ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Έρευνα για υπερσυμμετρία με τελικές καταστάσεις φωτονίων και εγκάρσιας ελλείπουσας ορμής και Αναζήτηση της ταυτόχρονης παραγωγής μποζονίου **Higgs** με ένα ζεύγος **top-anti-top quark** στην πλήρως αδρονική τελική κατάσταση με τον ανιχνευτή **CMS** στον **LHC** στο **CERN.**

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ
της
Γαρυφαλλιάς Πασπαλάκη
Πτυχιούχου Φυσικών Εφαρμογών του Εθνικού Μετσόβιου Πολυτεχνείου

Επιβλέπων:
Κωνσταντίνος Κουσουρής
Αναπληρωτής Καθηγητής ΕΜΠ

Αθήνα
19 Νοεμβρίου 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Έρευνα για υπερσυμμετρία με τελικές καταστάσεις φωτονίων και εγκάρσιας ελλείπουσας ορμής
και
Αναζήτηση της ταυτόχρονης παραγωγής μποζονίου **Higgs** με ένα ζεύγος **top-anti-top quark** στην πλήρως αδρονική τελική κατάσταση με τον ανιχνευτή **CMS** στον **LHC** στο **CERN.**

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ
της

Γαρυφαλλιάς Πασπαλάκη
Πτυχιούχου Φυσικών Εφαρμογών του Εθνικού Μετσόβιου Πολυτεχνείου

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:

1............Κωνσταντίνος Κουσουρής, Αν. Καθ. ΕΜΠ (Επιβλέπων)

2............Αριστοτέλης Κυριάκης, Ερευνητής Α. Ε.Κ.Ε.ΦΕ ΄Δημόκριτος΄ (Επιβλέπων)

3............Δημήτριος Λουκάς, Ερευνητής Α. Ε.Κ.Ε.Φ.Ε ΄Δημόκριτος΄

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

1...........Κωνσταντίνος Κουσουρής, Αν. Καθ. ΕΜΠ

2...........Αριστοτέλης Κυριάκης, Ερευνητής Α. Ε.Κ.Ε.Φ.Ε ΄Δημόκριτος΄

3...........Δημήτριος Λουκάς, Ερευνητής Α. Ε.Κ.Ε.Φ.Ε ΄Δημόκριτος΄

4...........Γεώργιος Τσιπολίτης, Καθ. ΕΜΠ

5...........Νικόλαος Τράκας, Καθ. ΕΜΠ

6...........Κωνσταντίνος Θεοφιλάτος, Αν. Καθ. ΕΚΠΑ

7...........Χρήστος Μάρκου, Ερευνητής Α. Ε.Κ.Ε.Φ.Ε ΄Δημόκριτος΄

Αθήνα, Νοέμβριος 2020

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF APPLIED MATHEMATICS
AND PHYSICAL SCIENCE

Search for supersymmetry in events with photons and large missing momentum and a search of the production of a standard model Higgs boson in association with a top quark pair (ttH) in the all-jet final state using large-radius jets with the CMS detector at CERN LHC.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
by

Garyfallia Paspalaki

Supervisor:
Konstantinos Kousouris
Assistant Professor, NTUA

Athens
19 of November 2020

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF APPLIED MATHEMATICS
AND PHYSICAL SCIENCE

Search for supersymmetry in events with photons and large missing momentum and a search of the production of a standard model Higgs boson in association with a top quark pair (ttH) in the all-jet final state using large-radius jets with the CMS detector at CERN LHC.

DOCTORAL OF PHILOSOPHY PHYSICS

by

Garyfallia Paspalaki

ADVISORY COMMITTEE:

1............ Konstantinos Kousouris,
As. Prof, NTUA (Supervisor)

2............Aristotelis Kyriakis, Director of Research,
NCSR ’Demokritos’ (Supervisor)

3............Dimitrios Loukas, Director of Research,
NCSR ’Demokritos’

EXAMINATION COMMITTEE:

1...........Konstantinos Kousouris, As. Prof, NTUA

2...........Aristotelis Kyriakis, Director of Research,
NCSR ’Demokritos

3...........Dimitrios Loukas, Director of Research,
NCSR ’Demokritos’

4...........Georgios Tsiploitis, Prof. NTUA

5...........Nikolaos Trakas, Prof. NTUA

6...........Konstantinos Theofilatos, As. Prof NKUA

7...........Christos Markou, Director of Research
N.C.S.R ’Demokritos’

Athens, November 2020

Ευχαριστίες

Η προετοιμασία για την διδακτορική διατριβή είναι ένα αποτέλεσμα πολλών χρόνων και εμπειριών με πολλές δυσκολίες αλλά παράλληλα προσφέρει ενθουσιασμό σε έναν νέο άνθρωπο που καλείται να παράγει έρευνα. Σε αυτή την προσπάθεια συντέλεσαν πολλοί άνθρωποι, από διάφορα μέρη του κόσμου και από διαφορετικές εθνικότητες.

Ένα πολύ μεγάλο ευχαριστώ οφείλω στον ακαδημαϊκό μου υπεύθυνο, αναπληρωτή καθηγητή κύριο Κωνσταντίνο Κουσουρή ο οποίος με εμπιστεύτηκε ως κύρια αναλύτρια της ανάλυσης $t\bar{t}H$ και με καθοδήγησε σε όλη την διάρκεια του διδακτορικού. Είμαι πολύ ευγνώμων για όλες τις επιστημονικές μας συζητήσεις που με βοήθησαν να θέτω τις σωστές ερωτήσεις και παράλληλα να είμαι προετοιμασμένη να απαντήσω στις ερωτήσεις των άλλων. Επίσης, τον ευχαριστώ για την στήριξη του για την μετέπειτα ακαδημαϊκή μου καριέρα και για την συνειδητοποίηση πως η έρευνα πρέπει να είναι ένα κομμάτι της ζωής μας που μας κάνει χαρούμενους.

Θα ήθελα επίσης να ευχαριστήσω τον καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κύριο Γιώργο Τσιπολίτη, ο οποίος ως καθηγητής μου στα προπτυχιακά μου βήματα με εισήγαγε στο τομέα της πειραματικής φυσικής υψηλών ενεργειών και κατά την διάρκεια της διδακτορικής μου διατριβής μου έδινε χρήσιμα σχόλια. Επίσης θα ήθελα να ευχαριστήσω όλη την υπόλοιπη ομάδα του CMS@NTUA: Γιάννη, Γιώργο, Άννα, Ρένα και Θοδωρή, οι οποίοι με βοήθησαν όποτε τους χρειάστηκα τόσο σε συζητήσεις φυσικής αλλά και σε πιο πρακτικά θέματα.

Ένα μεγάλο ευχαριστώ οφείλω στην ομάδα του CMS του Δημοκρίτου και συγκεκριμένα στους ερευνητές, Άρη Κυριάκη και Δημήτρη Λουκά οι οποίοι φρόντισαν να χρηματοδοτηθούν τα ταξίδια μου στο CERN καθώς και η παρουσία μου σε επιστημονικά συνέδρια. Θα ήθελα να ευχαριστήσω και τα υπόλοιπα μέλη του Ι.Π.Σ.Φ του Δημοκρίτου, Θεόδωρο Γέραλη, Γεώργιο Αναγνώστου και Γεώργιο Δασκαλάκη για την θερμή υποδοχή ως μέλος της ομάδας.

Θα ήθελα να ευχαριστήσω τον Δρ Ιάσονα Τόψη, ο οποίος, ακόμα ως διδακτορικός φοιτητής, με ανέλαβε στα πρώτα μου βήματα για την έρευνα για υπερσυμμετρία. Η στήριξή του ήταν καθοριστική για την μετέπειτα πορεία μου. Επίσης θέλω να ευχαριστήσω τους διδακτορικούς φοιτητές, Άννα και Δημήτρη με τους οποίους μοιραστήκαμε το γραφείο κατά την διάρκεια της παραμονής μου στον Δημόκριτο.

Ένα μεγάλο ευχαριστώ θα ήθελα να δώσω στο CMS collaboration για την υπέροχη δουλειά χωρίς την οποία δεν θα ήταν δυνατόν να πραγματοποιηθεί αυτό το διδακτορικό. Ένα σημαντικό κομμάτι του CMS collaboration είναι η δυνατότητα να δουλέψεις με ανθρώπους κυριολεκτικά από όλο τον κόσμο. Θα ήθελα λοιπόν να ευχαριστήσω τους φοιτητές που συνεργαστήκαμε στις δυο αναλύσεις, Allie και Fabio. Η συνεργασία μας ήταν άψογη, δουλέψαμε σαν ομάδα και αντιμετωπίσαμε όλες τις δυσκολίες που προέκυψαν. Επίσης θα ήθελα να ευχαριστήσω τους συντονιστές της ομάδας που ασχολείται με την ταυτοποίηση του b quark του CMS, Caroline, Ivan και Kirill, οι οποίοι με βοήθησαν στην πραγματοποίηση του service work μου για το πείραμα.

Τέλος, θέλω να ευχαριστήσω τους δικούς μου ανθρώπους. Αρχικά, τους φίλους μου, Λήδα, Χριστίνα, Ελένη, Ντενιάνα, Σπύρο, Μενέλαο και Πάνο. Ευχαριστώ πολύ την οικογένεια μου και συγκεκριμένα τις αδερφές μου Μύριαμ και Νίνα με τις οποίες η συγκατοίκηση μου έμαθε πολλά. Ένα ξεχωριστό ευχαριστώ θα ήθελα να δώσω στις δύο αγαπημένες μου θείες, Μαρία και Μπέλλα, χωρίς την στήριξη των οποίων δεν θα ήμουν εδώ.

Περίληψη

Η παρούσα διατριβή περιλαμβάνει την προσωπική μου δουλειά, η οποία πραγματοποιήθηκε στο Εθνικό Κέντρο Φυσικών Ερευνών (Ε.Κ.Ε.Φ.Ε) ΄Δημόκριτος΄ και στο Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ). Περιλαμβάνει δύο ανεξάρτητες αναλύσεις φυσικής οι οποίες χρησιμοποιούν δεδομένα από συγκρούσεις πρωτονίων ενέργειας κέντρου μάζας $\sqrt{s} =$ 13 TeV που συλλέχθηκαν το 2016 από τον ανιχνευτή CMS στο CERN και αντιστοιχούν σε ολοκληρωμένη φωτεινότητά 35.9 fb^{-1}. Το πρώτο μέρος περιλαμβάνει μια έρευνα για Υπερσυμμετρία (SUSY) σε μοντέλα με τελικές υπογραφές φωτονίων και υψηλής εγκάρσιας ελλείπουσας ορμής (p_T^{miss}). Σε αυτά τα μοντέλα συνήθως το σπάσιμο της υπερσυμμετρίας μεταφέρεται σε χαμηλότερες ενεργειακές κλίμακες μέσω διαμεσολαβητών βαθμίδας (gauge mediated Supersymmetry breaking - GMSB). Το ελαφρύτερο υπερσυμμετρικό σωματίδιο είναι το gravitino ($\widetilde{G}$) και το αμέσως επόμενο σωματίδιο είναι το νετραλίνο ($\widetilde{\chi}$). Η διατήρηση της $\mathcal{R}$ ομοτιμίας εγγυάται ότι το gravitino είναι σταθερό και αλληλεπιδρά ελαφρώς με τον ανιχνευτή με αποτέλεσμα να υπάρχει στο γεγονός ελλείπουσα εγκάρσια ορμή η οποία ορίζεται ως το αρνητικό διανυσματικό άθροισμα των ορμών όλων των σωματιδίων. Τα αποτελέσματα χρησιμοποιήθηκαν για να θέσουν όρια στις ενεργές διατομές των gluino και των squark. Συγκεκριμένα, μάζες gluino κάτω των 1.86 TeV και squark 1.59 TeV αποκλείονται σε διάστημα εμπιστοσύνης 95%.

Το δεύτερο κομμάτι της ανάλυσης αφορά την έρευνα για τη ταυτόχρονη παραγωγή ενός μποζονίου Higgs μαζί με ένα ζεύγος top-antitop (t$\bar{\text{t}}$H) στην πλήρως αδρονική κατάσταση. Μετά την ανακάλυψη του μποζονίου Higgs ένας από τους βασικούς στόχους του προγράμματος του LHC είναι να κατανοήσει σε βάθος τις ιδιότητες του και συγκεκριμένα τις συζεύξεις τους με τα σωματίδια του Καθιερωμένου Προτύπου (ΚΠ). Η σύζευξη του μποζονίου Higgs με τα μποζόνια βαθμίδας έχει προσδιοριστεί με αρκετή ακρίβεια. Όμως, υπάρχει μια σημαντική αβεβαιότητα στις συζεύξεις του με τα φερμιόνια. Η διεργασία t$\bar{\text{t}}$H είναι άκρως σημαντική στο CMS αφού χρησιμοποιείται για να μετρήσει τη σύζευξη Yukawa του top. Επιπροσθέτως, η διάσπαση $b\bar{b}$ έχει μεγαλύτερη πιθανότητα να συμβεί σε σχέση με τις άλλες διασπάσεις του Higgs και για αυτό τον λόγο συνεισφέρει στην γενικότερη έρευνα της διεργασίας t$\bar{\text{t}}$H. Σε σύγκριση με τις προηγούμενες έρευνες στην πλήρως αδρονική διάσπαση, αυτή η ανάλυση ερευνά μια καινούργια προσέγγιση επιλέγοντας jets με μεγάλο Lorentz boost. Σε αυτές τις περιπτώσεις το μποζόνιο Higgs και τα top quark μπορούν να δημιουργηθούν με υψηλό Lorentz boost και έτσι τα προϊόντα της διάσπασης τους μπορούν να ανακατασκευαστούν σε ένα μεγάλης ακτίνας jet. Αφού τα προϊόντα της διάσπασης είναι συγκολλημένα, κάποιος μπορεί να ανακατασκευάσει την μάζα του boosted jet η οποία αντιστοιχεί στην μάζα του κάθε υποψηφίου. Αξίζει να σημειωθεί πως η μεγαλύτερη φωτεινότητα που αναμένεται στην περίοδο του HL-LHC τα jets χαμηλής ορμής θα είναι δύσκολο να ανιχνευτούν από τους υπάρχοντες σκανδαλιστές. Αντιθέτως, τα boosted jets δεν θα επηρεαστούν από την αναποτελεσματικότητα των σκανδαλιστών, καθιστώντας αυτή την προσέγγιση του φασικού χώρου ιδανική. Για την επιτυχή ταυτοποίηση του Higgs μποζονίου και των top quark χρησιμοποιήθηκαν τεχνικές πολλών μεταβλητών (MVA). Κάθε boosted jet μπορεί να ταυτοποιηθεί ως Higgs, top ή jet προερχόμενο από διεργασίες QCD. Στην παρούσα διατριβή, αναπτύχθηκαν μέθοδοι για την εκτίμηση του υποβάθρου οι οποίοι βελτιστοποιήθηκαν ώστε να μεγιστοποιήσουν την ευαισθησία της ανάλυσης. Η ευαισθησία της ανάλυσης μελετήθηκε στα πλαίσια του παρατηρούμενου (αναμενόμενου) ορίου το οποίο υπολογίστηκε στις 9.4 ($7.6 < 10.4 < 14$) φορές τις προβλέψεις του καθιερωμένου προτύπου.

Abstract

This dissertation describes my PhD work that was carried out at National Centre of Scientific Research (N.C.S.R) "Demokritos" and National Technical University of Athens (N.T.U.A). The work consists of two independent analyses. Both analyses use proton-proton collision data at the center of mass energy $\sqrt{s} = 13$ TeV, collected in 2016 with the Compact Muon Solenoid (CMS) detector at CERN LHC and correspond to a total integrated luminosity of 35.9 fb^{-1}. The first part includes a search for gauge mediated supersymmetry breaking in events that involve photons and large missing transverse momentum p_T^{miss}. For the interpretation of the results a gauge mediated supersymmetry scenario (GMSB) was assumed. Supersymmetry is a popular extension of the standard model (SM) of particle physics. For this analysis, a gauge mediated supersymmetry scenario (GMSB) was assumed. In GMSB models, the lightest supersymmetric particle is the gravitino, and the next-to-lightest supersymmetric particle is often taken to be the neutralino. The conservation of R parity implies that the gravitino is stable and thus, it can not be detected. The resulting imbalance in the total observed transverse momentum is referred to as missing transverse momentum $\tilde{\mathrm{p}}_{\mathrm{T}}^{\text{miss}}$, defined as the negative vector sum of the transverse momenta of all visible particles in an event. Its magnitude is referred to as p_T^{miss}. If the NLSP is bino-like, its primary decay will be to a gravitino and a photon (γ), resulting in final states with significant p_T^{miss} and one or more photons. The results were used to set cross section limits on gluino and squark pair production in this framework. Gluino masses below 1.86 TeV and squark masses below 1.59 TeV are excluded at a 95% confidence level.

The second part of the thesis concerns a search of the production of a standard model Higgs boson in association with a top quark pair (ttH) in the all jet final state. After the Higgs boson discovery, one of the main goals of the LHC program is to understand in depth its properties and in particular its couplings with the Standard Model (SM) particles. The couplings of the Higgs boson to gauge boson have been established fairly precisely. However, there is a considerable uncertainty in the couplings to fermions. The associated production of the Higgs boson is of particular importance in CMS as it is used to measure the top Yukawa coupling. Furthermore, the $b\bar{b}$ decay mode has the largest branching fraction for the 125 GeV Higgs boson and therefore contributes a large proportion of the statistics in the context of the wider ttH search. Compared to previous searches in the fully jet final state, this analysis explores a novel approach by selecting events with highly Lorentz-boosted jets. The Higgs boson and top candidates can be produced with a large Lorentz boost and hence their decay products can be reconstructed in a large radius jet. Since the decay products are merged, one can fully reconstruct the mass of the "*boosted-jet*" which corresponds to the mass of each candidate. Notably, as luminosity leveling will be used extensively in HL-LHC, low-p_{T} jets will be more difficult to trigger. Jets that are reconstructed in the boosted regime will not suffer from this effect at the trigger level, making this approach favored in this high pile up environment. To successfully identify Higgs and top candidates, dedicated Multivariate Analysis Techniques (MVA) were developed. Each boosted jet can be identified as Higgs, top, or jet coming from QCD. The methods that were developed for the background estimation in order to optimize the sensitivity of the analysis are also presented. The analysis sensitivity was studied in terms of the observed (expected) limit which is found to be 9.4 ($7.6 < 10.4 < 14$) times the SM expectations.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Theoretical Overview

1.1 The Standard Model of particle physics

The Standard Model of Particle Physics (SM) is the theory that describes all fundamental particle and interactions. The first part of the SM is composed by the fundamental matter particles, the fermions, with spin -$^1\!/_2$. Fermions are classified as either quarks and leptons and are arranged in three generations of increasing mass. The particles of higher generations decay via weak interactions to particles of the first generation. Furthermore, there are six flavors of quarks known at present: the up (u), charm (c), top (t) quarks which carry a +$^2\!/_3$ electric charge and the down (d), strange (s), bottom (b) quarks that carry a -$^1\!/_3$ charge. There also six flavors of leptons, the electron (e), muon (μ), and the tau (τ) lepton which carry a +1 electric charge, each accompanied by their neutrino. The SM includes the electromagnetic, strong and weak forces and all their carrier particles. The elementary particles that carry the fundamental forces are known as gauge bosons. These forces work over different ranges and have different strengths. In order of decreasing force strength the relevant carriers are: the gluons for the strong force, the photon for the electromagnetic and the two W's and the Z for the weak force. The final elementary particle in the SM is the Higgs boson. The Higgs boson is a scalar boson resulting from the Higgs-mechanism that was introduced by Higgs, Englert and Brout in 1964. Figure 1.1 summarizes the particles of the standard model and shows some basic properties such as the mass, the charge and the spin of each particle. The main focus of particle physics is the study of elementary particles that constitute matter and their interactions. In the following sections the fundamental interaction will be analyzed in more detail.

1.1.1 The SM Lagrangian

The SM is a quantum field theory which is invariant under local transformations of its gauge group:

$$\mathrm{G_{SM}} = \mathrm{SU(3)_C} \otimes \mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y} \tag{1.1}$$

where:

- $\mathrm{SU(3)_C}$ is the non-abelian gauge symmetry group which describes the strong interactions [16, 17]. C stands for the color quantum number. Such a structure involves eight independent matrices, which are the generators of the group, reflecting the fact that the strong interaction is carried by eight vector bosons, the

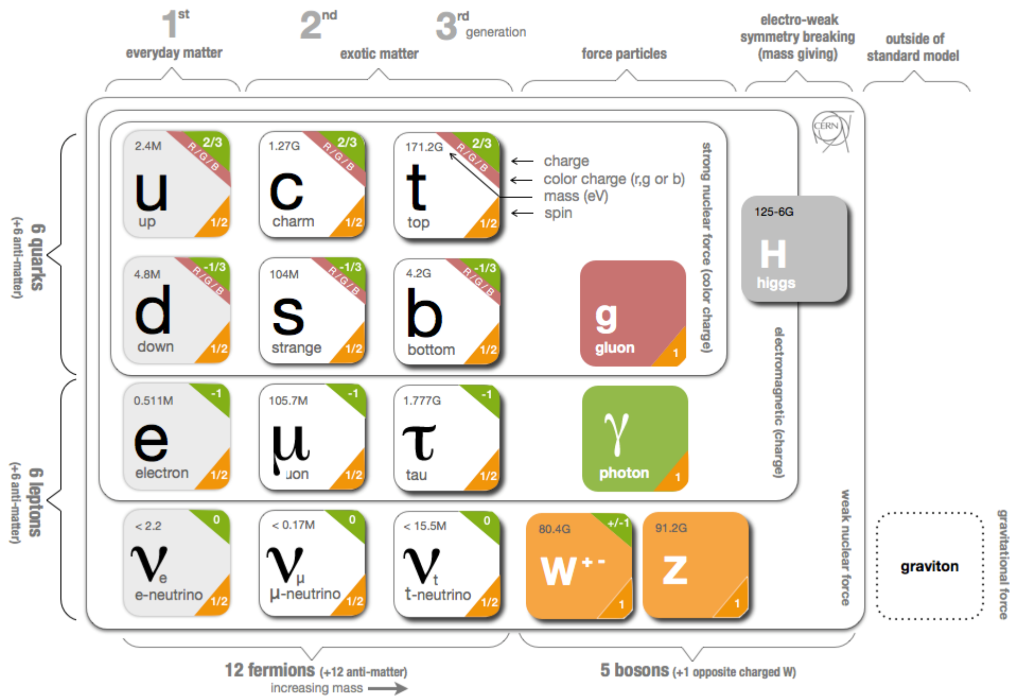

1st everyday matter
2nd exotic matter
3rd generation
force particles
electro-weak symmetry breaking (mass giving)
outside of standard model
6 quarks (+6 anti-matter)
6 leptons (+6 anti-matter)
2.4M 2/3 R/G/B u up 1/2
1.27G 2/3 R/G/B c charm 1/2
171.2G 2/3 R/G/B t top 1/2
charge
color charge (r,g or b)
mass (eV)
spin
4.8M -1/3 R/G/B d down 1/2
104M -1/3 R/G/B s strange 1/2
4.2G -1/3 R/G/B b bottom 1/2
g gluon 1
strong nuclear force (color charge)
125-6G H higgs
0.511M -1 e electron 1/2
105.7M -1 μ μon 1/2
1.777G -1 τ tau 1/2
γ photon 1
electromagnetic (charge)
< 2.2 0 νe e-neutrino 1/2
< 0.17M 0 νμ μ-neutrino 1/2
< 15.5M 0 ντ τ-neutrino 1/2
80.4G +/-1 W+- 1
91.2G Z 1
weak nuclear force
graviton
gravitational force
12 fermions (+12 anti-matter)
increasing mass
5 bosons (+1 opposite charged W)

Figure 1.1: The Standard Model of particle physics [1]

gluons. The gluons are massless, electrically neutral and carry the charge of strong interactions, known as "color". The strong interactions are well-described by the theory of quantum chromodynamics (QCD).

- $\mathrm{SU(2)_L \otimes U(1)_Y}$ is the weak isospin symmetry group which describes the electromagnetic and weak interactions together [18, 19] (electroweak interaction). L stands for the left-handed chirality (weak isospin) and Y for the hypercharge.

The SM can be described by the Lagrangian:

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{EWK}} + \mathcal{L}_{\mathrm{QCD}} + \mathcal{L}_{\mathrm{Higgs}} + \mathcal{L}_{\mathrm{Yukawa}} \tag{1.2}$$

where:

- $\mathcal{L}_{\mathrm{EWK}}$ and $\mathcal{L}_{\mathrm{QCD}}$ describe free fermions, free gauge bosons associated with the $\mathrm{SU(2)_L \otimes U(1)_Y}$ and $\mathrm{SU(3)_C}$ gauge symmetries, along with the interactions between fermions and gauge bosons as well as among gauge bosons

- $\mathcal{L}_{\mathrm{Higgs}}$ describe the Higgs particle and the electroweak symmetry breaking which will be described in section 1.2.1.

- $\mathcal{L}_{\mathrm{Yukawa}}$ describes the flavor physics.

1.2 The Standard Model Higgs boson

1.2.1 The mechanism of electroweak symmetry breaking

One of the characteristics of the SM Lagrangian, is that there are no explicit mass terms for the fermions and gauge fields, and thus it fails to explain why $\mathrm{W}^{\pm}$ and Z bosons are massive. The solution to this problem was given by Brought-Englert-Higgs mechanism [20, 21] that introduces a self-interacting complex scalar field ϕ which breaks spontaneously the $\mathrm{SU}(2) \otimes \mathrm{U}(1)$ symmetry. The SM scalar potential is

$$\mathrm{V}(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2 \tag{1.3}$$

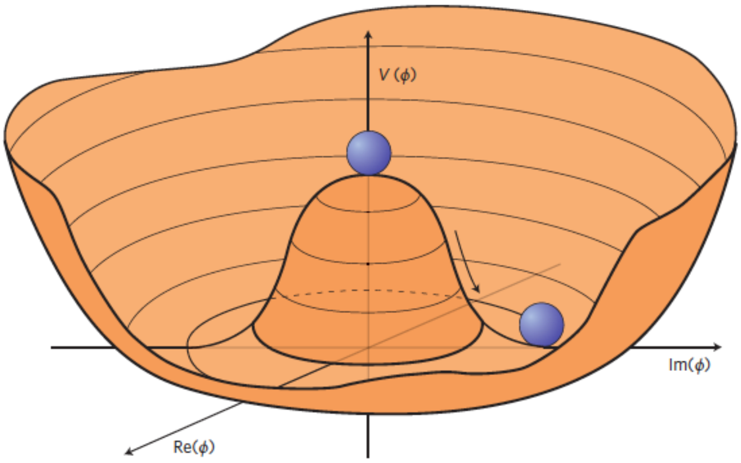


Figure 1.2: 'Mexican hat' potential that leads to 'spontaneous' symmetry breaking. The vacuum, i.e., the lowest-energy state, is described by a randomly-chosen point around the bottom of the brim of the hat [2].

where λ needs to be positive in order to have a finite minimum. If $\mu^2 > 0$, the potential has a unique minimum located at $\phi = (0, 0)$. If $\mu^2 < 0$, the potential has a finite set of minima defined by:

$$\phi^\dagger \phi = -\frac{\mu^2}{2\lambda} \tag{1.4}$$

This finite set of minima will result in a physical vacuum state. Without loss of generality, the vacuum state can be chosen to be

$$\langle \phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad \text{where} \quad v = \sqrt{-\frac{\mu^2}{\lambda}} \tag{1.5}$$

The choice of a specific vacuum results in symmetry breaking, although the potential itself respects the symmetry. The scalar field can be expanded around a vacuum expectation value (VEV) (figure 1.2) as

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + \mathrm{i}\phi_2 \\ v + h + \mathrm{i}\alpha^0 \end{pmatrix} \tag{1.6}$$

By introducing equation 1.6 to the standard model Lagrangian, three mass-less Goldstone bosons [22] are introduced, corresponding to ϕ_1 and ϕ_2 and α^0. By mixing the electroweak gauge fields they become the longitudinal components of the $W^{\pm}$ and Z gauge bosons. The Z and the W gauge bosons acquire masses,

$$M_W^2 = \frac{g^2 v^2}{4}, M_Z^2 = \frac{(g'^2 + g^2 v^2)}{4} \tag{1.7}$$

,where g and g' are the SU(2) and SU(1) gauge couplings, respectively. The remaining degree of freedom of the scalar field, h, is the physical Higgs boson with

$$m_H = \sqrt{-2\mu^2} = \sqrt{2}\lambda v \tag{1.8}$$

The fourth generator remains unbroken since it is the one associated to the conserved $\mathrm{U}(1)_{\mathrm{em}}$ gauge symmetry, and it's corresponding gauge field, the photon, remains massless. Similarly the eight color gauge bosons, the gluons, corresponding to the conserved $\mathrm{SU}(3)_{\mathrm{C}}$ gauge symmetry with eight unbroken generators, remain massless.

The electroweak symmetry breaking (EWSB) in the SM is responsible for generating mass for the W and Z gauge bosons. However, the fermions acquire mass through interactions with the Higgs field: the Yukawa interactions. The fermion mass m_f and the Yukawa coupling y_f is related by

$$m_f = \frac{1}{\sqrt{2}} y_f v \tag{1.9}$$

Although EWSB mechanism is very successful, it provides no additional insight on possible underlying reasons for the large variety of masses of the fermions, the so-called flavor hierarchy. The fermion masses, accounting for a large number of the free parameters of the SM, are simply translated into Yukawa couplings h_f. Figure 1.3 provides a schematic view of the Standard model interactions that were summarized in this section.

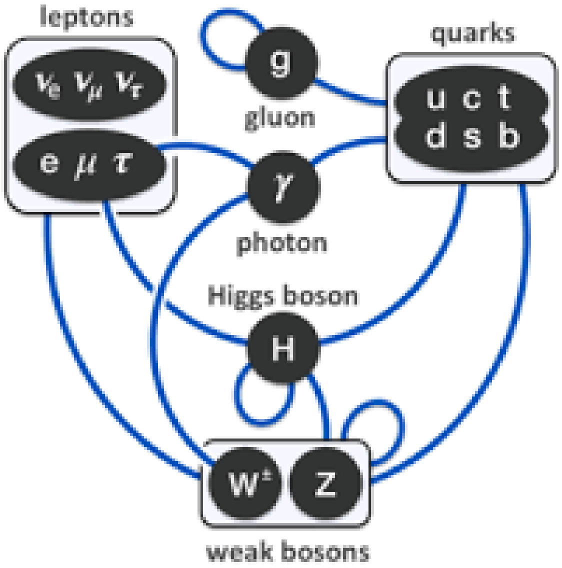


Figure 1.3: Standard Model Interactions

1.2.2 The Higgs boson properties

The SM Higgs boson is a CP-even scalar of spin 0. Its mass equals to $m_H = \sqrt{2\lambda}v$, where λ is the Higgs self-coupling parameter in $V(\Phi)$. In the SM, the quadratic coupling λ is a free parameter, whereas, the expectation value of the Higgs field, is fixed by the Fermi coupling $\mathrm{G_F}$ and is equal to $v = (\sqrt{2}\mathrm{G_F})^{-1/2} \approx 246$ GeV. Therefore, there is no a priori prediction for the Higgs mass. For a SM Higgs boson with mass $m_H \simeq 125$ GeV, the values of λ and $|\mu|$ are $\lambda \simeq 0.13$ and $|\mu| \simeq 88.8$ GeV. The interaction terms of the Higgs field [23], such as the couplings to gauge bosons and fermions and the Higgs boson self couplings, are summarized in the following Lagrangian:

$$\mathcal{L} = -g_{Hf\bar{f}}\bar{f}fH \; + \; \frac{g_{HHH}}{6}H^3 \; + \; \frac{g_{HHHH}}{24}H^4 \; + \; \delta_V V_\mu V^\mu (g_{HVV}H + \frac{g_{HHVV}}{2}H^2) \tag{1.10}$$

with,

$$g_{HHH} = \frac{3m_H^2}{u^2}, \quad g_{HHHH} = \frac{3m_H^2}{u^2}, \quad g_{Hf\bar{f}} = \frac{m_f}{u}, \quad g_{HVV} = \frac{2m_V^2}{u}, \quad g_{HHVV} = \frac{2m_V^2}{u^2},$$

where $V = \mathrm{W}^{\pm}$ or Z and $\delta_{\mathrm{W}} = 1$, $\delta_{\mathrm{Z}} = 1/2$.

The Higgs boson couplings to the fundamental particles are set by their masses. This type of interactions assumes a very weak interaction for light particles such as up and down quarks and electrons, but a strong one for heavy particles such as the W and Z bosons and the top quark. From equation 1.10, we observe that the Higgs couplings to fermions are linearly proportional to their masses, while the couplings to bosons are proportional to the square of the boson masses. As a result, the dominant mechanisms for Higgs boson production and decay involve the coupling of H to $\mathrm{W}^{\pm}$, Z and/or the heavier third-generation fermions (top and bottom quarks and the τ leptons.)

1.3 Higgs boson searches at the LHC

1.3.1 Higgs boson production

As stated in the previous section, the couplings between the Higgs boson and fermions or bosons is proportional to their mass. Thus, the most important production modes involves heavy particles like the vector bosons $\mathrm{W}^{\pm}$, Z_0 and the top quark. The four main Higgs boson production processes are:

- gluon-gluon fusion *(ggF)* through a heavy quark loop: $gg \rightarrow H$
- vector boson fusion *(VBF)*: $\mathrm{q_1q_2} \rightarrow \mathrm{V^*V^*} \rightarrow \mathrm{q_1^{'}q_2^{'}} + \mathrm{H}$
- associated production of the Higgs boson with a massive boson *(VH)*: $\mathrm{q\bar{q}} \rightarrow \mathrm{V^*} \rightarrow \mathrm{V} + \mathrm{H}$
- associated production of the Higgs boson with a pair of top quarks *(ttH)*: $\mathrm{gg} \rightarrow \mathrm{t\bar{t}} + \mathrm{H}$

The Feynman diagram of the dominant Higgs boson production modes at the LHC are shown in figure 1.4. The cross sections of these production processes are shown in figure 1.5 as a function of the center of mass energy, $\sqrt{s}$ and as a function of the Higgs boson mass for $\sqrt{s} = 13$ TeV.

