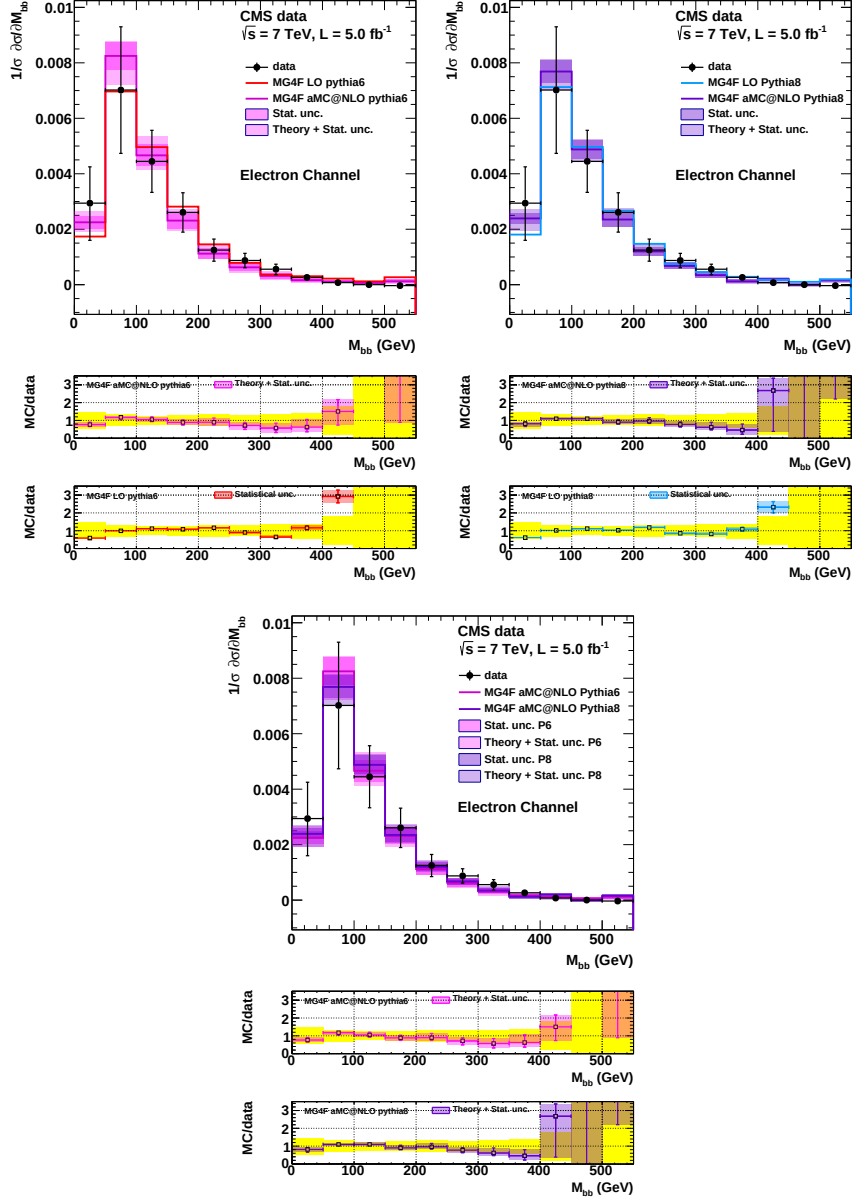


Figure D.39: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the M_{bb} in the electron channel.

• M_{bb} Muon Channel

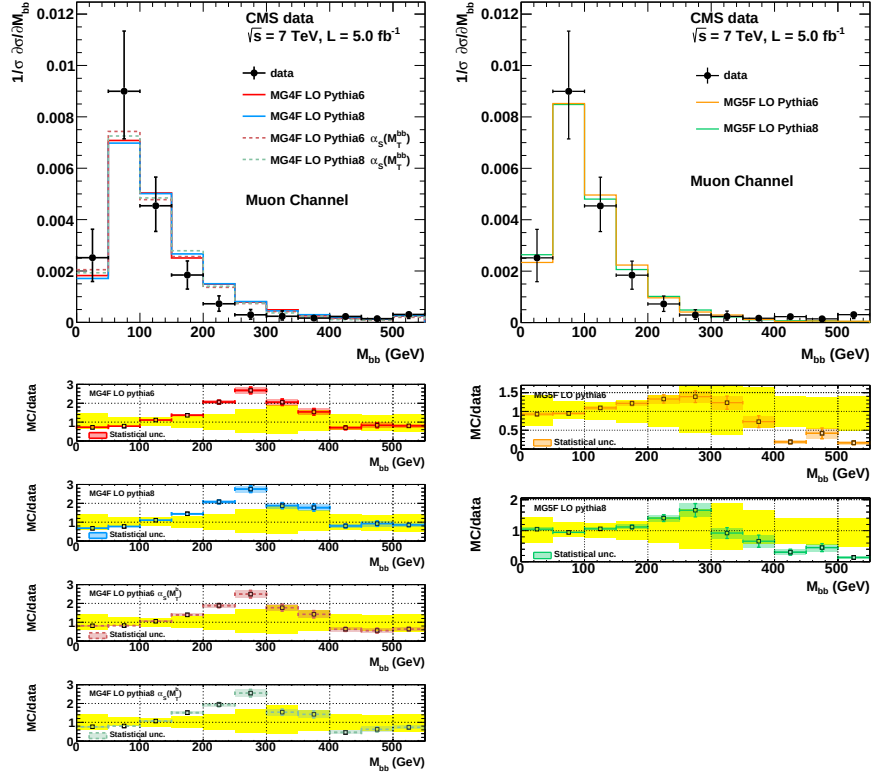


Figure D.40: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the M_{bb} in the muon channel.

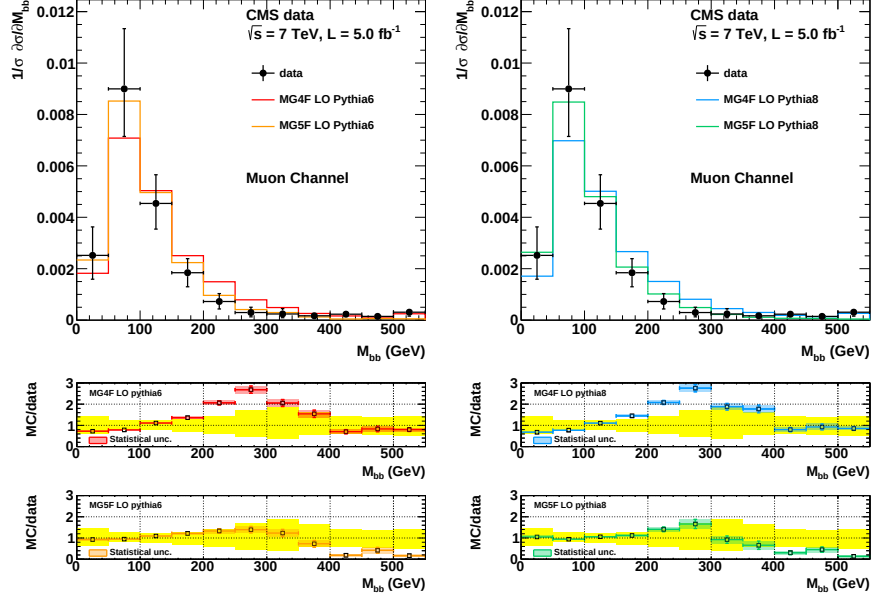


Figure D.41: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the M_{bb} in the muon channel.

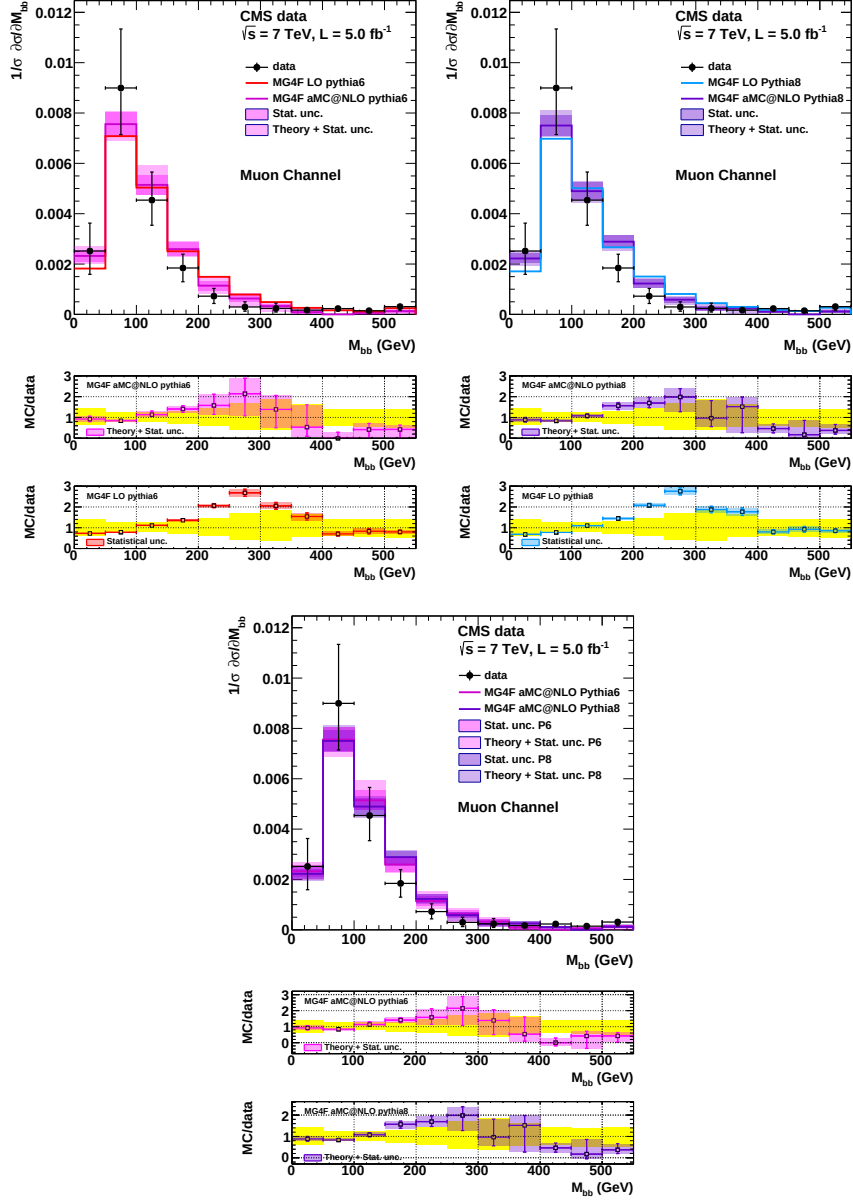


Figure D.42: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the M_{bb} in the muon channel.

• M_{bb} Combined Channel

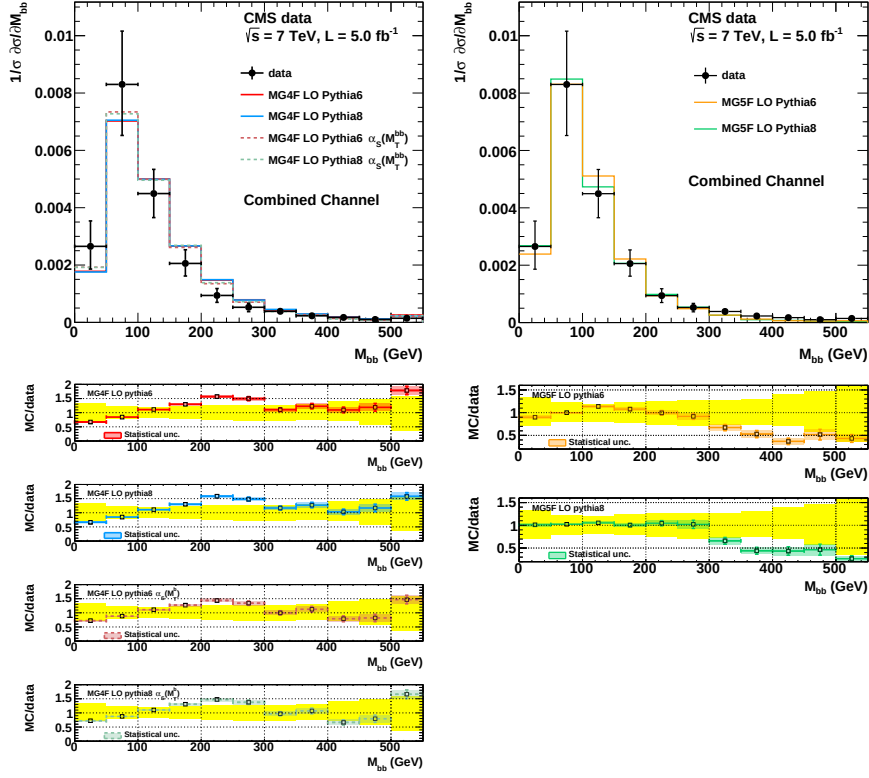


Figure D.43: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the M_{bb} in the combined channel.

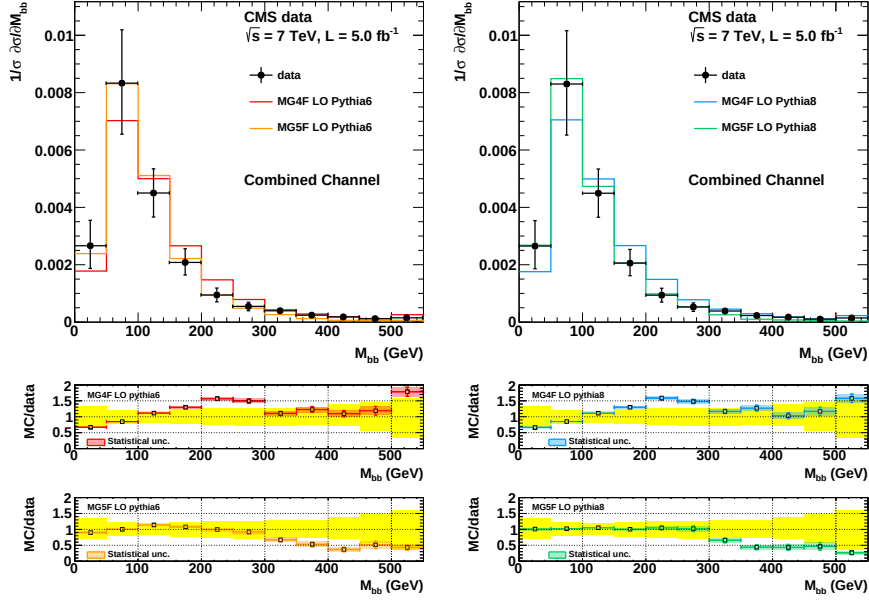


Figure D.44: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the M_{bb} in the combined channel.