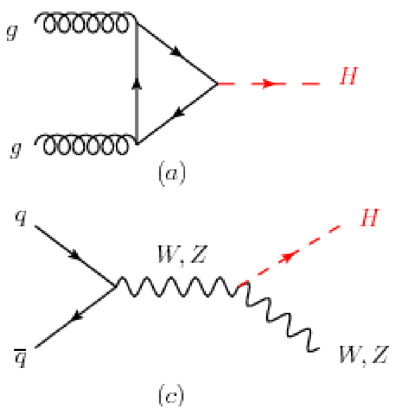


Figure 1.4: Generic Feynman diagrams contributing to the Higgs production in (a) gluon fusion, (b) weak-boson fusion, (c) Higgs-strahlung (or associated production with a gauge boson) and (d) associated production with top quarks

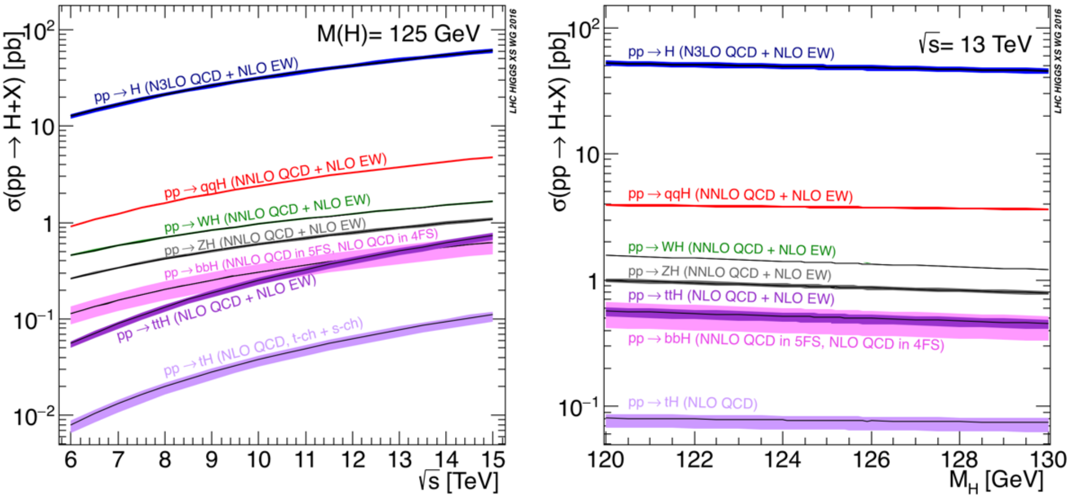


Figure 1.5: The SM Higgs boson production cross sections [3]: (left) as a function of the center of mass energy, $\sqrt{s}$ for pp collisions, (right) as a function of the Higgs boson mass for $\sqrt{s} = 13$ TeV. The theoretical uncertainties due to the higher order perturbative corrections are showed in the bands around the curves.

Gluon-gluon fusion

The dominant Higgs production mechanism at the LHC involves gluon fusion via an intermediate top-quark loop. As shown in figure 1.5, this particular processe's cross section is enhanced due to the fact that the Yukawa coupling between the Higgs boson and the heavy quarks present in the loop is high. The lowest order theoretical cross section is wellknown and used in many LHC studies to determine the experimental discovery sensitivity of the Higgs particle. Although in principle all quarks should be included in the loop, in practice the restriction to just the top quark suffices because the Higgs couples about 35 times more strongly to the top than to the next-heaviest fermion, the bottom quark, leading to a relative suppression of the bottom contribution by a factor 35^2. Therefore, the measurement of the ggF cross section also provides an indirect probe of the Higgs coupling to the top quark.

Vector Boson Fusion

The production of SM Higgs boson through vector-boson-fusion (VBF) mechanism features the second largest cross section among the Higgs production channels in hadronic collisions and, although smaller than the gluon-fusion one by about one order of magnitude, it still provides useful complementary information. The VBF process, illustrated in Figure 1.4 involves the radiation of a heavy vector boson from each incoming parton. Subsequently, the two vector bosons ”fuse” to produce a Higgs boson. The VBF production is very interesting as it provides special signatures for the Higgs boson identification. In particular, it features the presence of two forward quark jets, which can be exploited to identify such events. The process can be used to probe the Higgs coupling to the W and Z bosons.

Associated production with a heavy vector boson- Higgsstrahlung

This production mode is based on the annihilation of a quark couple into a virtual vector boson (off-shell) and the subsequent emission of a Higgs boson and of a real vector boson. Despite the smaller cross section compared to other production mechanism, the Higgsstrahlung can be exploited in the searches for the Higgs boson thanks to it’s clear signature. This is raised from the fact that the vector boson present in the final state can decay into leptons that are reconstructed very efficiently.

Associated production with top quarks pair

The measurement of the production cross section of a Higgs boson in association to a couple of heavy quarks (mainly with the top quark) can represent an excellent test of the Yukawa couplings. The on-shell top quarks are too heavy to be produced in a Higgs boson decay and thus this decay is kinematically forbidden. As a result the process $pp \to t\bar{t}H$ is the only way to directly constrain the top Yukawa couplings. At leading order this mechanism proceeds through a quark-antiquark annihilation into a couple of top-antitop quarks, where the Higgs boson is radiated from a top quark in the final state. This process is described in figure 1.6.

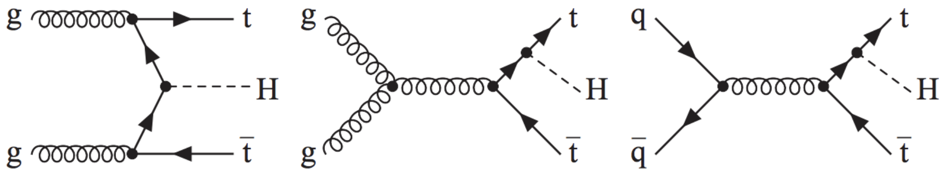


Figure 1.6: tree-level Feynman diagrams for the $pp \to ttH$ production process, with a gluon (g), a quark (q), a top quark (t), and a Higgs (H) boson

The associated Higgs boson production with a top quark-antiquark pair was observed in 2018 by the ATLAS [24] and CMS [25] collaborations. The observation was based on a combined analysis of proton-proton collision data that were collected by CMS and ATLAS detector in several run periods. To maximize the sensitivity, both analyses exploited the results of statistically independent searches for Higgs boson produced in conjunction with a top quark-antiquark pair and further decaying to W bosons, Z bosons, photons, τ leptons, or bottom quark jets. The overall agreement observed between the SM prediction and data is SM-like, since the quantum loops in this processes include top quarks. However, non-SM particles in the loops could introduce terms that compensate for, and thus mask, other deviations from the SM. Taking this into account, a measurement of the production rate of the tree level $\mathrm{t\bar{t}H}$ process can provide evidence for, or against, such new-physics contributions.

1.3.2 Higgs Boson decays-Higgs Boson discovery

The decay modes of a SM Higgs boson strongly depend on its mass m_H. For a low mass Higgs boson ($110\ \mathrm{GeV/c^2} < \mathrm{m_H} < 150\ \mathrm{GeV/c^2}$) it's natural width is only a few $\mathrm{MeV/c^2}$, thus the channels with the most important contribution are the $\mathrm{H} \to \mathrm{ZZ}^* \to 4\ell$, $\mathrm{H} \to \gamma\gamma$, $\mathrm{H} \to \mathrm{WW}^* \to 2\ell 2\nu$, $H \to \mathrm{b\bar{b}}$ and $\mathrm{H} \to \tau^+\tau^{-1}$. The sensitivity of the search for a given hypothesis in the Higgs boson mass depends on the Higgs boson production cross section, it's decay's branching ratio into a chosen final state, the signal selection efficiencies, the reconstructed mass resolution and the level of SM backgrounds in the final state. For the Higgs boson discovery the $\mathrm{H} \to \mathrm{ZZ}^* \to 4\ell$, and $\mathrm{H} \to \gamma\gamma$ played a crucial role since those decay modes give an excellent mass resolution of the reconstructed 4-lepton and di-photon final states. On the contrary the rest decay modes are less significant due to poor mass resolution coming from the presence of neutrinos in the $\mathrm{H} \to \mathrm{WW}^* \to 2\ell 2\nu$ and the presence large background for the $\mathrm{H} \to \mathrm{b\bar{b}}$ and $\mathrm{H} \to \tau^+\tau^{-1}$. In 2012 both ATLAS and CMS [26, 27] announced the discovery of SM like Higgs boson. Since then, many measurements were performed of its properties and its production cross sections in different channels, providing estimates of its cross section, decay rates and couplings [28], as well as measure its mass [29]. The production and decay rates are measured in terms of the signal strength μ. The signal strength μ is defined as the ratio of the measured production cross section or decay branching ratio to its SM prediction.

$$\mu = \frac{\sigma}{\sigma_{\mathrm{SM}}} \quad \text{or} \quad \frac{\Gamma}{\Gamma_{\mathrm{SM}}} \tag{1.11}$$

Similarly a coupling modifier κ, defined as the square root of the ratio of the measured cross section ($\kappa^2 = \sigma/\sigma_{\mathrm{SM}}$) or decay width ($\kappa^2 = \Gamma/\Gamma_{\mathrm{SM}}$) to the SM prediction is used to measure the Higgs boson couplings to bosons and fermions. Several individual search results and some combined results have already been published [30]. In the following part the Higgs decay modes are described in detail.

Higgs decays to vector bosons

As stated previously the decays to ZZ, where each Z boson further decays to two leptons gives a clean fully reconstructed resonant-mass peak, which can provide precision measurements of the Higgs boson mass [31]. The same stands for the $H \to \gamma\gamma$ process since it features a clear peak in a falling and well understood background [32]. The diphoton invariant mass distribution in the $H \to \gamma\gamma$ decay channel and the four lepton (4μ,$2\mathrm{e}2\mu$,4e) invariant mass distribution in the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ decay from the CMS collaboration are shown in Figure 1.7. Although the decay mode to WW has a large branching fraction due to the large mass of the W vector boson, this search can be challenging because of the presence of the neutrinos in the W decay products, resulting to a poor mass resolution. A $H \to WW$ analysis performed in CMS focused on the search for oppositely charged electron-muon pairs that was used to construct the dilepton invariant mass [33]. To extract the signal, the dilepton invariant mass was compared with the Higgs transverse mass which is calculated from the transverse momentum of the leptons and the missing transverse momentum.

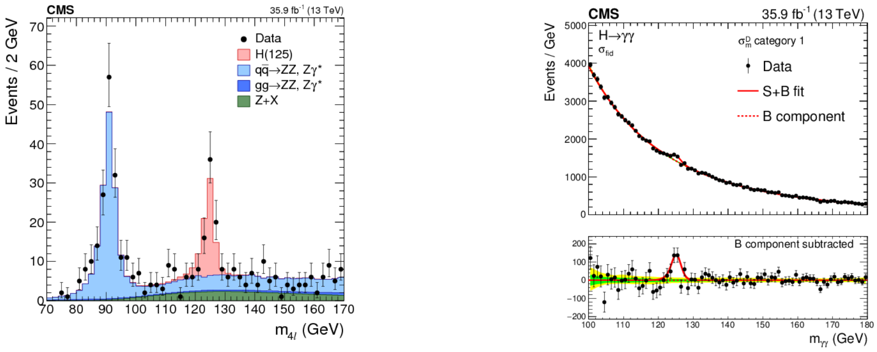


Figure 1.7: left: Four lepton (4μ,$2e2\mu$,4e) invariant mass distribution in the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ decay right: Diphoton invariant mass distribution in the $\mathrm{H} \to \gamma\gamma$ decay channel

Higgs decays to fermions

To establish the mass generation mechanism for fermions, it is necessary to probe the direct coupling of the Higgs boson to such particles. One of the most promising decay channels is the $\tau^+\tau^-$, because of the large event rate expected in the SM compared to the $\mu^+\mu^-$ channel and the smaller contribution from background events with respect to the $b\bar{b}$ decay channel. CMS collaboration performed a search [34] where events with both hadronic and leptonic decays of the tau leptons were considered. The presence of the neutrinos in the final state make the reconstruction of a clear $\tau\tau$ invariant mass impossible. However, the $\mathrm{m}_{\tau\tau}$ is estimated by a likelihood method along with the product of the final state.

The decay of a Higgs boson to a pair of b quarks ($\mathrm{H} \to \mathrm{b\bar{b}}$) has a predicted branching fraction of 58% for a standard model Higgs boson of mass $\mathrm{m_H} = 125\ \mathrm{GeV}$. A precise measurement of the rate of this process directly probes the Yukawa coupling of the Higgs boson to a down-type quark and provides a necessary test of the hypothesis that the Higgs field is the source of mass generation in the charged fermion sector of the SM. In the LHC the most sensitive production process is when the Higgs boson is produced in association with a vector boson (VH). In 2018 both ATLAS [35] and CMS [36] collaborations observed the $\mathrm{H} \to \mathrm{b\bar{b}}$ in associated production with the $\mathrm{W/Z}$ at 13 TeV with data collected in 2016 and 2017 corresponding to an integrated luminosity of 36.8 and $41.3\ \mathrm{fb}^{-1}$ respectively. The searches performed by CMS collaboration, focus on the leptonic decays of the weak vector boson and reconstruct the Higgs boson by selecting two b jets. A simultaneous binned-likelihood fit to the shape and yield of specific distributions for the signal and control regions for all channels combined is used to extract a possible Higgs boson signal.

1.3.3 Associated production of a standard model Higgs boson with a top quark-antiquark pair and its further decay to a pair of bottom quarks

A measurement of the associated production of a standard model Higgs boson with a top quark-antiquark pair $\mathrm{t\bar{t}H}$ and it's further decay to a pair of b quarks was performed both in ATLAS [37] and CMS [38] collaborations. The analysis performed by CMS, used data from proton-proton collisions that were collected by the CMS detector and correspond to an integrated luminosity of $41.5\ \mathrm{fb}^{-1}$. Those results were combined with $35.9\ \mathrm{fb}^{-1}$ data obtained in 2016. The combination of the two run years give a best-fit value of $\hat{\mu} = 1.15^{+0.15}_{0.15}(\mathrm{stat})^{+0.28}_{-0.25}(\mathrm{syst})$. The best fit of the $\mathrm{t\bar{t}H}$ signal strength relative to the standard model cross section ($\hat{\mu} = \sigma/\sigma_{\mathrm{SM}}$) comes from a combined fit of multivariate discriminant distributions is all categories. Events that are consistent with the production of the a top quark-antiquark pair with additional b quark jets are selected. Top quarks decay almost exclusively to a bottom quark and a W boson, and then the W boson can decay either into a charged lepton and a neutrino or into a pair of quarks. The decay of the W defines the final signatures recorded in the detector. All the $\mathrm{t\bar{t}}$ channels were considered. The fully hadronic channel, where both W bosons decay into quarks, the single-lepton channel, where one W boson decays into a charged lepton (electron or muon) and a neutrino and the other W boson decays into quarks, and the dilepton channel, where both W boson decay into a charged lepton (electron or muon) and a neutrino.

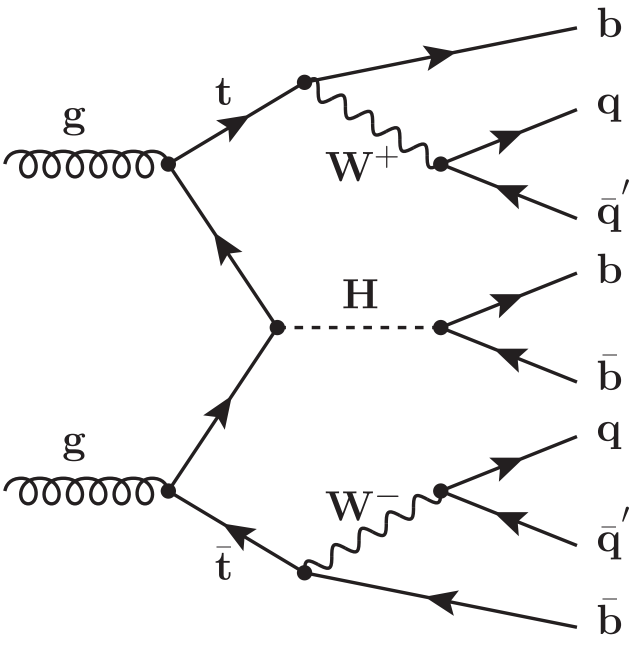


Figure 1.8: tree-level Feynman diagrams for the all-jet $ttH(H \to bb)$ process

The fully hadronic $\mathrm{t\bar{t}H(H \to b\bar{b})}$ final state

In the fully hadronic or all-jets channel, each W boson arising from the top quark decays, will decay into a pair of light quarks. In addition, the Higgs boson will decay to a pair of b quarks resulting in a final state with at least eight quarks, four of which are b quarks. All the eight final quarks will hadronize into jets. Those jets are typically produced at large angles with respect to the beam axis resulting in a relatively high transverse momentum. The feynman diagram of this process is showed in figure 1.8. This particular process is of high interest due to a very specific Higgs coupling space: all couplings are fermionic and restricted to the third-generation quarks only. This leads to an easier interpretation of results than those in other decay channels. The absence of leptons makes this final state challenging to target, however, it is possible to separate the signal from background by using sophisticated MVA techniques. A traditional strategy was performed by CMS [39] and ATLAS [40] collaborations in the so called resolved final state, where all the products for the Higgs and top decays could be reconstructed independently. This hypothesis is valid for Higgs transverse momenta up to 250 GeV. Compared to previous searches in the fully jet final state, the work presented in this thesis (Chapter 9) focuses on fully boosted and semi-boosted topologies. At higher $p_{\mathrm{T}}(\mathrm{p_T}/\mathrm{m} \approx 1)$ the Higgs boson and the top quark decay products are highly collimated ("*boosted*") and thus can no longer be reconstructed separately. In order to explore this phase space, hadronic tops and the Higgs are reconstructed in large-radius jets. In this analysis approach the Higgs boson is always reconstructed as a boosted jet.

1.4 Extensions of the Standard Model

1.4.1 Motivation for Supersymmetry

In the previous section the Standard Model (SM) and the Higgs physics was discussed. Although the SM gives a very good description of the elementary components of nature, it fails to describe it completely. For instance, it cannot explain the repeating pattern of the three observed generations in matter or to provide answers to questions like why the masses of the particles are what they are. Furthermore, it fails to embody a theory of the gravitational interactions. Experiments have already shown that with increasing energies, the effect of the strong force becomes weaker. This is a good indication that the electromagnetic, weak and strong forces could be merged to one single interaction in an incredibly high energy environment; the idea of Grand Unified Theories (GUTs). Supersymmetry (SUSY) is a theoretical favoured extension of the SM as it provides a mathematical framework that allows the strong and electroweak forces to unite and become a single interaction at a common energy. The existence of dark matter, which makes up approximately one quarter of the energy density of the universe, is another theoretical argument that cannot be explained by the SM of particle physics. Moreover, SUSY can explain some of the inconsistencies of the SM such as the infamous "hierarchy problem"; the large descrepancy between aspects of the weak nuclear force and gravity. The "hierarchy problem" is related to the Higgs mass stability under radiative corrections. The Higgs boson interacts with every particle with mass, and without an incredible fine-tuning of parameters, loop-order quantum effects from these particles would give enormous corrections to the Higgs mass, driving it from the electroweak scale to the Planck scale [41].

In Figure 1.9 the correction to the Higgs squared mass ($\mathrm{m_H^2}$) from a loop containing a Dirac fermion is given by:

$$\Delta \mathrm{m_H^2} = -\frac{|\lambda_{\mathrm{f}}|^2}{8\pi^2}\Lambda_{\mathrm{UV}}^2 + ... \tag{1.12}$$

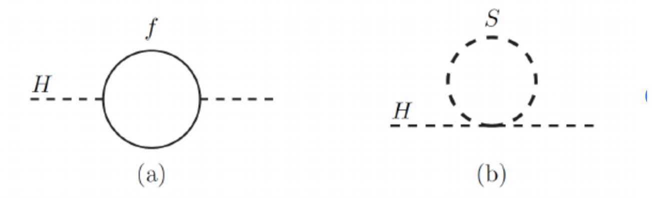


Figure 1.9: One-loop quantum corrections to the Higgs boson's squared mass $\mathrm{m_H^2}$, due to a Dirac fermion f (left) and a scalar S (right).

and the relevant corrections in case of the scalar S illustrated on the right part of Figure 1.9 is

$$\Delta \mathrm{m_H^2} = \frac{\lambda_\mathrm{S}}{16\pi^2}[\Lambda_\mathrm{UV}^2 - 2\mathrm{m_S^2}\ln(\Lambda_\mathrm{UV}/\mathrm{m_S}) + ...] \tag{1.13}$$

where λ_f is the fermion coupling to the Higgs boson and Λ_{UV} is an ultraviolet momentum cutoff used to regulate the loop integral. If Λ_{UV} is of the order of the reduced Planck scale, the correction $\Delta\mathrm{m_H^2}$ should be several orders of magnitude larger than the expected value for the Higgs mass. As a result, several fine-tuning of parameters, or in other words cancellations between the various contributions to $\Delta\mathrm{m_H^2}$ are required in order to adjust the electroweak scale to much smaller than the weak scale. Comparing equations 1.12 and 1.13 it naturally comes that a symmetry between fermions and bosons is needed due to the relative minus sign between fermion and boson loop contributions to the $\Delta\mathrm{m_H^2}$. Assuming $\lambda_\mathrm{S} = |\lambda_\mathrm{f}|^2$, then the Λ_{UV}^2 contributions will neatly cancel. The existence of a boson state corresponding to each fermion state would imply the presence of a symmetry. A solution scenario like this is given by Sypersymmetry (SUSY). Since no SUSY particle has been seen, SUSY is a broken symmetry, meaning that the SUSY particles (sparticles) are much heavier and less stable than their standard model counterparts. However, in case SUSY particles exist at energy scales compatible with the Large Hadron Collider (LHC) environment, we will be able to produce and study them.

1.5 Supersymmetry

Supersymmetry is one of the most well developed beyond the standard model theories that connect matter and force particles. This symmetry is succeeded by assigning to each SM particle a (super)partner with opposite sign spin; spin 1/2 for the force particles and spin 1 for the matter particles. The names of the superpartners are inspired from their SM counterparts. The matter superparterns are given a "s" for example selectron for electron, while force partners are given an "ino", i.g gluino for gluon. The symbols for supersymmetric particles are given simply by adding a tilde to the SM symbol. For example $\tilde{e}$ is the symbol for selectron. Figure 1.10 shows the particles of the Standard Models along with their superpartners.

The supersymmetric operator Q turns a bosonic state to a fermionic state.

$$Q|boson\rangle = |fermion\rangle, \qquad Q|fermion\rangle = |boson\rangle \tag{1.14}$$

The generator Q which changes the spin of a field by 1/2 is a fermionic spinor that carries an intrinsic angular momentum of 1/2. In this supersymmetric theory,

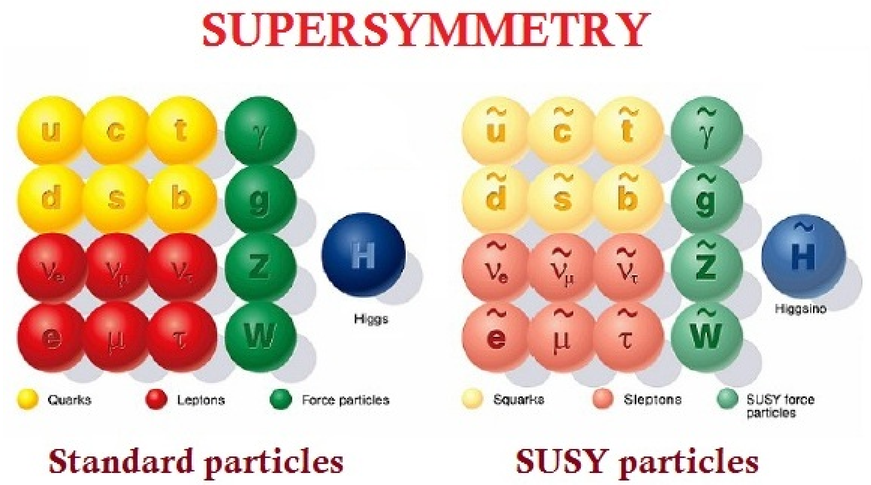


Figure 1.10: The Standard Model particles along with their Supersymmetric partners

single-particles are grouped into irreducible representations of the supersymmetric algebra, called supermultiplets. Each supermultiplet contains both fermion and boson states which are called superpartners of each other [41]. The supermultiplets can be divided into two categories; chiral supermultiplets and gauge supermultiplets which are listed in tables 1.1 and 1.2. Chiral supermultiplets 1.1 include the leptons, the quarks, the Higgs bosons and their superpartners. The SM fermions couple differently under different gauge fields resulting in separate scalar partners. The scalar partners of the SM fermions have the same gauge interactions as their partners. The Higgs boson is also described by a chiral multiple. In the Minimal Supersymmetric Standard Model (MSSM) in order to avoid gauge anomalies in the electroweak symmetry we need two Higgs chiral supermultiplets [42], one with hypercharge $Y = 1/2$ and one with $Y = -1/2$. The first H_u has the Yukawa coupling to give masses to the up-type quarks and leptons, whereas the H_d with $Y = -1/2$ will give masses to the down-type quarks and leptons. The 125 GeV standard model Higgs boson is a linear combination of H_u^0 and H_d^0. Gauge supermultiplets classify the gauge bosons and their superpartners 1.2. For a renormalizable theory, before the gauge symmetry is spontaneously broken, there should be a massless gauge boson. For a massless boson (spin 1) with two helicity states there is a superpartner which is a massless fermion of spin 1/2 with two helicity states. Gauge bosons and their fermionic partners must transform like the adjoint representation of the gauge group. Since the adjoint representation of a gauge group it's always its own conjugate, these fermions must have the same gauge transformation properties for left and right handed components. This combination of a spin 1/2 gaugino and spin1 gauge boson is called a gauge or vector supermultiplet.

Members of the same supermultiplet can be transformed to each other using the Q and $Q^\dagger$. Suppose two states $|b\rangle$ and $|f\rangle$, with masses m_b and m_f respectively that are members of the same multiplet. In the supersymmetric algebra [43] we have

$$\{Q, Q^\dagger\} = P^\mu \tag{1.15}$$

$$\{Q, Q\} = \{Q^\dagger, Q^\dagger\} = 0 \tag{1.16}$$

Particles		spin-0	spin-1/2	$SU(3)_C$	$SU(2)_L$	$U(1)_Y$
squarks quarks,	Q	$(\tilde{u}_L, \tilde{d}_L)$	(u_L, d_L)	3	2	$\frac{1}{6}$
3 families	$\bar{u}$	$\tilde{u}^*_R$	$u^\dagger_R$	$\bar{3}$	1	$-\frac{2}{3}$
	$\bar{d}$	$\tilde{d}^*_R$	$d^\dagger_R$	$\bar{3}$	1	$\frac{1}{3}$
sleptons, leptons	L	$(\tilde{\nu}\ \ \tilde{e}_L)$	νe_L	1	2	$-\frac{1}{2}$
3 families	$\bar{e}$	$(\tilde{e}^*_R)$	$e^\dagger_R$	1	1	1
Higgs, higgsinos	H_u	(H^+_u, H^0_u)	$(\tilde{H}^+_u, \tilde{H}^0_u)$	1	2	$\frac{1}{2}$
Higgs, higgsinos	H_d	(H^0_d, H^-_d)	$(\tilde{H}^0_d, \tilde{H}^-_d)$	1	2	$-\frac{1}{2}$

Table 1.1: The MSSM chiral supermultiplets

Particles	spin-1/2	spin-1	$SU(3)_C$	$SU(2)_L$	$U(1)_Y$
gluino, gluon	$\tilde{g}$	g	8	1	0
winos, W bosons	$\tilde{W}^{\pm}$, $\tilde{W}^0$	$W^{\pm}$, W^0	1	3	0
bino, B boson	$\tilde{B}^0$	B^0	1	1	0

Table 1.2: The MSSM gauge supermultiplets

$$[P^\mu, Q] = [P^\mu, Q^\dagger] \tag{1.17}$$

where P^μ is the four -momentum generator of spacetimes translations. Since $P^\mu P_\mu|b\rangle = m_b^2|b\rangle$ and $P^\mu P_\mu|f\rangle$ and considering the previous relations, we have:

$$P^\mu P_\mu Q|b\rangle = P^\mu P_\mu|f\rangle = m_f^2|f\rangle \tag{1.18}$$

$$P^\mu P_\mu Q|b\rangle = QP^\mu P_\mu|b\rangle = m_b^2 f\rangle \tag{1.19}$$

Combining the equations 1.18 and 1.19 we observe that $m_b = m_f$, which proves that superpartness must have equal masses. The supersymmetry generators Q, Q† also commute with the generators of gauge transformations leading to same electric charge, weak isospin and color degrees of freedom for particles that belong in the same multiple. Since no supersymmetric particle is observed in nature, Supersymmetry must be a broken symmetry that allows the superparticles to be heavier than the corresponding Standard Model ones. However, SUSY must be broken in a way that no ultraviolet divergences will appear in scalar masses. This is the so-called soft SUSY breaking that yields heavy superpartners in a natural way. In the SM, due to global gauge invariance, baryon (B) and lepton (L) numbers are conserved. However, those terms are violated in the SUSY framework. To construct renormalizable operators consistent with the SM gauge symmetries a new symmetry known as R-parity is introduced. R-parity is defined as:

$$P_R = (-1)^{2(B-L)+2s} \tag{1.20}$$

where s is the spin of the particle. The SM particles including the Higgs boson have an even R-parity ($P_R = +1$), while the sleptons, gauginos and higgsinos have odd parity ($P_R = -1$). If R-parity is conserved, no mixing is allowed between sparticles and the particles with even parity (SM particles). The conservation of R-parity results in some very important phenomenological observations:

- The lightest sparticle ($P_R = -1$) known as the lightest supersymmetric particle (LSP), has to be absolutely stable and produced at the end of the decay chain of a heavy unstable supersymmetric particle. If it is electrically neutral, it is weakly interacting with ordinary matter. As a result, the LSP could be a possible dark matter candidate. In collider experiments the LSP will result in an imbalance in the reconstructed momentum in the transverse plane of the detector,

- Each sparticle apart from LSP, must eventually decay into a state that contains an odd number of LSPs.

- In collider experiments, sparticles can only be pair produced.

1.5.1 Minimal Supersymmetric standard model

The Minimal Supersymmetric Standard Model (MSSM) is the most basic extension that incorporates SUSY. This framework is consistent with the SM and also provides a solution to the hierarchy problem as it stabilizes the weak scale. In MSSM R-parity is conserved and Poincare and gauge invariance is assured. As described in section 1.5, the supermultiplets of MSSM provide a partner for every SM particle with the a spin difference of 1/2. The Higgs boson in MSSM has a fermionic superpartner, the Higgsino.

The Lagrangian of the MSSSM is described as:

$$\mathcal{L}_{\text{MSSM}} = \mathcal{L}_{\text{SUSY}} + \mathcal{L}_{\text{Breaking}} \tag{1.21}$$

where $\mathcal{L}_{\text{SUSY}}$ is the SUSY generalization that contains all the gauge and Yukawa interactions and preserves supersymmetry invariance and $\mathcal{L}_{\text{Breaking}}$ describes the SUSY breaking.

The MSSM embodies supersymmetry into the Standard Model in a simple and elegant way by making "minimal" additions.

- adds superpartners to the gauge filed bosons (*gauginos*)

- adds superpartners to the fermions

- adds superpartners to the Higgs field (*higgsinos*)

- adds a second Higgs doublet.

The neutral higgsinos ($\tilde{H}_u^0$ and $\tilde{H}_d^0$) and the neutral electroweak gauginos ($\tilde{W}_0$ and $\tilde{B}_0$) form four mass eigenstates known as neutralinos $\tilde{\chi}^0$. The charged higgsinos ($\tilde{H}_u^+$ and $\tilde{H}_d^-$) and charged winos $\tilde{W}^{\pm}$ form four mass eigenstates known as charginos $\tilde{\chi}^{\pm}$. Mixing is allowed between gauginos and higgsinos with the same charge but forbidden for gluinos due to its color charge.

1.5.2 Gauge Mediated Supersymmetry Breaking (GMSB)

The symmetry breaking in SUSY can be achieved by introducing a hidden sector. This sector causes the breaking, which is then mediated to the MSSM (visible sector). In Gauge Mediated Supersymmetry Breaking (GMSB) [44, 45, 4] scenarios, the communication between hidden sector, where SUSY breaking takes place, and the visible MSSM sector (consisting of chiral supermultiplets shown in 1.1) is via the ordinary gauge interactions. The messangers communication between MSSM and the hidden sector also have a $\mathrm{SU(3)_C} \otimes \mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ interactions. General Gauge Mediation (GGM) scenario is one of the most popular and most robust ways of transmitting SUSY breaking to the MSSM. Compared with other SUSY breaking scenarios, in GMSB flavor changing neutral current processes and new sources of CP violation are naturally suppressed. The soft terms in MSSM come from loop diagrams involving these messengers, whose value is given by

$$\mathrm{m_{soft}} \sim \frac{\alpha_\mathrm{a}}{4\pi} \frac{\langle \mathrm{F} \rangle}{\mathrm{M_{mess}}} \tag{1.22}$$

where $\frac{\alpha_\mathrm{a}}{4\pi}$ is a loop factor for Feynman diagrams involving gauge interactions, F relates to the SUSY breaking scale and $\mathrm{M_{mess}}$ is the messenger mass scale. GMSB permits a significantly lower symmetry-breaking scale ($\langle \mathrm{F} \rangle$) than, for example gravity mediation, and therefore generically predicts that the gravitino ($\widetilde{\mathrm{G}}$) is the Lightest Supersymmetric Particle (LSP) whose mass is given by

$$m_{\widetilde{G}} \sim \langle F \rangle / M_P \sim keV \tag{1.23}$$

where $\mathrm{M_P}$ is the Planck scale.

1.5.3 Phenomenology of General Gauge Supersymmetry Breaking (GGMSB)

As mentioned previously, the gravitino $\widetilde{\mathrm{G}}$, is taken to be the LSP. That means that it is considered a stable particle that weakly interacts with the detector, resulting in missing transverses momentum. The next-to-the-lightest supersymmetric particle (NLSP) is the neutralino $\widetilde{\chi}^0$ or the chargino ($\widetilde{\chi}^{\pm}$). The decay modes of the NLSP are decided by the manner in which bino (the superpartner of the $\mathrm{U(1)}$), wino (the superpartner of the $\mathrm{SU(2)}$) and higgsino components mix, and hence define the nature of this mass eigenstate [46]. In case of gaugino-like NLSP, the neutralino consists predominantly of either the bino or the wino gauge field.

- A **bino like** NLSP decays predominantly into a gravitino ($\widetilde{\mathrm{G}}$) and a photon (γ) with a branching fraction $\sim \cos\theta_\mathrm{w}^2$, while the decay to a gravitino and a Z boson is sub-dominant. The left plot on figure 1.11 shows the branching fraction of a bino-like NLSP as a function of its mass. Experimentally this kind of scenarios can be targeted using collider events with final signatures of $\gamma\gamma + p_T^{\mathrm{miss}}$ or $\gamma + p_T^{\mathrm{miss}}$ final states, where p_T^{miss} is the magnitude of missing transverse momentum.

- for a **wino-like** NLSP, the splitting between the charged and the neutral wino is small resulting in neutral and charged winos to become co-NLSPs. The charged wino will decay directly into the gravitino ($\widetilde{\mathrm{G}}$) and a $W^{\pm}$ as well, while the neutral will decay dominantly to a gravitino ($\widetilde{\mathrm{G}}$) and a Z boson ($\sim \cos\theta_w{}^2$) and sub-dominantly to a gravitino ($\widetilde{\mathrm{G}}$) and a photon (γ) ($\sim \sin\theta_\mathrm{w}^2$). The right plot on

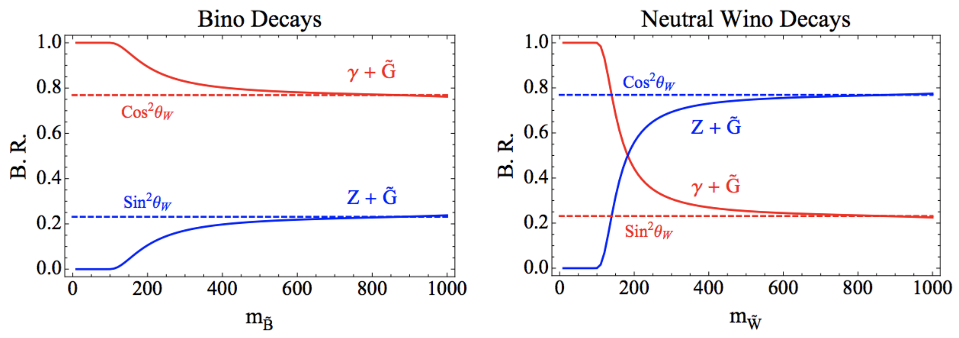


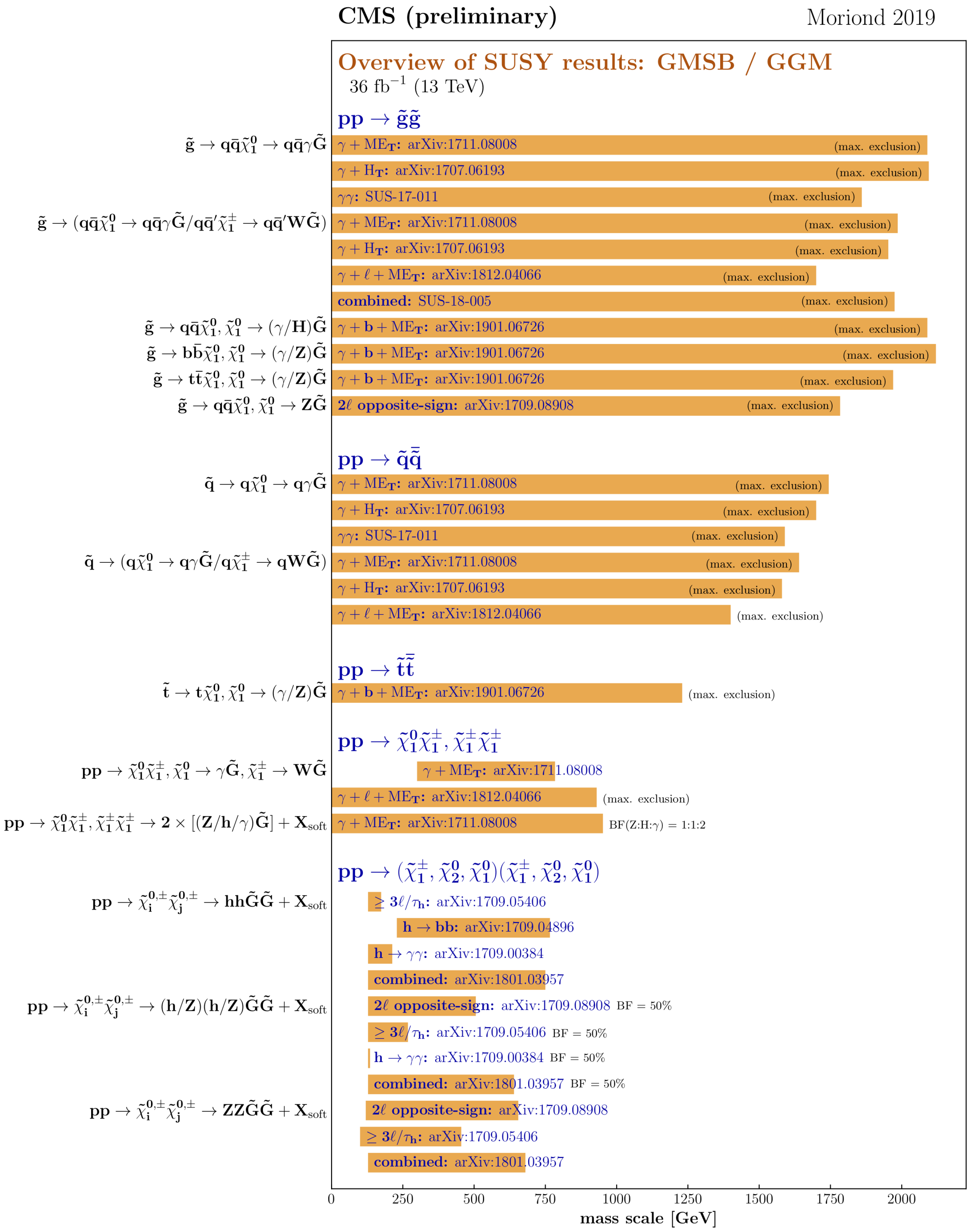
Figure 1.11: The bino and neutral wino NLSP branching fractions to Z or γ plus gravitino [4]. The branching fraction is determined by the weak mixing angle, and, at low mass, by the phase space suppression of decays to Z's.

figure 1.11 shows the branching fraction of a wino-like NLSP as a function of its mass. These scenarios can result in signatures with lepton $+\gamma + p_T^{\rm miss}$.

- a **higgsino like** NLSP, will decay preferably to a Higgs boson or a photon (γ) and a gravitino ($\widetilde{\rm G}$) and subdominately to a Z boson or a photon (γ) and a gravitino ($\widetilde{\rm G}$). Models with higgsino like NLSP may result in $b\bar{b}$, coming from H decay, and $\gamma + p_T^{\rm miss}$ final states.

1.5.4 General Gauge Mediation (GGM) simplified models

As mentioned in section 1.5.2, GMSB provides several advantages. However, even with the inherent simplicity of gauge mediation, there is a plethora of models with a wide variety of features. Thus, it is difficult to predict which exact set of parameters is realized in nature. In addition, many of these possible parameter sets can lead to identical experimental signatures. Such scenarios include several event topologies that requires the presences of one or more leptons, photons, jets and of course high $p_T^{\rm miss}$ coming from the undetected LSP. In order to cover as much phase space as possible, SUSY searches at the LHC are based on those signatures. On top of this, certain assumptions are made, leading to the so-called "Simplified Model Scenarios" (SMS) [47, 48]. The SMS can be well described by a small number of parameters directly related to collider physics observables: particle masses (and their decay widths, which can sometimes be neglected), production cross-sections, and branching fractions. In addition, constraints on a wide variety of models can be deduced from limits on simplified models [49]. As a result, the plethora of proposed models leads to wide experimental searches. Most SUSY searches at the LHC target a signal on the production of new, heavy particles that decay into SM particles and a stable undetected LSP which result in high $p_T^{\rm miss}$. However, large $p_T^{\rm miss}$ can be produced from SM processes such as leptonically decays of top quarks, weak gauge bosons and heavy flavor production. Moreover, $p_T^{\rm miss}$ can arise from instrumental effects. The studies to address those issues are discussed in detail later. In figure 1.12 the latest results for simplified models in the context of gauge-mediated supersymmetry breaking from CMS collaboration are summarized. On the x-axis the mass scale reach of each analysis is shown. The searches, based on final signatures, are categorized as all-hadronic, single lepton plus jets, opposite-sign and same-sign dileptons, multileptons, inclusive searches and searches with photons.



Selection of observed limits at 95% C.L. (theory uncertainties are not included). Probe **up to** the quoted mass limit for light LSPs unless stated otherwise. The quantities ΔM and x represent the absolute mass difference between the primary sparticle and the LSP, and the difference between the intermediate sparticle and the LSP relative to ΔM, respectively, unless indicated otherwise.

Figure 1.12: The Mass reach for simplified models in the context of gauge-mediated supersymmetry breaking [5].

Search for GGMSB in the Diphoton final state

A part of this thesis concerns a search of GGMSB that involves photons. More specifically, a bino-like neutralino ($\widetilde{\chi}^0$) is assumed that further decays to a gravitino ($\widetilde{G}$) and a photon (γ), resulting in events with two photons and significant missing energy. The Feynman diagrams of processes with this signature are shown in figure 6.1. These simplified model scenarios assumes:

- R-parity conservation
- In case of sparticles production (pair of gluinos), their production is completely defined by the QCD theory and parton distribution functions
- All sparticles, except gluino, $\widetilde{\chi}^0$ and $\widetilde{\mathrm{G}}$ are very heavy and not accessible at the LHC.
- Branching fractions of the decay of sparticles to various channels are simplistically decided. The models assume a 100% branching fraction for the gluinos and squarks to decay as shown in figure 6.1. The squarks can be either first or second generation. We assume a 100% branching fraction for the NLSP neutralino to decay to a nearly massless gravitino and a photon, $\widetilde{\chi}^0 \rightarrow \widetilde{\mathrm{G}}\gamma$.

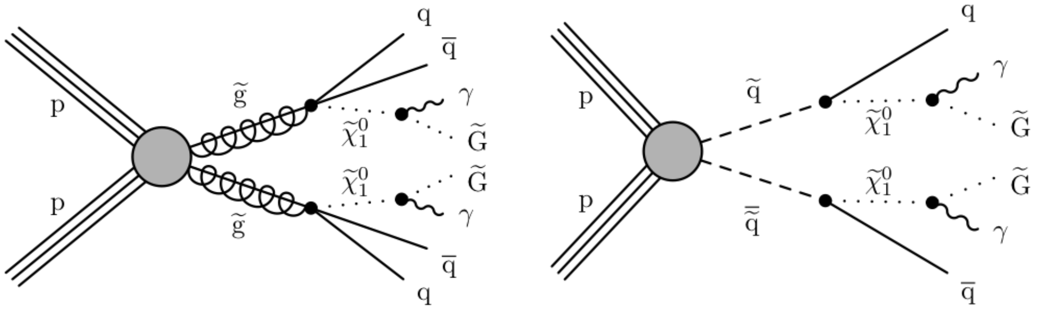


Figure 1.13: Diagrams showing the production of signal events in the collision of two protons. In gluino $\tilde{g}$ pair production (left), the gluino decays to an antiquark $\tilde{q}$, quark q, and a neutralino $\widetilde{\chi}^0$. In squark $\tilde{q}$ pair production (right), the squark decays to a quark and a neutralino $\widetilde{\chi}^0$. In both cases, the neutralino $\widetilde{\chi}^0$ subsequently decays to a photon γ and a gravitino $\widetilde{\mathrm{G}}$.

Chapter 2

Experimental Setup

2.1 The Large Hadron Collider (LHC)

The Large Hadron Collider (LHC) [50] is a two-ring hadron accelerator and collider of 26.7 km circumference tunnel at European Organization for Nuclear Research (CERN). It lies beneath the Franco-Swiss border near Geneva, Switzerland at a depth ranging from 45 to 170 m below the earth's surface. It is the largest and most powerful particle accelerator ever built and is the gem of the CERN accelerator complex, shown in figure 2.1. The LHC was build to answer questions about fundamental particle physics, at the TeV energy scale.

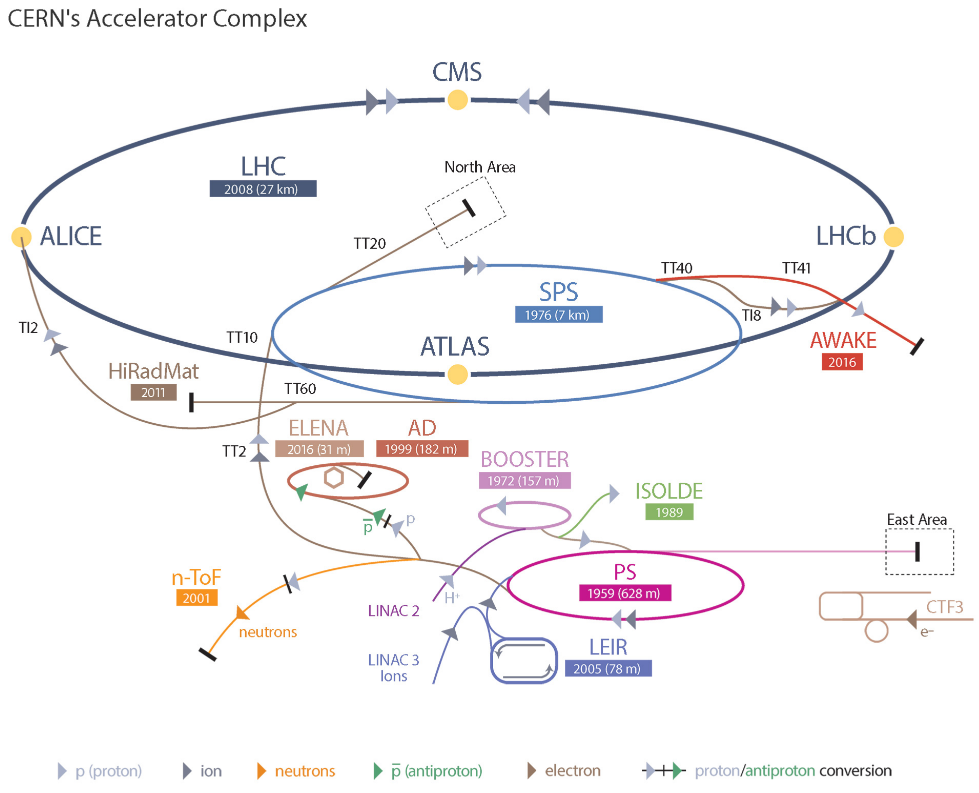


Figure 2.1: Schematic overview of the CERN accelerator complex. [6]

The LHC is supplied with protons acquired by stripping electrons from hydrogen atoms. The beams (protons or heavy ions) are accelerated in stages. First, the linear accelerator (LINAC2) generates 50 MeV protons and then the Proton Synchrotron Booster (PSB) increases their energy to 1.4 GeV. After that, the beam is fed to the Proton Synchrontron (PS) where it is accelerated to 25 GeV, before they reach the Super Proton Synchroton (SPS) where the proton energy reaches the 450 GeV. Protons leaving the SPS are eventually injected into the LHC main ring to be accelerated up to their maximum energy. The protons arrive in the LHC in bunches of approximately 10^{11} protons, with a bunch spacing of $25\ ns$, resulting in 2808 bunches in total. Proton beams orbit the LHC in two metal pipes with a very high vacuum, the beam pipes. Once the beams reach the desirable energy, the optics are changed to squeeze the beams at the interaction points, and the magnets separating the beams are squeezed off, resulting in collisions.

One of the key parameters of a collider is the instantaneous luminosity L. This is directly related to the observed rate of an interaction process with,

$$\frac{dN}{dt} = \sigma L \tag{2.1}$$

where, σ is the cross section of the process. The instantaneous luminosity L of a collider is given by:

$$L = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi\epsilon_\eta \beta^*} F \tag{2.2}$$

where N_b is the number of particles per bunch, n_b is the number of bunches per beam, f_{rev} is the revolution fequency, $\gamma_r = E/m$ is the relativistic gamma factor, ϵ_η is the normalized transverse beam emmitance, β^* is the β function at the collision point, and F is the geometric luminosity reduction factor due to the crossing angle at the interaction point. LHC is designed to reach an instantaneous luminosity of $L = 10^{34} cm^{-2} s^{-1}$, while in 2017 a twice the design value instantaneous luminosity was achieved. Figure 2.2 shows the integrated luminosity for the the run years of Run2. The total proton-proton cross section at a centre-of-mass energy of $\sqrt{s} = 13$ TeV is expected

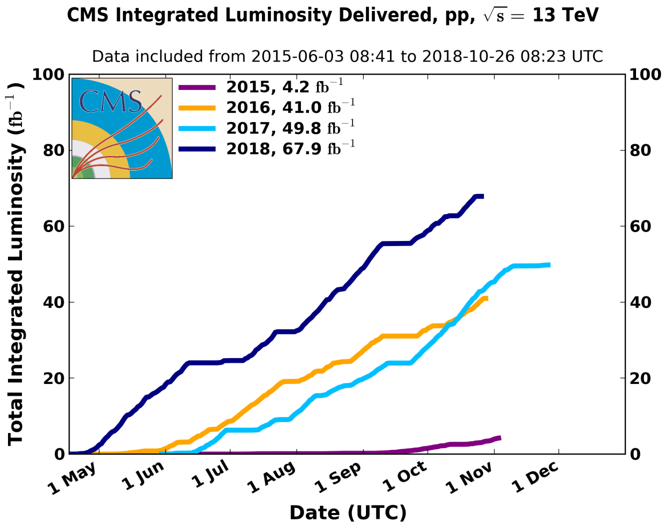


Figure 2.2: CMS integrated luminosity for Run2 [7].

to be approximately 70 mb, which means that around 20 p-p collisions will occur at each bunch crossing at the design luminosity.

2.2 The CMS experiment

The Compact Muon Solenoid (CMS) [51, 10] detector is a multi-purpose detector apparatus operating at the LHC. CMS is located at the LHC point 5, in the village of Cessy in France, about 100 meters underground. CMS, weights 14000 tonnes, which makes it the heaviest detector among the LHC detectors. Despite it's high weight, its dimensions 22x15 m make it relatively *compact*. One of the distinctive properties of the CMS experiment is the precise, high efficiency *muon* measurement, enabled by large gas chamber detectors. The detector is constructed inside and around a large superconducting *solenoid* magnet, that provides a strong magnetic field of 3.8 T in the inner part of the detector, bending the trajectories of charged particles for precise measurement of momentum and charge. An overview of the CMS detector is shown is figure 2.3. Going from the beam pipe to the solenoid, there is a tracker measuring the momenta of charged particles, an electromagnetic calorimeter to measure the energies of photons and electrons, and a hadronic calorimeter for measuring the energies of charged and neutral hadrons. Outside the solenoid, there are muon chambers measuring momenta of muons.

To describe the CMS detector, both right handed Cartesian coordinates and polar coordinates are used, with the nominal collision point as the origin in both cases. For the Cartesian coordinates, the x axis and y axis are in the transverse plane pointing along the inward radial direction of the LHC ring and along the upward vertical direction, respectively, while the z-axis is parallel to the beam. For the polar coordinates, ϕ represent the azimuthal angle from the x axis in the transverse plane, r the radial distance in the plane, and θ the polar angle from the z-axis in the y- z plane. Instead of θ another spacial coordinate is used called *pseudorapidity*, which is defined as

$$\eta = -\ln(\tan\frac{\theta}{2}) \tag{2.3}$$

For highly relativistic particles with $\mathrm{m} << \mathrm{p}$, η is equal to the *rapidity* defined as

$$\mathrm{y} = \frac{1}{2}\ln\frac{\mathrm{E}+\mathrm{p_Z}}{\mathrm{E}-\mathrm{p_z}} \tag{2.4}$$

In hadron collider physics, *pseudorapidity* is preferred over the polar angle θ, since particle production is constant as a function of rapidity. Moreover, differences in rapidity are Lorentz invariant under boosts along the longitudinal axis and hence, a measurement of a rapidity difference $\Delta\mathrm{y}$ between particles is not dependent on the longitudinal boost of the reference frame (such as the laboratory frame). This is an important feature for hadron collider physics, where the colliding partons carry different longitudinal momentum fractions x, which means that the rest frames of the parton-parton collisions will have different longitudinal boosts.

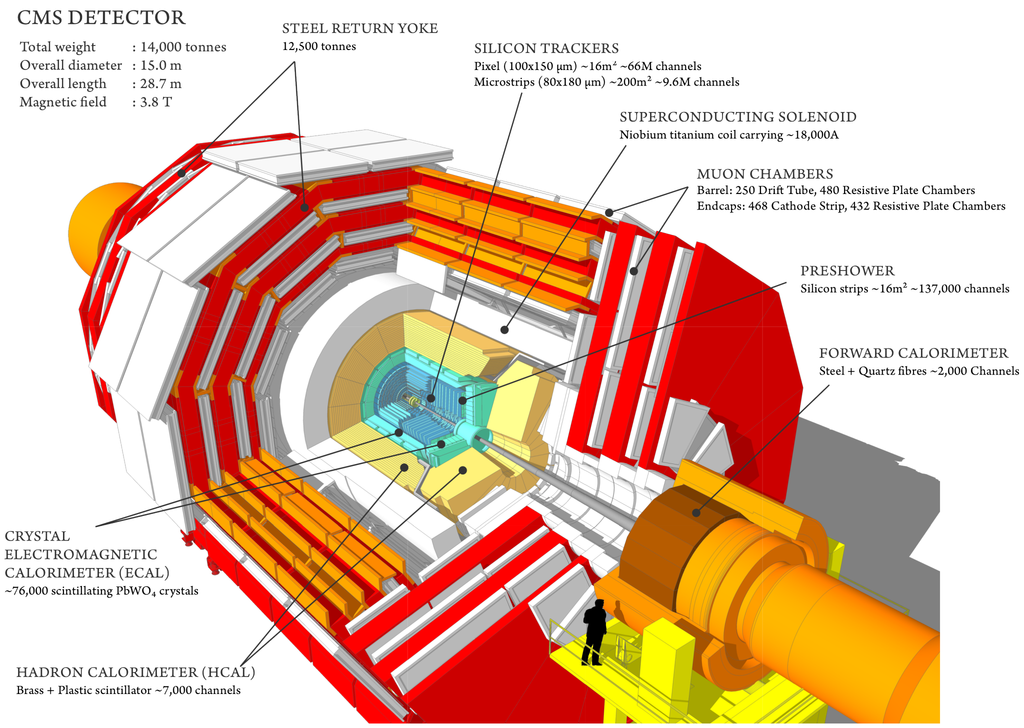

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 2.3: Overview of the Compact Muon Solenoid (CMS) detector. The general characteristics of each part of the detector are shown here along with relevant magnitudes and the x, y and z coordinate system [8].

2.3 Superconducting Solenoid

One of the most important components of the CMS detector is the superconducting solenoid that provides the bending power necessary to precisely measure the momentum of all charged particles produced in the collision. The magnet is located between the calorimeters and the muon system. It is 12.5 m long and has an inner diameter of 6m. The magnet produces an magnetic field of 3.8 T and has a stored energy of 2.6 GJ. A 12,000 ton steel yoke made up of 5 wheels in the barrel and 3 endcap disks serves to return the magnetic flux. The solenoid is suspended in a vacuum cryostat and cooled to 4.5 K with liquid helium.

2.4 The inner tracking system

The CMS inner tracking system consists the most inner part of the CMS detector, surrounding the interaction point. It is designed to provide precise and efficient measurements of the trajectories of charged particles (electrons, muons and hadrons) and precise reconstruction of primary and secondary interaction vertices. The structure of the tracker system is illustrated in figure 2.4 where different modules of the inner tracker are shown: pixel, inner barrel (*TIB*), outer barrel (*TOB*), inner disks (*TID*) and endcaps (*TEC*). The CMS tracker [52] is composed of a pixel detector with three barrel layers at radii 4.4 cm, 7.3 cm and 10.2 cm and a silicon strip tracker with 10 barrel detection layers extending outwards to a radius of 1.1 m. Each system is completed by endcaps which consist of two disks in the pixel detector and 3 plus (up to) 9 disks in

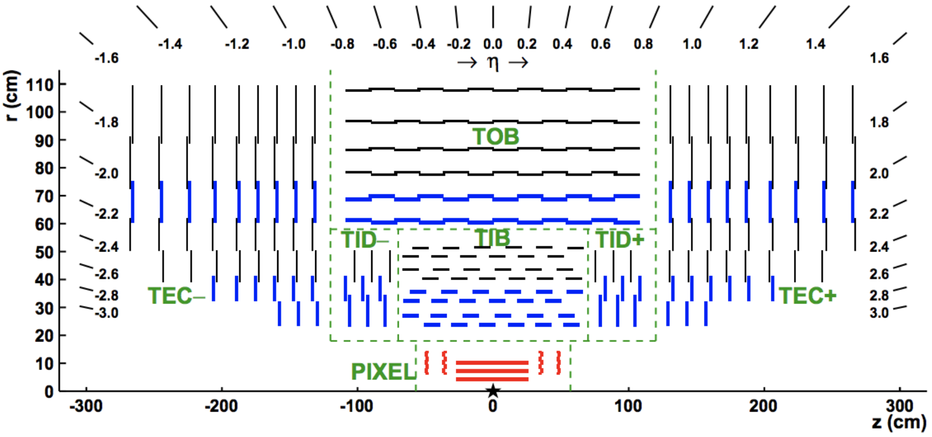


Figure 2.4: Schematic cross section through the CMS tracker in the $r - z$ plane [9]. The tracker is symmetric about the horizontal line r = 0, so only the top half is shown here.

the strip tracker on each side of the barrel, extending the acceptance of the tracker up to a pseudorapidity of $|\eta| < 2.5$.

2.5 Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) [53, 54] of the CMS detector is a hermetic homogeneous calorimeter. ECAL is a radiation tolerant and total absorption calorimeter made up of lead tungstate crystals (PbWO_4). It's purpose is to measure the energy of electromagnetic objects such as photons and electrons as well as the EM components of jets and hadrons that deposit their energy in the crystals of the ECAL. It has a large dynamic range coupled with excellent linearity up to 1 TeV.

An electromagnetic shower for a photon starts as an electron-positron pair production, while for an electron starts as Bremsstrahlung radiation. Both develop to a cascade of electrons, positrons and photons through repeating processes of pair productions and Bremsstrahlung. The CMS ECAL is designed to measure the photon and electron energies with high resolution, which is essential for analyses that include photons and electrons. For this purpose, it is made by materials ideal for scintillation. Lead tungstate is an inorganic scintillator. The passage of charged particles produce electron-hole pairs in the conduction and valence bands. In addition, (PbWO_4) has a Moliére radius which is defined as the radius of a cylinder containing 90% of the shower's energy depositions, that is only 2.2 cm. This results to an excellent position resolution and separation between showers. This property is of particular importance since it allows to distinguish photons from isolated neutral pions ($\pi^0 \to \gamma\gamma$ decays). It is also able to provide triggering information and aids particle identification. ECAL is divided into the barrel covering a pseudorapidity range of $|\eta| < 1.479$ and two endcaps at $1.479 < |\eta| < 3.0$. The endcap part includes a preshower detector covering the region $1.65 < |\eta| < 2.6$. Figure 2.5 shows the ECAL geometry with different components.

The preshower, is installed in front of each ECAL endcap. It is made of two layers of lead radiation followed by silicon strip sensors. Its principal aim is to identify neutral pions in the endcaps within a fiducial region $1.653 < |\eta| < 2.6$. Furthermore, it helps

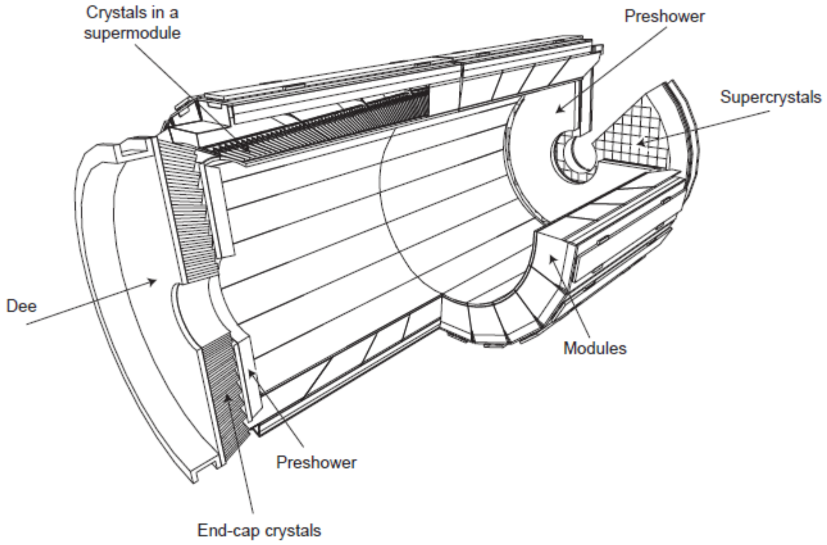


Figure 2.5: CMS ECAL geometry schema showing different components[10].

the identification of electrons against minimum ionizing particles and improves the position determination of electrons and photons with high granularity. The intrinsic energy resolution of the ECAL barrel is measured to be:

$$\left(\frac{\sigma}{\mathrm{E}}\right)^2 = \left(\frac{\mathrm{S}}{\sqrt{\mathrm{E}}}\right)^2 + \left(\frac{\mathrm{N}}{\mathrm{E}}\right)^2 + \mathrm{C}^2 \qquad (2.5)$$

where S is the stochastic term, N is the noise term and C is a constant term. The stochastic term describes the event-to-event fluctuations in the lateral shower containment, the photostatistics contribution (2.1%) and the fluctuations in the energy deposit in the preshower absorbed with respect to what is measured in the preshower silicon detector. The small stochastic term ensures that the photon energy resolution is excellent in the typical range of photons in jets (1-50 GeV). The noise term includes contributions of electronics, digitization and pileup noise, while contributions to the constant term comes from the non-uniformity of the longitudinal light collection, intercalibration errors and leakage of energy from the back of the crystal. Using test beam data, a typical energy resolution was found to be:

$$\left(\frac{\sigma}{\mathrm{E}}\right)^2 = \left(\frac{2.8\%}{\sqrt{\mathrm{E}}}\right)^2 + \left(\frac{12\%}{\mathrm{E}}\right)^2 + 0.3\% \qquad (2.6)$$

ECAL reconstruction

A crucial part of the SUSY Diphoton analysis is the successfully reconstruction of photons, thus the ECAL is the most important sub-detector part. In this section the reconstruction based on a clustering algorithms is described. The energy deposited in the ECAL crystals is generally spread out over a few neighbouring crystals, so the total energy that is measured involves several crystals. CMS developed a specific clustering algorithm able to measure the energy and direction of stable neutral particles such as photons and neutral hadrons and also to separate them from charged hadron deposits. Another purpose of the cluster algorithm is to efficiently identify and reconstruct electrons and accompanying bremsttrahlung photons. Lastly, the clustering algorithm should be able to supplement the energy measurement of charged hadrons which cannot be accurately measured by the tracker. The reconstruction algorithm consists of three steps. First, cluster seeds are identified as local caloremeter cell (ECAL crystall) energy maxima above a given energy. Second, topological clusters are grown from the seeds by aggregating crystals with at least one side in common with a cell already in the cluster, and with an energy in excess of a given threshold. These thresholds represent about two standard deviations of the electronic noise in the ECAL (i.e. 80 MeV in the barrel and up to 300 MeV in the endcaps). In the third step, the so called particle flow (PF) Clusters are dynamically merged into superclusters. Dynamic superclustersing allows good energy containment, robustness against pileup and automatically takes into account the detector geometrical variations with η.

2.6 Hadron Calorimeter

The hadron calorimeter (HCAL) is radially restricted between the outer extend of the ECAL and the inner extent of the magnet coil. The hadron calorimeters are very important for the measurement of the hadron jets and neutrinos or other exotic particles resulting in apparent missing transverse energy. In addition, HCAL can provide supplementary measurements to the ECAL and the muon systems. The primary hadrons that have sufficiently long lifetime to traverse the CMS calorimeter are pions, kaons, protons and neutrons. These hardons traverse the ECAL quiet transparently, but form complex showers in the brass absorber. As mentioned previously, the electromagnetic showers consists of a cascade of photons conversions $\gamma \rightarrow e^+e^-$ and Bremsstrahlung radiation $e \rightarrow e + \gamma$. The hadronic showers on the other hand, proceeds through an increasing number of primary strong interactions with many particle types including electromagnetic components via neutral pion decays $\pi_0 \rightarrow \gamma\gamma$. The fraction of the hadronic shower energy transferred to an electromagnetic cascade depends on the shower energy, with the fraction about 50% for a 100 GeV shower and 70% for a 1 TeV shower. The energy resolution of the HCAL has been measure to be:

$$\frac{\sigma}{E} = \frac{110\%}{\sqrt{E}} \oplus 9\% \tag{2.7}$$

where E is expressed in GeV. The leading contribution to the HCAL energy resolution comes by effects from not fully containing the hadronic shower, while a stochastic noise term S of 110% and a constant term of 9%. Figure 2.6 shows a view of the HCAL in the $y - z$ plane. The HCAL is composed of a barrel (*HB*) and endcap (*HE*) component, which are both contained inside the solenoidal magnet, and the outer (HO) and forward (HF), which are located outside the solenoid. The (*HB*) covers the range $|\eta| < 1.3$ while the HE covers $1.3 < |\eta| < 3.0$. Due to the limited space available for the HCAL within the solenoid, the HO is included outside the solenoid in order to increase the total interaction length.

The HCAL is the key subdetector to precisely measure the energy of particles constituents of jets and in particular the energy of b jets that are crucial for the second part of the thesis. By combining the HCAL response with information from the individual parts of the detector as detailed described in section 3.1, the jet energy resolution achieved is about ten percent.

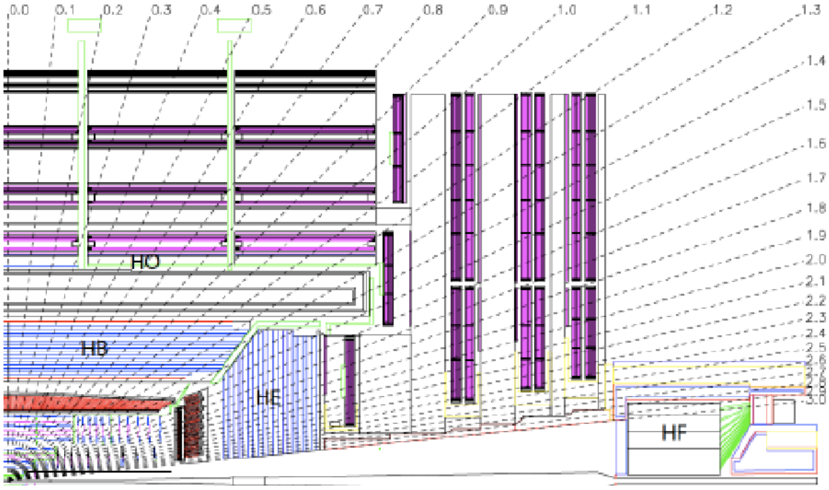


Figure 2.6: View of the CMS hadronic calorimeter in the y-z plane.

2.7 Muon detectors

Muon detection is a powerful tool for recognizing signatures of interesting processes over a high background rate that is expected in the high luminosity environment at the LHC. Therefore, precise and robust muon measurements are critical. The muon system [55] is the outermost component of the CMS detector, consisting of a barrel ($|\eta| < 1.2$) and two endcaps ($0.9 < |\eta| < 2.4$) which are illustrated in figure 2.7. It is designed to perform three task; muon identification, momentum measurement and triggering. The reconstruction of the momentum and the charged muon is achieved over the entire kinematic range and consists of three types of gaseous detectors; drift tubes (*DT*), the cathodes strip chambers (*CSC*) and the resistive plate chambers (*RPC*). The first two gaseous detectors can also provide independatly trigger information based on the p_T of muons with high efficiency and background rejection. The RPCs are installed both in the barrel and in the endcaps covering a region of $|\eta| < 1.6$ and providing a fast, independent, and highly-segmented trigger with a sharp p_T threshold. They also help to resolve ambiguities in attempting to make tracks from multiple hits in a chamber.