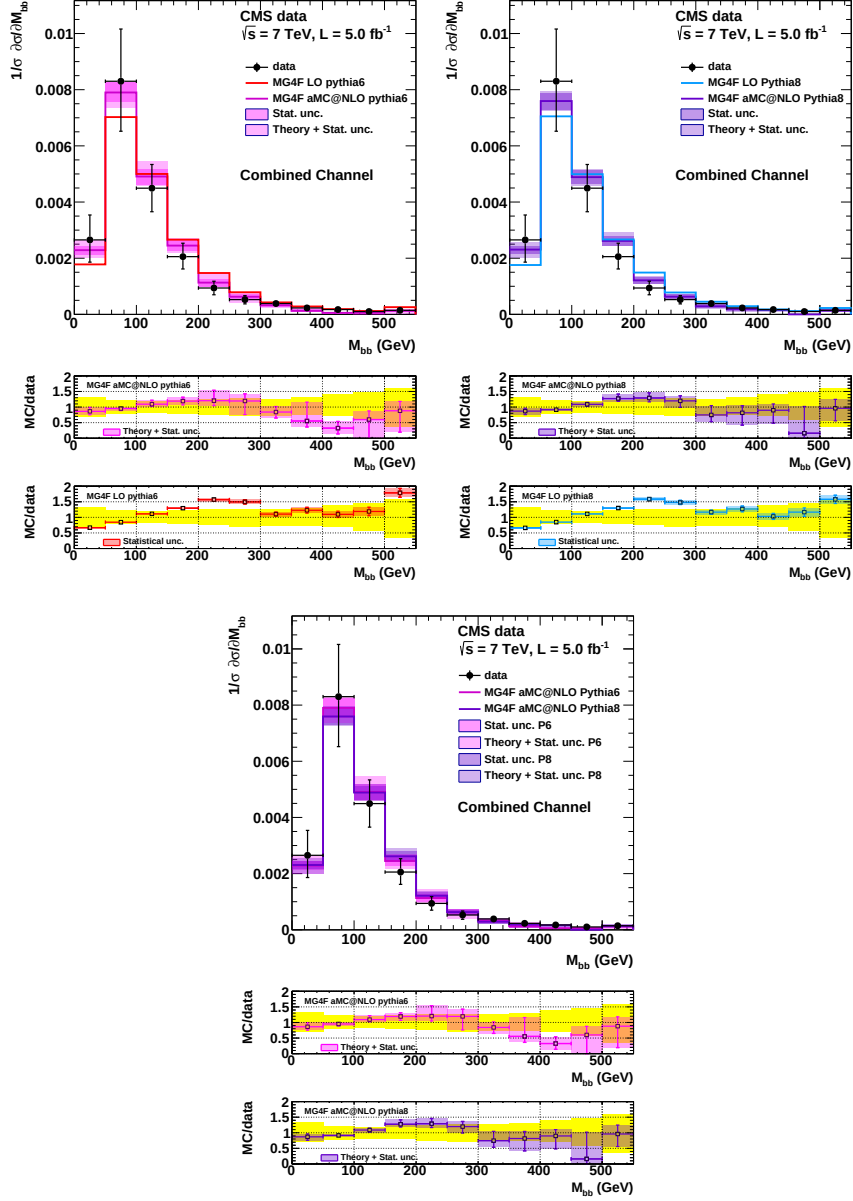


Figure D.45: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the M_{bb} in the combined channel.

- Invariant mass of the $Zb\bar{b}$ system
- $M_{Zb\bar{b}}$ Electron Channel

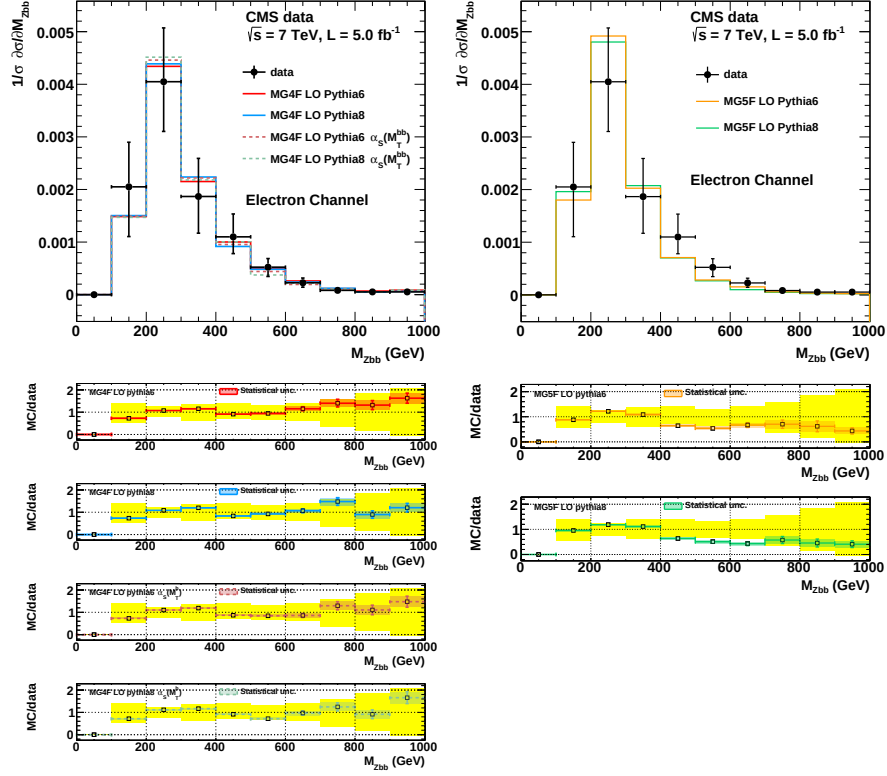


Figure D.46: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the $M_{Zb\bar{b}}$ in the electron channel.

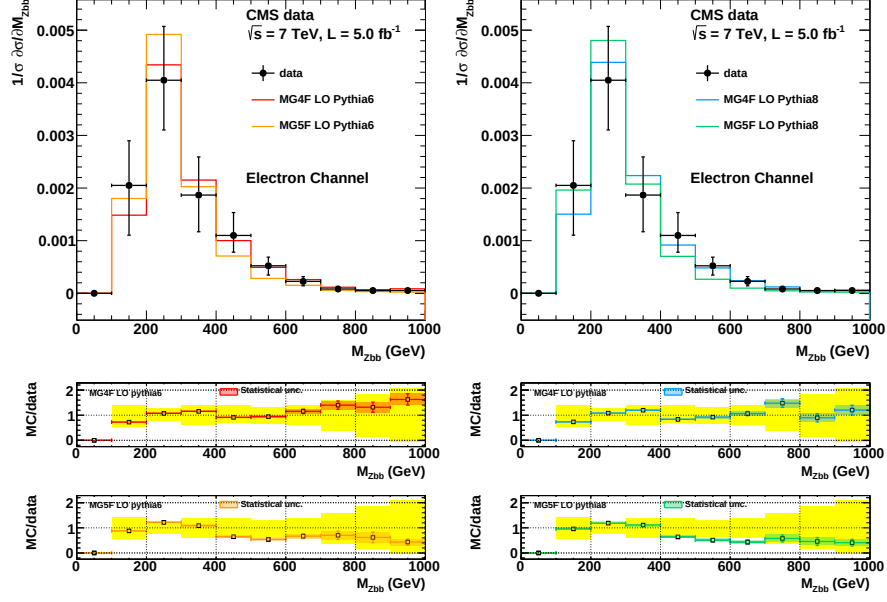


Figure D.47: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the M_{Zbb} in the electron channel.

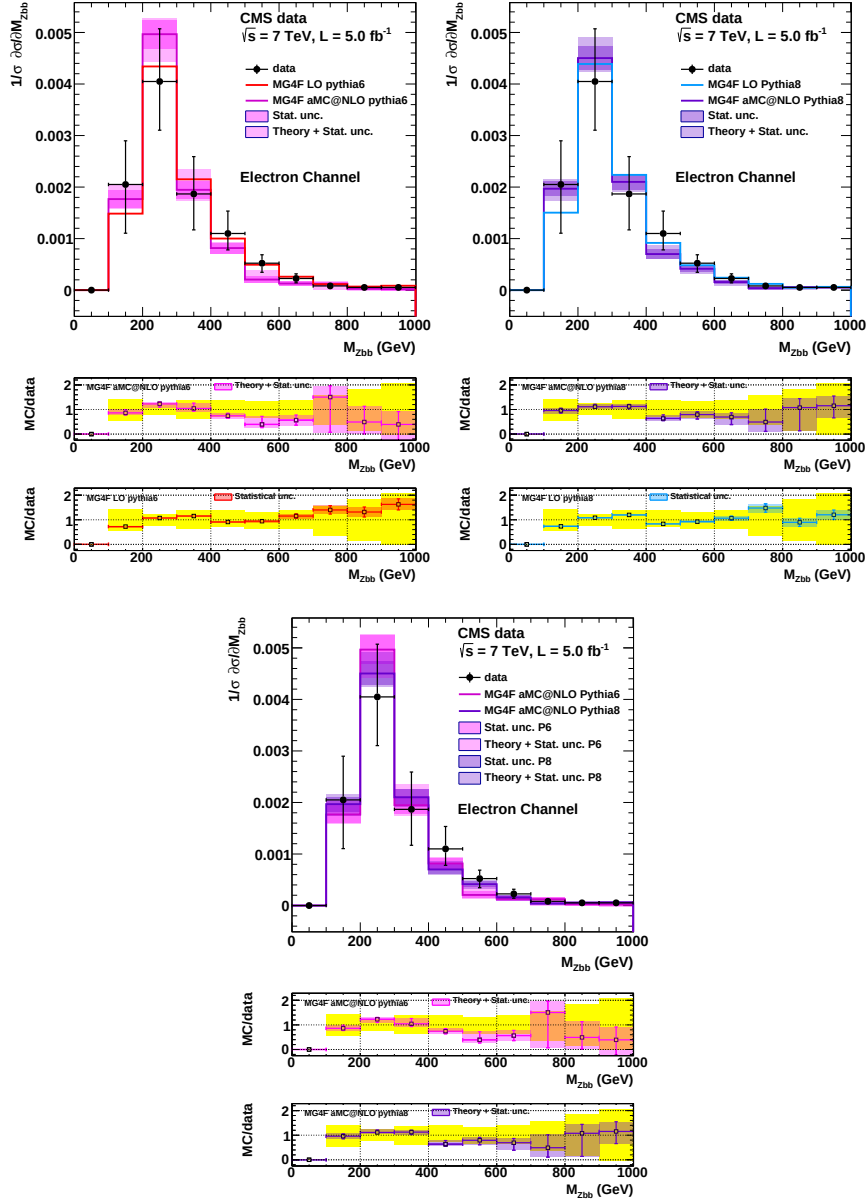


Figure D.48: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the M_{Zbb} in the electron channel.

• M_{Zbb} Muon Channel

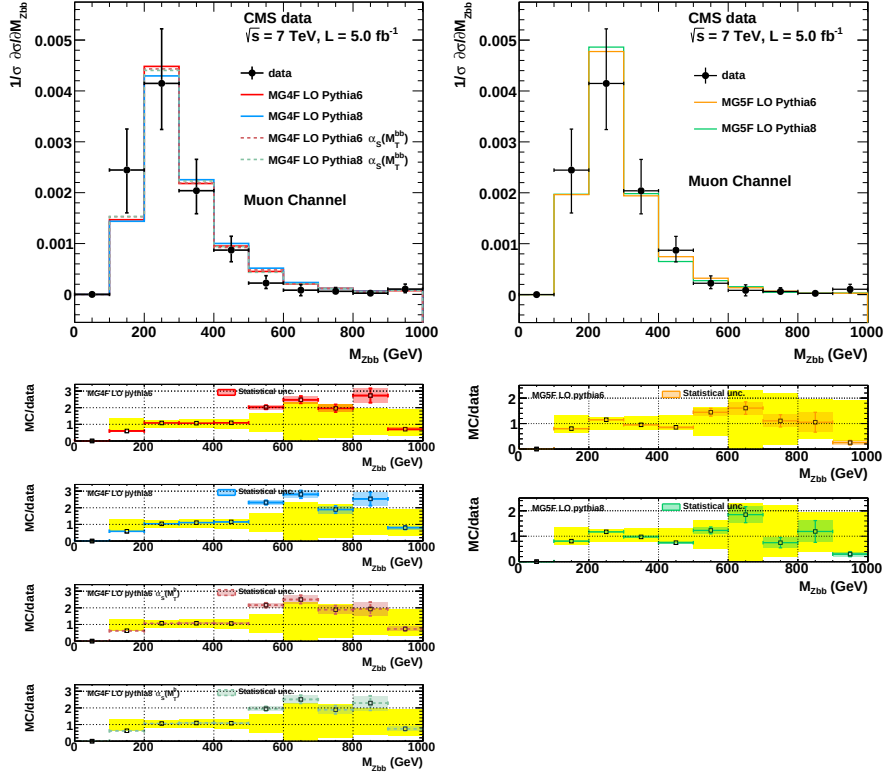


Figure D.49: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the M_{Zbb} in the muon channel.

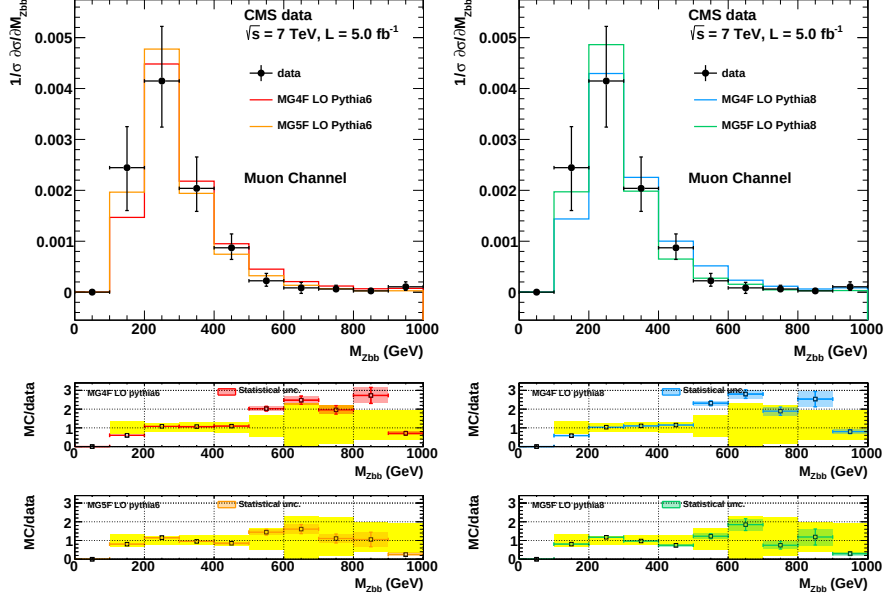


Figure D.50: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the M_{Zbb} in the muon channel.

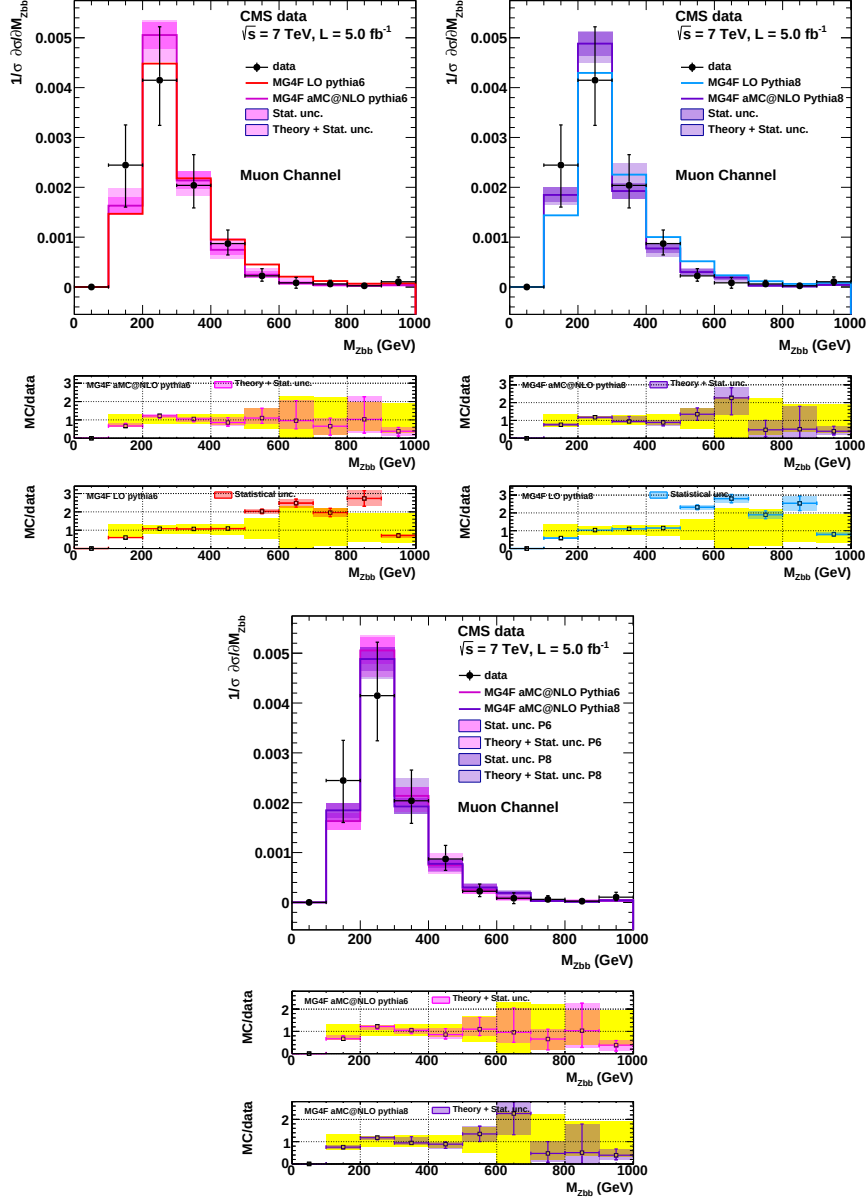


Figure D.51: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the M_{Zbb} in the muon channel.

• M_{Zbb} Combined Channel

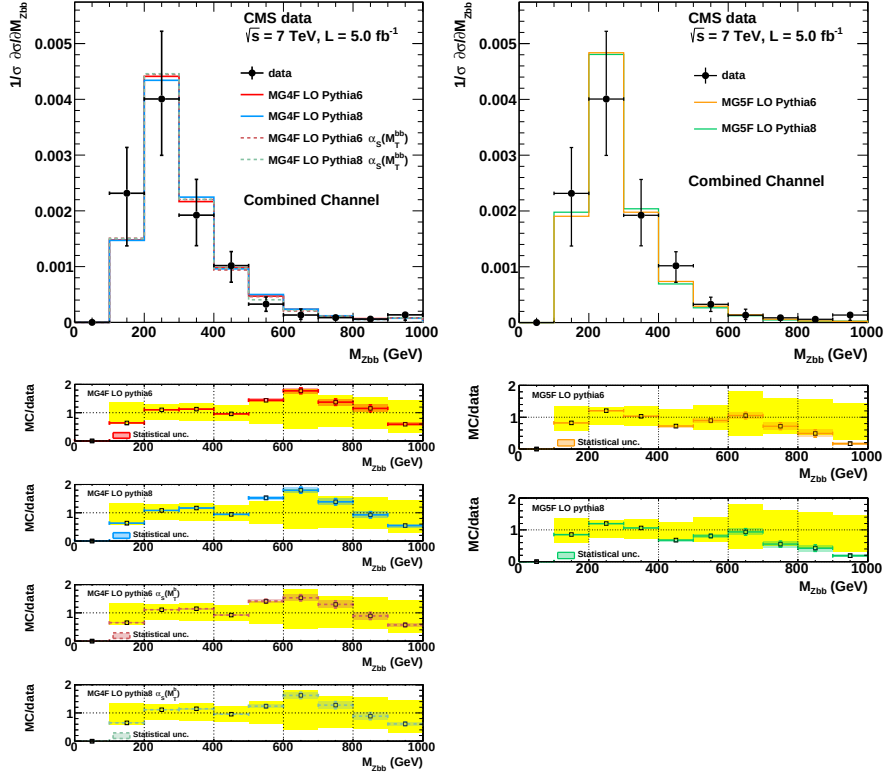


Figure D.52: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the M_{Zbb} in the combined channel.

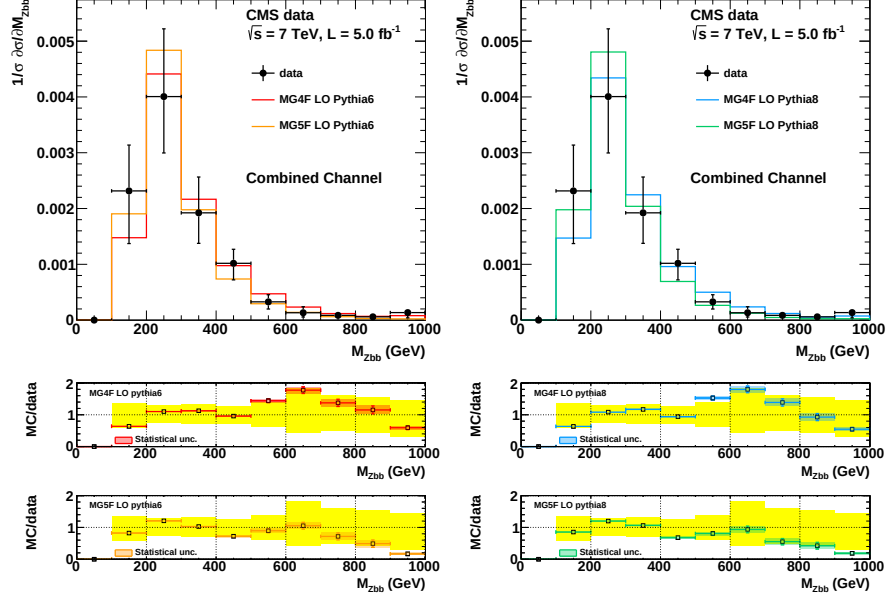


Figure D.53: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the M_{Zbb} in the combined channel.

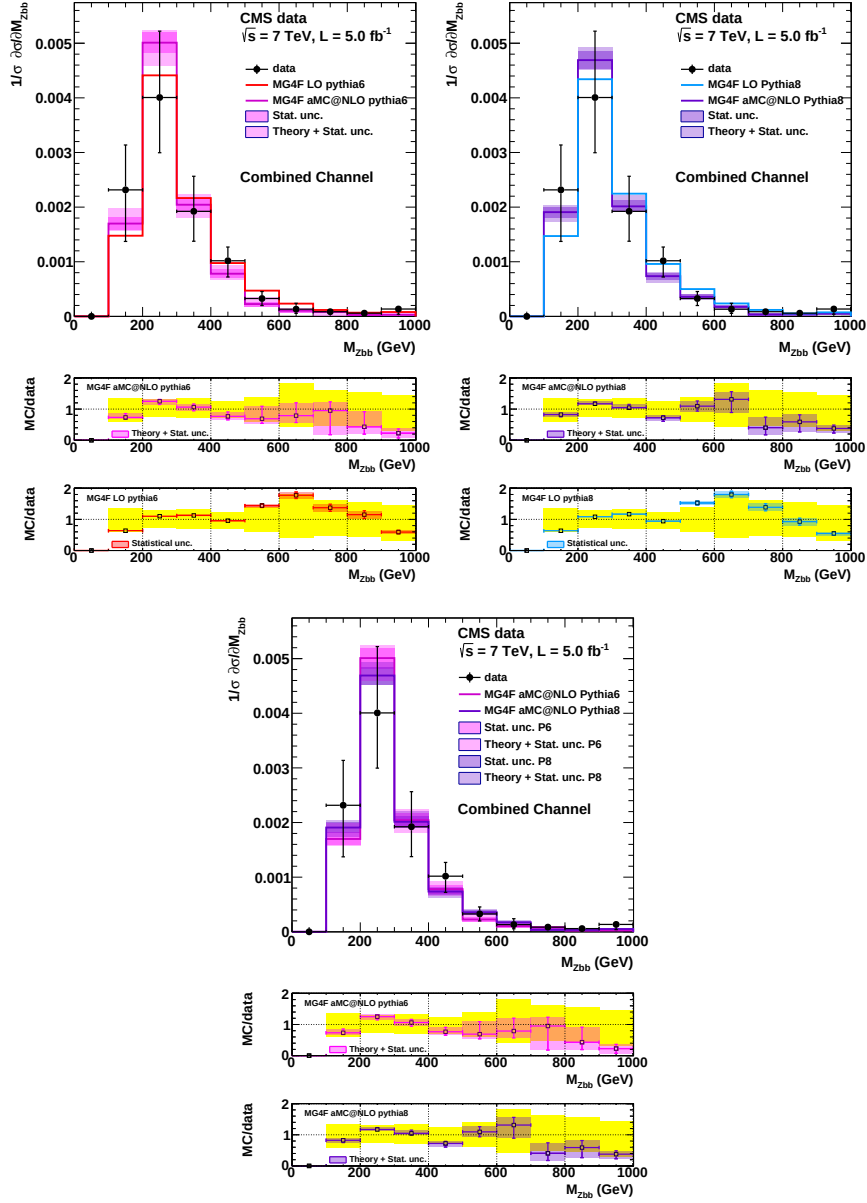


Figure D.54: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the M_{Zbb} in the combined channel.

- Difference ΔR between the two b jets
- $\Delta R_{b,b}$ Electron Channel

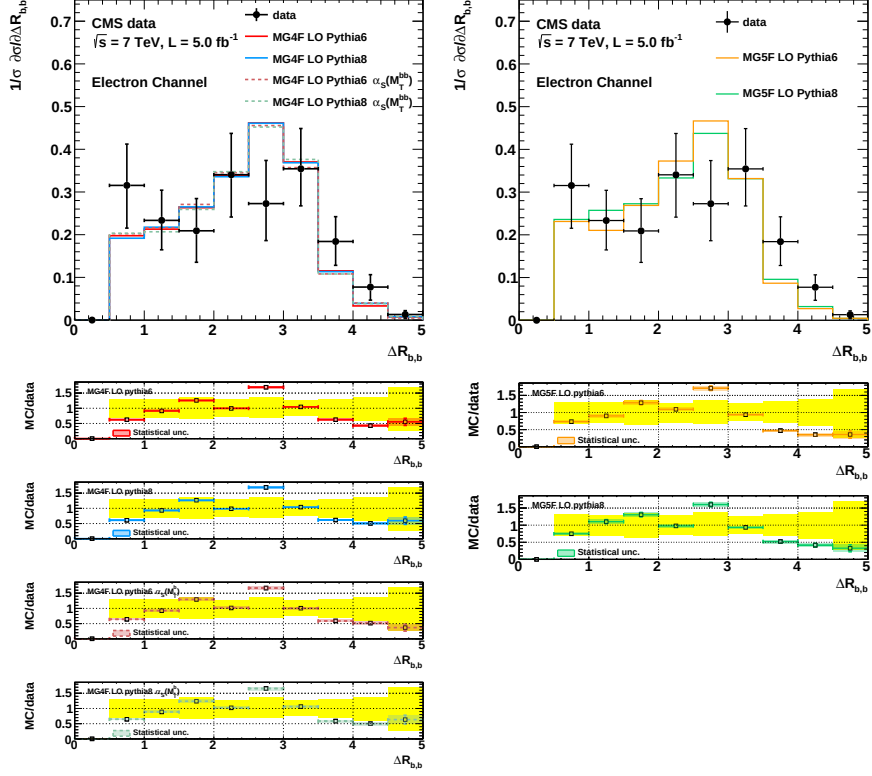


Figure D.55: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the $\Delta R_{b,b}$ in the electron channel.

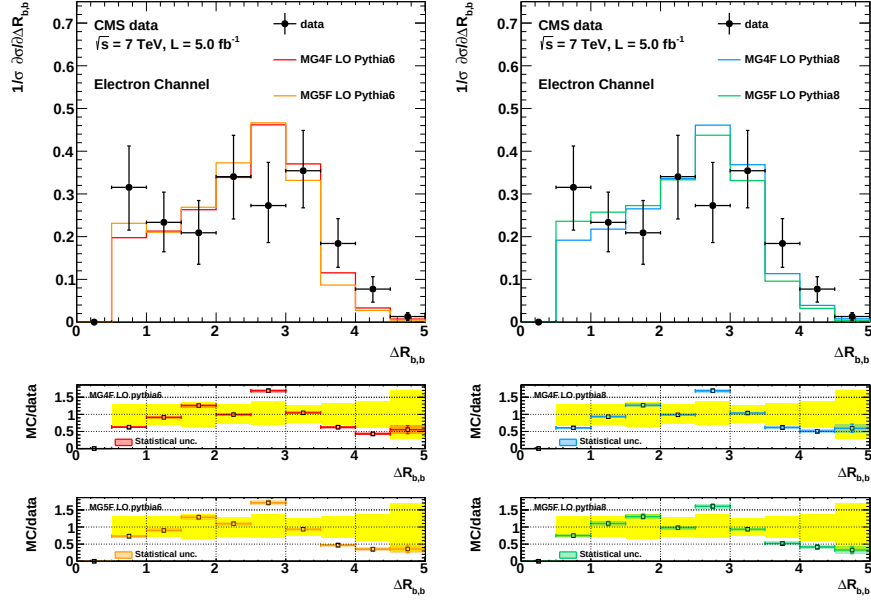


Figure D.56: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the $\Delta R_{b,b}$ in the electron channel.