2.8 CMS trigger system

The LHC provides proton-proton collisions every 25 ns, corresponding to a crossing frequency of $40\ \mathrm{MHz}$. At the designed centre-of-mass energy and luminosity ($10^{34}\mathrm{cm}^2\mathrm{s}^{-1}$) of the LHC, with a proton-proton cross section of $\approx 70\mathrm{mb}$ around 700 million proton-proton collisions are expected per second. This corresponds to an event rate of $\approx 700\mathrm{MHz}$. For an accurate measurement of the event, all the information of the subdetector systems is necessary. Collecting this information in a small given time is feasibly impossible. In addition, only a small fraction of events are interesting for physics analyses. Therefore, it is crucial to drastically reduce the event rate by selecting events that are more likely to be important. This task is performed by a two-tiered trigger system [56], which is the start of the physics event selection process. The first level (L1) is composed of custom hardware and uses partial, fast response data from the calorimetres and muon systems. It is able to identify and select events that contain objects like muons, electrons, photons or jets at a rate up to 100kHz within $4\mu s$. The L1 trigger has

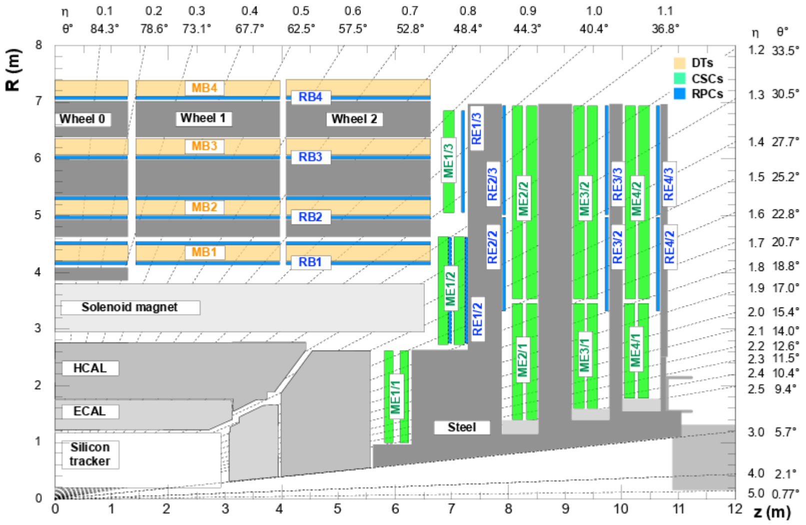


Figure 2.7: Cross-sectional view of a quadrant of the CMS detector in the (r, z) plane, showing the layout of the muon detector [11].

local, regional and global components. The local triggers; Trigger Primitive Generator (*TPG*), identify energy deposits in calorimeter trigger towers and track segments or hit patterns in muon chambers. Regional triggers determine ranked and sorted trigger objects such as an electron or muon candidates, by combining their information and using pattern logic. The Global Calorimeter (*GCT*) and Global Muon Triggers (*GMT*) determine the highest-rank calorimeter and muon objects across the entire experiment and transfer them to the global trigger (*GT*) which decides the rejection or the acceptance of an event for further evaluation by the High Level Trigger (HLT). A schematic view of the L1 trigger is illustrated in figure 2.8.

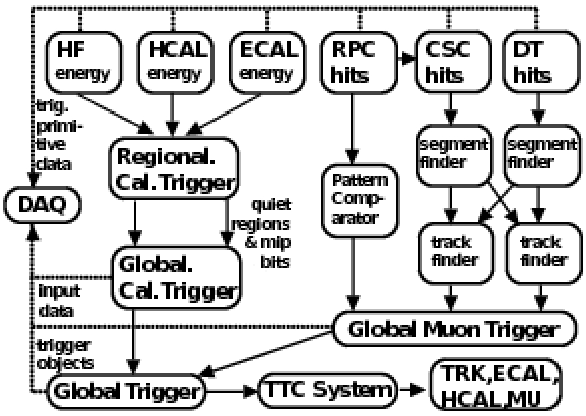


Figure 2.8: The architecture of the CMS Level-1 trigger system in Run 2 [12]

.

The second level also known as high-level trigger (*HLT*) runs a version of the full event reconstruction software optimized for fast processing on a farm of commercial processors, and reduces an event rate to around 1kHz. HLT has access to the complete read-out data and can therefore perform complex calculations similar to those made by the off-line analysis software. It is also divided into internal "steps", named L-2, L-2.5, L-3, where L-2 uses inputs from the calorimeter and muon detectors, L-2.5 is referred to algorithms that use partial tracker information such as pixel hits and L-3 refers to a selection that includes the reconstruction of full tracks in the tracker. The HLT is able to make more sophisticated trigger decisions than the L1. That can include information from sophisticated b-tagging discriminators and reconstructed jets. After the events pass the HLT trigger, they are stored to disk for offline processing that can take several seconds per event and performs a much more computational intensive event reconstruction which is then used for data analysis.

2.9 Upgrades

The increasing luminosity through the years arise the need of system and subdetector upgrades. The first CMS *PhaseI upgrade* program started at the end of Run1 and it is planned to continue until the start of LHC Run3 in 2021. It included a complete upgrade of the Level-1 trigger system, an upgrade of the photon sensors of the HCAL and HF scintillation system, as well as, an installation of a new pixel detector that replaced the old in early 2017. The new pixel detector [57] contains four barrel layers and three discs in the endcaps, with double pixels. It is expected to be be more robust against high pile-up and to have a faster readout system with an increased number of channels. After the end of Run3, the CMS detector will go under the *HL-LHC* or *PhaseII upgrade* which will take place in 2024-2026 [58]. This upgrade will prepare the detector to address the challenges of the high-luminosity LHC conditions. The proposed operating scenario is to level the instantaneous luminosity at $5 \times 10^{34}\mathrm{cm}^2\mathrm{s}^{-1}$ from potential peak value $2 \times 10^{34}\mathrm{cm}^2\mathrm{s}^{-1}$ at the beginning of the fills, and to deliver $250\ \mathrm{fb}^{-1}$ per year for a further 10 year of operations. Under these conditions the event pile up will rise substantially and thus, the *PhaseII upgrade* is crucial for the smooth operation of the detector.

Chapter 3

Object Reconstruction

3.1 Particle flow algorithm

As mentioned in the previous chapter, the CMS detector is a general purpose detector that consists of many detector layers nested around the beam axis. A schematic view of all the detector's layers is illustrated in figure 3.1. Each detector layer is able to provide individual information about several physics objects such as photons, electrons, muons and jets. However, a significantly improvement of event description can be achieved by correlating the basic elements from all the detector layers to identify each final-state particle, and by combining the corresponding measurements to reconstruct the particle properties on the basis of this identification. This holistic approach is called *particle-flow (PF)* reconstruction [59]. The design of the CMS detector is ideal for a particle flow event reconstruction. This is due to the high tracking efficiency while keeping a low fake rate, the full detector coverage provided by the ECAL and finally the ability to reconstruct high purity muons. In this chapter the basic reconstruction steps are discussed.

3.2 Track and vertex reconstruction

The track reconstruction [60] from charged particles is performed iteratively with the combinatorial track finder (CFT) algorithm. CFT is based on Kalman filtering and it is used to build tracks from hits in the pixel and strip layers of the inner tracker. Several iterations are performed. In each iteration, initial seeds are first generated with only a few hits compatible with a charged-particle trajectory. After that, a KF-based [61] track process is used to collect hits form all tracker layers that are compatible with the extrapolated charged-particle trajectory. Then a fit to all hits is performed to determine the full track parameters and their uncertainties. Finally, to reject low-quality tracks, several selection criteria are applied. This iterative tracking approach benefits from high track finding efficiency while keeping a low misrecontruction rate. The Compact Muon Solenoid (CMS) tracker is expected to be traversed by about 1000 charged particles at each bunch crossing, produced by an average of more than twenty proton–proton (pp) interactions. One of the most important tasks is the correct identification of the primary vertex coming directly from p-p collisions. The collision vertices are reconstructed from particle tracks by extrapolating them from the tracker region towards the interaction point. The most probable vertex is estimated by a deterministic annealing algorithm.

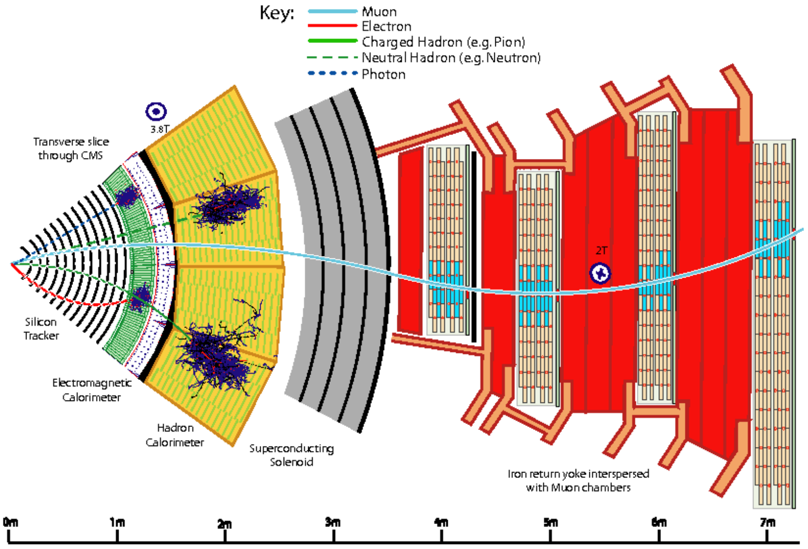

Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
3.8T
Transverse slice through CMS
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
2T
Iron return yoke interspersed with Muon chambers
0m
1m
2m
3m
4m
5m
6m
7m

Figure 3.1: A sketch of the specific particle interactions in a transverse slice of the CMS detector, from the beam interaction region to the muon detector. All the detector layers are shown.

For high-p_T objects, the primary vertex is selected using the so called *track jets*. First, the track jets are selected by applying the anti-k_T jet clustering algorithm [62] to the tracks associated with each vertex. After that, the position of each vertex is fitted using an adaptive vertex filter. Second, the missing transverse momentum is constructed as the negative vector sum of $\vec{p_T}$ of the track jets. Finally, the vertex with the highest summed of p_T^2 of all physics objects is selected as the primary vertex. The physics objects used in this calculation are produced by a jet-finding algorithm applied to all charged-particle tracks associated to the vertex, plus the corresponding p_T^{miss} computed from those jets.

Muon Tracks

Muon tracks can be reconstructed using information from both the inner tracker and the outer muon detector. Depending on the different roles of each subdetector system, different types of muons can be reconstructed. For instance, when the reconstruction comes only from the muon detector, then we have the *standalone muon*. If the track parameters of a standalone muon are compatible with a track of the inner tracker, then both tracks are used to form a *global muon* track. If at least one muon segment matches with the extrapolated track from inner tracker, then the corresponding muon is called *tracker muon*. The global muons have a better performance compared to standalone and tracker muons since they combine information from both detectors. The reconstructed muons are fed into the PF algorithm where several quality criteria are applied on the track fit χ^2, number of hits/tracks in muon chambers and also the compatibility with primary vertex.

Electrons Tracks

Electron tracks are reconstructed by combining information from the inner tracker and the ECAL. For the successfully reconstruction of the electron track two independent approaches can be used [63]. The first, also known as, ECAL-based approach, is based on the ECAL measurements and uses energetic ECAL clusters with $\mathrm{E_T} > 4\ \mathrm{GeV}$. The cluster energy and position are used to infer the position of the hits expected in the innermost tracker layers under the assumptions that the cluster is produced either by an electron or by a positron. Bremsstarhlung photons, due to the interaction between the electron and the tracker material must be taken into account. The energy of the electron and of possible bremsstrahlung photons is collected by grouping the ECAL clusters that are reconstructed in a small window in η and an extended window in ϕ around the electron direction into a *supercluster*. This is motivated by the fact that, bremstrahhlung photons are not affected by the magnetic field and thus, their energy is more spread in the ϕ direction. To reconstruct the electrons missed by the ECAL-based approach, a tracker-based electron seeding method was developed in the context of PF reconstruction. This approach has better performance for identifying low p_T electrons or electrons inside jets that are less isolated. Tracks from the iterative tracking are used as potential seeds for electrons, if their p_T exceeds $2\ \mathrm{GeV}$. The large probability for electrons to radiate in the tracker material can be exploited to disentangle electron tracks from charged hadron tracks. The radiation will result in increasing energy loss and thus, the track will contain fewer hits or have a higher fit χ^2. A preselection based on the number of hits and the fit χ^2 is therefore applied, and the selected tracks are fit again with a Gaussian-sum filter (*GSF*) [64]. This is more suitable than the KF for electrons as it allows for sudden and substantial energy losses along the trajectory. The final selection for the tracker-based electrons is based on a boosted-decision-tree (*BDT*) classifier that combines several variables such as the number of hits and the information obtained from the GSF and KF track fit. The electron seeds obtained with both the ECAL-based and the tracker-based methods are merged and fed to the PF algorithm.

3.3 Calorimeter deposit clustering

The clustering algorithm is the same for each component of the calorimeter system: ECAL barrel and endcaps, HCAL barrel and endcaps, and the two preshower layers and it is performed separately. The basic reconstruction principles of the ECAL are described in section 2.5. In this section a more detailed overview is given. The electromagnetic showers in the ECAL and the hadron showers in the HCAL are wider than a single ECAL crystal or HCAL module. Therefore, clustering of the energy deposits in these crystals and modules, commonly referred to as *calorimeter-cells*, is needed to determine the energies of the particles that initiate the showers. The clustering is seeded by cells where deposited energy exceeds a certain threshold. The neighboring cells with an energy above the twice the noise level are associated with the seeds forming topological clusters. An expectation-maximization (EM) algorithm based on a Gaussian-mixture model is then used to disentangle the clusters within a topological cluster. The energy deposition of a topological cluster is modeled as a sum of N Gaussian energy deposits, where N corresponds to the number of seeds. Dedicated methods are developed to calibrate the calorimeter clusters, as an accurate measurement of their energy is crucial for a consistent global event description. These calibrations are complemented by a residual correction to the cluster energies, derived using GEANT4 simulation with the full CMS detector. The correction accounts for the effect of clustering thresholds (up to 20% corrections to cluster energies) and the shadowing of the ECAL endcap crystals by the preshower (corrections up to 40%). After these corrections, the calibrated energies and the simulated energies of photons typically agree at the percent level. The corrections are validated by studying data events with two energetic photons originating from π^0 decays. This is done by reconstructing the invariant mass of the diphoton system and comparing it to the nominal mass of π^0. The ECAL supercluster energies are corrected and calibrated using an MVA regression algorithm trained with simulated events. The small residual differences between data and simulation are corrected by comparing the invariant mass peaks reconstructed in $\mathrm{Z} \rightarrow \mathrm{e}^+\mathrm{e}^-$ events.

3.4 Particle Identification and Reconstruction

3.4.1 The linking algorithm

One of the main challenges in CMS, is to link together the multiple PF elements that are produced in different subdetectors in order to extract a global event description. These elements need to be linked together to fully reconstruct the particle, while avoiding double-counting from different subdetectors. The fundamental core of the PF reconstruction is the link algorithm which provides the connection between different PF elements. The links between the different PF elements are established by extrapolating the trajectories reconstructed in the tracker while taking into account the effect of the magnetic field. The elements are matched geometrically in the (η, ϕ) plane which provides a metric to quantify the quality of the link. As bremsstrahlung can cause sudden changes in the curvature of electron trajectories, the standard CTF method can lead to lost hits and poor fit quality. In order to collect the energy of Bremsstrahlung photons emitted by electrons, tangents to the tracks are extrapolated to the ECAL from the intersection points between the track and each of the tracker layers. ECAL and HCAL clusters, as well as, preshower clusters are also linked, by checking if the cluster position in the more granular calorimeter is within the boundary of a cluster in the other calorimeter. Finally the segments reconstructed in the muon detector are linked with the inner tracker. The output of the link algorithm is a collection of PF blocks, with each block consisting of elements associated either by a direct link or by an indirect link through common elements. The particle flow blocks are in the form of charge tracks, calorimeter clusters and muon tracks. The algorithms treats it's block separately. First, muons are identified if the muon's combined momentum is compatible with the determined one from the tracker. Then the muon is called *particle-flow-muon* and the corresponding track is removed. Each of the remaining tracks in the block gives rise to a *particle -flow-neutral-hadron*, the momentum and energy of which are taken directly from the track momentum. If the calibrated calorimetric energy is compatible with the track momentum, the charged-hadron momenta are redefined by a fit of the measurements in the tracker and the calorimeters. The *particle-flow-electrons* are identified and reconstructed if the calibrated calorimetric energy is compatible with the track momentum. On the contrary, when the calibrated energy of the closest ECAL and HCAL clusters linked to the track(s) is significantly larger than the total associated charged-particle momentum or, the ECAL and HCAL clusters are not linked to any track at all, then *particle-flow-photons*, and possibly *particle-flow-neutral-hadron* are reconstructed. In the final step, the algorithm is revisited in a post-processing step that aims to rectify misidentified objects such as misreconstructed high-p_T muons or charged hadrons.

3.4.2 Muons

Particle flow muons are reconstructed from the inner tracker and the outer muon detector. Additional identification criteria are applied to standalone, global and tracker muon candidates. Those are based on the various quality parameters referring to muon isolation and reconstruction and are used to suppress misidentified muons, mostly charged hardons, while preserving a high efficiency for both isolated muons and muons inside jets. If the muon p_T of the inner tracker is less than 200 GeV, then this momentum is chosen. Otherwise, the momentum is determined by the fit with the smallest χ^2 of the following: tracker muon only; tracker and first muon detector plane; global muon; and global excluding the muon detector planes featuring a high occupancy.

3.4.3 Electrons and Photons

Electrons and photons interact in a similar way with the ECAL. An electron candidate is seeded from a GSF track, provided that the corresponding ECAL cluster is not linked to 3 or more additional tracks. A photon candidate, on the other hand, is seeded from a ECAL energy cluster with transverse energy, $\mathrm{E_T} > 10$ GeV when there is no corresponding link to GSF track. Within the tracker acceptance ($|\eta| < 2.5$), all ECAL clusters not linked to a track are considered photons, while all HCAL clusters without a linked track are considered neutral hadrons.

Photon Reconstruction

Photons showers deposit their energy in several crystals in the ECAL. Approximately 94% of the incident energy of photon is contained in 3×3 crystals and 97% in 5×5 crystals. The presence of material in front of the calorimeter results in bremstahlung and photon conversions. Due to the strong magentic field, the energy that reaches the calorimeter is spread in ϕ. The energy is therefore clustered at the electromagnetic calorimeter level by building a cluster of clusters (Super Cluster or SC), which is extended in ϕ. The topological variable $\mathrm{R_9} = \mathrm{E_{3\times3}}/\mathrm{E_{SC}}$ defined as the energy sum of 3×3 crystals centered on the most energetic, divided by the energy of the Super Cluster, is used to discriminate between unconverted ($\mathrm{R_9} > 0.94$) and converted ($\mathrm{R_9} < 0.94$) photons.

A loose preselection is applied to all prompt and non-prompt photons within the fiducial region of the ECAL. In order to avoid misidentifying an electron to a photon a conversion-safe veto is applied. In addition, to make sure the photon identification is performed in a region where simulation can properly describe the data, a selection is applied to keep the common phase space between data passing the trigger and the MC where no trigger requirement is applied. The variables used for preselection are built to be similar to ones used in the trigger and in the electromagnetic filter applied to simulated QCD background at generation level. This specific filter requires the presence in the event of at least two particles that can produce an energy deposit in the ECAL sufficient to mimic a photon "*fake photons*". The variables that are used for the preselection cuts are listed below. Those cuts results to an efficiency of 80% of successfully identifying a photon.

H/E: The ratio of the energy deposited in the HCAL tower directly behind the ECAL supercluster to the supercluster energy is required to be less than 0.0396.

$\boldsymbol{\sigma_{i\eta i\eta}}$: The energy weighted (single crystal energy over the Super Cluster energy) standard deviation of single crystal η within the 5×5 crystals centered at the crystal with maximum energy.

Particle Flow Charged Isolation : The p_T sum of all PF charged hadrons within a hollow cone of $0.02 < \Delta\mathrm{R} < 0.3$ around the supercluster is required to be less than 0.441 GeV, where $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

Particle Flow Photon Isolation : The p_T sum of all photons within a cone of $\Delta\mathrm{R} < 0.3$, excluding a strip in η of 0.015 around the supercluster, is required to be less than $2.5718 + (0.0047 \times \mathrm{p}_\mathrm{T}^{\gamma})$.

Particle Flow Neutral Isolation : The p_T sum of all neutral hadrons within a cone of $\Delta\mathrm{R} < 0.3$ around the supercluster is required to be less than
$2.725 + (0.0148 \times \mathrm{p}_\mathrm{T}^{\gamma}) + (0.000017 \times (\mathrm{p}_\mathrm{T}^{\gamma})^2)$.

A series of corrections are applied to the supercluster energy to account for variations in the shower containment and any potential losses for photons that convert in the tracker. The shower containment can vary for several reasons, including variations in the longitudinal depth at which the shower passes through the sides of the crystal and is lost in the inter-crystal dead space. The corrections are derived in simulation as a function of η, E_T, R_9, and the spread of the cluster in ϕ. After the corrections, the photon energy resolution is better than 3% for photons in the barrel [65].

Photon Identification

The photon identification is performed using a set of cuts on six discriminating variables. To maximize the sensitivity, cut values are optimized separately in four categories with significantly different levels of background and mass resolution. The following variables are used to distinguish isolated photons originating from the primary interaction from the background that comes from low multiplicity jets with high electromagnetic content:

- Relative combined isolation using selected event vertex. To compute this variable first an isolation sum is constructed as:

$$\sum \mathrm{Iso} = \mathrm{Iso}^{\mathrm{track}} + \mathrm{Iso}^{\mathrm{ECAL}} + \mathrm{Iso}^{\mathrm{HCAL}} \tag{3.1}$$

  where:

  - $\mathrm{Iso}^{\mathrm{track}}$ is the scalar sum of the transverse momenta of tracks which are consistent with originating from the primary vertex
  - $\mathrm{Iso}^{\mathrm{ECAL}}$ is computed as the transverse energy sum of ECAL energy deposits in crystals located within a cone of size $\Delta\mathrm{R} < 0.3$.
  - $\mathrm{Iso}^{\mathrm{HCAL}}$ is the sum of the energies of HCAL towers whose centers lie within a ring-shaped region of outer radius $\Delta\mathrm{R} = 0.4$ and inner radius $\Delta\mathrm{R} = 0.15$, centered on the ECAL Supercluster position.

  Each isolation sum contains a significant contribution from pile-up and underlying events. To maintain high efficiency under high pile up conditions, the contribution to $\sum Iso$ from pile-up and the underlying event is estimated on an event by event basis as the product of the measured energy density ρ for the event determined using the FastJet algorithm, and an effective area $\mathrm{A_{eff}}$ corresponding to the isolation cone excluding veto regions. $\mathrm{A_{eff}}$ is determined empirically as the ratio of the slopes of linear fits to the mean value of Iso vs nPV and to the value of ρ vs nPV in Z events. The pile-up corrected isolation sum is equal to:

$$\sum \mathrm{Iso}^{\mathrm{PUcorr}} = \sum \mathrm{Iso} - \rho \mathrm{A_{eff}} \tag{3.2}$$

The isolation sum is then scaled by $\mathrm{p_T^{pho}}/50\ \mathrm{GeV}$, where $\mathrm{p_T^{pho}}$ is the transverse energy of the photon determined using the selected primary vertex. The relative isolation is thus given by:

$$\mathrm{Iso}^{\mathrm{rel}} = \frac{\sum \mathrm{Iso}^{\mathrm{Pucorr}}}{\mathrm{p_T^{pho}}/50\ \mathrm{GeV}} \tag{3.3}$$

- Relative combined isolation using event vertex giving highest $\mathrm{Iso}^{\mathrm{track}}$. The relative combined isolation using event vertex is computed similarly with the difference that $\mathrm{Iso}^{\mathrm{track}}$ is computed for each reconstructed primary vertex and the largest value is used. This definition of isolation accounts for the case where a wrong primary vertex is selected and thus, it is complementary to the previous definition.

- Relative track isolation using selected event vertex. Since track isolation is the most discriminating of the three sub-detector isolation variables, a cut is additionally applied on relative isolation defined using track isolation only:

$$\mathrm{Iso}^{\mathrm{rel,track}} = \frac{\sum \mathrm{Iso}^{\mathrm{track}}}{\mathrm{p_T^{pho}}/50\ \mathrm{GeV}} \tag{3.4}$$

  where $\mathrm{Iso}^{\mathrm{track}}$ is the one computed using the selected primary vertex. No pileup subtraction is required since only tracks consistent with the selected primary vertex are included in the sum.

- H/E. The ratio of hadronic energy to electromagnetic energy is calculated as the ratio of the sum of HCAL tower energies within a cone of size $\Delta R < 0.15$ centered on the ECAL Super Cluster position, to the energy of the Super Cluster.

- $\sigma_{\eta\eta}$. The transverse shape of the electromagnetic cluster is computed with logarithmic weights and is defined as:

$$\sigma_{\eta\eta}^2 = \frac{\sum_{\mathrm{i}}^{5\times5} \mathrm{w_i}(\eta_{\mathrm{i}} - \bar{\eta}_{5\times5})^2}{\sum_{\mathrm{i}}^{5\times5} \mathrm{w_i}}; \quad \mathrm{w_i} = \max \quad (0, 4.7 + \ln \frac{\mathrm{E_i}}{\mathrm{E_{5\times5}}}) \tag{3.5}$$

  where $\mathrm{E^i}$ and η^{i} are the energy and pseudorapidity of the $\mathrm{i^{th}}$ crystal within the 5×5 electromagnetic cluster and $E_{5\times5}$ and $\bar{\eta}_{5\times5}$ are the energy and η of the entire 5×5 cluster. The value of $\sigma_{\eta\eta}^2$ tends to be smaller for single isolated photons (including converted photons, since the cluster is spread in the ϕ direction only), than for the background which is dominated by jets consisting of $\pi^0 s$ decaying to two photons.

- Minimum threshold on $\mathrm{R_9}$. A minimum threshold on $\mathrm{R_9}$ is applied to photons in the ECAL endcaps in order to exclude very poorly reconstructed photons.

The set of cuts in the above variables that corresponds to the *Photon ID* are chosen in order to have a high signal efficiency and a low fake rate.

3.4.4 Jets

Jets are the experimental signatures of quarks and gluons produced in high-energy processes such as head-on proton-proton collisions. They arise from the hadronization of quarks and consists of many closely spaced particles which are detected as charged hadrons ($\pi^{\pm}$, $K^{\pm}$, or protons), neutral hadrons (K_L^0, or neutrons), non isolated photons from π^0 decays and less often additional muons or electrons from decays of charged hadrons. In CMS, jets are typically reconstructed by clustering the PF candidates with the anti-k_T [62] algorithm implemented in the FastJet package [66]. The anti-k_T algorithm works as follows: The distance between a PF candidate i is obtained for each surrounding particle j in terms of *rapidity* y and angle ϕ as $\Delta R_{ij}^2 \equiv (y_i - y_j)^2 + (\phi_i - \phi_j)^2$. R is a distance parameter related to the radius of the jet. Then, the variable d_{ij} which describes the distance parameters for each particles is determined as

$$d_{ij} = \min(p_{T,i}^{-2}, p_{T,j}^{-2}) \frac{\Delta R_{ij}^2}{R^2} \tag{3.6}$$

For each particle i, the beam distance d_{iB} is defined as $d_{iB} = p_{T,i}^{-2}$. The clustering proceeds by finding the minimum d_{min} of all the particles distances d_{ij} and beam distance d_{iB}. If the smallest distance is d_{ij}, then particles i and j are combined to form a single entity. On the other hand, if the smallest distance is d_{iB}, then entity i is considered a jet and removed from the list of entities. The above steps are repeated until there are no particles left. The anti-kT algorithm outperforms similar algorithms (k_T and Cambridge/Aachen) and iterative-cone algorithms, and thus it is used as the standard jet clustering algorithm in CMS. This algorithm typically results in particles with high p_T since all low- p_T particles are accumulated within a cone of radius R. The distance parameter R is chosen based on the use cases of the resulting jet collection. The default jet collection in CMS consists of jets with small radius of $R = 0.4$. This collection is referred as "Ak4" jet collection. Due to the small R, each Ak4 jet typically corresponds to a single quark or gluon from the hard-scattering process. Jet collections that corresponds to larger R values, such as $R = 0.8$ (Ak8) and $R = 1.5$ (Ak15) are used to reconstruct hadronically decaying particles with large Lorentz-boost. Those particles can be hadronically decaying top quarks, W, Z and Higgs bosons. High-p_T particles decaying products can be collimated and thus, they can be reconstructed in a single-R jet. Additional energy attributed to the jets which comes from pp interactions other than the hard-scatter event at the primary vertex (PV), can contaminate the reconstructed jets. The additional energy from these events is often called pileup and CMS has developed algorithms to account on this effect. If a charged hadron is reconstructed in the tracker and is identified as originating from a pileup vertex, it is removed from the collection of particles used to form physics objects. This procedure is widely used in jet reconstruction and is referred to *charged hadron subtraction* (*CSH*) [67].

An event-by-event jet area–based correction is applied to the jet four-momentum to remove the remaining (neutral-particle) energy from pileup interactions. An alternative way to mitigate pileup is the *pileup per particle identification* also known as *PUPPI* [68]. In this approach, each reconstructed particle is assigned with a weight that reflects the probability that this particle is originating from pile up interactions. This probability is based on local energy distribution around the particle, event pileup properties and tracking information (for charged particles) and uses this estimation to scale the four-momentum of each PF candidate, before clustering them into jets.

When a jet is reconstructed, the jet energies are calibrated in terms of the jet energy scale (*JES*) and jet energy resolution (*JER*) [69]. The JES corrections are necessary to correct the measured jet energy to match its true value. They are implemented in stages and applied both in Monte Carlo (MC) simulated events and data. In each step, corrections account for the offset energy coming from multiple proton-proton collisions in the same and adjacent beam crossings (pileup), the detector response to hadrons, and residual differences between data and MC simulation as a function of the jet pseudorapidity η and transverse momentum p_T, are obtained. Finally a jet-flavour correction is applied to account for differences in the quark-gluon composition of jets. In the final step, the uncertainties assigned to those corrections are determined. Measurements showed that the JER observed in data is worse than MC and thus simulated jets should be smeared to better describe the data. The JER correction for each jet is calculated as a function of the p_T and η of the jet and the p_T of the clustered generator-level particles, if they are matched to the jet particles. In case the jets are not matched with a generated particle, a stohastic smearing is applied based on the η of the jet and the resolution of its p_T. The uncertainty in the derivation of the JER is also considered in the event selection and final result.

3.5 Heavy flavour jets

In simulated events the jet flavor is defined through the principle of *ghost matching* [70]. In this approach one can add to the collection of particles the generator-level b and c hadrons and then perform the jet clustering. In order for the reconstructed jet momentum to stay intact, these b and c hadrons have almost negligible momenta (i.e. ghost hadrons) and so, only their direction is considered in clustering algorithms. The algorithm gives rise to the following jet flavors:

b jets: the jet is labelled as a b jet if at least one b ghost hadron is clustered inside the jet.

c jets: if at least one c ghost hadron is clustered inside the jet while there is no b hadron, the jet is labelled a c jet.

light jets: If no b or c hadrons are clustered inside the jet, it is labeled as a *light flavour* (LF) jet.

pile up jet: If a jet has no matched generator–level jet, it is automatically labelled a pileup jet, independent of whether a b or c ghost hadron is found.

The successfully identification of jets originating from the hadronization of bottom quarks (b jets) or charm quarks (c jets) is crucial to efficiently select events with these quarks in the final state. Dedicated b or c tagging algorithms [71] have been developed in CMS. Those algorithms profit from the relative long lifetime of b and c hadrons. More specifically, the hadronization of b and c quarks leads to the formation of short-lived heavy-flavour hadrons, with lifetimes of the order of 1ps or less. Depending on the hadron p_T, this corresponds to flight distances from a few mm up to $1\ \mathrm{cm}$. Therefore, the decay products of these hadrons do not point back directly to the primary interaction vertex (*PV*), given that their origin lies at the secondary decay vertex (*SV*). This results in reconstructed tracks which are displaced with respect to the PV. This displacement can be characterized by the *impact parameter* IP, which is defined as the distance of closest approach between a track and the PV. Moreover, the relative higher mass of b (5 GeV) and c (2 GeV) hadrons compared to the mass of light flavour jets LF $(\mathrm{u,s,d})$ can provide additional information for a reconstruction of such a SV. The increased (semi)leptonic branching fraction of b and c hadrons compared to light hadrons results in the presence of low–energy (soft) electrons or muons in about 20% (10%) of the b (c) jets. The presence of a soft muon allows a very pure selection of HF jets and also provides a way to distinguish them from LF jets. In figure 3.2, the discussed properties to separate HF from LF jets are visualized.

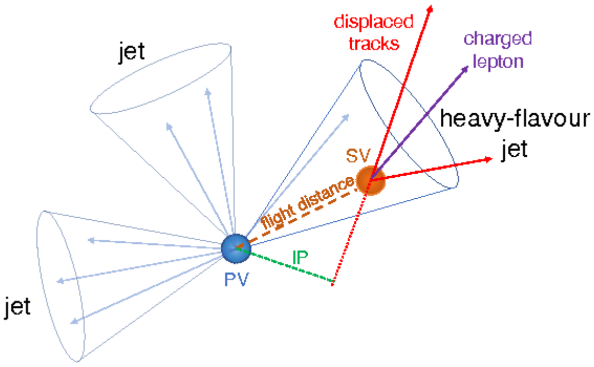


Figure 3.2: illustration of the flight distance defined as the distance between the primary vertex (PV) and the secondary vertex (SV), and the impact parameter (IP) defined as the distance from the PV to the track at its closest point of approach.

Track selection

The identification of HF jets strongly relies on the performance of the tracker. As can been seen from figure 3.3, the fraction of b (c) hadron tracks in b (c) jets accounts for only $\sim 30\%$ $(\sim 15\%)$ of the total track content and a non–negligible fraction of pileup $(\sim 40\%)$ and fake $(\sim 5\%)$ tracks is present. Therefore, a preselection is made for the tracks that are used in the HF tagging algorithms. The requirements imposed to the tracks are:

- $p_T > 1$ GeV, which reduces the fraction of misreconstructed or poorly reconstructed tracks

- at least 8 tracker hits must be associated with the track

- the normalized χ^2 is required to be less than 5 to ensure a good quality fit

- the absolute value of the transverse and longitudinal impact parameter of the track must be < 0.2 and < 17 cm, respectively, to reject charged particle tracks having their origin from sources with large displacement from the primary vertex (e.g. photon conversions and nuclear interactions in the beam pipe or the first layers of the pixel detector)

- tracks must be associated to jets in a cone $\Delta R < 0.3$ around the jet axis, where the jet axis is defined by the primary vertex and the direction of the jet momentum

- the point of closest approach between the track trajectory and the jet axis, must be within 5 cm of the primary vertex

The effects of this selection are showed in the right part of figure 3.3.

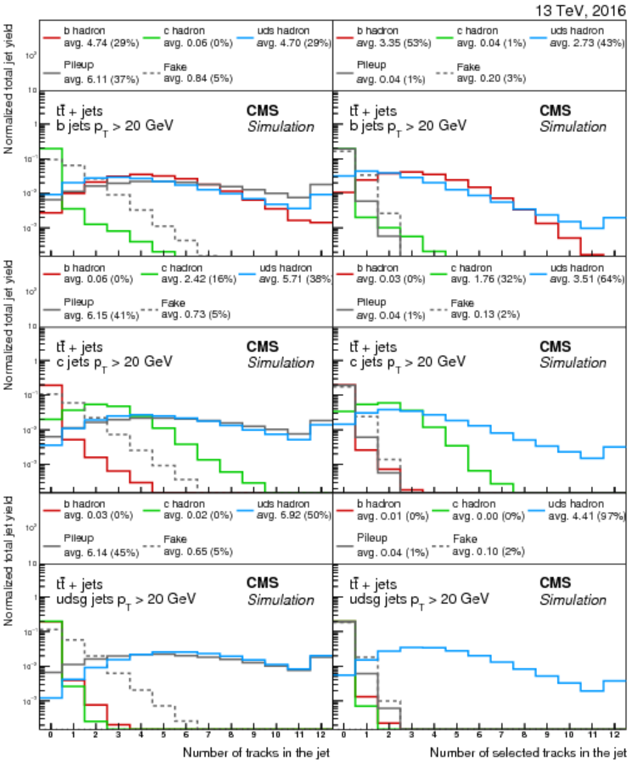


Figure 3.3: Fraction of tracks from different origins before (left) and after (right) applying the track selection requirements on b (upper), c (middle), and light-flavour (lower) jets in $t\bar{t}$ events.