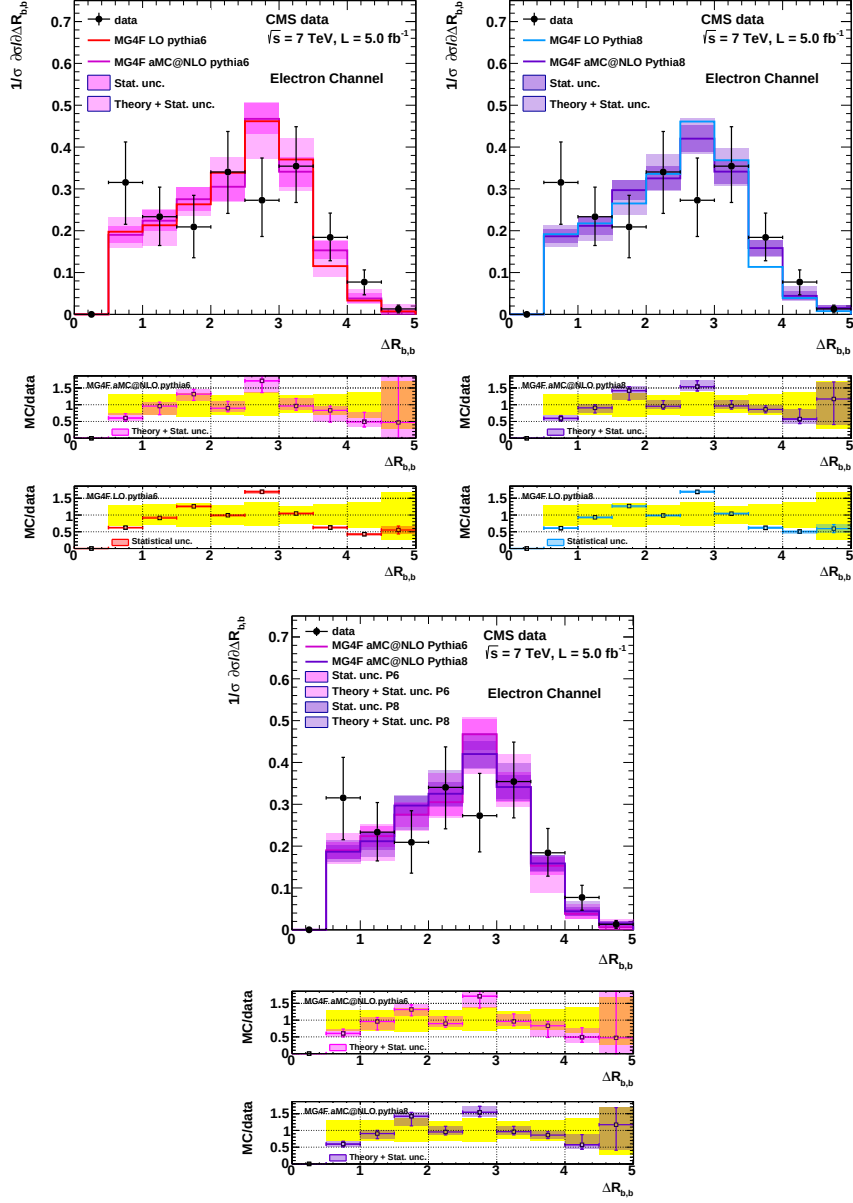


Figure D.57: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the $\Delta R_{b,b}$ in the electron channel.

• $\Delta R_{b,b}$ Muon Channel

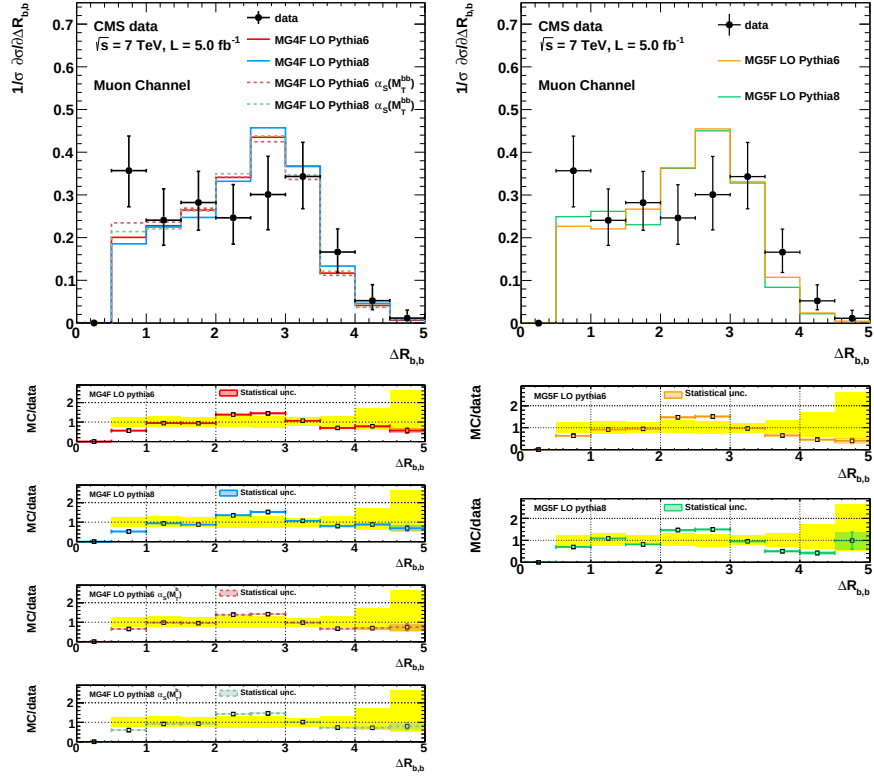


Figure D.58: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the $\Delta R_{b,b}$ in the muon channel.

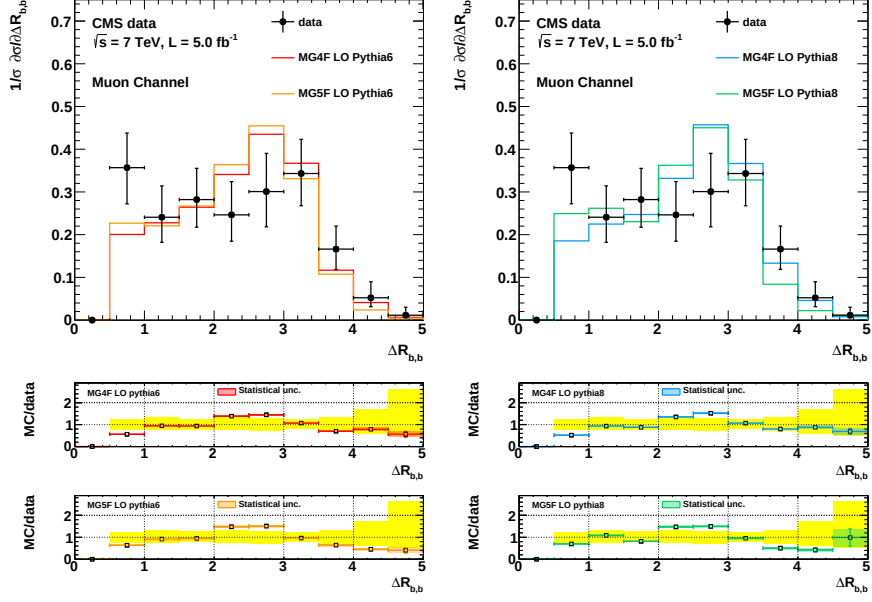


Figure D.59: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the $\Delta R_{b,b}$ in the muon channel.

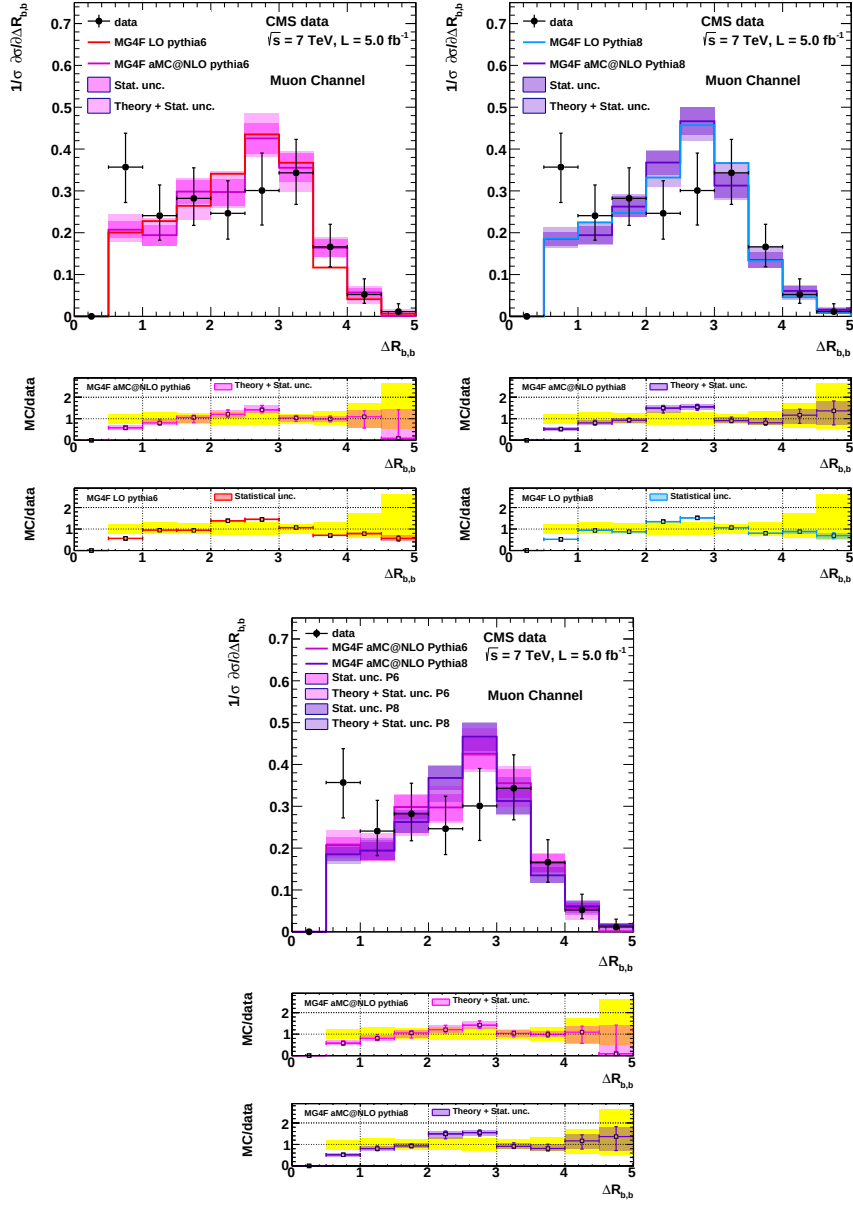


Figure D.60: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the $\Delta R_{b,b}$ in the muon channel.

• $\Delta R_{b,b}$ Combined Channel

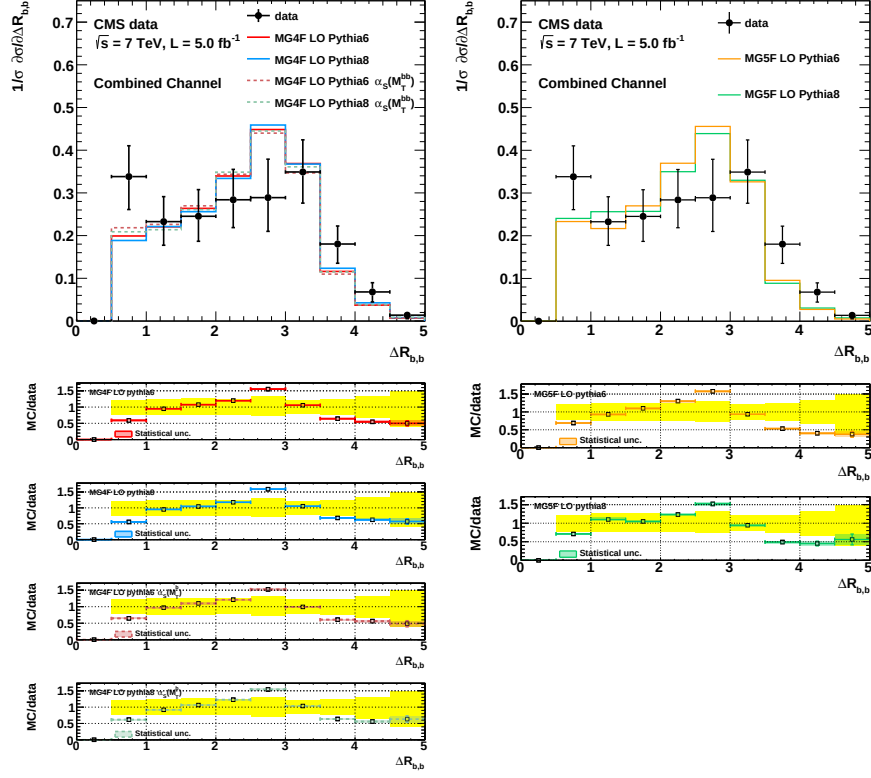


Figure D.61: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the $\Delta R_{b,b}$ in the combined channel.

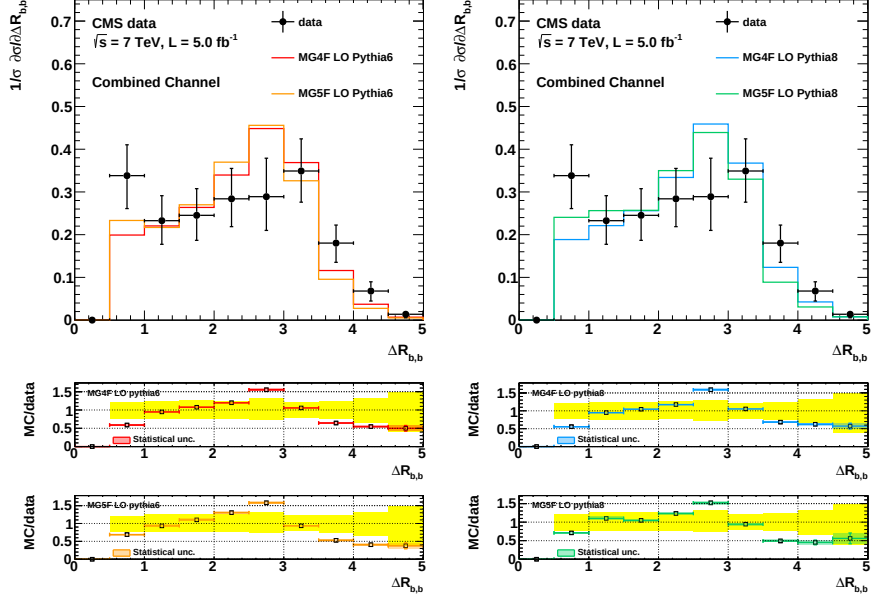


Figure D.62: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the $\Delta R_{b,b}$ in the combined channel.

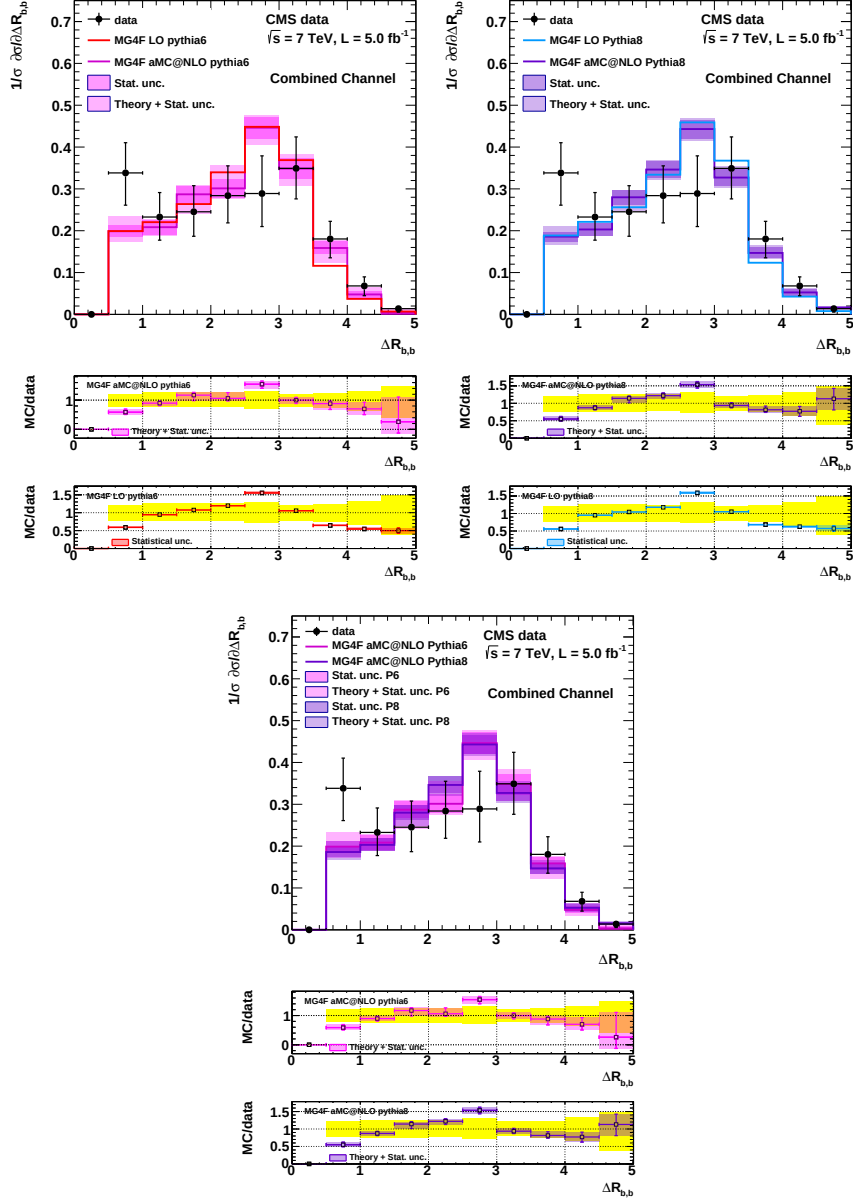


Figure D.63: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the $\Delta R_{b,b}$ in the combined channel.

- Azimuthal angular difference between the Z and the bb system
- $\Delta\Phi_{Z,bb}$ Electron Channel

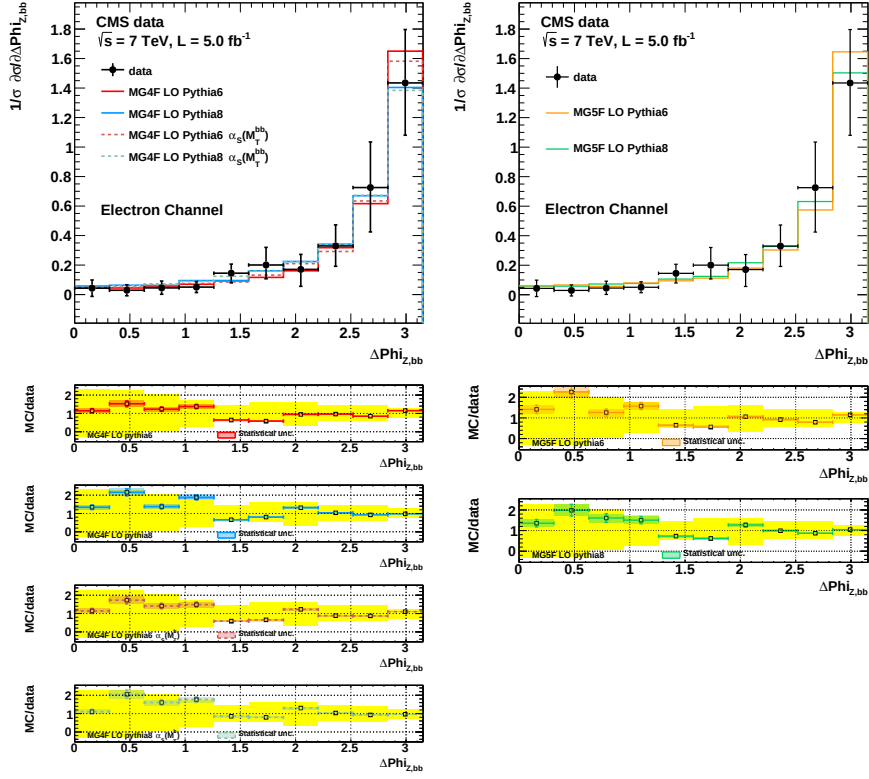


Figure D.64: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the $\Delta\Phi_{Z,bb}$ in the electron channel.

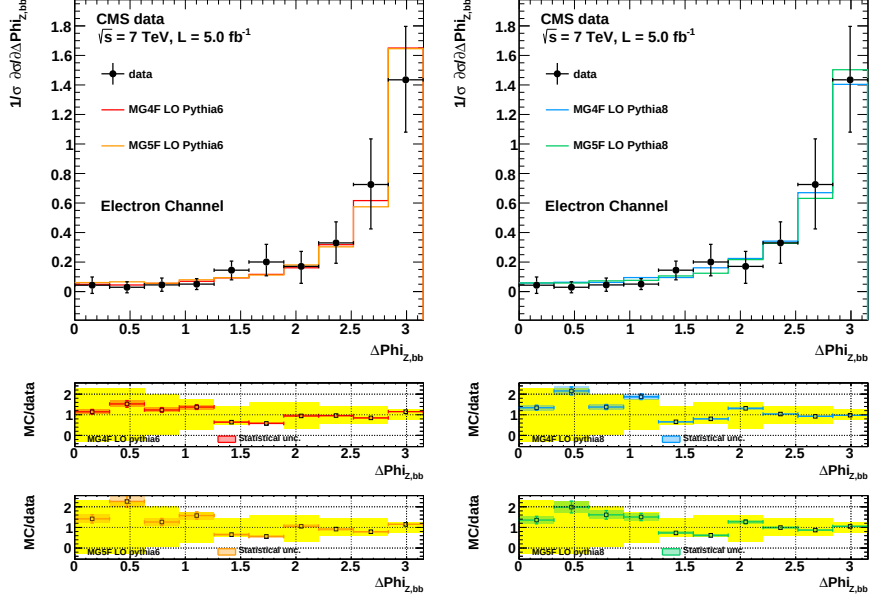


Figure D.65: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the $\Delta\Phi_{Z,bb}$ in the electron channel.

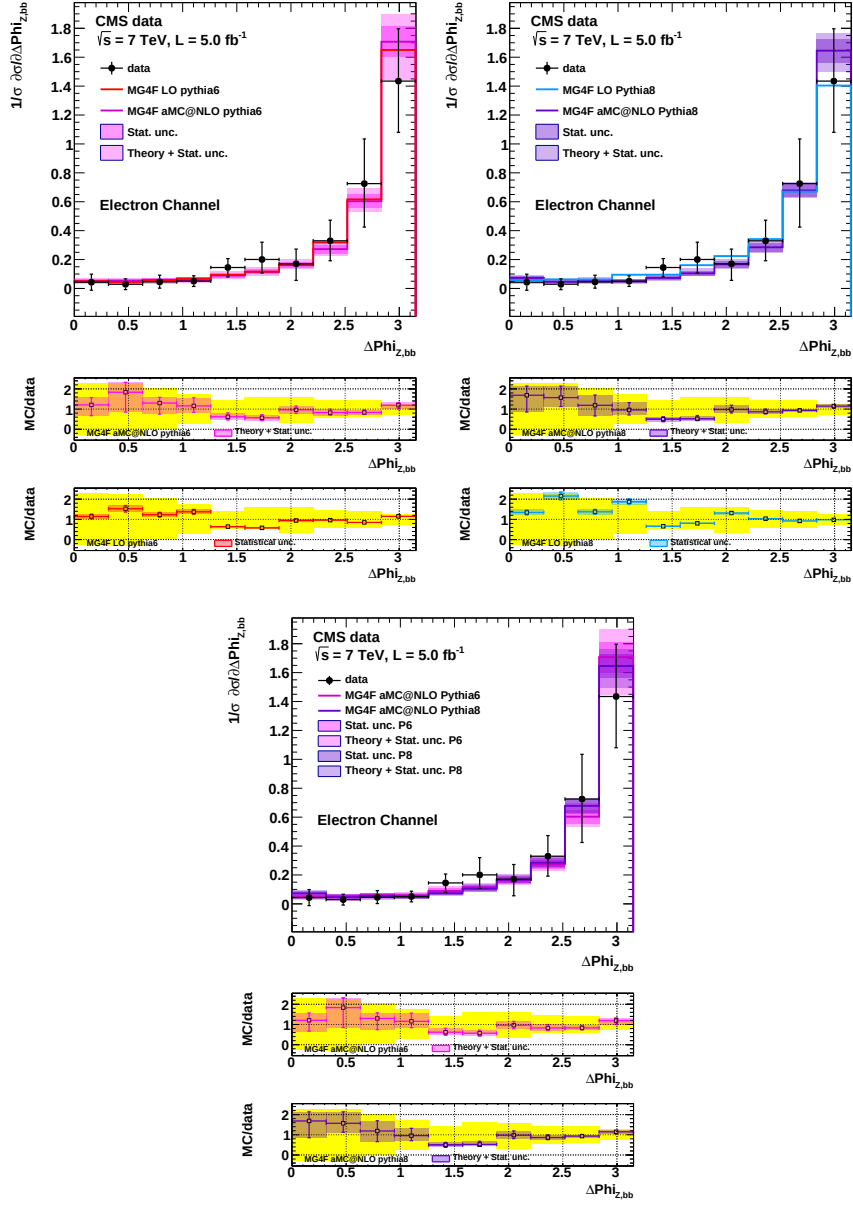


Figure D.66: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the $\Delta\Phi_{Z,bb}$ in the electron channel.

• $\Delta\Phi_{Z,bb}$ Muon Channel

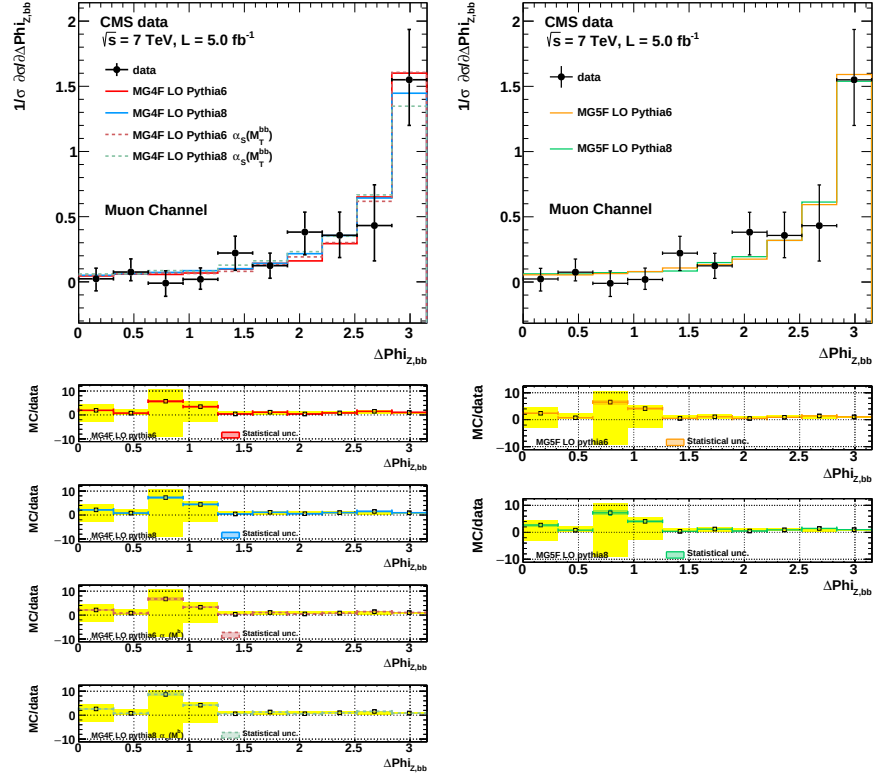


Figure D.67: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the $\Delta\Phi_{Z,bb}$ in the muon channel.