Secondary vertex

As previously discussed, the lifetime of b and c hadrons gives rise to displaces secondary vertices at flight distances of a few mm up to the order of a cm from the interaction point. Therefore, the presence of one or more SVs in a jet is a good indication that the jet originates from a b and c hadron, and thus CMS has developed algorithms dedicated to their reconstruction. The Inclusive Vertex Finder algorithm (IVF) for instance, uses as input the collection of reconstructed tracks in the event and selects tracks with $p_T > 0.8$ GeV and a longitudinal IP < 0.3 cm. Compared to another SV algorithm, the Adaptive Vertex Reconstruction (*AVR*), IVF is not seeded from tracks associated to the reconstructed jets. The selected tracks are then used to identify clusters of nearby tracks based on their minimum distance and the angles between them. The clusters are fit with the adaptive vertex fitter and a cleaning procedure is applied. At this stage, tracks can appear in multiple vertices and therefore, one of the vertices is removed based on the number of shared tracks and distance between the vertex and another one. Furthermore, tracks in the secondary vertex compatible with the primary vertex are removed. When there are at least 2 tracks associated to the secondary vertex after the track arbitration, the vertex is refit and similar selection criteria are applied. IVF outperforms the alternative AVR algorithm. The SV finding efficiency, defined as the number of jets with a reconstructed SV assigned to them divided by the total number of jets, is found to be $\sim 75\%$ for b jets and $\sim 38\%$ for c jets, whereas, only around 10% of LF jets has a secondary vertex assigned to them by mistake. The properties of the a SV provides further information that can be used to distinguish HF from LF jets. One of the most powerful discrimination variable is the corrected invariant mass of the SV defined as:

$$\mathrm{M_{SV}^{corrected}} = \sqrt{\mathrm{M_{SV}^2} + \mathrm{p_{SV}^2}\sin(2\theta)} + \mathrm{p}\sin(\theta) \tag{3.7}$$

where, $\mathrm{M_{SV}}$ and $\mathrm{p_{SV}}$ is the invariant mass and momentum of the summed tracks that are associated to the SV and θ is the angle between the secondary vertex momentum and the vector pointing from the primary vertex to the secondary vertex, which is referred to as the secondary vertex flight direction. The correction applied in the SV accounts for the the observed difference between its flight direction and its momentum. Moreover, particles that were not reconstructed or failed to be associated with the secondary vertex are also considered in the $\mathrm{M_{SV}^{corrected}}$. Figure 3.4 shows the discriminating power between the various jet flavours for the IVF secondary vertex mass (left) and 2D flight distance significance (right). The secondary vertex mass for b jets peaks at higher values compared to that of the other jet flavours which is expected due to higher mass of the b hadron.

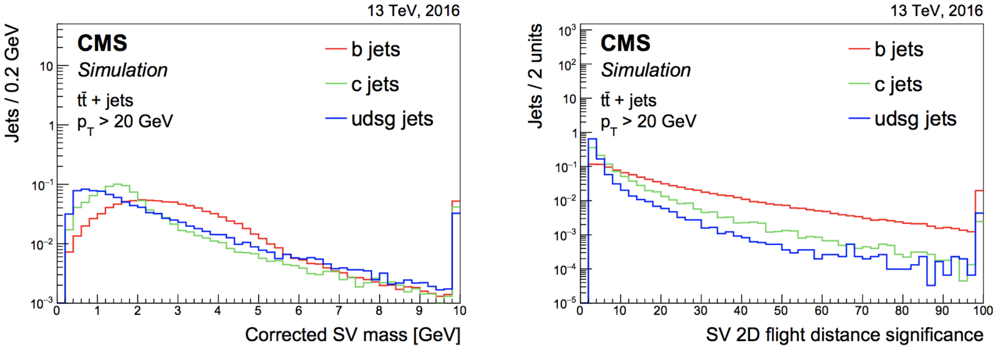


Figure 3.4: Distribution of the corrected secondary vertex mass (left) and of the secondary vertex 2D flight distance significance (right) for jets containing an IVF secondary vertex.

Displaced tracks

The decay products of heavy flavour jets (b and c) will result in displaced tracks relative to the position of the PV. The impact parameter (IP) of the track which measures the point of closest approach between the reconstructed tracks and PV is used to parametrize this replacement. This parameter is defined either in the full three–dimension space (3D), in the transverse plane (2D) or as a one dimensional projection along the beam line (longitudinal). The vector pointing from the PV to the point of closest approach with the track is referred to as the IP vector. The IP value can either be positive or negative, depending on whether the angle between the IP vector and the jet axis is smaller or larger than $\pi/2$ respectively. LF jets are expected to have an IP value close to zero, whereas, b and c jets are expected to have a much larger positive tail. This is illustrated in figure 3.5. Other variables that contribute significantly in discrimination of HF jets are the jet multiplicity, where more SV are expected from HF events.

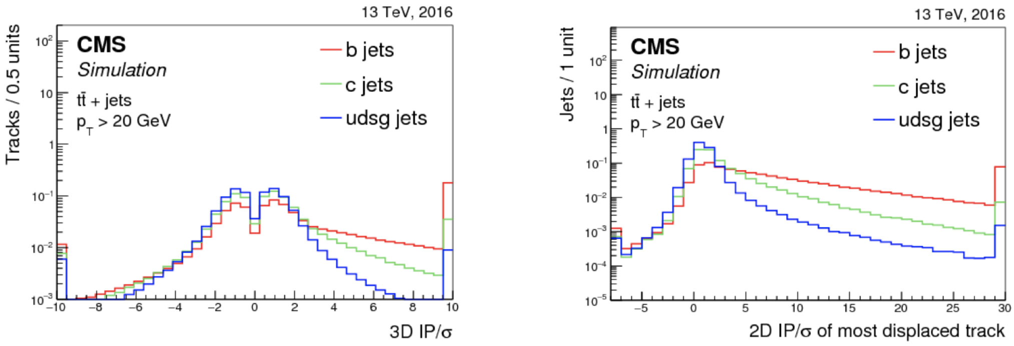


Figure 3.5: Distribution of the 2D (left) and 3D (right) impact parameter significance for the track with the highest 2D (3D) impact parameter significance for jets of different flavours in $t\bar{t}$ events.

Soft Leptons

Although the presence of an electron or muon is precent only for 20% for b and 10% for c jets, the presence of low-energy nonisolated "soft lepton" (*SL*) allows the selection of a pure sample of heavy flavour jets. Discriminating variables using soft lepton information are typically similar to the variables based on track information alone as shown in figure 3.6.

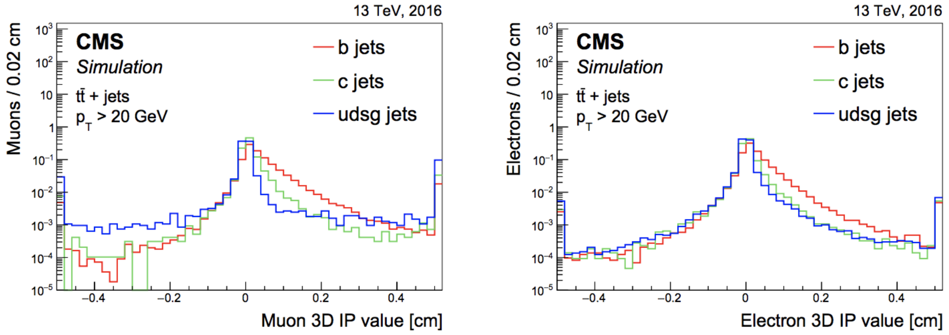


Figure 3.6: Distribution of the 3D impact parameter value for soft muons (left) and soft electrons (right) for jets of different flavours in $t\bar{t}$ events.

3.5.1 Heavy-flavour jet identification algorithms

The main purpose of a heavy flavor identification algorithm is to provide a jet-based observable that can discriminate b and c jets from light jets. If the algorithm is used to identify b jets, then it is referred as a b-tagger. In order to successfully identify a b hadron, we need to fully exploit all the variables described previously. For this reason a Multivariate (MVA) approach is used. Several Machine learning (ML) classifiers are trained, that profit from the availability of the large-scale simulated events and the up-to-date hardware and software developments. Each classifier provides a discriminator value above which a jet is characterized as a b jet. Such a threshold is often referred as a *working point* (WP) and determines the average tagging efficiency of the jet flavor of interest, as well as, a misidentification probability to tag a jet of another flavor. In addition, analyses can also extract information from the full discriminator shape. Since those discriminators have strong separation power between b and c or light jets, they can be used to preform a fit to the data. In addition, the full distribution of the discriminator can also be used as an input to another ML-based algorithm.

In RunI, the so called *Jet Probability* (JP) and *Combined Secondary Vertex* (CSV) taggers were used [72]. The JP tagger uses the track variables described previously and assigns a likelihood for the jet to originate from the primary vertex. In RunII, the JP tagger was used only for calibration measurements of other taggers and CSV was retrained (CSVv2). CSVv2 combines the information of displaced tracks with the information of secondary vertices associated to the jet. Jets are divided in three vertex-dependent exclusive categories:

> **RecoVertex**: at least one reconstructed secondary vertex in the jet
>
> **PseudoVertex**: no secondary vertex is found, but there is at least two tracks with impact parameter significance larger than 2
>
> **No Vertex**: Containing jets not assigned to one of the previous two categories. Only the information of the selected tracks is used.

The advantage of this algorithm is that it does not suffer from inefficiencies in the secondary vertex reconstruction, since cases with none reconstructed vertex are also considered. Compared to the first version of CSV, the CSVv2 uses a different vertex reconstruction algorithm (IVF) and more variables as input to the tagger. The discriminating variables that are combined in the tagger are the following

- the corrected SV mass
- track multiplicity from SV
- the ratio of the energy carried by tracks at the vertex with respect to all tracks in the jet
- the ΔR between the flight direction of the secondary vertex with the smallest uncertainty on its flight distance and the jet axis for jets in the RecoVertex category, or the ΔR between the total summed four-momentum vector of the selected tracks for jets in the PseudoVertex category
- the 3D signed impact-parameter significance for each track in the jet

- The "track $p_{T,rel}$", defined as the track p_T relative to the jet axis, i.e. the track momentum perpendicular to the jet axis, for the track with the highest 2D impact parameter significance.
- The "$\Delta R(\text{track, jet})$", defined as the ΔR between the track and the jet axis for the track with the highest 2D impact parameter significance.
- The "track $p_{T,rel}$ ratio", defined as the track p_T relative to the jet axis divided by the magnitude of the track momentum vector for the track with the highest 2D impact parameter significance.
- The "track distance", defined as the distance between the track and the jet axis at their point of closest approach for the track with the highest 2D impact parameter significance.
- The "track decay length", defined as the distance between the primary vertex and the track at the point of closest approach between the track and the jet axis for the track with the highest 2D impact parameter significance.
- The "summed tracks ET ratio", defined as the transverse energy of the total summed four-momentum vector of the selected tracks divided by the transverse energy of the jet.
- The "$\Delta R(\text{summed tracks, jet})$ ", defined as the ΔR between the total summed four-momentum vector of the tracks and the jet axis.
- The "first track 2D IP significance above c threshold", defined as the 2D impact parameter significance of the first track that raises the combined invariant mass of the tracks above 1.5 GeV. This track is obtained by summing the four-momenta of the tracks adding one track at the time. Every time a track is added, the total four momentum vector is computed. The 2D impact parameter significance of the first track that is added resulting in a mass of the total four-momentum vector above the aforementioned threshold is used as a variable. The threshold of 1.5 GeV is related to the c quark mass.
- The number of selected tracks.
- The jet p_T and η

CSVv2 was the recommended tagger for analyses that used 2016 data, as the analysis of this thesis. However, in 2017 a new, optimal, version of CSVv2 tagger, "DeepCSV", was developed that takes advantage of the evolving field of deep machine learning. DeepCSV uses the same input variables and the same secondary vertex reconstruction as CSVv2, and combines them with a deep neural network with more hidden layers and more nodes per layer. Then, a simultaneous training is performed in all vertex categories and for all jet flavours. The outcome of the algorithm is four independent output classes $P(b/bb/c/udsg)$ each of which accounts as a probability for a certain jet flavour category. $P(b)$ and $P(bb)$ are defined according to whether the jet contains exactly one or two b hadrons respectively, $P(c)$ is defined when exactly one c hadron is found and finally, $P(udsg)$ is defined when none of the above is found. When the sum of the $P(b) + P(bb)$ of the DeepCSV discriminator is above a given threshold then the jet is tagged as a b jet. DeepCSV outperforms the other taggers as illustrated in figure 3.7, where the ROC curves of CSVv2, DeepCSV and another tagger, cMVAv2, are compared. The ROC curve is defined as the efficiency of tagging a b jet versus the misidetification probability of tagging a c, or light flavour jet. The curve closest to the right lower corner corresponds to the best performing tagger.

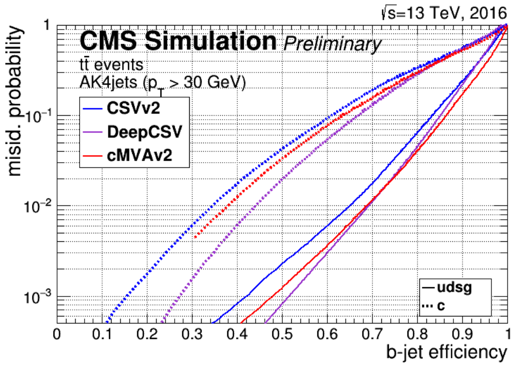


Figure 3.7: Performance of the b jet identification efficiency algorithms demonstrating the probability for non-b jets to be misidentified as b jet as a function of the efficiency to correctly identify b jets.

An improvement in the identification of HF jets can be achieved by introducing more low-level inputs from all charged and neutral PF candidates that are clustered inside the jet, and combine them with all the track and SV based variables. This is done by the more advanced algorithm called DeepFlavor or DeepJet. The output of the discriminator is six independent classifiers $P(b/bb/blep/c/udc/g)$, that similarly to DeepCSV, they account on the probability that a certain jet flavor is found. A new classifier $P(blep)$ is introduced which considers the leptonic b hadron decays. For DeepFlavor tagger the discriminator is defined as the sum of $P(b) + P(bb) + P(blep)$. DeepFlavor is the best performing tagger as showed in figure 3.8. In this figure DeepCSV is compared to the old (three–layer) pixel detector geometry used in 2016 (green line). It is also clear that the upgraded (four-layer) pixel detector results in better performance of the taggers, since it provides more accurate information about tracking related variables and also SV reconstruction is done more efficiently.

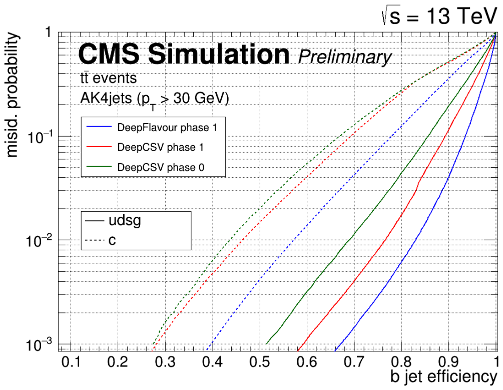


Figure 3.8: Performance of the DeepCSV and DeepFlavour b jet identification algorithms demonstrating the probability for non-b jets to be misidentified as b jet, as a function of the efficiency to correctly identify b jets.

3.5.2 Commissioning of the taggers-Comparison studies in data and simulation

As already described in detail, different b-tagging techniques are developed in CMS which benefit from the long life time, high mass and large momentum fraction of the b-hadron produced in b-quark jet. In order to validate the tagger performance it is necessary to compare the simulated input variables and the tagger distributions with the data. Constant monitoring of the basic quantities provided to the high-level taggers is fundamental to ensure a good tagging performance and to spot potential issues during the data taking. The variables of interest and the distributions of the taggers are compared with data in different event topologies with different flavour composition. Those topologies are:

Inclusive multijet sample: This sample is enriched in light and pileup jets. The events are required to have at least one Ak4 jet with $p_T > 40$ GeV. Data and simulated multijet events are compared using jets with $50 < p_T < 250$ GeV.

Muon-enriched jet sample: This sample requires a muon and therefore this topology is dominated by jets containing heavy-flavour hadrons. Events are considered if they satisfy an online selection with at least two Ak4 jets with $p_T > 40$ GeV of which at least one contains a muon with $p_T > 5$ GeV. Jets with $50 < p_T < 250$ GeV and a muon from simulated muon-enriched multijet are selected and compared to the data.

Dilepton $t\bar{t}$ sample: This sample is enriched in b jets from top quark decays. Isolated muons and electrons with $p_T > 25$ GeV are selected. In this topology there is also a small contribution from pileup jets due to the relatively low threshold on jet p_T.

Single-lepton $t\bar{t}$ sample: A higher fraction of c jets is expected that comes from the hadronically decay of the W boson. Events with exactly one isolated electron or muon are selected. The electron (muon) is required to have a $p_T > 40(30)$ GeV and $|\eta| < 2.4$.

The comparison of simulated events and data that were collected in 2016 is illustrated in figure 3.9 to figure 3.11. The comparison is made for discriminating variables (figure 3.9, figure 3.10) such as the 3D IP significance of tracks, the corrected SV mass and the distribution of the CSVv2 and DeepCSV taggers (figure 3.11). Overall there is a good agreement between data and simulation. The observed discrepancy around zero is explained by the sensitivity of this variable to the tracker alignment and the uncertainty in the track parameters.

Comparison studies can as well be performed between two different data periods. This way, any discrepancies due to the data taking conditions can be spotted and understood. An example of such a study is illustrated in figure 3.12, where the full 2017 dataset is compared to a subset of events reconstructed using 2018 data. For comparison purposes, the integrals of all distributions are normalized to unity.

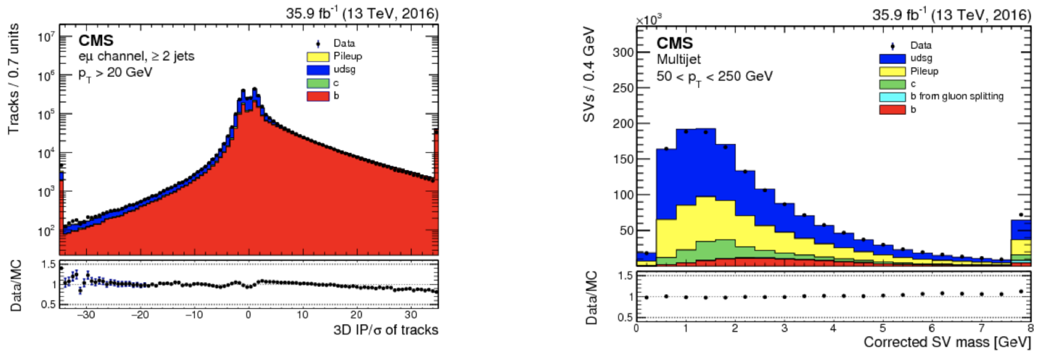


Figure 3.9: Examples of input variables used in heavy-flavour tagging algorithms in data compared to simulation. Impact parameter significance of the tracks in jets from the dilepton $t\bar{t}$ sample (upper left), corrected secondary vertex mass for the secondary vertex with the smallest uncertainty in the 3D flight distance for jets in an inclusive multijet sample (upper right),

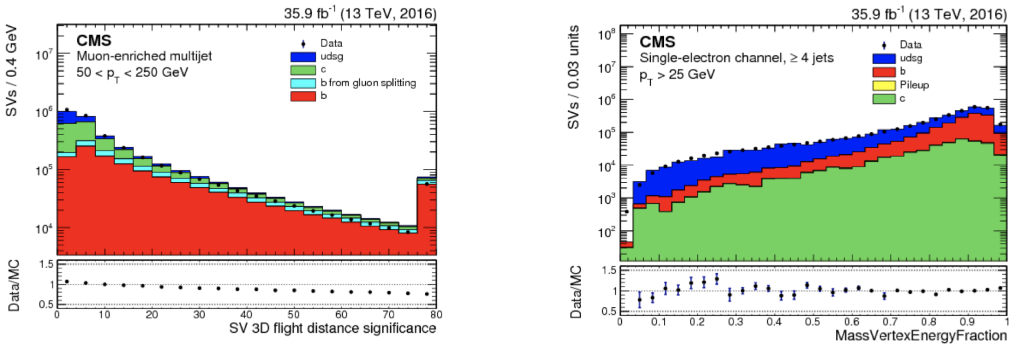


Figure 3.10: secondary vertex flight distance significance for jets in a muon-enriched jet sample (left), and distribution of the massVertexEnergyFraction variable single-lepton $t\bar{t}$ sample.

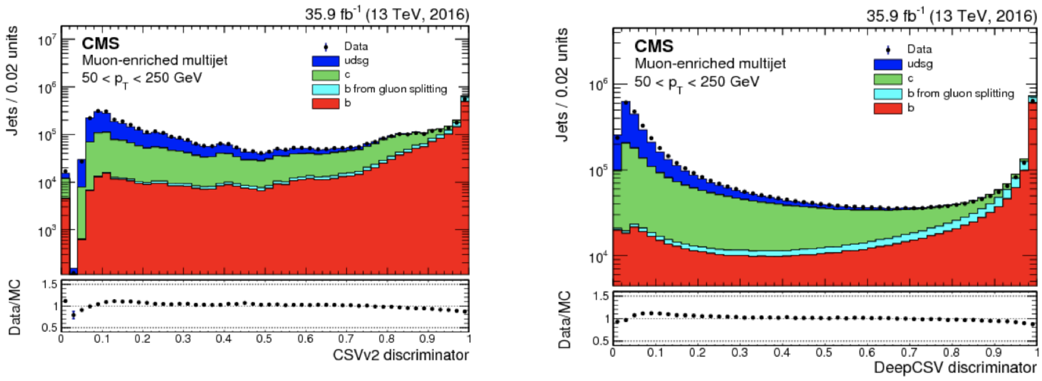


Figure 3.11: Examples of discriminator distributions in data compared to simulation. The CSVv2 (left) and DeepCSV (right) discriminators for jets in the muon-enriched multijet sample are compared with data

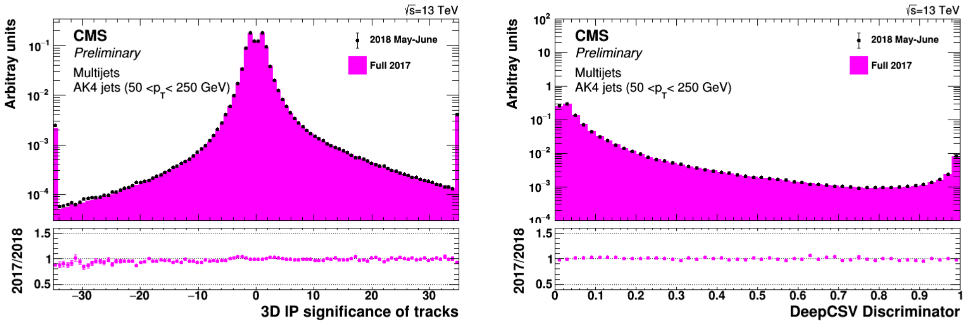


Figure 3.12: The impact parameter significance of the tracks in jets in an inclusive multijet sample (left) and the DeepCSV discriminator distribution in jets in an inclusive multijet sample (right). For both plots the black dots correspond to data recorded in 2018, compared to the distribution from 2017 data [13].

3.6 Missing Transverse Momentum

CMS is a full coverage hermetic detector capable to interact and reconstruct stable or long-lived particles produced in pp collisions. However, particles such as neutrinos or hypothetical neutral weakly-interacting particles do not leave a signal to the detector, and thus their presence can only be inferred through the visible momentum imbalance in the transverse plane. This quantity is known as missing transverse momentum ($\vec{p_T}^{miss}$) and its magnitude is denoted as p_T^{miss} [73]. The $\vec{p_T}^{miss}$ is defined as the negative vector sum of the $\vec{p_T}$ of all reconstructed PF candidates. The definition of $\vec{p_T}$ includes all the physics objects (muons, electrons, photons, τ_h candidates and jets) that are reconstructed from the PF candidates as described in the previous sections, but also the unclustered energy, defined as the energy of all the PF candidates not clustered into any physics object. The precise measurements of p_T^{miss} is crucial for analyses that include neutrinos is the final states such as leptonic decays of the W boson. Moreover, p_T^{miss} is one of the key variables for beyond the standard model searches where hypothetical particles leave the detector without interacting. However, p_T^{miss} reconstruction is sensitive to the resolutions and mis-measurements of the reconstructed particles, and also to detector artifacts. In addition p_T^{miss} reconstruction is highly affected by pile-up events. All the above issues should be well understood and CMS has performed studies of the performance of p_T^{miss} in data and simulation [73]. The p_T^{miss} is calibrated by propagating the effect of the jet energy corrections as:

$$\vec{p_T}^{miss} = \vec{p_T}^{miss,\text{uncorrected}} - \sum_{jets}(\vec{p_T}^{\text{corrected}} - \vec{p_T}^{\text{uncorrected}}) \tag{3.8}$$

where $\vec{p_T}^{\text{uncorrected}}$ and $\vec{p_T}^{\text{corrected}}$ is before and after applying the jet energy corrections. To mitigate pile up events, only jets with the corrected p_T above 15 GeV are considered in the sum. Jets that corresponds to electromagnetic showers from electrons and photons are removed by excluding jets that have $> 90\%$ of the jet energy deposit in the ECAL. The same stands for jets that contain global and standalone muons.

Chapter 4

Simulation of Collision events

4.1 Event generation

One of the main challenges of LHC physics is to provide accurate theoretical predictions on the observable quantities which are expected from the particle detectors. The ability of understanding the complexity of high energy collisions rely on how well those quantities can be modelled in simulation. These simulations are often not based on exact analytical calculations, but rather rely on Monte Carlo sampling techniques [74]. The Monte Carlo method uses random sampling applied to a theoretical model to predict its expected behavior in realistic conditions. It relies on computer simulations and can give correct solutions especially in cases where a deterministic solution cannot be derived. Examples in high energy physics include event simulation, where particles are produced in random direction and position, but obey some theoretical constraints, and detector simulation, where the detector behavior is modeled precisely by taking into account several parameters.

The relatively poor understanding of the strong force (QCD) that acts among the colliding protons makes this field of research highly non trivial. The workflow used to produce simulated pp collisions events can be factorized into separate steps: first, the *generation of the hard process* (using perturbative QCD) is followed by the forward and backward evolution of *parton showers* and then, the *hadronization* of the partons is produced. Moreover, the partons that do not participate in the hard scattering can still undergo soft scattering processes *(underlying event)*. A schematic representation of a typical pp collision is shown in figure 4.1, indicating each of the steps in the chain that is briefly described below: [75].

Hard scattering: When two protons collide, only one of the partons (quarks or gluons) inside each proton participates in the main interaction of interest. The underlying structure of the proton is essential since it determines the probability for each parton to participate in this hard scatter.

Parton shower: The radiation of gluons (QCD radiation) or electromagnetic radiation can result in production of additional particles. This phenomenon is described by the parton shower (PS).

Hadronization and fragmentation: Due to the principle of color confinement, only color neutral hadrons are observed. Consequently the colored partons after

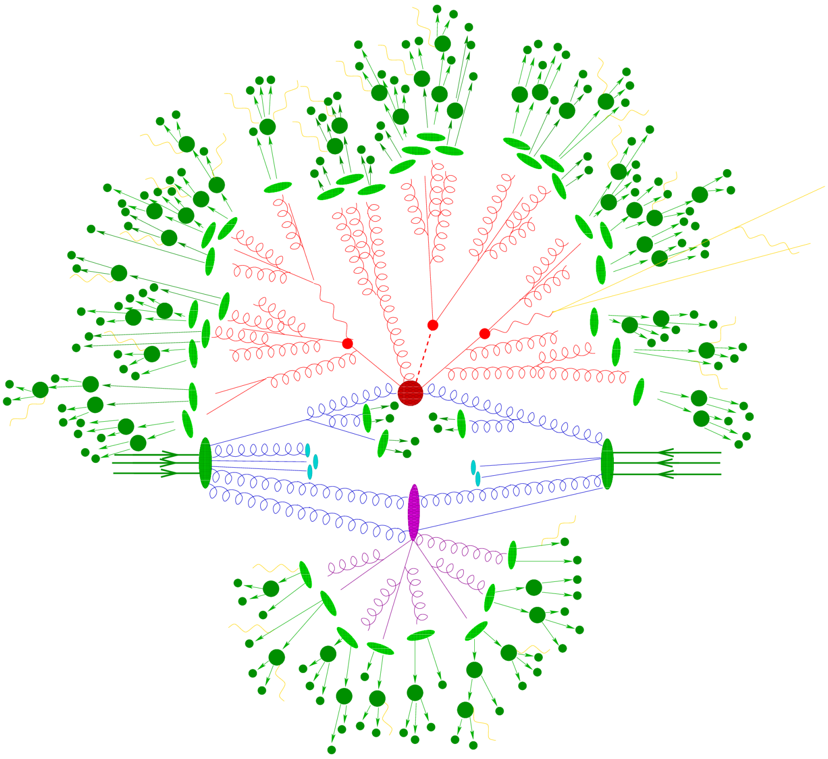


Figure 4.1: Sketch of a hadron-hadron collision as simulated by a Monte-Carlo event generator. The red blob in the center represents the hard collision, surrounded by a tree-like structure representing Bremsstrahlung as simulated by parton showers. The purple blob indicates a secondary hard scattering event. Parton-to-hadron transitions are represented by light green blobs, dark green blobs indicate hadron decays, while yellow lines signal soft photon radiation [14].

the PS undergo a hadronization process. These hadrons can then further decay into stable particles that make their way through the detector.

Underlying event: The part of the proton that did not participate in the hard interaction can still undergo soft scattering processes. As a result multi–parton interactions can occur- the so–called underlying event (*UE*). Those interactions are no longer described by perturbative QCD. Therefore, nonperturbative multiple-parton interaction models and diffraction models with tunable parameters are necessary.

Detector simulation: Finally, the stable particles will interact with the detector's material to leave their experimental signature. A proper simulation of the detector response is the final step before the event reconstruction can start.

4.1.1 Hard process

The first and most fundamental step of the simulation chain consists of the accurate calculation of the matrix element of the process of interest. In proton-proton collisions, the centre of mass energy of the partons in the hard collisions varies from event to event according to parton density functions (PDFs). The PDF $f(\chi, Q^2)$ describes the probability that a certain type of quark or gluon is found carrying a fraction χ of the total momentum of the incoming proton when it is probed at an energy scale Q^2. Those PDFs are determined by collaborations such as NNPDF, MSTW or CTEQ by fitting these functions to data observed in deep inelastic scattering, Drell-Yan and multijet processes. During the LHC Run 2, a recent set of PDFs produced by the NNPDF Collaboration, known as NNPDF3.0 [76], is widely used in event generation. These PDFs are based on a large variety of experimental results from different experiments operating at different energy scales, and they are scrutinized with several closure tests. Some of those PDFs are illustrated in figure 4.2.

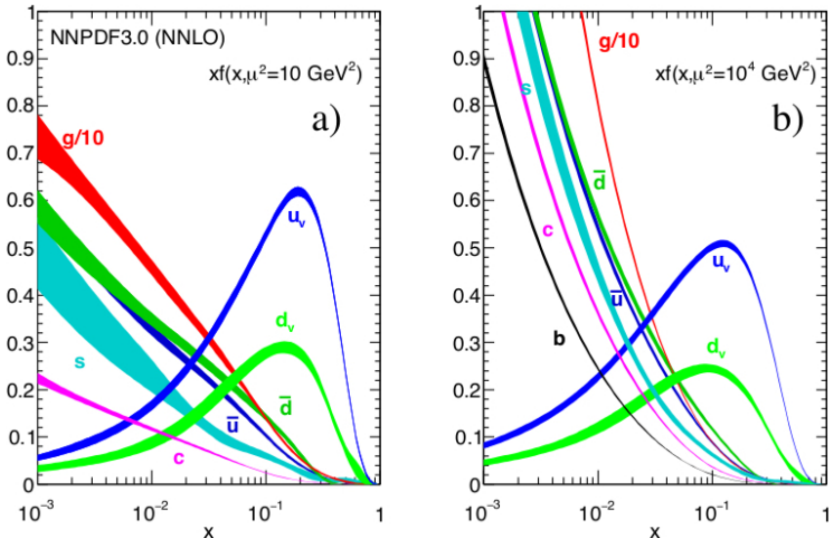


Figure 4.2: Examples of NNPDF3.0 parton distribution functions, shown as a function of χ at low momentum transfer of 10 GeV^2 (left) and at high momentum transfer of 10^4 GeV^2 (right), with $\alpha_s(M_Z^2\,)$ set to 0.118. [15]

A collision between two partons, one from each side, gives the hard process of interest, which can be due to an interaction described within or beyond the standard model. Using the incoming partons as input, the simulation of the hard process is performed by the event generator. It produces hypothetical events with the distributions and rates predicted by theory based on the cross section formulae of the physics process of interest. The cross section can be calculated by means of the so-called $\mathrm{factorization\ theorem}$. According to the theorem, the hadron itself is described by the whole particle composition interacting on a soft binding energy scale, whereas the collisions occur between the partons on a hard energy scale with large transverse momenta. Therefore, the total cross section can be factorized through the convolution of the partonic cross section $\hat{\sigma}_{\mathrm{i,j}} \rightarrow X$ of partons i and j and their corresponding PDFs $\mathrm{f_i(x_i, Q^2)}$ and $\mathrm{f_j}(\chi_\mathrm{j}, \mathrm{Q}^2)$. This convolution is integrated over the fractional momenta $\mathrm{x_i}$ and $\mathrm{x_j}$ and summed over all possible initial-state partons that may result in the final-state X of interest. This way all possible initial state configuration are taken into account as:

$$\sigma_{\mathrm{pp\rightarrow X}} = \sum_{\mathrm{i,j}} \int \mathrm{dx_i} \int \mathrm{dx_i f_i(x_i, Q^2 f_j(x_j, Q^2)} \hat{\sigma}_{\mathrm{i,j}} \rightarrow \mathrm{X} \tag{4.1}$$

To determine the partonic cross section ($\hat{\sigma}_{\mathrm{i,j}} \rightarrow \mathrm{X}$) the matrix element (ME) of the phase–space integration of the final–state X along with factorized from the proton PDFs are taken into account [15]. This is done with matrix element generation software that provides automatic calculations of the matrix elements up to a given order perturbation theory which uses Monte Carlo sampling techniques to generate a set of simulated events in the desired phase space. The ME used in CMS are POWHEG (LO_NLO), MADGRAPH (LO), MC_AMC@NLO (NLO) and PYTHIA (LO). All ME generators need as input a model that describes the particle content, couplings, interactions and other constants in order to calculate the matrix element. However, since they do not include hadronization, they have to be interfaced with other generators in order to produce the full event.