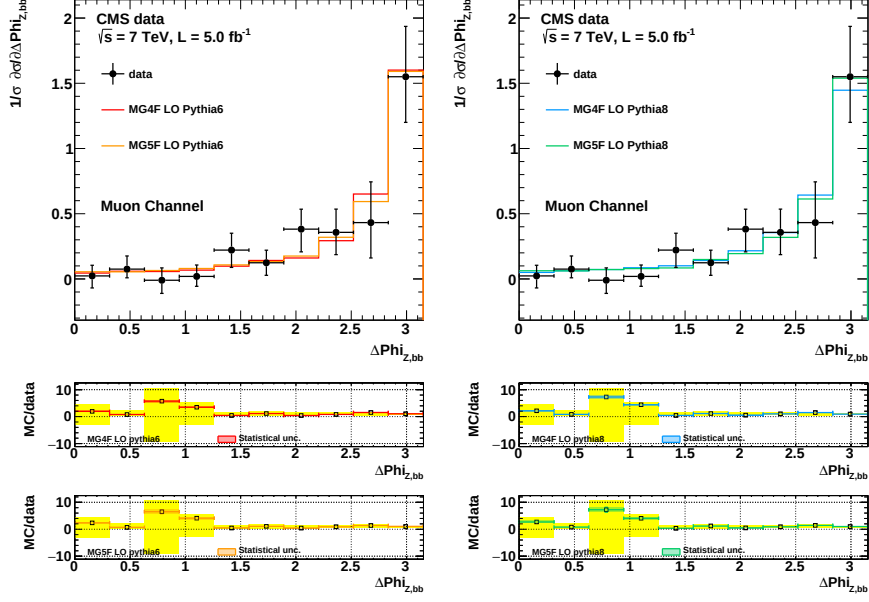


Figure D.68: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the $\Delta\Phi_{Z,bb}$ in the muon channel.

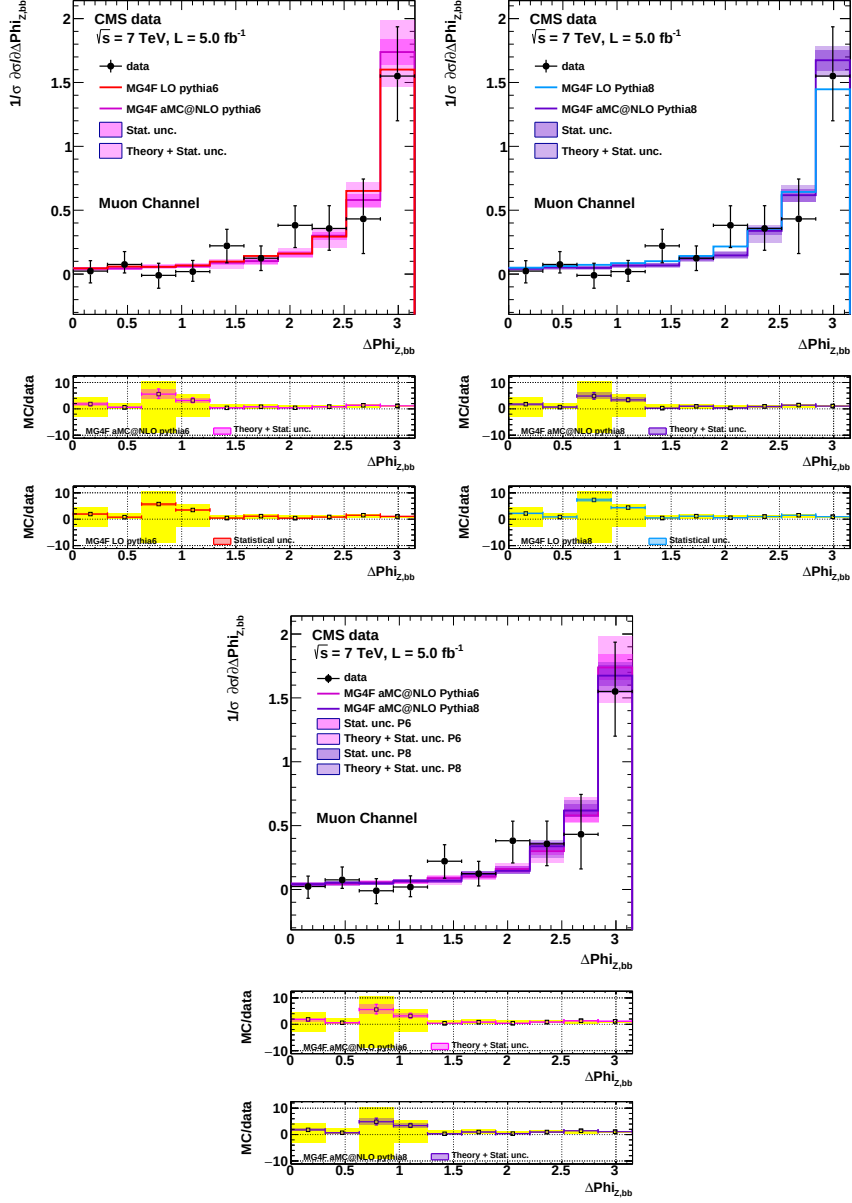


Figure D.69: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the $\Delta\Phi_{Z,bb}$ in the muon channel.

• $\Delta\Phi_{Z,bb}$ Combined Channel

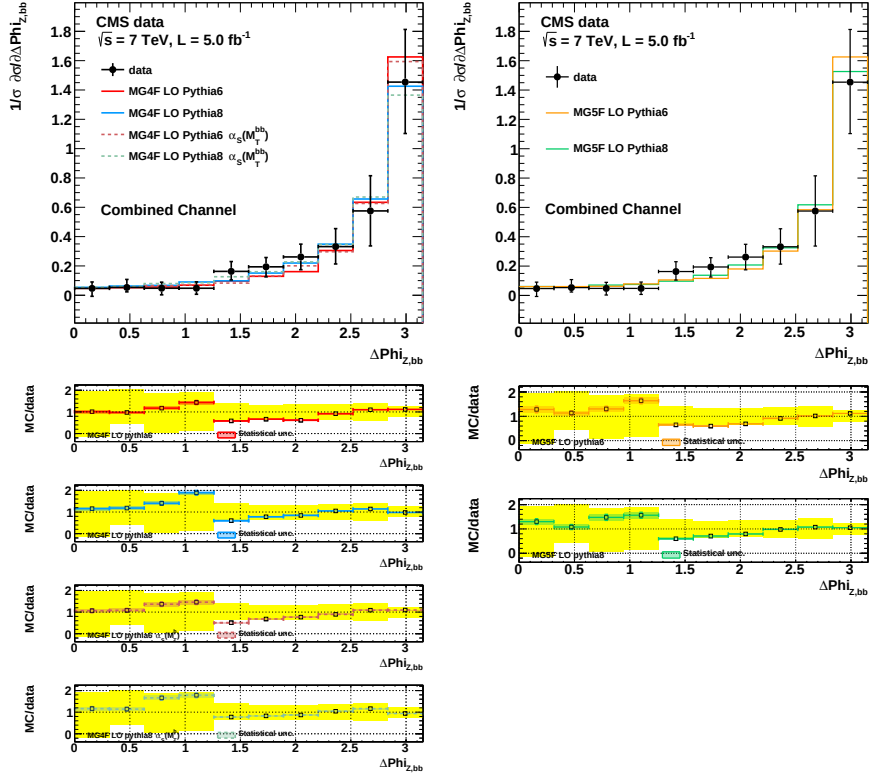


Figure D.70: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions, hadronised with PYTHIA 6 and PYTHIA 8, for the distribution of the $\Delta\Phi_{Z,bb}$ in the combined channel.

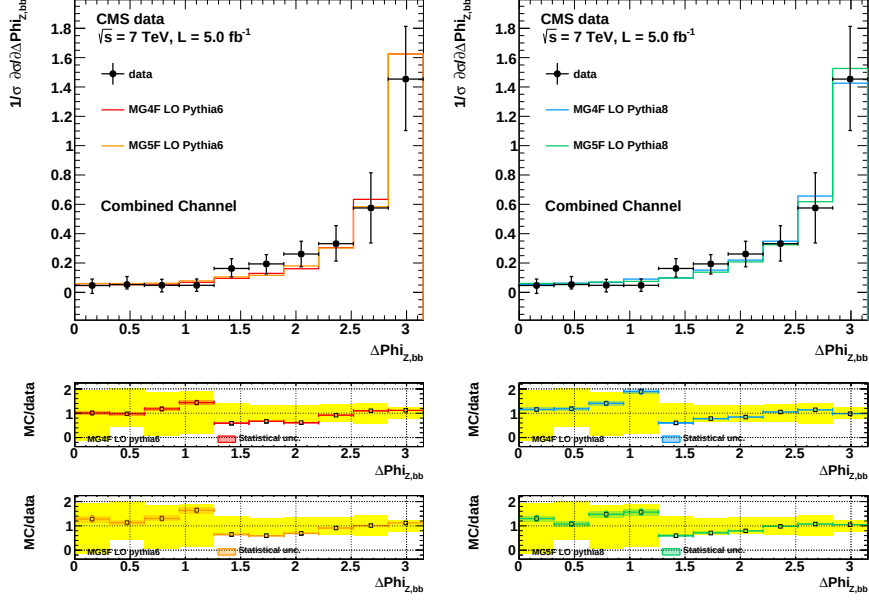


Figure D.71: Comparison of unfolded data with events generated with MG4F and MG5F generators, and hadronised with PYTHIA 6 (left) and PYTHIA 8 (right), for the distribution of the $\Delta\Phi_{Z,bb}$ in the combined channel.

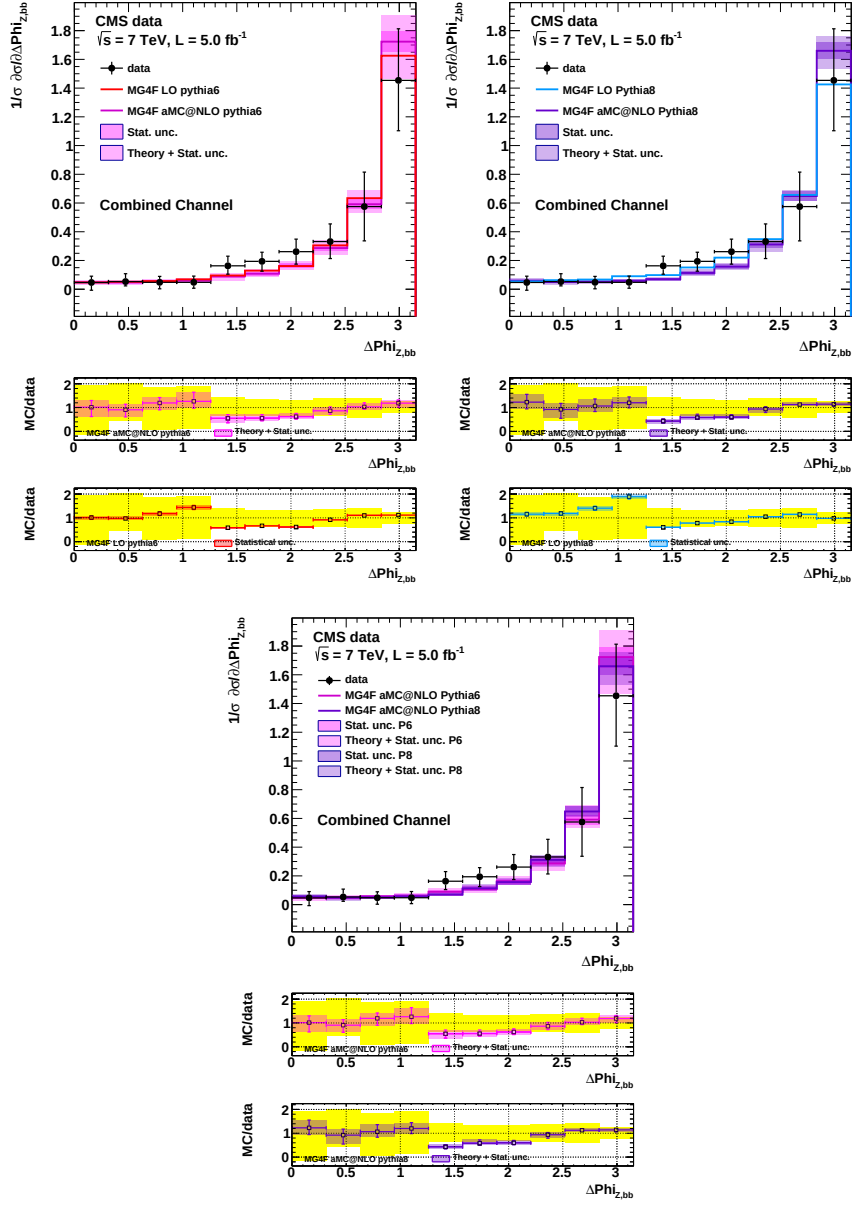


Figure D.72: Comparison of unfolded data with events generated with MG4F LO and MG4F aMC@NLO, hadronised with PYTHIA 6 (left) and PYTHIA 8 (right); and with MG4F aMC@NLO hadronised with PYTHIA 6 and PYTHIA 8 (middle), for the distribution of the $\Delta\Phi_{Z,bb}$ in the combined channel.

• PYTHIA 8 study of Q_{match} variation: p_T^Z Combined Channel

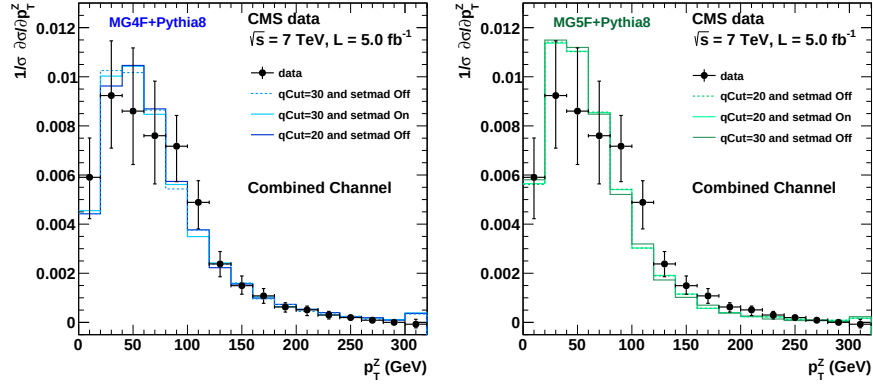


Figure D.73: Comparison of unfolded data with events from the MG4F (left) and MG5F (right) LO generator productions and hadronised with PYTHIA 8, for the distribution of the p_T of the Z boson in the electron channel. Distributions with a different Q_{match} (qCut) than the default one, and the default one are superimposed.

Bibliography

- [1] Empedocles. On Nature. *Simplicius, Physics*, Frag. B17:157–159, -450 BC.
- [2] Particle Data Group Collaboration. Review of Particle Physics. *Chin.Phys.*, C38:090001, 2014.
- [3] J.J Thomson. Cathode rays. *Phil.Mag.*, 44:293–316, 1897.
- [4] E. Rutherford. The scattering of alpha and beta particles by matter and the structure of the atom. *Phil.Mag.*, 21:669–688, 1911.
- [5] J. Chadwick. Possible Existence of a Neutron. *Nature*, 129:312, February 1932.
- [6] W. Gerlach and O. Stern. Experimental test of the applicability of the quantum theory to the magnetic field. *Z.Phys.*, 9:349–352, 1922.
- [7] E. Fermi. Versuch einer Theorie der β -Strahlen. I. *Z.Phys.*, 88:161–177, 1934.
- [8] F. Reines and C. L. Cowan. The Neutrino. *Nature*, 178:446–449, September 1956.
- [9] S.H. Neddermeyer and C.D. Anderson. Note on the Nature of Cosmic Ray Particles. *Phys.Rev.*, 51:884–886, 1937.
- [10] I. V. Anicin. The Neutrino - Its Past, Present and Future. March 2005, [arXiv:physics/0503172](https://arxiv.org/abs/physics/0503172).