4.1.2 Parton Showering

The colored particles that participate in the hard scattering process can undergo a chain of soft radiation or branchings into other particles. As a result additional final-state particles are produced that are not included in the initial matrix element calculation. This additional radiation is described as a parton shower (PS) [75]. To model the PS an approximate higher–order correction to the hard scattering is needed in the limits of either very soft radiation of gluons $(q \to qg)$ or very collinear splitting of a gluon into a quark–antiquark pair $(g \to qq)$ or into another pair of gluons $(g \to gg)$. In case this additional radiation happens through the initial–state partons it is referred to as initial–state radiation (*ISR*), as opposed to final–state radiation (*FSR*) describing the parton shower for the final–state particles. The parton shower proceeds by considering for each of the partons in the event the probability that it undergoes a branching into two daughter particles. Examples of frequently used parton shower simulators are PYTHIA and Herwig++ that are interfaced with more precise ME generators at higher order. The parton shower continues until a fixed energy scale Λ_{QCD} is reached, above which the perturbative description of QCD is not longer valid.

4.1.3 Hadronization and fragmentation

Above the Λ_{QCD} scale, the strong coupling constant grows to values that do not any longer allow for a perturbative expansion with reliable predictions at fixed order. The colored partons, therefore, due to color confinement they need to be combined into colored-neutral states. This process is known as hadronization and the non–perturbative nature of QCD at these scales forces us to resort to phenomenological descriptions of these processes based on models such a the Lund string model and the cluster model [77, 78].

Finally, fragmetantions functions are used to describe how the momentum of an initial parton is distributed among the final–state particles that result from the partons after the PS and hadronization. Those functions are used in order the predicted energy distributions to actually match the observed data.

4.1.4 The underlying event

The appearance of multiple proton interactions in the same bunch-crossing (pileup) results in low energy activity which is spatially separated from pp interactions responsible for the hard scattering process. The remaining proton remnants consisting of partons that did not take part in the hard scattering process contribute in the event activity from the hard pp interactions. This additional low-energy activity is known as the underlying event (*UE*). The UE could result not only from proton remnants, but also from multi-parton interactions and color reconnections. In the latter case, the partons in the proton remnant are not independent from those participating in the hard scattering as there exist also color connections between them and thus can cause interference effects between the hard scattering and the UE. Since UE can affect the measurement of several quantities it is necessary to efficiently model it [79]. This additional event activity is not described by another hard scattering process, but rather with a large set of parameters tuned to the data.

4.1.5 Detector Simulation

The end of the simulation chain consists of the the full description of the CMS detector. A full simulation of the CMS detector has been integrated in the GEANT4 toolkit [80] that includes detailed descriptions of each single detector module and each detector layer described in chapter 2. A detailed description of the geometry and materials of all components of CMS is implemented in GEANT4, including both the active detector elements and the passive material such as cables and cooling systems. The software traces the particles through the detector in small steps, using Monte Carlo simulation to impose the particles to different stochastic processes according to their probabilities. GEANT4 includes models to describe a variety of interactions with the detector material, including the effects of electric and magnetic fields, bremsstrahlung, photon conversions, multiple scattering, ionization, and interactions between hadrons and nuclei ranging from MeV-scale elastic scattering of neutrons to GeV or even TeV scale hadron showers. Finally, also pileup interactions are modeled and added at this stage as they may interfere with the signals from the hard scattering processes as the particles pass through the detector. To fully simulate all the detector subsystems, substantial computational power and time is needed. Instead, CMS has developed a fast detector simulation package known as *fastsim*. The *fastsim* [81] serves as a fast, less accurate but reliable alternative to the detailed *fullsim* simulation. The *fastsim* implementation uses simplified parameterizations of the reconstruction efficiencies for several physics objects to avoid a full simulation of all the interactions of the particles with the detector layers. As increasing LHC luminosity and pile up will require ever higher numbers of events, *fastsim* is soon expected to find wider usage, starting with the upgrade studies.

Chapter 5

Statistical Methods

The goal of a statistical analysis is to characterize the observed data, under some theoretical hypothesis. It should be able to provide answers to questions like, "If there is no significant excess corresponding to the presence of a signal, how large signals can be excluded based on the observed data?". Or, "if there is an excess in the data, how likely is it to originate from the signal modeled by a given signal model?". In both cases, the parameter of interest in the analysis is the amount of signal, represented by the *signal strength modifier* μ. The signal strength modifier μ is defined as a parameter that varies the signal yield, thus representing different signal hypotheses. If $s(b)$ is the expected event yield for signal (background) events, the expected total yield is $\mu s + b$. The procedure of hypothesis testing starts with the definition of the null hypothesis and the construction of a suitable *test statistics*. Then, the observed value of the test statistics is calculated from data and compared to the expected distribution of the test statistic. In this chapter the basic aspects of statistical analysis that are applied in the analysis are discussed. In traditional "*cut-and-cound*" experiments the test statistic was defined simply based on the expected and observed event yields, obtained after online and offline selections. In addition, there is a more powerful approach where the summary statistic is calculated from the selected events and used to derive the test statistic. The summary statistic can be any distribution that discriminates between the background and signal events, such as a reconstructed mass distribution or output of an MVA classifier. Those distributions of the summary statistics on a "*shape analysis*" are referred as *templates*. In a shape analysis, the test statistic incorporates both the expected event yield $\mu \mathrm{s_i} + \mathrm{b_i}$ and the observed yield $\mathrm{n_i}$ in each bin of the summary statistic. The normalization of the signal in the templates $(\mathrm{s_i})$ can be based on a specific theoretical model, or it can be arbitrary, as it is only an initial value for the fit to data. If the production cross section (σ) and the branching fraction to the final state are well known, the signal can be normalized accordingly and the signal strength modifier represents deviation from the theory expectation. In case of a more generic search it is more convenient to normalize the signal templates to an arbitrary initial value, such as $\sigma = 1\ \mathrm{pb}$, and 100% branching fraction.

5.1 Likelihood Construction

The construction of the likelihood can be motivated by considering a counting experiment. The probability to count a number of x events follows a Poisson distribution:

$$\mathrm{Poisson}(\mathrm{x}|\lambda) = \frac{\lambda^{\mathrm{x}}}{\mathrm{x}!}\mathrm{e}^{-\lambda} \tag{5.1}$$

where λ denotes the expected number of events. This approach can be extended to multiple N bins by multiplying each Poisson probability,

$$\mathrm{p}(\mathrm{x}|\lambda) = \prod_{\mathrm{i}}^{\mathrm{N}} \mathrm{Poisson}(\mathrm{x_i}|\lambda_{\mathrm{i}}) \tag{5.2}$$

where $\mathrm{x_i}$ and λ_{i} are the measured and expected event counts, respectively, in bin i. In practice, the set of N bins comprises all histogram bins of a selected variable distribution in several measurement categories and $\mathrm{p}(\mathrm{x}|\lambda)$ denotes the conditional probability to measure x events given a particular statistical model. In the context of high-energy physics analyses, this model contains our knowledge about production mechanisms and rates of certain processes, the amount of recorded and selected collision events in terms of luminosity, detector acceptance, and reconstruction efficiencies. Uncertainties of the model, both systematical and statistical, are incorporated as "*nuisance*" parameters $\boldsymbol{\theta}$ that are, in Bayesian terms, subject to prior probability distributions, often estimated through theoretical reasoning or external measurements [82]. Thus, the number of expected events in bin i can be written as

$$\lambda_{\mathrm{i}} = \mu \cdot \mathrm{s_i}(\boldsymbol{\theta}) + \mathrm{b_i}(\boldsymbol{\theta}), \tag{5.3}$$

with the numbers of signal and background events $\mathrm{s_i}$ and $\mathrm{b_i}$, respectively.
For a specific, fixed measurement x, the likelihood as a function of a model parameters $(\mu, \boldsymbol{\theta})$ [83] can be expressed as:

$$\mathrm{L}(\mu, \boldsymbol{\theta}|\ \mathrm{x}) := \mathrm{p}(\mathrm{x}|\ \mu, \boldsymbol{\theta}) = \prod_{\mathrm{i}}^{\mathrm{bins}} \mathrm{Poisson}\ (\mathrm{x_i}|\ \mu \cdot \mathrm{s_i}(\boldsymbol{\theta}) + \mathrm{b_i}(\boldsymbol{\theta})). \tag{5.4}$$

In general, the nuisances $\boldsymbol{\theta}$ may also depend on the bin i as well as on particular processes that contribute to $\mathrm{b_i}(\boldsymbol{\theta})$. This dependence can be written as:

$$\mathrm{b_i}(\boldsymbol{\theta}) = \sum_{\mathrm{p}}^{\mathrm{processes}} \mathrm{b_{p,i}} \cdot \prod_{\mathrm{n}}^{\dim(\boldsymbol{\theta})} \pi_{\eta,\mathrm{p,i}}(\theta_{\mathrm{n}}) \tag{5.5}$$

where $\mathrm{b_{p,i}}$ is the nominal, expected number of events in bin i contributed by process p, and $\pi_{\eta,\mathrm{p,i}}$ is the prior probability distribution of nuisance n evaluated at θ_{n}.
Different sources of uncertainty, corresponding to different nuisance parameters, can be treated as fully correlated (100% correlation), anti-correlated (-100%), or independent (0%). The correct assignment of correlations depends on the specific uncertainties in hand. Partially correlated uncertainties are treated by splitting them to fully (un)correlated sub-components. In CMS analysis there are typically three different types of nuisance's on processes and bins.

> **Rate-changing nuisances**: Those uncertainties affect the overall normalization of one or more processes. Such uncertainties are for example, the uncertainty on the theoretical cross sections. Up and down variations of these uncertainties, describing central 68% confidence intervals, do not lead to shape changes of a template, but rather vary the total number of expected events, i.e., the integral of the template, equally in all its bins. Depending on the likelihood of the nuisance

parameter θ_n, normal Gaussian prior probability distributions with mean one can be used to predict the impact on the event rate of the uncertainty. However, for large uncertainties the normal distribution is rather broad and thus, must be truncated at zero in order to avoid negative values when computing the logarithm of the likelihood function. In order not to systematically overestimate the event yields, a log-normal distribution is employed which is parametrized as:

$$\pi_\theta = \frac{1}{\theta} \cdot \frac{1}{\sqrt{2\pi\sigma}} \exp\Big[- \frac{(\ln\theta)^2}{2\sigma^2} \Big] \tag{5.6}$$

where, σ is the width of the distribution and corresponds to the relative uncertainty, which is estimated a-priori on the basis of theoretical calculations or external measurements.

Shape-changing nuisances: On the contrary to the rate uncertainties that only affect the template normalization, some uncertainties introduce variations of expected event yields that are correlated among bins of a template. To estimate the effect on those uncertainties, the templates are recreated where the model parameter in question is varied according to the boundaries of its central 68% confidence interval. The pdf is constructed as a continuous set of variations by morphing the nominal and varied templates using a suitable morphing function. In some cases, simple linear interpolation can be sufficient, but usually a more sophisticated technique known as horizontal morphing [84] is used. This results in two variants of the template that characterize the change of its shape given a discrete variation of the underlying parameter by ± 1 standard deviation.

Statistical uncertainties: This kind of uncertainties reflect the amount of available statistics of simulated events that estimate the expected yield for each process and bin. One possibility to model statistical uncertainties is to introduce one separate nuisance per process and bin with a prior probability following a Poisson distribution. In case of a sufficiently large number of simulated events per bin, the sum of Poisson distributions can be approximated by a single Gaussian distribution with reasonable accuracy. In general, the precondition on the minimal bin content is done by binning optimization.

5.2 Statistical inference methods

In the content of searches for new physics processes, the statistical inference methods are based on a hypothetical signal that might exist in a well-defined phase space in addition to a well-known background. If there is no potential interference effects, a null hypothesis is defined as the non-existence of the signal process ($\mu = 0$) and it is referred as "background-only hypothesis". On the other hand, the alternative hypothesis postulating the presence of signal ($\mu > 0$) is called “signal hypothesis”. Independent of the specific signal model, a discovery is usually claimed when the measured data is incompatible with the background-only hypothesis. From the Neyman-Pearson lemma [85], the most powerful test statistic Q to evaluate two contrary hypotheses is given by their likelihood ratio. To construct the test statistic, the likelihood function is normalized by its maximum-likelihood value and the profile likelihood ratio for the likelihood from equation 5.4 is defined as:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} \tag{5.7}$$

where λ depends on the measurement x and $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ in the denominator are the $\text{maximum likelihood estimators}$ (MLEs). The MLEs are the parameters that maximize L and minimize $-\ln(\text{L})$, respectively. $\hat{\hat{\boldsymbol{\theta}}}(\mu)$ in the numerator denotes the conditional MLE that would maximize L for a prespecified value of the parameter of interest μ. Therefore, the likelihood ratio λ depends on μ and is, in particular, independent of the nuisance parameters θ. This construction is called "profiling".

In case of limited statistics of signal-like events in a measurement x and $\hat{\mu}$ can become negative. In order to avoid an artificial constraint on μ being positive, which would lead to calculational complications [86], the test statistic is adjusted to

$$\widetilde{\lambda}(\mu) = \begin{cases} \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}, & \hat{\mu} \geq 0 \\ \\ \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0, \hat{\hat{\boldsymbol{\theta}}}(0))}, & \hat{\mu} < 0 \end{cases} \tag{5.8}$$

which is equivalent to requiring $\hat{\mu} \geq 0$ In case of certain regularity conditions [87] while the sample size is sufficiently large, the transformed test statistic is equal to:

$$\widetilde{q}(\mu) = -2\ln\widetilde{\lambda}(\mu) \tag{5.9}$$

and is asymptotically converges towards a non-central χ^2 distribution with one degree of freedom.

5.3 Limit setting

The test statistic in equation 5.8 can be used to differentiate the hypothesis that a signal process is produced at a certain rate μ from the alternative stating that the production rate is smaller. This can be reflected by a case distinction. $\hat{\hat{x}}$

$$\widetilde{q}_\mu = \begin{cases} -2\ln\widetilde{\lambda}(\mu), & \hat{\mu} \leq \mu \\ 0, & \hat{\mu} > \mu \end{cases} \tag{5.10}$$

A single value of the test statistics $\widetilde{q}_{obs}$ is obtained for every measurement x. To determine the probability of measuring a value of $\widetilde{q}_{obs}$ or greater, its underlying probability distribution $f(\widetilde{q}_\mu|\mu)$ is required. As $\widetilde{q}$ is constructed from the profile likelihood ratio, it is independent of the nuisance parameters $\boldsymbol{\theta}$ so that the conditional MLE $\hat{\hat{\boldsymbol{\theta}}}(\mu)$ provides a good estimate for the determination of $f(\widetilde{q}_\mu|\mu)$. This can be achieved by sampling from the full statistical model to create a "toy" dataset, or by using a single "Asimov" dataset [87]. In the latter case, the observed data is replaced by the Asimov data set, formally defined such that when it is used to evaluate the estimator for a parameter, one obtains the true parameter value.

After determining the $f(\widetilde{q}_\mu|\mu)$, the p-value describing the compatibility with the hypothetical signal strength μ is

$$CL_{S+B} = \int_{\widetilde{q}_{obs}}^{\infty} f(\widetilde{q}_\mu|\mu) d\widetilde{q}_\mu \tag{5.11}$$

Solving the above equation 5.11 for $CL_{S+B} = 0.05$ for μ, the frequentist upper limit is derived on μ at 95% confidence level. In case of statistical downward fluctuations, this upper limit would become arbitrarily small without being able to exclude $\mu = 0$ [88]. To account on this issue, one defines the ratio:

$$CL_S = \frac{CL_{S+B}}{CL_B} \tag{5.12}$$

where the denominator is the $p-$value of the consistency with the background-only hypothesis,

$$CL_B = \int_{\tilde{q}_{obs}}^{\infty} f(\tilde{q}_\mu|0) d\tilde{q}_\mu \tag{5.13}$$

The p-value for the observed data represents the probability of finding data of equal or greater incompatibility with the null hypothesis, under the assumption that the null hypothesis is true. Therefore if the p-value is smaller than a predefined threshold α, the null hypothesis can be considered as excluded at a confidence level (C.L.) of $1-\alpha$. From equation 5.13, the value μ, obtained by setting $CL_S = 0.05$, denotes the observed, frequentist upper limit above which the signal strength is excluded by the measurement at 95% confidence. When the analysis is "blinded", then two variants of upper limit can be used. The first, the "injected", which follows the above prescription, except that the measured data are replaced with simulation. On the contrary, the "expected" limit $f(\tilde{q}|\mu)$ is sampled from a statistical model that assumes the absence of signal. A value larger than one states that the background alone could explain the measurement, regardless of whether or not a signal exists. Therefore, the expected limit constitutes a sensitivity measure for blinded analyses.

5.4 Exclusion significance

The claim of an observation is predicated on the exclusion of the background-only hypothesis $\mu = 0$. Hence, the test statistic is redefined to [87]

$$\tilde{q}_0 = \begin{cases} -2\ln\tilde{\lambda}(\mu), & \hat{\mu} > 0 \\ 0, & \hat{\mu} \leq 0 \end{cases} \tag{5.14}$$

The $p-$value describing the probability of the measurement assuming $\mu = 0$ would be:

$$\int_{\tilde{q}_{obs}}^{\infty} f(\tilde{q}_0|0) d\tilde{q}_0 \tag{5.15}$$

The exclusion significance is often expressed in of Gaussian standard deviations:

$$p_0 = \int_{s}^{\infty} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx \tag{5.16}$$

solved for s. A significance greater $s = 3$ ($p_0 = 1.3 \times 10^{-3}$) is called an “evidence” for the existence of a new physics process, $s = 3$ ($p_0 = 1.3 \times 10^{-3}$) while marks the threshold for claiming an “observation”.

5.5 Profile likelihood fit

To determine the parameter of interest μ under a signal hypothesis, the negative log-likelihood (NLL) ,$-2\mathrm{lnL}(\mu,\boldsymbol{\theta}|\ x)$, normalized by $L(\hat{\mu},\hat{\boldsymbol{\theta}})$ is minimized. In a statistical analysis of a physics experiment this minimization is often described by a fitting procedure, with the MLE of μ referred as "best fit value".

In case of multiple substantial nuisances, parameter profiling can be applied. For a fixed value $\mu = \mu'$, the nuisance parameter values that would minimize the NLL, i.e., the conditional MLEs $\hat{\hat{\boldsymbol{\theta}}}(\mu')$, are determined and the corresponding minimum of the NLL is retained. This procedure is repeated for different values of μ' and represents a scan over possible values of μ, where the overall minimal NLL value yields the best fit value $\hat{\mu}$,

$$\hat{\mu} = \mathrm{arg}_{\mu}\mathrm{min} - 2\mathrm{lnL}(\mu, \hat{\theta}(\mu)|\mathrm{x}) \tag{5.17}$$

An example of a profile likelihood scan is shown in figure 5.1. The minimum of the resulting curve is shifted to zero as one is only interested in its position in terms of the parameter of interest. Therefore, and owing to the asymptotic χ^2-like behavior as discussed above, the generally asymmetric uncertainty on the best fit value is extracted through the intersection of the scanned curve with -$2\Delta\mathrm{lnL} = 1$.

In addition, the best fit values and uncertainties of arbitrary nuisance parameters can be calculated by hypothetically considering them as the parameter of interest in the scan and profiling out all other parameters, while optionally fixing the actual parameter of interest μ to its best fit value. An alternative, but computationally more demanding approach is the scan in multiple dimensions. The resulting a-posteriori, or “post-fit” nuisances θ_{post}, and especially nuisance pulls, i.e., the change of central nuisance values and their probability distribution widths $\Delta\theta$ with respect to the a-priori, or “pre-fit” expectation θ_{pre},

$$\mathrm{pull} := \frac{\theta_{post} - \theta_{pre}}{\Delta\theta_{pre}} \tag{5.18}$$

provide valuable insights to assess the validity of the underlying statistical model.

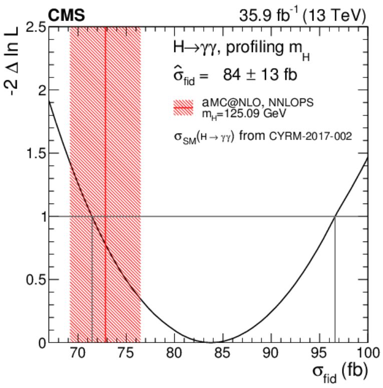


Figure 5.1: Likelihood scan (black curve) for the fiducial cross section measurement, where the value of the SM Higgs boson mass is profiled in the fit. The measurement is compared to the theoretical prediction (vertical red line), shown with its uncertainty (red hatched area), and it is found in agreement within the uncertainties.

Part II

Search for GGM in final states with two photons and missing transverse momentum

Chapter 6

Analysis strategy

6.1 Introduction

This chapter describes a search for GGMS breaking in events with exactly two photons and large p_T^{miss}. The target production mechanism is described by simplified models of gluino (T5gg) and squark (T6gg) pair production. In both models, the lightest supersymmetric particle (LSP) is the Gravitino ($\widetilde{G}$), which is taken to be nearly massless. The next-to-lightest supersymmetric particle (NLSP) is the neutralino, $\widetilde{\chi}_1^0$. Those models ensure R-parity conservation resulting the lightest supersymmetric particle (LSP) to be stable and weakly interacting. Therefore, proton-proton (p p) collisions that produce SUSY particles will have an imbalance in the total observed transverse momentum, referred to as missing transverse momentum p_T^{miss}. The models assume a 100% branching fraction for the NLSP neutralino to decay to a gravitino and a photon ($\tilde{\chi}_1^0 \to \tilde{G}\gamma$) resulting in characteristic events with large p_T^{miss} and two photons as shown in figure 6.1.

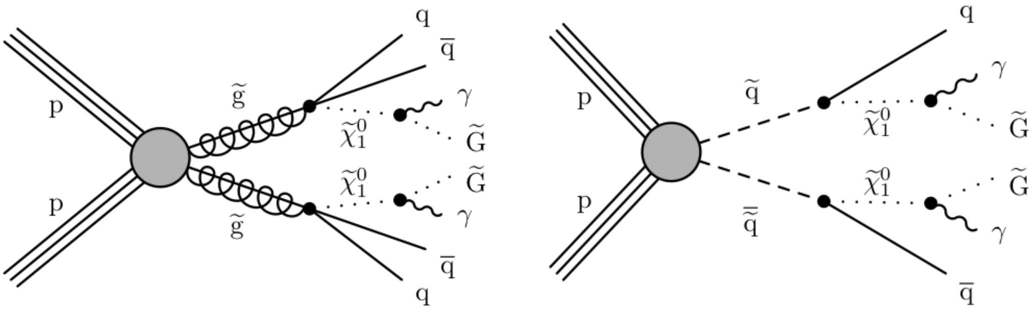


Figure 6.1: Diagrams showing the production of signal events in the collision of two protons. In gluino $\tilde{g}$ pair production (left), the gluino decays to an antiquark $\tilde{q}$, quark q, and a neutralino $\tilde{\chi}_1^0$. In squark $\tilde{q}$ pair production (right), the squark decays to a quark and a neutralino $\tilde{\chi}_1^0$. In both cases, the neutralino $\tilde{\chi}_1^0$ subsequently decays to a photon γ and a gravitino $\tilde{G}$.

6.2 Data and simulated samples

6.2.1 Data sets

The analysis was performed using data collected by the CMS detector in 2016 that correspond to an integrated luminosity of 35.9 fb^{-1} of pp collisions at a center-of-mass energy $\sqrt{s} = 13$ TeV. The data sets used in the analysis are listed in table 6.1. The triggers used to select interesting events for analysis are described in section 6.3 Those samples were reprocessed in February 2017 to take into account improved calibrations and performance corrections. After validating the quality of the data, CMS released a so–called "Golden JSON" [1] which is a summary of all run numbers which are good for data analysis.

The CMS experiment has developed a new analysis object format (”Mini-AOD”) [89] targeting approximately 10% of the size of the Run 1 AOD format. This format profits from sufficient information to serve about 80% of CMS analysis, while dramatically simplifying the needed disk and I/O resources. Both MC and data samples used in this analysis are in MINIAOD data format and include object candidates such as photons, electrons, jets, etc that come form the reconstruction algorithm described in chapter 3.

2016 data sets
/DoubleEG/Run2016B-03Feb2017_ver2-v2/MINIAOD
/DoubleEG/Run2016C-03Feb2017-v1/MINIAOD
/DoubleEG/Run2016D-03Feb2017-v1/MINIAOD
/DoubleEG/Run2016E-03Feb2017-v1/MINIAOD
/DoubleEG/Run2016F-03Feb2017-v1/MINIAOD
/DoubleEG/Run2016G-03Feb2017-v1/MINIAOD
/DoubleEG/Run2016H-03Feb2017_ver2-v1/MINIAOD
/DoubleEG/Run2016H-03Feb2017_ver3-v1/MINIAOD

Table 6.1: list of data sets that used in the diphoton SUSY search.

6.2.2 Monte Carlo simulation samples

To simulate the expected SUSY signal distributions, two signal Monte Carlo simulation samples were produced. The T5gg simulated sample was produced in bins of gluino and neutralino mass. The simulated signal samples were generated with MADGRAPH_aMC@NLO and generated with either two gluinos or two squarks and up to two additional partons in the matrix element calculation. The parton showering,

[1] This analysis is based on the follow golden JSON:
Cert_271036-284044_13TeV_23Sep2016ReReco_Collisions16 JSON.txt.

hadronization, multiple-parton interactions, and the underlying event were described by the PYTHIA8 program with the CUETP8M1 [90] generator tune.

For each gluino and neutralino mass bin, a total of 40,000 events were produced, except for bins with gluino or squark masses above 2.0 TeV, where only 20,000 events were produced per bin. For gluino masses from 1,400 to 2,500 GeV, events were generated in bins of 50 GeV. In the T6gg simulated sample, the squark masses ranged from 1,400 GeV to 2,050 GeV in bins of 50 GeV. The neutralino masses ranged from 10 GeV up to the mass of the gluino or squark and were binned in 100 GeV segments. Finer binning was used in the compressed region where $M_{\tilde{\chi}_1^0}$ is within 300 GeV of $M_{\tilde{g}}$ or $M_{\tilde{q}}$, and in the region with low $M_{\tilde{\chi}_1^0}$. These mass ranges were selected to overlap and expand upon the mass ranges excluded by previous searches. The parton distribution functions are obtained form NNPDF3.0. The cross sections are calculated at next-to-leading order (NLO) accuracy including the resummation of soft gluon emission at next-to-leading-logarithmic (NLL) accuracy, with all the unconsidered sparticles assumed to be heavy and decoupled [91]. Apart from signal simulation, background simulation processes were used as well, to validate the analysis performance and to estimate subdominant backgrounds. The full list of MC samples used in the analysis are summarized in table 6.2 for the signal and in table 6.3 for background.

The detector response to particles produced in the simulated collisions is modeled with the GEANT4 detector simulation package for SM processes. Because of the large number of SUSY signals considered, with various gluino, squark, and neutralino masses, the detector response for these processes is simulated with the CMS fast simulation that generally agree with the results from the full simulation up to 10$\%$ [92].

SUSY signal samples
/SMS-T5Wg_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/ RunIISpring16MiniAODv2-PUSpring16Fast_80X_mcRun2_asymptotic_2016 miniAODv2_v0-v1/MINIAODSIM
/SMS-T5Wg_mGo2150To2500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/ RunIISpring16MiniAODv2-PUSpring16Fast_80X_mcRun2_asymptotic_2016 miniAODv2_v0-v1/MINIAODSIM
/SMS-T6Wg_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/ RunIISpring16MiniAODv2-PUSpring16Fast_80X_mcRun2_asymptotic 2016_miniAODv2_v0-v1/MINIAODSIM
/SMS-T6Wg_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/ RunIISpring16MiniAODv2-PUSpring16Fast_80X_mcRun2_asymptotic 2016_miniAODv2_v0-v1/MINIAODSIM

Table 6.2: Signal MC samples

6.3 Analysis triggers

The analysis phase space is partly driven by the availability of photon triggers. In particular two diphoton triggers developed for the search of Higgs boson decaying

Background MC samples
/GJet_Pt-40toInf_DoubleEMEnriched_MGG-80toInf_TuneCUETP8M1_13TeV_Pythia8/RunIISummer16MiniAODv2- PUMoriond17_80X_mcRun2 asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM
/ZGGToNuNuGG_5f_TuneCUETP8M1_13TeV-amcatnlo-pythia8/RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM
/DYJetsToLL_M-50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/RunIISpring16MiniAODv2-PUSpring16_80X_mcRun2_asymptotic_2016_miniAODv2_v0_ext1-v1/MINIAODSIM

Table 6.3: Background MC samples

into a pair of photons ($H \rightarrow \gamma\gamma$) are used. The primary trigger used for both signal and control regions, requires the diphoton invariant mass to be greater than 90 GeV. In addition, a control trigger that was designed to collect $Z \rightarrow ee$ events is used to construct control regions with electrons. The triggering strategies and studies of their efficiencies are discussed in the following sections.

6.3.1 Trigger requirements

The list of triggers used to accumulate the events in the 35.9 fb^{-1} data sample is showed in table 6.4. Since the analysis uses only $|\eta| < 1.4442$ photons only the trigger requirements in the barrel will be discussed. The variables used in the trigger are defined in previous sections 3.2, 3.4.3 and are described below.

Trigger paths
signal trigger: HLT_Diphoton30_18_R9Id_OR_IsoCaloId_AND_HE_R9Id_Mass90_v*
control trigger: HLT_Diphoton30_18_R9Id_OR_IsoCaloId_AND_HE_R9Id_DoublePixelSeedMatch_Mass70_v*

Table 6.4: Trigger paths for 2016 diphoton analysis

Diphoton_30_18: The triggers require two photons with an asymmetric E_T cut of 30 GeV and 18 GeV on the leading and subleading photons, respectively. The transverse energy E_T of a photon is defined as the magnitude of the projection of the photon momentum on the plane perpendicular to the beams.

R9Id: Photons satisfy $R_9 > 0.85$, where R_9 is a measure of the overall transverse spread of the shower. (details can be found in section 3.4.3)

IsoCaloId:

- $\sigma_{i\eta i\eta} < 0.015$: Photons are required to have a log-fractional weighted shower width in iη-space less than 0.015.
- ECAL isolation $< (6 + 0.012 * \text{Photon}\ \ E_T)$
- Track isolation $< (6 + 0.002 * \text{Photon}\ \ E_T)$

HE_R9Id:

- $R_9 > 0.5$
- $H/E < 0.1$

All photons are required to pass the H/E and loose R_9 cuts in HE_R$_9$Id, and either the tighter R_9 cuts in $R_9 Id$ or the isolation and shape cuts in IsoCaloId. The leading leg of the filter requires the photon candidate to be matched to an L1 seed. It can be matched to one of several SingleEG and DoubleEG L1 filters, but the largest contribution comes from the lowest unprescaled triggers: namely, SingleEG40 and DoubleEG_22_15. Both photons must satisfy the trailing filter, which is unseeded. In addition to the cuts listed above, the invariant mass of the diphoton system is required to be greater than 90 GeV.

The control trigger shares all of the same requirements as the primary trigger, with two exceptions: the invariant mass of the two electromagnetic objects is required to be greater than 70 GeV rather than 90 GeV, and both electromagnetic objects are required to be matched to a pixel seed. A pixel seed is defined as at least two hits in the pixel detectors that are consistent with the location of the energy deposit in the ECAL.

6.3.2 Trigger efficiency

In order to exploit a fully efficient trigger, the estimation of the overall trigger efficiency is necessary. The ECAL has a similar response in electrons and photons and thus, the trigger efficiency can be calculated from $Z \rightarrow ee$ events in data using the tag-and-probe method. In this method, two electron candidates are required. One serves as the “tag” and is required to pass looser photon identification criteria compared to the second electron candidate. The second electron candidate serves as the “probe” and has to satisfy the same selection criteria as our offline photon identification (see section 3.4.3). In order to ensure a high purity of electromagnetic objects, the invariant mass of the di-electron system must be between 75 and 105 GeV. The single electron control trigger HLT_Ele27_WPTight_Gsf is used as the trigger to collect the "tag" events. The efficiency ϵ of the HLT path or trigger filters that is being studied is given by the following equation,

$$\epsilon_{trig} = \frac{N_{pass}}{N_{total}} \tag{6.1}$$

where N_{total} is the total number of tag and probe pairs passing all requirements, and N_{pass} is the number of tag and probe pairs in which the probe passes the trigger filter.

The L1 seed is a combination of several prescaled and unprescaled L1 filters that evolved differently during the 2016 data period. Therefore, the L1 efficiency is difficult to calculate. Instead, the efficiency of the seeded leg was calculated without explicitly requiring the probe to match to an L1 seed. This way, a total efficiency with which photon candidates pass both the L1 seed and the leading leg of the HLT path is derived.

The efficiency is calculated as a function of photon p_T as showned in the upper part of figure 6.2. The efficiency was fit to an error function to calculate the overall efficiency at the plateau. For photon $p_T > 40$ GeV, the leading filter is 98.2% efficient.

Tag and probe objects for the trailing leg efficiency must pass the same set of requirements as those used in the leading leg efficiency calculation, with the additional requirement that the tag must pass the leading filter. This requirement arises from the way HLT modules are structured. Filters are processed sequentially, and if an event fails one filter, the subsequent filters are skipped. The lower part of figure 6.2 shows the efficiency of the trailing filter as a function of photon p_T. For $p_T > 40$ GeV, the trigger is 99.8% efficient.

The efficiency of the trigger is calculated as well with respect to the diphoton invariant mass. For this calculation, we required two photons passing our analysis selection criteria, two photons satisfying the trailing leg of the trigger, and one photon passing the leading leg of the trigger. The efficiency was given by the number of diphoton events passing the full HLT path over the total number of diphoton events passing our requirements. The efficiency of the trigger as a function of invariant mass s shown in Figure 6.3. For $m\gamma\gamma > 100$ GeV, the trigger is 99.4% efficient.

The efficiency of the trigger as a whole is the product of all three efficiencies. Two factors of the trailing leg efficiency are needed since both photons are required to pass that leg:

$$\epsilon_{trig} = \epsilon_{lead} \times \epsilon_{trail}^2 \times \epsilon_{mass} = 97.2\% \tag{6.2}$$

In the double electron trigger, a pixel seed cut is applied only to the trailing leg of the trigger. This additional requirement results in a significantly lower overall efficiency for this leg. This results in an overall trigger efficiency of 79.8% for the control trigger.

6.4 Object definition

The analysis search region require large p_T^{miss}, and two energetic photons. Apart from that, a collection of electrons and jets that mimic the response of the photon referred as "fakes" are also constructed. In this section the object definition and the offline cuts to achieve high purity, are described.