- [11] G. Danby, J.-M. Gaillard, K. Goulianos, L. M. Lederman, N. Mistry, M. Schwartz, and J. Steinberger. Observation of High-Energy Neutrino Reactions and the Existence of Two Kinds of Neutrinos. *Physical Review Letters*, 9:36–44, July 1962.
- [12] M. L. Perl et al. Evidence for anomalous lepton production in e^+e^- annihilation. *Physical Review Letters*, 35:1489–1492, December 1975.
- [13] DONUT Collaboration. Observation of tau neutrino interactions. *Phys.Lett.*, B504:218–224, 2001, [arXiv:hep-ex/0012035](#).
- [14] C.M.G. Lattes, G.P.S. Occhialini, and C.F. Powell. Observations on the Tracks of Slow Mesons in Photographic Emulsions. 1. *Nature*, 160:453–456, 1947.
- [15] Hideki Yukawa. On the interaction of elementary particles. *Proc. Phys. Math. Soc. Jap.*, 17:48–57, 1935.
- [16] M. Gell-Mann. A schematic model of baryons and mesons. *Physics Letters*, 8:214–215, February 1964.
- [17] G. Zweig. An SU(3) model for strong interaction symmetry and its breaking. Version 1. 1964, CERN-TH-401.
- [18] Elliott D. Bloom, D.H. Coward, H.C. DeStaebler, J. Drees, Guthrie Miller, et al. High-Energy Inelastic $e\ p$ Scattering at 6-Degrees and 10-Degrees. *Phys.Rev.Lett.*, 23:930–934, 1969.
- [19] Martin Breidenbach, Jerome I. Friedman, Henry W. Kendall, Elliott D. Bloom, D.H. Coward, et al. Observed Behavior of Highly Inelastic electron-Proton Scattering. *Phys.Rev.Lett.*, 23:935–939, 1969.
- [20] S.L. Glashow, J. Iliopoulos, and L. Maiani. Weak Interactions with Lepton-Hadron Symmetry. *Phys.Rev.*, D2:1285–1292, 1970.
- [21] Makoto Kobayashi and Toshihide Maskawa. CP Violation in the Renormalizable Theory of Weak Interaction. *Prog.Theor.Phys.*, 49:652–657, 1973.
- [22] Haim Harari. A New Quark Model for Hadrons. *Phys.Lett.*, B57:265, 1975.
- [23] S.W. Herb, D.C. Hom, L.M. Lederman, J.C. Sens, H.D. Snyder, et al. Observation of a Dimuon Resonance at 9.5 GeV in 400 GeV Proton-Nucleus Collisions. *Phys.Rev.Lett.*, 39:252–255, 1977.

- [24] CDF Collaboration. Observation of top quark production in $\bar{p}p$ collisions. *Phys.Rev.Lett.*, 74:2626–2631, 1995, [arXiv:hep-ex/9503002](#).
- [25] D0 Collaboration. Search for high mass top quark production in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV. *Phys.Rev.Lett.*, 74:2422–2426, 1995, [arXiv:hep-ex/9411001](#).
- [26] Albert Einstein. Zur Elektrodynamik bewegter Körper. *Annalen Phys.*, 17:891–921, 1905.
- [27] Max Planck. Über das Gesetz der Energieverteilung im Normalspektrum. *Annalen Phys.*, 4:553, 1901.
- [28] F. Halzen and Alan D. Martin. *Quarks and leptons: an introductory course in modern particle physics*. ISBN-9780471887416, 1984.
- [29] E. Fermi. Versuch einer Theorie der β -Strahlen. I. *Zeitschrift für Physik*, 88:161–177, March 1934.
- [30] T.D. Lee and Chen-Ning Yang. Question of Parity Conservation in Weak Interactions. *Phys.Rev.*, 104:254–258, 1956.
- [31] C.S. Wu, E. Ambler, R.W. Hayward, D.D. Hoppes, and R.P. Hudson. Experimental Test of Parity Conservation in Beta Decay. *Phys.Rev.*, 105:1413–1414, 1957.
- [32] R.P. Feynman and Murray Gell-Mann. Theory of Fermi interaction. *Phys.Rev.*, 109:193–198, 1958.
- [33] E.C.G. Sudarshan and R.e. Marshak. Chirality invariance and the universal Fermi interaction. *Phys.Rev.*, 109:1860–1860, 1958.
- [34] Gargamelle Neutrino Collaboration. Observation of Neutrino Like Interactions Without Muon Or Electron in the Gargamelle Neutrino Experiment. *Phys.Lett.*, B46:138–140, 1973.
- [35] UA1 Collaboration. Experimental Observation of Lepton Pairs of Invariant Mass Around 95 GeV/ c^2 at the CERN SPS Collider. *Phys.Lett.*, B126:398–410, 1983.
- [36] UA2 Collaboration. Evidence for $Z^0 \rightarrow e^+e^-$ at the CERN anti-p p Collider. *Phys.Lett.*, B129:130–140, 1983.
- [37] UA1 Collaboration. Experimental Observation of Events with Large Missing Transverse Energy Accompanied by a Jet Or a Photon(s) in p anti-p Collisions at $\sqrt{s} = 540$ GeV. *Phys.Lett.*, B139:115, 1984.

- [38] UA2 Collaboration. Observation of Single Isolated Electrons of High Transverse Momentum in Events with Missing Transverse Energy at the CERN anti-p p Collider. *Phys.Lett.*, B122:476–485, 1983.
- [39] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964.
- [40] Peter W. Higgs. Broken symmetries, massless particles and gauge fields. *Phys.Lett.*, 12:132–133, 1964.
- [41] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys.Lett.*, B716:30–61, 2012, [arXiv:1207.7235](#).
- [42] 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.*, B716:1–29, 2012, [arXiv:1207.7214](#).
- [43] Albert Einstein. The Foundation of the General Theory of Relativity. *Annalen Phys.*, 49:769–822, 1916.
- [44] ATLAS, 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. 2015, [arXiv:1503.07589](#).
- [45] H1 Collaboration. Inclusive Deep Inelastic Scattering at High Q^2 with Longitudinally Polarised Lepton Beams at HERA. *JHEP*, 1209:061, 2012, [arXiv:1206.7007](#).
- [46] ZEUS Collaboration. Measurement of high- Q^2 neutral current deep inelastic e^+p scattering cross sections with a longitudinally polarized positron beam at HERA. *Phys.Rev.*, D87(5):052014, 2013, [arXiv:1208.6138](#).
- [47] CMS Collaboration. Measurements of differential and double-differential Drell-Yan cross sections in proton-proton collisions at $\sqrt{s} = 8$ TeV. 2014, [arXiv:1412.1115](#).
- [48] CMS Collaboration. Measurement of the muon charge asymmetry in inclusive $pp \rightarrow W + X$ production at $\sqrt{s} = 7$ TeV and an improved determination of light parton distribution functions. *Phys.Rev.*, D90(3):032004, 2014, [arXiv:1312.6283](#).
- [49] J. Pumplin, D.R. Stump, J. Huston, H.L. Lai, Pavel M. Nadolsky, et al. New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, 0207:012, 2002, [arXiv:hep-ph/0201195](#).

- [50] Hung-Liang Lai, Marco Guzzi, Joey Huston, Zhao Li, Pavel M. Nadolsky, et al. New parton distributions for collider physics. *Phys.Rev.*, D82:074024, 2010, [arXiv:1007.2241](#).
- [51] A.D. Martin, W.J. Stirling, R.S. Thorne, and G. Watt. Parton distributions for the LHC. *Eur.Phys.J.*, C63:189–285, 2009, [arXiv:0901.0002](#).
- [52] Richard D. Ball, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Jose I. Latorre, et al. A first unbiased global NLO determination of parton distributions and their uncertainties. *Nucl.Phys.*, B838:136–206, 2010, [arXiv:1002.4407](#).
- [53] V.N. Gribov and L.N. Lipatov. Deep inelastic $e p$ scattering in perturbation theory. *Sov.J.Nucl.Phys.*, 15:438–450, 1972.
- [54] Yuri L. Dokshitzer. Calculation of the Structure Functions for Deep Inelastic Scattering and e^+e^- Annihilation by Perturbation Theory in Quantum Chromodynamics. *Sov.Phys.JETP*, 46:641–653, 1977.
- [55] Guido Altarelli and G. Parisi. Asymptotic Freedom in Parton Language. *Nucl.Phys.*, B126:298, 1977.
- [56] The Durham HepData Project - PDF plotter. <http://hepdata.cedar.ac.uk/pdf/pdf3.html>. Accessed: 2014-12-13.
- [57] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 0605:026, 2006, [arXiv:hep-ph/0603175](#).
- [58] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput.Phys.Commun.*, 178:852–867, 2008, [arXiv:0710.3820](#).
- [59] G. Corcella, I. G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, Peter Richardson, Michael H. Seymour, and Bryan R. Webber. HERWIG 6: an event generator for hadron emission reactions with interfering gluons (including supersymmetric processes). *JHEP*, 01:010, 2001, [arXiv:hep-ph/0011363](#).
- [60] Johan Alwall, Michel Herquet, Fabio Maltoni, Olivier Mattelaer, and Tim Stelzer. MadGraph 5 : Going Beyond. *JHEP*, 1106:128, 2011, [arXiv:1106.0522](#).
- [61] Fabio Maltoni and Tim Stelzer. MadEvent: Automatic event generation with MadGraph. *JHEP*, 02:027, 2003, [arXiv:hep-ph/0208156](#).

- [62] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, et al. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 1407:079, 2014, [arXiv:1405.0301](#).
- [63] Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD computations with Parton Shower simulations: the POWHEG method. *JHEP*, 0711:070, 2007, [arXiv:0709.2092](#).
- [64] Stefano Frixione, Fabian Stoeckli, Paolo Torrielli, Bryan R. Webber, and Chris D. White. The MC@NLO 4.0 Event Generator. 2010, [arXiv:1010.0819](#).
- [65] Paolo Nason and Bryan Webber. Next-to-Leading-Order Event Generators. *Ann.Rev.Nucl.Part.Sci.*, 62:187–213, 2012, [arXiv:1202.1251](#).
- [66] V.V. Sudakov. Vertex parts at very high-energies in quantum electrodynamics. *Sov.Phys.JETP*, 3:65–71, 1956.
- [67] M.L.Mangano. Merging multijet matrix elements and shower evolution in hadronic collisions. <http://cern.ch/~mlm/talks/lund-alpgen.pdf>. Accessed: 2015-02-28.
- [68] J. Alwall et al. Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions. *Eur. Phys. J. C*, 53:473, 2008, [arXiv:0706.2569](#).
- [69] Stephen D. Ellis and Davison E. Soper. Successive combination jet algorithm for hadron collisions. *Phys.Rev.*, D48:3160–3166, 1993, [arXiv:hep-ph/9305266](#).
- [70] S. Catani, F. Krauss, R. Kuhn, and B.R. Webber. QCD matrix elements + parton showers. *JHEP*, 0111:063, 2001, [arXiv:hep-ph/0109231](#).
- [71] Rikkert Frederix and Stefano Frixione. Merging meets matching in MC@NLO. *JHEP*, 1212:061, 2012, [arXiv:1209.6215](#).
- [72] Bo Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand. Parton Fragmentation and String Dynamics. *Phys.Rept.*, 97:31–145, 1983.
- [73] P. Bartalini, E.L. Berger, B. Blok, G. Calucci, R. Corke, et al. Multi-Parton Interactions at the LHC. 2011, [arXiv:1111.0469](#).
- [74] CMS Collaboration. Measurement of the Underlying Event Activity at the LHC with $\sqrt{s} = 7$ TeV and Comparison with $\sqrt{s} = 0.9$ TeV. *JHEP*, 1109:109, 2011, [arXiv:1107.0330](#).

- [75] John M. Campbell, R. Keith Ellis, F. Maltoni, and S. Willenbrock. Associated production of a Z Boson and a single heavy quark jet. *Phys.Rev.*, D69:074021, 2004, [arXiv:hep-ph/0312024](#).
- [76] John M. Campbell, R. Keith Ellis, F. Maltoni, and S. Willenbrock. Production of a Z boson and two jets with one heavy-quark tag. *Phys.Rev.*, D73:054007, 2006, [arXiv:hep-ph/0510362](#).
- [77] Fabio Maltoni, Thomas McElmurry, and Scott Willenbrock. Inclusive production of a Higgs or Z boson in association with heavy quarks. *Phys. Rev. D*, 72:074024, 2005, [arXiv:hep-ph/0505014](#).
- [78] J. Pumplin, D.R. Stump, J. Huston, H.L. Lai, Pavel M. Nadolsky, et al. New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, 0207:012, 2002, [arXiv:hep-ph/0201195](#).
- [79] Fabio Maltoni, Giovanni Ridolfi, and Maria Ubiali. b-initiated processes at the LHC: a reappraisal. *JHEP*, 1207:022, 2012, [arXiv:1203.6393](#).
- [80] S. Dittmaier, S. Dittmaier, C. Mariotti, G. Passarino, R. Tanaka, et al. Handbook of LHC Higgs Cross Sections: 2. Differential Distributions. 2012, [arXiv:1201.3084](#).
- [81] 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(1):012003, 2014, [arXiv:1310.3687](#).
- [82] W. J. Stirling. Private communication. <http://www.hep.ph.ic.ac.uk/~wstirling/plots/crosssections2012.pdf>. Accessed: 2015-03-08.
- [83] G.C. Branco, P.M. Ferreira, L. Lavoura, M.N. Rebelo, Marc Sher, et al. Theory and phenomenology of two-Higgs-doublet models. *Phys.Rept.*, 516:1–102, 2012, [arXiv:1106.0034](#).
- [84] Stephen P. Martin. A Supersymmetry primer. *Adv.Ser.Direct.High Energy Phys.*, 21:1–153, 2010, [arXiv:hep-ph/9709356](#).
- [85] J.-M. Gerard and M. Herquet. A Twisted custodial symmetry in the two-Higgs-doublet model. *Phys.Rev.Lett.*, 98:251802, 2007, [arXiv:hep-ph/0703051](#).
- [86] Simon de Visscher, Jean-Marc Gerard, Michel Herquet, Vincent Lemaitre, and Fabio Maltoni. Unconventional phenomenology of a minimal two-Higgs-doublet model. *JHEP*, 0908:042, 2009, [arXiv:0904.0705](#).