6.4.1 Photons

In order to guarantee a fully efficient trigger, photons with $p_T > 40$ GeV are considered. Moreover, the considered SUSY events are expected to be produced in the central region of the detector, and thus only photons in the fiducial region of the ECAL barrel, $|\eta| < 1.4442$, are taken into account. A photon candidate is identified as a photon, if it pass the cut-based photon ID derived by the CMS e/γ Physics Object Group (EGM POG). For this analysis we use the medium working point which has an efficiency of 80% to reconstruct a real photon. The variables and the set of cuts that consist the medium working point are described in detail in section 3.4.3. In addition, photons must satisfy $R_9 > 0.5$. This is done to ensure that the trigger, which has an R_9 requirement, is fully efficient. As already mentioned, electrons and photons have similar detector response. Therefore, to distinguish photon candidates from electrons, photons are required to pass a pixel seed veto (PSV), which means that photons must not

ID Requirement	Photons	Electrons	Fakes
Pixel seed veto	Applied	Reversed	Applied
$\sigma_{i\eta i\eta}$	< 0.01022		$0.01022 < \sigma_{i\eta i\eta} < 0.015$
Charged hadron isolation	< 0.441		$0.441 < Iso < 25$
Photon Isolation	$< 2.571 + 0.0047 p_T$		$< 15 + 0.0047 p_T$
Neutral hadron Isolation	$< 2.2725 + 0.0148\ p_T$		$< 15 + 0.0148 p_T + 0.000017 p_T^2$
R_9	> 0.5		$0.5 < R_9 < 0.9$
H/E	< 0.0396		< 0.0396

Table 6.5: Photon, electron and fake requirements.

be matched to a pixel seed. The pixel seed is defined as at least two hits in the pixel detectors consistent with a charged particle trajectory which would arrive at the ECAL.

6.4.2 Electrons

Due to the similarities between photons and electrons, the latter provides an excellent candidate to construct control regions. The electron collection satisfies all the photon requirements but fails the PSV. This way, the reconstructed control region is orthogonal to the signal region with no bias between the control region and the candidate diphoton sample we are trying to model.

6.4.3 Fakes

In addition to photons and electrons, a third, orthogonal category referred to as “fake” photons is defined. Fakes are primarily electromagnetically-rich jets that have been misidentified as photons. This collection is used to predict the QCD background from the data. To construct the fake orthogonal collection a set of cuts that describe the photon ID sideband is used. More specifically, fakes are required to fail either $\sigma_{i\eta i\eta}$ or the charged hadron isolation requirement of the photon ID. In order for the fake definition to be more "signal-like", upper bounds (0.015 and 25 for $\sigma_{i\eta i\eta}$ and charged hadron isolation, respectively) are set on both variables. In addition, fakes must satisfy $0.5 < R_9 < 0.9$. This particular cut is motivated by the fact that real photons in SUSY evens will have an R_9 close to unity, and thus this cut will prevent any possible SUSY signal to enter in our control region. In order to counteract the effect this cut and to ensure enough statistics on the control sample the neutral hadron isolation and photon isolation cuts are loosened significantly. The requirements of all the objects are summarized in table 6.5.

6.4.4 Object cleaning

To avoid double counting particles, all objects are cross cleaned. Since the muons are reconstructed with a higher purity than any other particle, they are used in the first step of the cleaning procedure: any electromagnetic object (photon, electron, or fake) that is within $\Delta R < 0.3$ of a muon candidate is removed. After that, any photon that overlaps within $\Delta R < 0.3$ of an electron is removed and finally, if a fake overlaps with an electron or photon candidate within $\Delta R < 0.4$, the fake candidate is removed. Fakes are primarily jets reconstructed using the anti-kT algorithm with a distance parameter of 0.4, and thus a larger ΔR separation is necessary.

6.4.5 Lepton Veto

The targeted SUSY signatures do not require the presences of a lepton. Therefore, any event that contains additional muons or electrons is vetoed. This cut does not hurt the signal sensitivity, and makes this SUSY analysis orthogonal to other CMS searches for gauge-mediated supersymmetry breaking with photons and leptons in the final state.

6.4.6 Signal and Control regions

After every object in each event is identified, then the event could be classified into one of four mutually exclusive categories based on the electromagnetic objects with the highest p_T in the event. Events with two photons passing the above criteria consist the $\gamma\gamma$ sample. Events with two electrons, one electron and one photon, or two fakes are categorized as ee, $e\gamma$, or ff events, respectively. In each case, the two electromagnetic objects are required to be separated by $\Delta R > 0.6$. The $\gamma\gamma$, ff, and eγ samples are required to pass the primary trigger. To fulfil the trigger requirements, the invariant mass of the two electromagnetic objects is required to be greater than 105 GeV. For the ee sample, the control trigger was used to collect events. In order to succeed a high purity electron sample, the invariant mass of the di-electron pair had to be in the window of 75 and 105 GeV.

6.5 Photon selection efficiency

6.5.1 Photon ID scale factor

The photon identification efficiency is calculated via a tag and probe method using $Z \to ee$ events. The procedure is similar to the one used for the trigger efficiency calculation described in section 6.3.2. $Z \to ee$ are collected using a single electron trigger. One electron is served as the tag and the other as the probe. The efficiency is then given by the number of probes that pass the photon ID over the total number of tag and probe pairs. The efficiency is derived for both data and simulation and the ratio of the two efficiencies is referred to as a scale factor. The official scale factors calculated by the EGM POG for Moriond 2017 were used [93]. Those scale factors are calculated in bins of photon p_T and η. Figure 6.4 shows the scale factors and the corresponding uncertainties in bins of p_T and η.

Instead of using the full map of scale factors, a weighted average over all photons passing our selection criteria in each SUSY signal mass point is chosen. The average scale factors and uncertainties are shown in figure 6.5.

The photon ID used in this analysis differs from the official POG recipe in one aspect: we apply an $R_9 > 0.5$ requirement on top of the medium ID due to the presence of an R_9 cut in our analysis trigger. An additional check was performed to see the effect on this cut. The data/MC scale factors were all consistent with unity and thus, the official scale factors were used.

The final value used in the analysis was:

$$\text{Photon Scale Factor} = 1.002 \pm 0.013 \tag{6.3}$$

6.5.2 Scale factor for pixel seed veto

The prescription for photon identification differs on the electron identification only of the presence or absence of a seed track on the pixel detector. For this reason, the efficiency of the pixel seed veto for photons cannot be determined from the tag-and-probe method described above and must be obtained from photons in $Z \to \mu\mu\gamma$ events. The official scale factor calculated by the EGM POG is :

$$\text{Pixel Seed Factor} = 0.998 \pm 0.013 \tag{6.4}$$

Since our candidate sample requires two photons in the final state, two factors of both values are used.

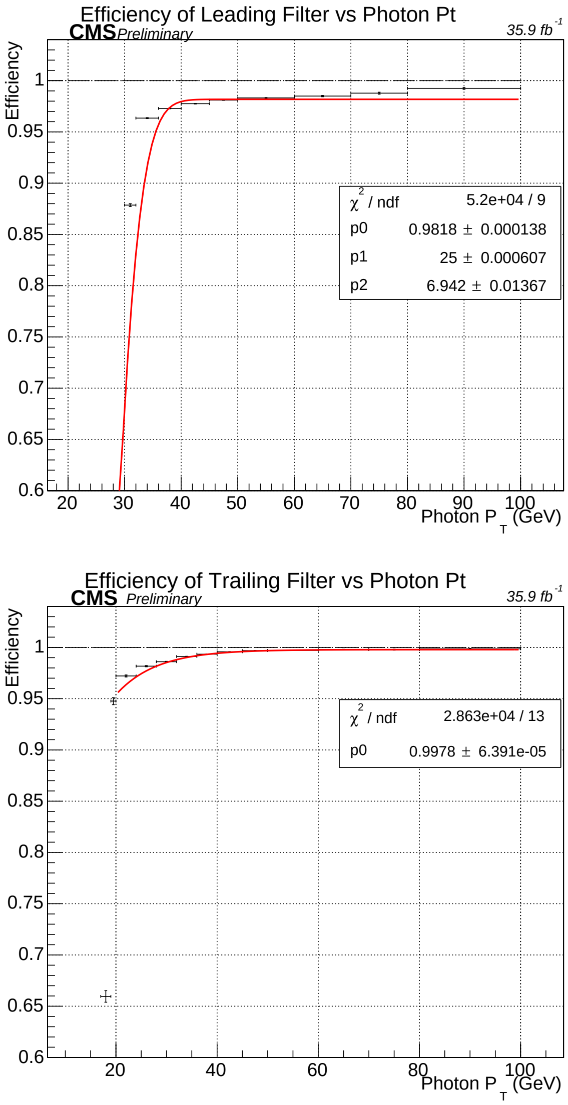


Figure 6.2: Efficiency of the leading (upper part) and sub-leading (down part) legs as a function of photon p_T. The leading leg includes the efficiency of the L1 seed requirement. For photon p_T above 40 GeV, the leading leg is 98.2% efficient and the sub-leading leg is 99.8% efficient),

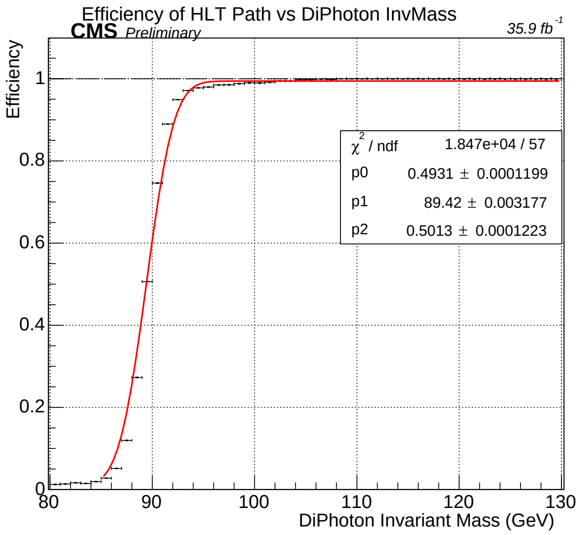


Figure 6.3: Trigger efficiency as a function of the invariant mass of the diphoton system. Above 100 GeV, the trigger is 99.4% efficient.

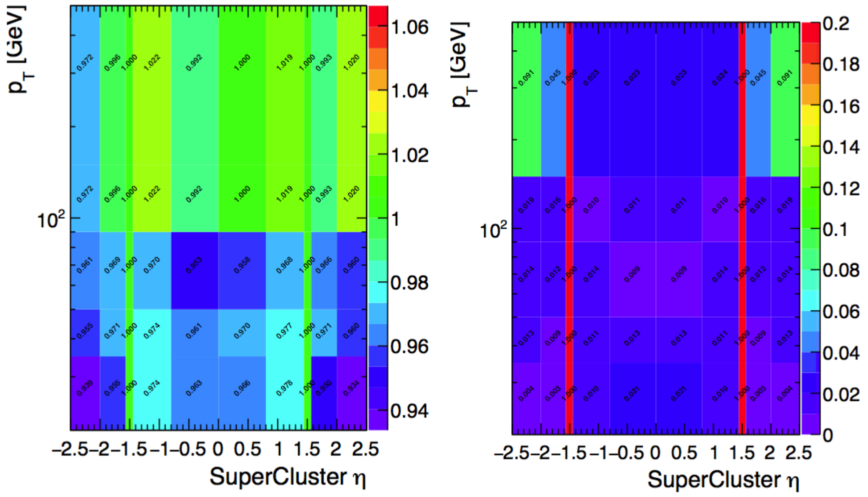


Figure 6.4: Derived scale factor (left) and uncertainties (right) in bins of photon p_T and η

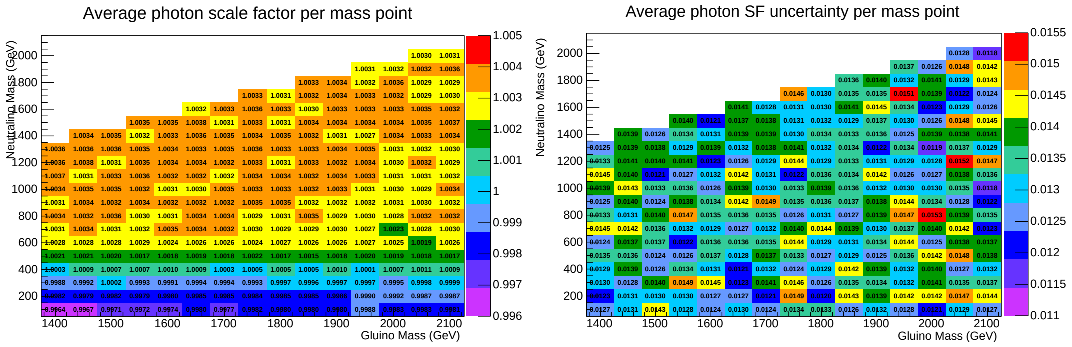


Figure 6.5: Scale factors (left) and uncertainties (right) averaged over all photons in each bin in the neutralino versus gluino mass plane.

Chapter 7

Susy analysis

7.1 Signal region

This analysis is based on a search for an excess of events with two energetic photons and significant p_T^{miss}. Therefore, the signal region is made up of events with $p_T^{miss} > 100$ GeV and two photons passing the criteria described in section 6.4.1. The signal region is divided into six p_T^{miss} bins, with bin boundaries at 100, 115, 130, 150, 185, and 250 GeV. The last bin includes all events with $p_T^{miss} > 250$ GeV.

7.2 Background overview

Accurate estimation of the background from the Standard Model processes is an essential part of the analysis, as we aim to distinguish a possible small excess caused by signal events from this background. There are several standard model processes that can mimic our signal events. The dominant one is by far from quantum chromodynamics (QCD). Those processes include primarily multi-jet events, where electromagnetically-rich jets are misidentified as photons. Other processes include events with true photons either from associated photon production or initial-state radiation. The main characteristic of this processes is that they lack of genuine p_T^{miss}, where its appearance comes from instrumental effects. This background is derived directly from the data (data-driven) using a control region from a side-band of the photon ID. Another background contribution comes from the electroweak (EWK) background. This background include $W\gamma$ or $W + jet$ events where W decays leptonically ($W \to e\nu$). Those processes have inherent p_T^{miss} from the neutrino and can mimic the signal topology when an electron is misidentified as a photon. This background is estimated as well, in a data-driven way by using an $e\gamma$ control sample to measure the misidentification rate in data. Finally, there is an irreducible background from $Z\gamma\gamma \to \nu\nu\gamma\gamma$ events. This contribution is small, and thus is modeled from simulation.

7.3 QCD background

As described above, direct diphoton production with initial state radiation and multijet events lack genuine p_T^{miss}, but can emulate GGM signal topologies if the hadronic activity in the event is poorly measured. In the latter case, photons may be reconstructed in the event as a result of the misidentification of electromagnetically rich jets. Due to its large cross section the most significant background for this analysis consist of those events. The contribution of QCD background in the signal region is estimated by a control region of "fake" objects as described in section 6.4.3. The fake collection is orthogonal to the nominal photon collection and thus this sideband can be used to construct control regions. More specifically, the p_T^{miss} tail of the QCD background is modeled using a "fake-fake" (ff) control sample made up of events with two fakes. In addition, a $\gamma + fake$ sample was used to perform a data-driven closure test on the QCD prediction.

7.3.1 Ratio Method

To model the QCD background we rely on the observation that we expect the relative fraction of $\gamma\gamma$ and ff events to vary smoothly with p_T^{miss}, so that their ratio can be modeled with a simple function. The functional form for the $\gamma\gamma$/ff ratio with respect to p_T^{miss} can be extrapolated from the control region into the signal region to get an estimate of the QCD background.

In case the ff control sample has the same composition as the candidate $\gamma\gamma$ sample, no sensitive dependence on p_T^{miss} is expected, and thus the ratio will be flat as a function of p_T^{miss}. However, the R_9 requirement on the fake objects results in a lower purity in the ff control sample than the candidate sample. This leads to more ff events at high p_T^{miss} relative to the $\gamma\gamma$ sample. This is clearly illustrated in figure 7.1. Different functional forms were investigated to model the p_T^{miss} dependence, and an exponential function was found to describe best the data. The $\gamma\gamma$ to ff ratio is fitted with the function of the form $p_0 e^{-p_1 x}$ in the $p_T^{miss} < 100$ GeV control region as shown in figure 7.1. The predicted number of QCD background events (N^i_{QCD}) in bin i of the signal region is then given by the following equation.

$$N^i_{QCD} = g^i_{ave} N^i_{ff} \tag{7.1}$$

where N^i_{ff} is the number of observed ff events and g^i_{ave} is the average value of the fit function $g(p_T^{miss})$ in that bin.

7.3.2 Validation methods and Closure test

To validate and to assign a systematic uncertainty on the ratio method, several cross checks were done. The additional checks on the data rely on different "fake" object definitions that are used to estimate the QCD background. In addition, different closure tests, in the data and simulation, were performed to check the behavior of the $\gamma\gamma$/ff ratio. All those tests are described in detail in the following sections.

Cross check using high-R_9 ff sample

The first cross check uses an alternative fake definition to derive a second QCD background prediction, by noting that the p_T^{miss} distribution of the ff control sample is dependent on the R_9 requirement on the misidentified photons. More specifically, a control sample with two fakes satisfying all the normal requirements, except the R_9 which is required to be greater than 0.9, is used. The so-called high-R_9 ff control sample has a purity which is more representative of the composition in the $\gamma\gamma$ candidate sample. However, this control sample cannot be used as the nominal control sample

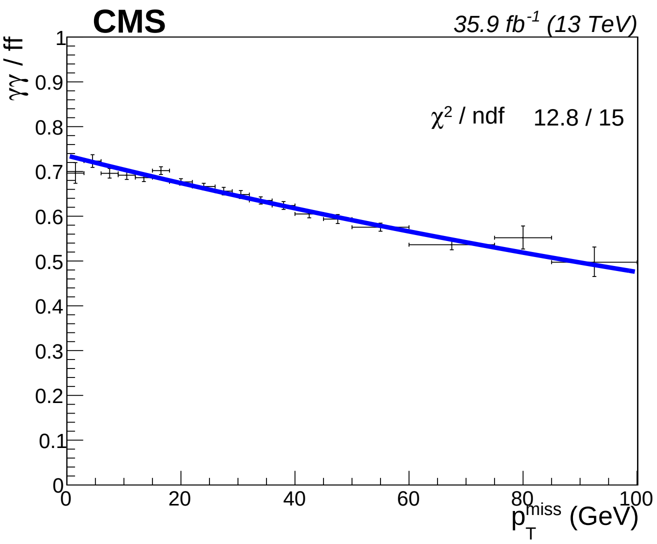


Figure 7.1: Ratio of the number of events in the candidate $\gamma\gamma$ sample to the ff control sample as a function of in the low-p_T^{miss} control region. The ratio is fit to an exponential function of the form $p_0 e^{-p_1 x}$.

due to the expected signal contamination in the signal region. This is rather used to derive a correction factor to account for the effect that the R_9 requirement on the shape of the p_T^{miss} distribution. Firstly, the ratio of the nominal (low-R_9 ff) and the high-R_9 ff control sample is fitted with an exponential function with respect to p_T^{miss} as showed in figure 7.2 for the $p_T^{miss} < 100$ GeV control region. After deriving the corrected ff sample, we construct the ratio of $\gamma\gamma$ to corrected ff as a function of p_T^{miss}. This ratio is then fitted with a constant in region $p_T^{miss} < 100$ GeV as shown in figure 7.3. The size of the correction is between 20 and 40% in the p_T^{miss} signal region.

Table 7.1 compares the QCD predictions derived using both methods. The uncertainties include the 1 σ fit uncertainties and the ff statistical uncertainties. The two methods agree within uncertainties in all bins. The difference between the two methods or the uncertainty on that difference (whichever is larger) is taken as a systematic uncertainty on the primary QCD prediction.

p_T^{miss}(GeV)	Ratio Method	High-R_9 correction method
100-115	$99.0^{+7.4}_{-7.0}$	$93.1^{+7.0}_{-6.6}$
115-130	$32.8^{+4.3}_{-3.8}$	$30.3^{+4.0}_{-3.5}$
130-150	$18.8^{+3.2}_{-2.8}$	$17.1^{+2.9}_{-2.5}$
150-185	$9.9^{+2.3}_{-1.9}$	$8.9^{+2.0}_{-1.7}$
185-250	$3.1^{+1.3}_{-0.9}$	$2.7^{+1.1}_{-0.8}$
≥ 250	$1.0^{+0.8}_{-0.5}$	$0.8^{+0.7}_{-0.4}$

Table 7.1: QCD background predictions using the Ratio Method and the High-R_9 correction method.

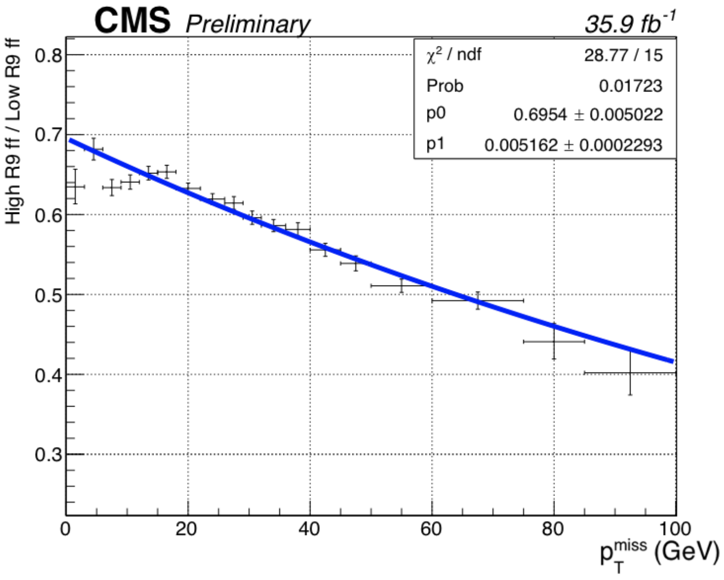


Figure 7.2: Ratio of the number of events in the high-R_9 ff sample to the nominal low-R_9 ff control sample in the low-p_T^{miss} control region. The ratio is fit to an exponential function of the form $p_0e^{-p_1x}$. The function is then used to correct the nominal ff sample to account for the effect of the R_9 cut on the p_T^{miss} shape

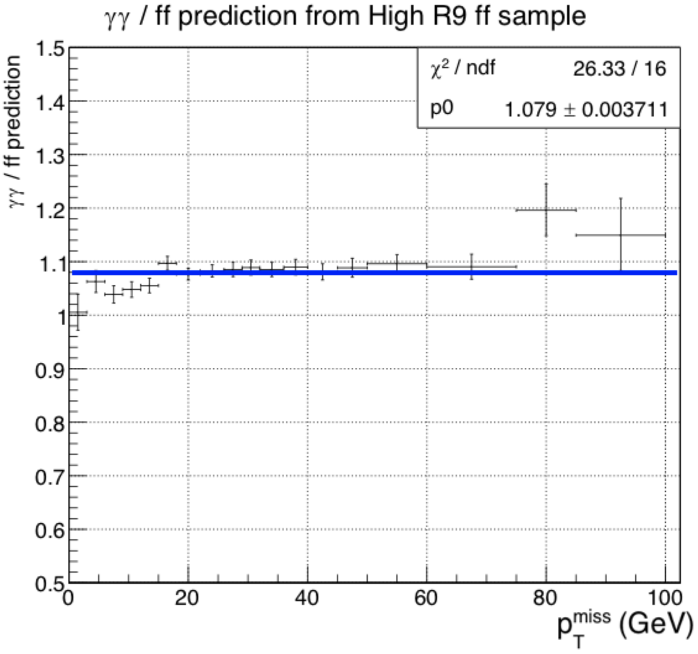


Figure 7.3: Ratio of the number of events in the candidate $\gamma\gamma$ sample to the corrected ff sample as a function of p_T^{miss} in the p_T^{miss} control region. The ratio is fitted to a constant function and used to derive an alternative QCD background estimate

Cross check using mixed-R_9 ff sample

Due to SUSY signal contamination of the high-R_9 ff sample in the signal region, we are forced to perform the above test only in the low-p_T^{miss} control region. In order to get an estimation of the R_9 requirement in the signal region, we need a different control sample without any signal contamination. Therefore, a third sample referred to as the mixed-R_9 ff sample is constructed from events with one fake object satisfying $R_9 > 0.9$ and one fake object satisfying $R_9 < 0.9$. The signal contamination of this sample is negligible and thus we are able to look at the full distribution of p_T^{miss} as showed in figure 7.4. The low-p_T^{miss} control region is fitted as previously with the function of the form $p_0 e^{-p_1 x}$, however, the exponential fit continues to describe the shape of the ratio well in the signal region.

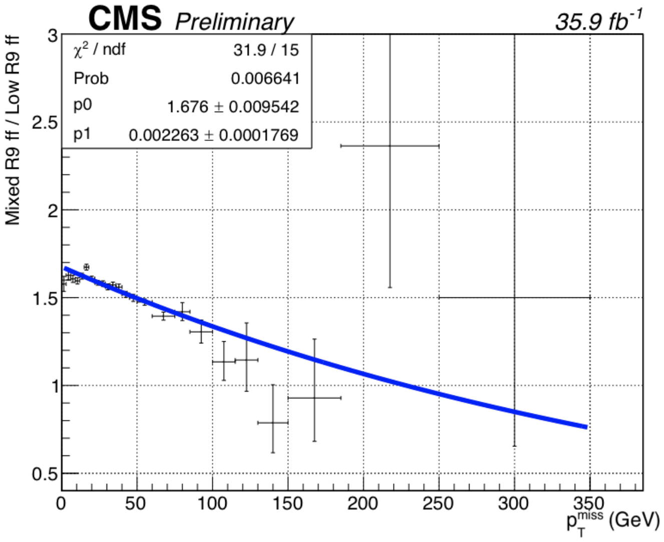


Figure 7.4: Ratio of the number of events in the mixed-R_9 ff sample to the nominal (low-R_9) ff control sample as a function of p_T^{miss}. The ratio is fit to an exponential function of the form $p_0 e^{-p_1 x}$ for p_T^{miss} GeV. Because the signal contamination is negligible in both samples, the ratio is shown for the full p_T^{miss} distribution. The exponential fit continues to describe the ratio well in the signal region.

A comparison in the signal region between the observed number of events in the mixed-R_9 ff sample and the predicted number of events derived by applying the exponential fit is showed in table 7.2. The uncertainties on the prediction include the fit and statistical uncertainties, and the uncertainty on the observed values are statistical only.

7.3.3 Closure test using $\gamma+$ fake control sample

Another closure test of the method was done by using a substitute for the diphoton candidate sample. This control sample consists of one photon and one fake satisfying the normal $R_9 < 0.9$ requirement. This control sample includes processes with true p_T^{miss}. The principal contributions are $W+jet$ events where the W decays to an electron and a neutrino and the electron is misidentified as photon. To subtract this contribution

p_T^{miss}(GeV)	Mixed -R_9 prediction	Mixed-R_9 observed
100-115	$282.5^{+21.0}_{-19.6}$	$245^{+16.7}_{-15.6}$
115-130	$96.5^{+12.5}_{-11.2}$	$87^{+10.4}_{-9.3}$
130-150	$57.4^{+9.7}_{-8.4}$	$37^{+7.1}_{-6.1}$
150-185	$32.1^{+7.3}_{-6.1}$	$26^{+6.2}_{-5.1}$
185-250	$11.3^{+4.5}_{-3.4}$	$20^{+6.4}_{-5.2}$
≥ 250	$3.8^{+3.0}_{-1.8}$	$6^{+3.6}_{-2.4}$

Table 7.2: Closure test with Mixed-R_9 ff events. A comparison in the signal region between the observed number of events in the mixed-R_9 ff sample and the predicted number of events derived by applying the exponential fit.

from the $\gamma f p_T^{miss}$ distribution we apply the electron-photon misidentification rate to an electron+fake control samples. The $\gamma f/ff$ ratio is shown in figure 7.5. The last bin in the signal region has been omitted due to potential SUSY signal contamination. In Table 7.3, the number of observed and predicted events for the γf sample up to $p_T^{miss} = 250$ GeV are shown. The uncertainties on the prediction include the fit uncertainties and the statistical uncertainties.

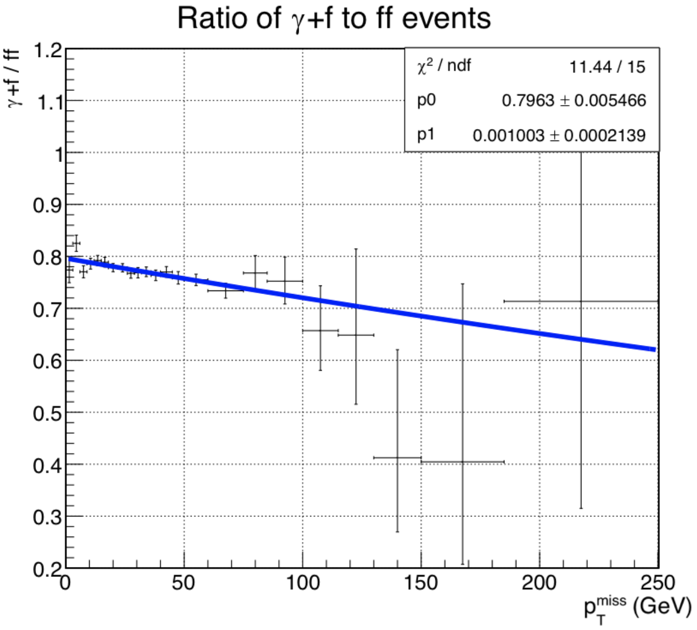


Figure 7.5: Ratio of the number of events in the γ f sample to the nominal (low-R_9) ff control sample as a function of p_T^{miss}. The ratio is fit to an exponential function of the form $p_0 e^{-p_1 x}$ for $p_T^{miss} < 100$ GeV and used to predict the number of γ f events in the signal region. The last bin is omitted due to potential signal contamination in the γ f sample.

p_T^{miss}(GeV)	Predicted	Observed
100-115	$154.4^{+10.5}_{-11.2}$	141.9
115-130	$53.5^{+6.9}_{-6.1}$	49.3
130-150	$32.5^{+5.5}_{-4.7}$	19.4
150-185	$18.8.1^{+4.3}_{-3.5}$	11.3
185-250	$7.3^{+2.9}_{-2.2}$	7.8

Table 7.3: Closure test using γf sample.

7.3.4 $\gamma\gamma/ff$ ratio in simulation

The behavior of the $\gamma\gamma/ff$ ratio as a function of p_T^{miss} was also checked in simulation using Double EM Enriched QCD and GJet MC samples. In order to have enough statistics, samples were built from events with exactly one photon or exactly one fake plus a jet. This ratio is illustrated in figure 7.6 where there is not the same dependence on p_T^{miss} as what is observed in data. Instead, the ratio is flat as a function of p_T^{miss}, which is consistent with what we would expect in the case where the ff distribution has the same composition as the candidate sample. The R_9 variable is difficult to model, and it is unsurprising that the effect of R_9 on p_T^{miss} is not captured in simulation.

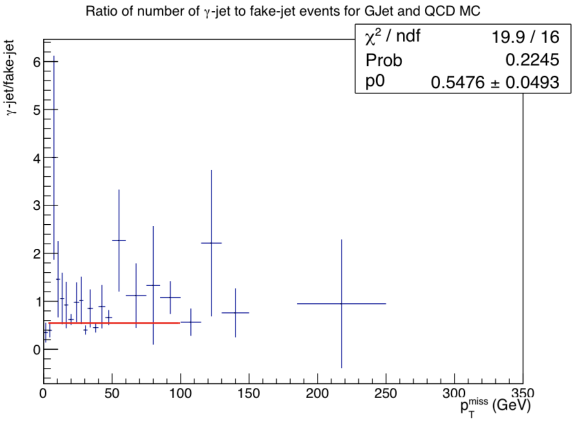


Figure 7.6: Ratio of the number of γ + jet events to fake + jet events in QCD and GJet simulation. The ratio has been fit to a constant.

7.4 Electroweak background

As mentioned in previous sections, isolation cone energies and shower shape in the calorimeters are similar for photons and electrons. The distinguishable feature used to separate between electrons and photons is the tracker activity. Photons do not interact with the silicon detectors and hence leave no track. Since electron are ionizing particles, they are registered by the tracking system by producing electron-hole pairs in the silicon sensors. In this analysis, the number of pixel seeds is used to discriminate between electrons and photons. Few real photons are discarded this way, but some electrons may be misidentified as photons by the inefficiency of the pixel detector. Those kind of events consist the subdominant background for this analysis also known as electroweak (EWK) background. The EWK background comes from primarily $W\gamma$ and W+jet events where W decays leptonically and the electron is misidentified as a photon. Unlike the QCD background, the p_T^{miss} is inherent and comes from the neutrino that remains undetected. To estimate this effect, the probability of an electron faking a photon referred to as fake rate is measured in a control region. This is done by comparing the invariant mass peak in a double electron sample (ee) with the invariant mass peak in a sample of events with one electron and one photon $(e\gamma)$. The composition of the $e\gamma$ sample and the calculation of the misidetification rate are described in detail in the following sections.

7.4.1 Composition of the $e\gamma$ sample

The $e\gamma$ control sample consists of γ+jets or $W\gamma$ events. In figure 7.7 a comparison is shown for data and simulation. The distribution of the data is fitted using γ +jet and $W\gamma$ simulated templates. From the fit, we derive that 80% of the $e\gamma$ events observed in data are from γ +jet processes and the remaining consist of $W\gamma$ events. However, in the signal region where $p_T^{miss} > 100$ GeV, $W\gamma$ events dominate. For $100 < p_T^{miss} < 130$ GeV, 12.1% are γ + jet events. For $p_T^{miss} > 130$ GeV, γ + jet events contribute only 2.6%.

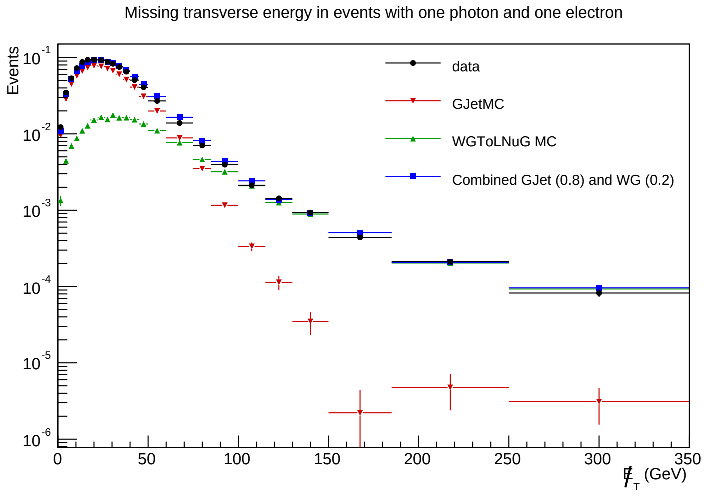


Figure 7.7: Data versus Monte Carlo comparison of the eγ control sample p_T^{miss} distribution. To determine the relative contributions of the γ+jets and $W\gamma$ processes, the data distribution was fit using the γ+jets and $W\gamma$ shapes as templates. The data are shown in black, and the total MC prediction is shown in blue. The γ+jets MC (red) was scaled to 80% of the observed events in data, and the $W\gamma$ MC (green) was scaled to 20% of the data distribution. The data correspond to an integrated luminosity of 35.6 fb^{-1}.

7.4.2 Determination of Fake Rate

To estimate the electroweak background, we need to extract the rate at which electrons are misidentified as photons $f_{e\to f}$. The number of observed $Z \to ee$ events in the $e\gamma$ and ee mass spectra is expressed as a function of $f_{e\to f}$ as:

$$\begin{aligned} N^Z_{e\gamma} &= f_{e\to f}(1 - f_{e\to f})N'_Z \\ N^Z_{ee} &= (1 - f_{e\to f})^2 N'_Z \end{aligned} \tag{7.2}$$

where N'_Z is the actual number of $Z \to ee$ events.