- [87] ALEPH, DELPHI, L3, OPAL, SLD, LEP Electroweak Working Group, SLD Electroweak Group, SLD Heavy Flavour Group Collaboration. Precision electroweak measurements on the Z resonance. *Phys.Rept.*, 427:257–454, 2006, [arXiv:hep-ex/0509008](#).
- [88] ALEPH, DELPHI, L3, OPAL, LEP Electroweak Collaboration. Electroweak Measurements in Electron-Positron Collisions at W-Boson-Pair Energies at LEP. *Phys.Rept.*, 532:119–244, 2013, [arXiv:1302.3415](#).
- [89] Lyndon Evans and Philip Bryant. LHC Machine. *JINST*, 3:S08001, 2008.
- [90] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008.
- [91] ALICE Collaboration. The ALICE experiment at the CERN LHC. *JINST*, 3:S08002, 2008.
- [92] CMS Collaboration. The CMS experiment at the CERN LHC. *JINST*, 3:S08004, 2008.
- [93] LHCb Collaboration. The LHCb Detector at the LHC. *JINST*, 3:S08005, 2008.
- [94] F. Marcastel. CERN’s Accelerator Complex. CERN Graphic Design service, Oct 2013.
- [95] Mike Lamont. The First Years of LHC Operation for Luminosity Production. *Proceedings, 4th International Particle Accelerator Conference (IPAC 2013)*, page MOYAB101, 2013.
- [96] CMS Luminosity - Public Results. <https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults>. Accessed: 2015-04-13.
- [97] ECFA High Luminosity LHC Experiments Workshop: Physics and Technology Developments Summary submitted to ECFA. 96th Plenary ECFA meeting. January 2015.
- [98] Tai Sakuma and Thomas McCauley. Detector and event visualization with sketchup at the cms experiment. *Journal of Physics: Conference Series*, 513(2):022032, 2014.
- [99] CMS Collaboration. The performance of the CMS muon detector in proton-proton collisions at $\sqrt{s} = 7$ TeV at the LHC. *JINST*, 8:P11002, 2013, [arXiv:1306.6905](#).

- [100] CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *JINST*, 9(10):P10009, 2014, [arXiv:1405.6569](#).
- [101] CMS Collaboration. Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in pp Collisions at $\sqrt{s} = 7$ TeV. *JINST*, 8:P09009, 2013, [arXiv:1306.2016](#).
- [102] CMS Collaboration. CMS physics: Technical Design Report Volume 1: Detector Performance and Software. 2006. CERN-LHCC-2006-001.
- [103] CMS Collaboration. Measurements of the production cross sections for a Z boson and one or more b jets in pp collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 1406:120, 2013, [arXiv:1402.1521](#).
- [104] CMS Collaboration. Measurements of the Z+b-jet cross section at $\sqrt{s} = 7$ TeV. *JHEP*, 1206:126, 2012, [arXiv:1204.1643](#).
- [105] CMS Collaboration. Absolute Calibration of Luminosity Measurement at CMS: Summer 2011 Update. CMS Physics Analysis Summary CMS-PAS-EWK-11-001, 2011.
- [106] CMS Collaboration. Measurements of inclusive W and Z cross sections in pp collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 01:080, 2011, [arXiv:1012.2466](#).
- [107] R. Field. Early LHC Underlying Event Data - Findings and Surprises. [arXiv:1010.3558](#). 2010.
- [108] CMS Collaboration. Measurement of the underlying event activity at the LHC with $\sqrt{s} = 7$ TeV and comparison with $\sqrt{s} = 0.9$ TeV. *JHEP*, 09:109, 2011, [arXiv:1107.0330](#).
- [109] GEANT4 Collaboration. GEANT4—a simulation toolkit. *Nucl. Instrum. Meth. A*, 506:250, 2003.
- [110] K. Melnikov and F. Petriello. Electroweak gauge boson production at hadron colliders through $\mathcal{O}(\alpha_s^2)$. *Phys. Rev. D*, 74:114017, 2006, [arXiv:hep-ph/0609070](#).
- [111] J. M. Campbell and R. K. Ellis. MCFM for the Tevatron and the LHC. *Nucl.Phys.Proc.Suppl.* 205-206, 2010.
- [112] J. M. Campbell, R. K. Ellis, and C. Williams. Vector boson pair production at the LHC. *JHEP*, 07:018, 2011, [arXiv:1105.0020](#).
- [113] CMS offline software. <https://github.com/cms-sw/cmssw>. Accessed: 2014-11-04.

- [114] Standard Model Cross Sections for CMS at 7 TeV. <https://twiki.cern.ch/twiki/bin/view/CMS/StandardModelCrossSections?rev=27>. Accessed: 2015-04-13.
- [115] CMS Collaboration. Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV. *JINST*, 8:P09009, 2013, [arXiv:1306.2016](#).
- [116] CMS Collaboration. Commissioning of the Particle-Flow Event Reconstruction with the First LHC collisions recorded in the CMS detector. CMS Physics Analysis Summary CMS-PAS-PFT-10-001, 2010.
- [117] CMS Collaboration. Commissioning of the particle-flow event reconstruction with leptons from J/Y and W decays at 7 TeV. CMS Physics Analysis Summary CMS-PAS-PFT-10-003, 2010.
- [118] CMS Collaboration. Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and Missing ET. CMS Physics Analysis Summary CMS-PAS-PFT-09-001, 2009.
- [119] CMS Collaboration. Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV. CMS Physics Analysis Summary CMS-PAS-PFT-10-002, 2010.
- [120] CMS Collaboration. Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV. *JINST*, 7:P10002, 2012, [arXiv:1206.4071](#).
- [121] M. Cacciari, G.P. Salam, and G. Soyez. The anti-kt jet clustering algorithm. *JHEP*, 04:063, 2008.
- [122] Matteo Cacciari and Gavin P. Salam. Dispelling the N^3 myth for the k_t jet-finder. *Phys. Lett. B*, 641:57, 2006, [arXiv:hep-ph/0512210](#).
- [123] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet user manual. *Eur. Phys. J. C*, 72, 2012, [arXiv:1111.6097](#).
- [124] CMS Collaboration. Determination of jet energy calibration and transverse momentum resolution in CMS. *JINST*, 6:P11002, 2011, [arXiv:1107.4277](#).
- [125] CMS Collaboration. Jet Energy Scale performance in 2011. CMS Detector Performance Note CERN-CMS-DP-2012-006, 2012.
- [126] Matteo Cacciari and Gavin P. Salam. Pileup subtraction using jet areas. *Phys. Lett. B*, 659:119, 2008, [arXiv:0707.1378](#).

- [127] CMS Collaboration. Commissioning of b-jet identification with pp collisions at $\sqrt{s} = 7$ TeV. CMS Physics Analysis Summary CMS-PAS-BTV-10-001, 2010.
- [128] CMS Collaboration. Identification of b-quark jets with the CMS experiment. *JINST*, 8:P04013, 2011, [arXiv:1211.4462](#).
- [129] CMS Collaboration. Missing transverse energy performance of the CMS detector. *JINST*, 6:P09001, 2010, [arXiv:1106.5048](#).
- [130] Freya Blekman et al. Missing transverse energy significance and algorithm descriptions. CMS Internal Analysis Note CMS-AN-10-04, 2010.
- [131] CMS Collaboration. CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update. CMS Physics Analysis Summary LUM-13-001, 2013.
- [132] CMS Collaboration. Measurement of the inclusive production cross sections for forward jets and for dijet events with one forward and one central jet in pp collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 6:36, 2012, [arXiv:1202.0704](#).
- [133] ATLAS Collaboration. Measurement of the Inelastic Proton-Proton Cross-Section at p $\sqrt{s} = 7$ TeV with the ATLAS Detector. *Nat. Commun.*, 2:463, 2011, [arXiv:1104.0326](#).
- [134] TOTEM Collaboration. First measurement of the total proton-proton cross section at the LHC energy of $\sqrt{s} = 7$ TeV. *EPL*, 96, 2011, [arXiv:1110.1395](#).
- [135] R. Castello et al. Leptons efficiency and jets selection for the measurement of Z+b-jet ($Z \rightarrow \ell^+ \ell^-$) cross section. CMS Internal Analysis Note CMS-AN-11-264, 2011.
- [136] CMS Collaboration. Performance of CMS muon reconstruction in pp collisions at $\sqrt{s} = 7$ TeV. CMS Physics Analysis Summary CMS-MUO-10-004, 2011.
- [137] CMS Collaboration. b-Jet Identification in the CMS Experiment. CMS Physics Analysis Summary CMS-PAS-BTV-11-004, CERN, Geneva, 2012.
- [138] CMS Collaboration. Measurement of the cross section and angular correlations for associated production of a Z boson with b hadrons in pp collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 12:039, 2013, [arXiv:1310.1349](#).

- [139] CMS Collaboration. Measurement of the ZZ production cross section and search for anomalous couplings in $2\ell 2\ell'$ final states in pp collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 01:063, 2013, [arXiv:1211.4890](#).
- [140] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B*, 716:30, 2012, [arXiv:1207.7235](#).
- [141] CMS Collaboration. Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *JHEP*, 06:081, 2013, [arXiv:1303.4571](#).
- [142] CMS Collaboration. Study of the Mass and Spin-Parity of the Higgs Boson Candidate via Its Decays to Z Boson Pairs. *Phys. Rev. Lett.*, 110:081803, 2012, [arXiv:1212.6639](#).
- [143] CMS Collaboration. Performance of b tagging at $\sqrt{s} = 8$ TeV in multijet, $t\bar{t}b\bar{a}$ and boosted topology events. CMS Physics Analysis Summary BTV-13-001, 2013.
- [144] CMS Collaboration. Study of double parton scattering using $W + 2$ -jet events in proton-proton collisions at $\sqrt{s}=7$ TeV. *JHEP*, 03:032, 2014, [arXiv:1312.5729](#).
- [145] Louis Lyons, Duncan Gibaut, and Peter Clifford. How to combine correlated estimates of a single physical quantity. *Nucl. Instrum. Meth. A*, 270:110, 1988.
- [146] ATLAS Collaboration. Measurement of differential production cross-sections for a Z boson in association with b -jets in 7 TeV proton-proton collisions with the ATLAS detector. *JHEP*, 1410:141, 2014, [arXiv:1407.3643](#).
- [147] F. Febres Cordero, L. Reina, and D. Wackerroth. W - and Z -boson production with a massive bottom-quark pair at the Large Hadron Collider. *Phys. Rev. D*, 80:034015, 2009, [arXiv:0906.1923](#).
- [148] Hung-Liang Lai, Marco Guzzi, Joey Huston, Zhao Li, Pavel M. Nadolsky, Jon Pumplin, and C.-P. Yuan. New parton distributions for collider physics. *Phys. Rev. D*, 82:074024, 2010, [arXiv:1007.2241](#).
- [149] CMS Collaboration. Measurement of the $Z/\gamma^* + b\bar{b}$ -jets cross section in pp collisions at $\sqrt{s} = 7$ TeV with the CMS detector. CMS Physics Analysis Summary CMS-PAS-SMP-12-003, 2012.

- [150] CMS Public Webpage: Measurements of the production cross sections for a Z boson and one or more b jets in pp collisions at $\sqrt{s} = 7$ TeV. <https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSMP13004>. Accessed: 2015-02-13.
- [151] CMS Collaboration. Search for the standard model Higgs boson decaying to bottom quarks in pp collisions at $\sqrt{s} = 7$ TeV. *Phys. Lett. B*, 710:284, 2012, [arXiv:1202.4195](#).
- [152] CMS Collaboration. Search for the standard model Higgs boson produced in association with a W or Z boson and decaying to bottom quarks. *Phys. Rev. D*, 89, 2014, [arXiv:1310.3687](#).
- [153] G. D’Agostini. A Multidimensional unfolding method based on Bayes’ theorem. *Nucl. Instrum. Meth.*, A362:487–498, 1995.
- [154] K. Bierwagen. Bayesian Unfolding. *Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, Geneva, Switzerland:260–263, 17–20 January 2011, CERN–2011–006.
- [155] Andreas Hocker and Vakhtang Kartvelishvili. SVD approach to data unfolding. *Nucl. Instrum. Meth.*, A372:469–481, 1996, [arXiv:hep-ph/9509307](#).
- [156] V. Kartvelishvili. Unfolding with Singular Values Decomposition. *Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, Geneva, Switzerland:264–269, 17–20 January 2011, CERN–2011–006.
- [157] G. D’Agostini. Improved iterative Bayesian unfolding. 2010, [arXiv:1010.0632](#).
- [158] Tim Adye. Unfolding algorithms and tests using RooUnfold. pages 313–318, 2011, [arXiv:1105.1160](#).
- [159] The RooUnfold package and documentation. <http://hepunix.rl.ac.uk/~adye/software/unfold/RooUnfold.html>. Accessed: 2014-10-30.
- [160] V. Blobel. Unfolding Methods in High-energy Physics Experiments. 1984, DESY-84-118.
- [161] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, NJ 07632, USA, 1974.

- [162] R. Brun and F. Rademakers. ROOT: An object oriented data analysis framework. *Nucl.Instrum.Meth.*, A389:81–86, 1997.
- [163] The TUnfold package documentation. <http://www.desy.de/~sschmitt/tunfold.html>. Accessed: 2014-10-30.
- [164] Stefan Schmitt. TUnfold: an algorithm for correcting migration effects in high energy physics. *JINST*, 7:T10003, 2012, [arXiv:1205.6201](#).
- [165] R. D. Cousins. Interim recommendations on unfolding. http://cousins.web.cern.ch/cousins/unfolding_14mar13.pdf. Accessed: 2014-11-30.
- [166] Mikael Kuusela. A Statistician’s Perspective on Unfolding. Private communication at SMP MC Forum: Unfolding: Pitfalls and Recommendations. <https://indico.cern.ch/event/320358/>. Accessed: 2014-05-26.
- [167] The RIVET project. <https://rivet.hepforge.org/>. Accessed: 2014-11-30.
- [168] Johan Alwall, A. Ballestrero, P. Bartalini, S. Belov, E. Boos, et al. A Standard format for Les Houches event files. *Comput.Phys.Commun.*, 176:300–304, 2007, [arXiv:hep-ph/0609017](#).
- [169] Peter Zeiler Skands. Tuning Monte Carlo Generators: The Perugia Tunes. *Phys.Rev.*, D82:074018, 2010, [arXiv:1005.3457](#).
- [170] Tune comparisons. <http://rivet.hepforge.org/tunecmp/index.html>. Accessed: 2015-01-30.
- [171] CMS Collaboration. Search for H/A decaying into Z and A/H, with $Z \rightarrow \ell\ell$ and $A/H \rightarrow b\bar{b}$ or $A/H \rightarrow \tau\tau$. CMS Physics Analysis Summary CMS-PAS-HIG-15-001, 2015.