From a sample of $N'_{e\gamma}$ events with a real photon and a real electron, the number of events $N_{e\gamma}$ that will actually get reconstructed as having one photon and one electron is given by:

$$N_{e\gamma} = (1 - f_{e\to\gamma})N'_{e\gamma} \tag{7.3}$$

The fraction of $N'_{e\gamma}$ that ends up in the candidate $\gamma\gamma$ ($N_{\gamma\gamma}$) sample can be written as:

$$N_{\gamma\gamma} = f_{e\to f}N'_{e\gamma} = N_{e\gamma}\frac{f_{e\to\gamma}}{1 - f_{e\to\gamma}} \tag{7.4}$$

From equation 7.2 it is clear that the last factor is the ratio $R_{\gamma/e} = N^Z_{e\gamma}/N^Z_{ee}$. Thus the EWK background prediction is given by $R_{\gamma/e}$ times the number of observed $e\gamma$ events.

7.4.3 Fake rate calculation

The fake rate is calculated with a tag and probe method with $Z \to ee$ events. In particular, the $Z \to ee$ invariant mass peak is plotted for the di-electron (ee) and the $e\gamma$ control sample which contains one electron and one photon passing the basic selection. Both samples are collected using a single electron trigger. The number of of $Z \to ee$ events in the invariant mass peak of both samples ($N_{e\gamma}$ and N_{ee}) are found using an extended likelihood fit, where the signal shape is modeled by second-order polynomials convoluted with gausian distribution with floating mean and width. This is illustrated in figure 7.8. The fake rate was calculated in data as a function of various kinematic variables, including the track multiplicity of the primary vertex, the vertex multiplicity of the event, the probe's $\sigma_{i\eta i\eta}$ value, and the p_T^{miss} in the event as shown in figure 7.9. To cover all the dependencies, a 30% systematic is assigned to the fake rate. The final value used in the analysis is:

$$f_{e\to\gamma} = (2.63 \pm 0.80)\% \tag{7.5}$$

7.4.4 Electroweak background estimation

As mentioned on section 7.4.2 the fraction of $N_{e\gamma}$ that ends up in the candidate $\gamma\gamma$ sample is given by $N_{\gamma\gamma} = f_{e\to\gamma}N_{e\gamma}$. The estimate of events from the electroweak background in the signal $p_T^{miss} > 100$ GeV region is given in table 7.4. The uncertainties include the statistical uncertainties and the systematic uncertainties from the fake rate calculation.

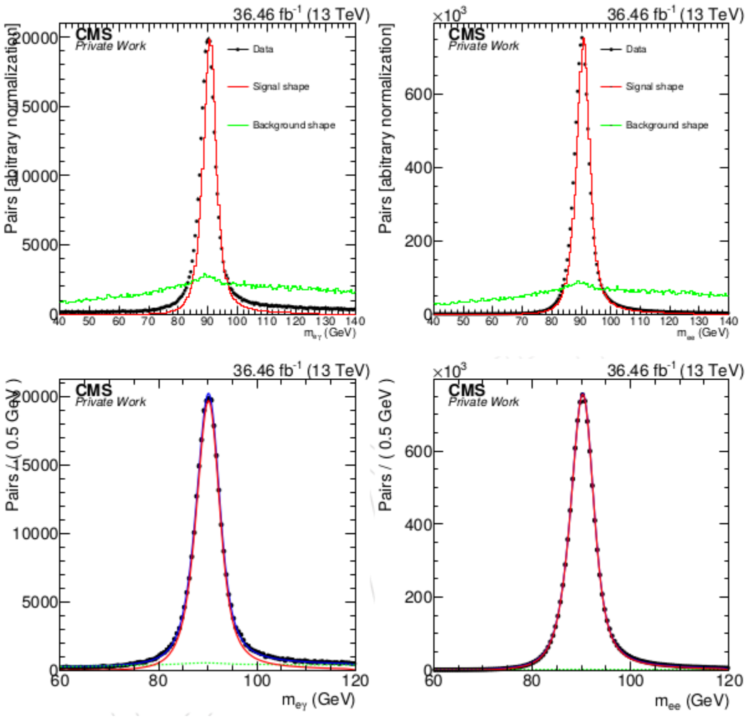


Figure 7.8: Invariant mass distributions for the $e\gamma$ (left) and ee (right) samples. For every plot the data (black markers), the signal template (red), and the background template (green) are shown.

p_T^{miss}(GeV)	Prediction of the Electroweak background
100-115	13.7 ± 4.2
115-130	9.0 ± 2.7
130-150	7.4 ± 2.3
150-185	6.1 ± 1.9
185-250	5.8 ± 1.8
≥ 250	3.3 ± 1.0

Table 7.4: Estimation of the total EWK background for $p_T^{miss} > 100$ GeV.

7.5 Irreducible $Z\gamma\gamma$ background

The last and smaller background contribution comes from $Z\gamma\gamma \rightarrow \nu\nu\gamma\gamma$ events. Since this process has true p_T^{miss} it is not included in the QCD background, but since it has two real photons rather than electrons misidentified as photons, it is not included in the EWK background either. This irreducible background is modeled via simulation and a flat uncertainty of 50% is assigned to cover any potential mismodeling. Table 7.5 shows the prediction of the irreducible $Z\gamma\gamma$ background.

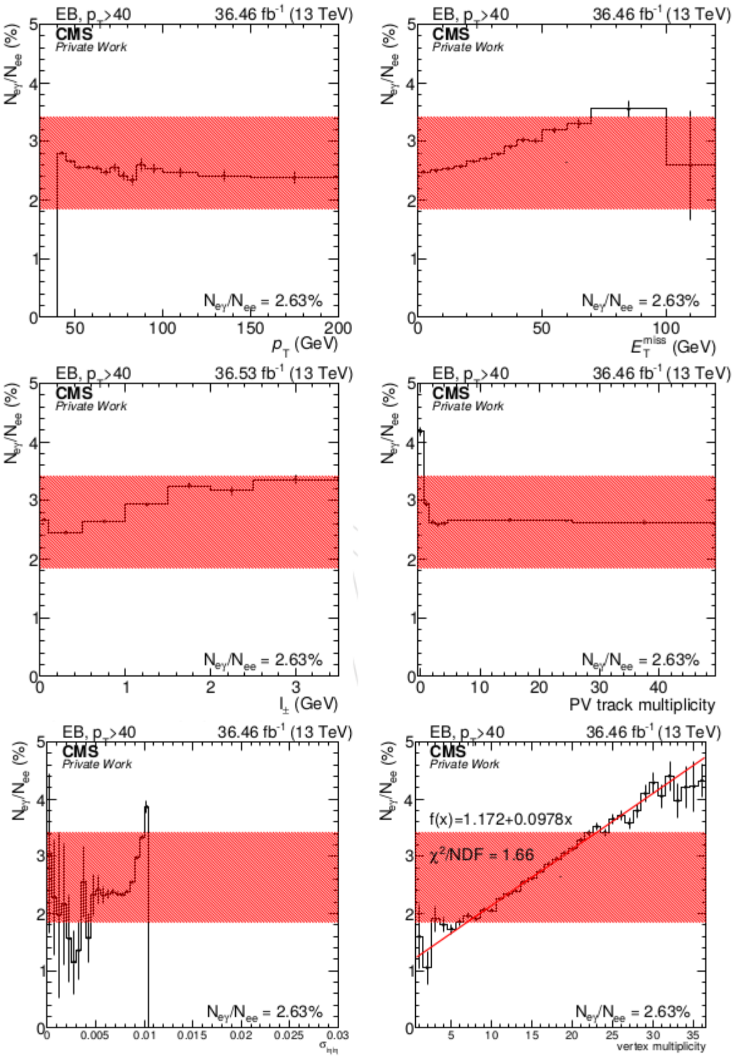


Figure 7.9: Value of $R_{\gamma/e}$ for probes in the barrel as a function of various kinematic variables: probe p_T, p_T^{miss}, isolation $I_{\pm}$ and track multiplisity of the primary vertex (PV). The red band corresponds to a 30% uncertainty on $R_{\gamma/e}$.

p_T^{miss}(GeV)	Expected number of $Z\gamma\gamma$ events for $p_T^{miss} > 100$ GeV
100-115	1.3 ± 0.6
115-130	1.1 ± 0.6
130-150	1.1 ± 0.6
150-185	1.3 ± 0.7
185-250	1.3 ± 0.6
≥ 250	1.1 ± 0.6

Table 7.5: Background contribution from $Z\gamma\gamma$ events for $p_T^{miss} > 100$ GeV.

7.6 Systematic uncertainties

There are two main sources of systematic uncertainties.: those associated with one of the background estimates and those associated with the expected signal yields. In the following section the sources of systematic uncertainties will be described in detail.

7.6.1 Uncertainties on background estimates

The larger uncertainties on the total background estimate come from the QCD prediction method. The QCD background is made up of events that do not have true p_T^{miss}. To estimate this background we make use of a ff control sample. There are three sources of uncertainty on the QCD prediction. First, there is the statistical uncertainty from the limited number of events in the ff control sample. Second, there is a systematic uncertainty from the 1σ uncertainties on the fit parameters which are propagated through the final estimate. Finally, there is a systematic uncertainty from the difference between the primary prediction and the cross check prediction. The uncertainty is given by the difference between the two predictions or the uncertainty on that difference, whichever is larger. Those contributions are summarized in table 7.6.

p_T^{miss}(GeV)	Expected QCD	Stat. Uncert	Fit Uncert	Cross Check Uncert
100-115	99.0	$+7.2, 6.7$	± 1.8	± 9.9
115-130	32.8	$+4.2, 3.7$	± 0.7	± 5.5
130-150	18.8	$+3.2, 2.7$	± 0.5	± 4.0
150-185	9.9	$+2.3, 1.9$	± 0.3	± 2.8
185-250	3.1	$+1.3, 0.9$	± 0.1	± 1.5
≥ 250	1.0	$+0.8, 0.5$	± 0.1	± 0.8

Table 7.6: Systematic and Statistical Uncertainties on the QCD Background Estimate

There are two main sources of uncertainties for the EWK background estimation. As the QCD background, the limited statistics on the control sample contributes a large portion of the overall uncertainty. However, the larger contribution comes from the method used to calculate the rate at which electrons are misidentified as photons. As described in section 7.4.3, a 30% uncertainty is assigned to the EWK background to cover all potential dependencies of the misidentification rate on the event kinematics.

For the $Z\gamma\gamma$ background, the only uncertainty considered is a 50% uncertainty to cover any potential mismodeling of the p_T^{miss} shape or uncertainty on the $Z\gamma\gamma \to \nu\nu\gamma\gamma$ cross section.

7.6.2 Other sources of systematic uncertainties

In addition to the systematic uncertainties arising from the background estimation techniques described above, there are other systematic uncertainties that affect the final analysis sensitivity. These include the uncertainty in the Parton Distribution Functions (PDF's) and the variation of cross section ratios (K factors) between leading order PDF's to next-to-leading order PDF's. The PDF uncertainties are taken from the NNPDF30 variations. Other uncertainties on the signal include finite MC statistics and the photon data/MC scale factor. There is also a 2.5$\%$ uncertainty on the integrated luminosity of the data sample. Moreover, the jet energy scale uncertainty is included by recalculating the expected signal yields using the 1σ fluctuations of the p_T^{miss}. The various uncertainties included in the calculation of the exclusion region are summarized in Table 7.7. Ranges of uncertainty arise when there are different values for the uncertainty for different signal points.

Systematic Uncertainty	[$\%$]
Integrated luminosity	2.3
Photon Data/MC scale factor	2.4
Jet energy scale	up to 23
Finite MC statistics	up to 26
PDF error on cross section	13-22

Table 7.7: Summary of systematic uncertainties included in the determination of the expected exclusion contours.

Chapter 8

Results and Interpretations

8.1 Final background prediction and observation

The data-driven background estimation methods are applied on all data recorded at $\sqrt{s} = 13$ TeV. The analysis was optimized in the "blinded" control region of $p_T^{miss} < 100$ GeV to avoid any influences by possible signal or deviation. After performing several checks, discussed in chapter 7, to ensure the robustness of the analysis, the background prediction was compared with the observed data in the signal region and a max likelihood fit is performed for the six p_T^{miss} signal bins. The p_T^{miss} distributions for the full background prediction and the unblinded data prior to the fit are shown in figure 8.1. Two benchmark signal models are also shown, corresponding to the T5gg gluino pair production simplified model with gluino mass equal to 1700 GeV or 2000 GeV and neutralino mass 1000 GeV. The expected and observed numbers of events for each bin in the signal region prior to the fit are shown in table 8.1. Table 8.2 shows the expected and observed numbers of events for each bin in the signal region for the post fit distributions. Notably, in the last bin we observe 12 events and expect $5.4^{+1.6}_{-1.6}$ background events (pre-fit). The significance of the observed data after the fit across all six bins of the signal region is calculated using the likelihood ratio test for each mass pair value of $m_{\tilde{\chi}_1^0}$ versus $m_{\tilde{g}}$ or $m_{\tilde{\chi}_1^0}$ versus $m_{\tilde{q}}$ for the T5gg and T6gg models respectively. The significance does not strongly depend on the SUSY masses, and for all masses in both models, the significance is found to correspond to between 2.35 and 2.45 standard deviations. Several studies were performed to characterize the fit and the excess in the final p_T^{miss} bin and to ensure that the statistical treatment of the data is robust. In particular, the pre- and postfit distributions were checked to make sure that the pulls are consistent within the uncertainties.

8.2 Signal acceptance and efficiency

The generated SUSY signal events described in section 6.2.2, are required to pass the same selection criteria described in chapter 6. For every mass point, the number of events expected in each of the six signal region bins is calculated. In addition, the

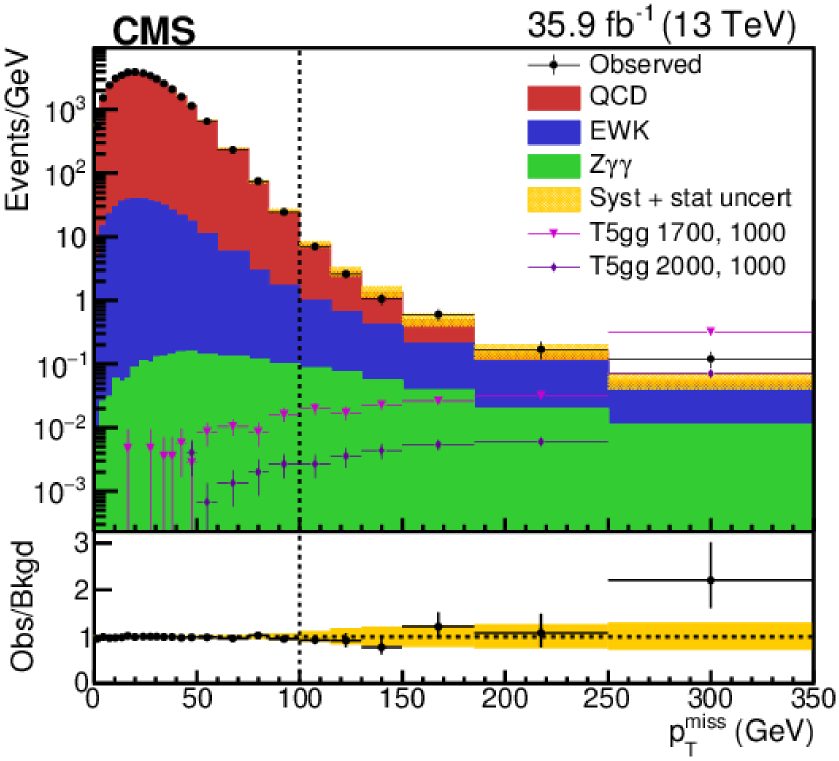


Figure 8.1: The top panel shows the observed p_T^{miss} distribution in data (black points) and predicted background distributions prior to the fit described in the text. The vertical line marks the boundary between the control region ($p_T^{miss} < 100$ GeV) and the signal region ($p_T^{miss} > 100$ GeV). The last bin is the overflow bin and includes all events with $p_T^{miss} > 250$ GeV. The QCD background is shown in red, the EWK background is shown in blue, and the $Z\gamma\gamma$ background is shown in green. The p_T^{miss} distribution shown in pink (purple) corresponds to the T5gg simplified model with $m_{\tilde{g}} = 1700(2000)$ GeV and $m_{\tilde{\chi}_0^1} = 1000$ GeV. The p_T^{miss} distributions from the T6gg simplified model are similar to the T5gg distributions shown here. The bottom panel shows the ratio of observed events to the expected background. The error bars on the ratio correspond to the statistical uncertainty in the number of observed events. The shaded region corresponds to the total uncertainty in the background estimate.

overall acceptance × efficiency ($A \times \epsilon$), defined as the number of events passing the full $\gamma\gamma$ selection divided by the total number of generated events, is calculated across the full 2D mass plane. The $A \times \epsilon$ distributions for the T5gg and T6gg simplified model frameworks is shown in the up and down pad of figure 8.2. It is worth noticing that at low neutralino mass, the $A \times \epsilon$ decreases because the photons are softer and thus they will probably fail the $p_T > 40$ GeV cut or the $m_{\gamma\gamma} > 105$ GeV cut.

8.3 Limits and interpretations

From table 8.1 it is clear that for the first five signal bins the number of observed events is consistent with the total number of expected background events in each bin in the signal region. An excess of events corresponding to 2.4 standard deviations is observed in the last signal bin. In the absence of a signal, these results can be

p_T^{miss}(GeV)	QCD	EWK	$Z\gamma\gamma$	Total background	Observed
$100-115$	99 ± 12	13.7 ± 4.2	1.3 ± 0.6	114 ± 13	105
$115-130$	$32.8^{+7.0}_{-6.7}$	9.0 ± 2.7	1.1 ± 0.6	$42.9^{+7.4}_{-7.3}$	39
$130-150$	$18.8^{+5.1}_{-4.9}$	7.4 ± 2.3	1.1 ± 0.6	$27.3^{+5.6}_{-5.4}$	21
150-185	$9.9^{+3.6}_{-3.4}$	6.1 ± 1.9	1.3 ± 0.7	$17.4^{+4.1}_{-3.9}$	21
185-250	$3.1^{+1.9}_{-1.7}$	5.8 ± 1.8	1.3 ± 0.6	$10.2^{+2.7}_{-2.6}$	11
≥ 250	$1.0^{+1.1}_{-0.9}$	3.3 ± 1.1	1.1 ± 0.6	$5.4^{+1.6}_{-1.5}$	12

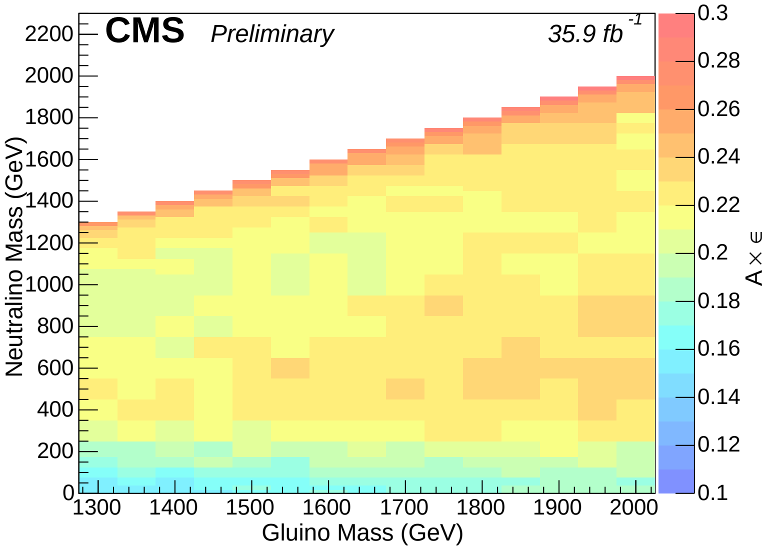
Table 8.1: Number of expected background and observed data events with 35.9 fb^{-1} of 13 TeV data in the signal region prior to the fit. The uncertainty in each expected background yield includes the statistical uncertainty and all of the systematic uncertainties described in Section 7.6 added in quadrature.

p_T^{miss}(GeV)	QCD	EWK	$Z\gamma\gamma$	Total background	Observed
$100-115$	92.7 ± 7.9	15.9 ± 3.8	1.6 ± 0.8	110.1 ± 7.4	105
$115-130$	29.7 ± 4.4	10.4 ± 2.5	1.4 ± 0.7	41.5 ± 3.9	39
$130-150$	16.0 ± 3.2	8.5 ± 2.1	1.3 ± 0.7	25.9 ± 3.1	21
150-185	9.3 ± 2.7	7.1 ± 1.8	1.6 ± 0.8	18.1 ± 2.6	21
185-250	2.6 ± 1.2	6.7 ± 1.6	1.6 ± 0.8	10.9 ± 1.8	11
≥ 250	0.7 ± 0.8	4.0 ± 1.0	1.4 ± 0.7	6.0 ± 1.2	12

Table 8.2: Number of expected background and observed data events with 35.9 fb^{-1} of 13 TeV data in the signal region after the fit. The stated uncertainties are the post-fit uncertainties in each expected background yield.

used to set limits on the allowed squark, gluino, and neutralino masses in the T5gg and T6gg simplified models that were described in section 6.1. In both models, the next-to-lightest supersymmetric particle is the neutralino, which decays with a 100% branching fraction to a photon and a gravitino, the lightest supersymmetric particle. The first simplified model assumes gluino pair production, with each gluino decaying to a neutralino and quarks. The second simplified model assumes squark pair production, with each squark decaying to a quark and a neutralino. In this analysis, we use the standard CLs method that was explained in detail in section 5.3 to determine 95% confidence level intervals for the cross section times branching fraction in each mass bin.

In figure 8.3 the expected reach of the analysis as a function of the gluino or squark and neutralino masses are shown. The predicted NLO + NLL cross section is used for each signal point, and the median expected mass exclusion includes a band representing the 1σ variation of the experimental uncertainties. For typical values of neutralino mass, we expect to exclude gluino masses out to 2.02 TeV and squark masses out to 1.74 TeV. The observed limits were 1.86 TeV for gluino masses and 1.59 TeV for



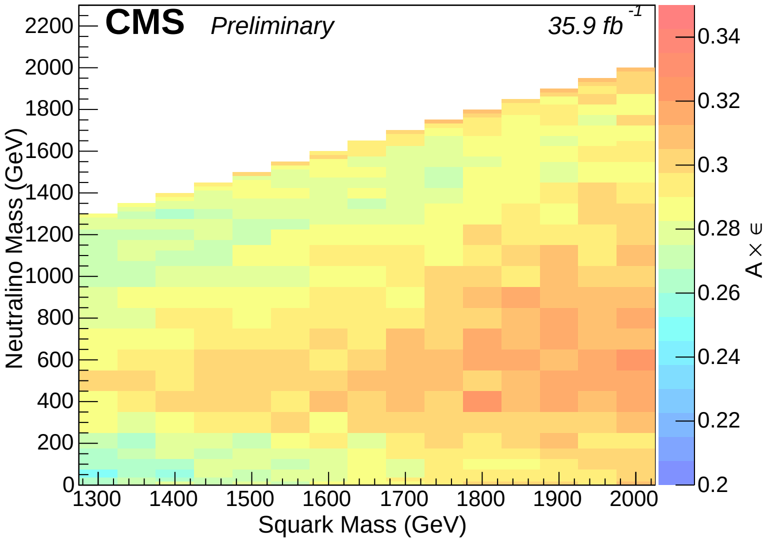


Figure 8.2: Acceptance × efficiency as a function of gluino and neutralino masses (up) for the T5gg simplified model and as a function of squark and neutralino masses (bottom) for the T6gg simplified model.

squark masses. This is an increase in sensitivity of more than 300 GeV for each model with respect to the analysis performed with 2.3 fb^{-1} of integrated luminosity collected using the CMS detector in 2015 [94]. The observed exclusions are for gluino masses less than 1.86 TeV and squark masses less than 1.59 TeV, where the difference between the expected and observed exclusions is driven by the excess observed in the data. The analysis described in this dissertation improves the observed limits by 210 GeV for gluino masses and 220 GeV [95] for squark masses with respect to the previous CMS result. In addition, the results are comparable with similar searches performed by the ATLAS collaboration[96, 97].

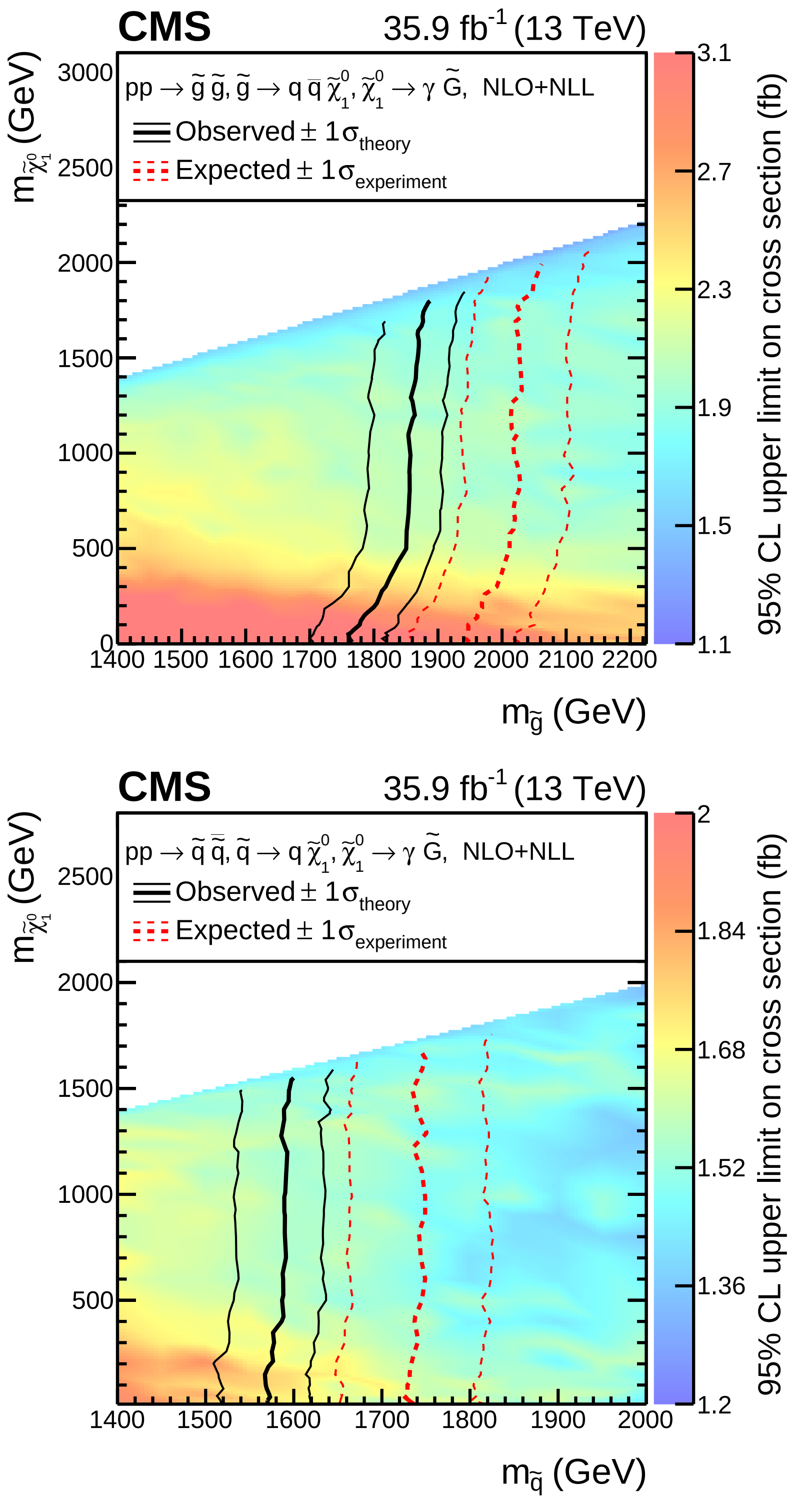


Figure 8.3: The 95% confidence level upper limits on the gluino (up) and squark (bottom) pair production cross sections as a function of gluino or squark and neutralino masses. The contours show the observed and expected exclusions assuming the NLO+NLL cross sections, with their one standard deviation uncertainties

8.4 GGMB combination analysis

The analysis in this dissertation was part of an effort for a combined search of new physics that involves final states with at least one photon and large missing transverse momentum, motivated by generalized models of gauge-mediated supersymmetry (SUSY) breaking [98]. Those signature include events with at least one photon and large missing transverse momentum and are categorized into events with two isolated photons [95], events with a lepton and a photon [99], events with additional jets [100], and events with at least one high-energy photon [101]. Combining the various photonic final states could be challenging since the analysis channels are not exclusive by construction. For that reason, an optimized veto strategy is applied to remove any overlap. Another important point to consider is the correlation in the background estimation of different channels. In contrast to the interpretation of a single analysis, additional correlations between the uncertainties used in the contributing searches have to be taken into account in the combination. The comparison between the observed yield and the background prediction after the overlap subtraction for all search bins is shown in figure 8.4. The majority of the search bins show a good agreement between the observation and the standard model prediction. Thus, no evidence for physics beyond the standard model is found, and the results of the combination are used to set upper exclusion limits with respect to several SUSY scenarios.

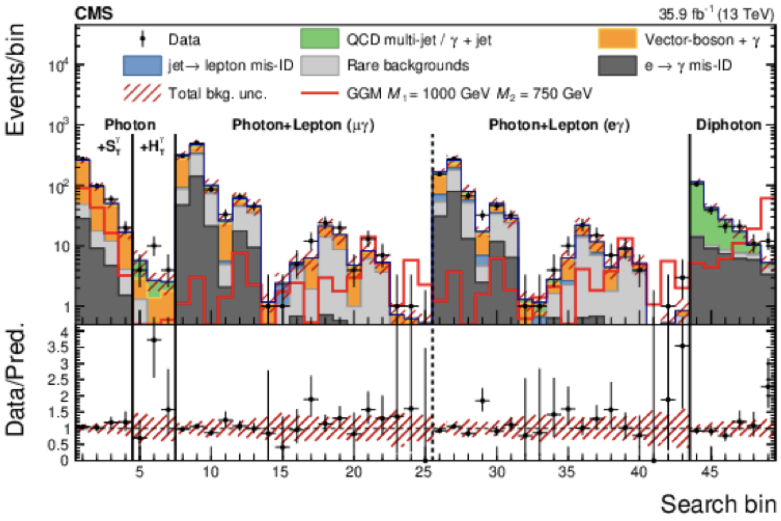


Figure 8.4: Predicted pre-fit background yields, where the values are not constrained by the likelihood fit, and observed number of events in data for all search bins used in the combination.

The results of the combination are interpreted in terms of the GGM scenario shown in figure 8.5. The main focus lies on the $\tilde{\chi}_1^0$ decay, since this decay corresponds to the only source of photons, which are essential for each of the combination analysis. Besides the $\tilde{\chi}_1^0$ decaying to γ and G, the photon can also be substituted by a Z boson leading to final states with additional fermions. The corresponding branching fractions of the $\tilde{\chi}_1^0$ for these two decays are determined by the composition of the neutralino and depend, similar to the abundance of the production processes, on the GGM input parameters.

The observed and the expected exclusion limits for the physical mass plane are shown in figure 8.6, where the phase space between the colored lines and the black line is excluded. In the physical mass plane only signal points with a mass difference above 120 GeV are shown to enable a precise projection of the physical masses from the GGM model parameters. In case of the observed limit, this gain can only be found at low neutralino masses, while at high neutralino masses the observed limit of the combination is slightly exceeded by the search that includes photons and leptons. This feature is mainly caused by the excess in the Diphoton signal region that was discussed in section 8.1. Thus, the combination results in an observed (expected) limit on the chargino mass up to 850 (1050) GeV depending on the neutralino mass which results in a gain of the order of 100 GeV.

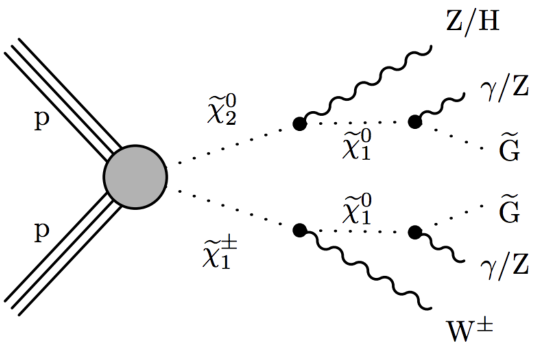


Figure 8.5: GGM diagram of $\tilde{\chi}_1^\pm \tilde{\chi}_2^0$ that used to set upper limits on the combination analysis

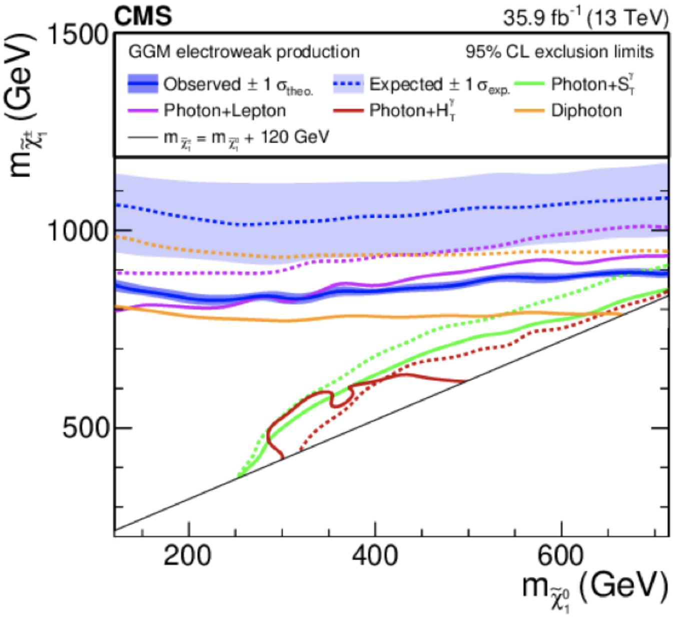


Figure 8.6: Exclusion contours at 95% CL of the individual searches and the combination for the sparticle mass planes.

Part III

Search of the production of a standard model Higgs boson in association with a top quark pair in the all-jet final state using large-radius jets

Chapter 9

Analysis strategy

9.1 Introduction

This chapter presents an overview of the strategy applied in this analysis of proton-proton collisions recorded by the CMS experiment in 2016 to search for $\mathrm{t\bar{t}H}(H \rightarrow (b\bar{b}))$ production in the all-jet final state. More specifically, we consider a phase space where the W bosons from both top quarks decay to light quarks, resulting in final states with at least eight quarks, four of which are b quarks. The experimental signature of this signal consists of jets that are primarily produced at large angles with respect to the beam axis. As a result, they are expected to have relatively high transverse momentum. At higher p_T ($p_T/m \approx 1$), the Higgs boson and the top quark decay products are highly collimated ("boosted") and thus they can be reconstructed at large radius jets. Due to the high boost considered in this analysis ($p_T > 300$ GeV), the Higgs boson and the top quarks have a large probability to be reconstructed as large-radius jets. Therefore, final states that consists of one or more of such jets are considered, where at least one boosted jet is tagged as Higgs. Several challenges have to be addressed in order to measure the rare signal process with considerable sensitivity in the presence of overwhelming standard model backgrounds. While these challenges stem from physical considerations, they directly affect the pursued overall strategy and the technical design of the analysis. First and foremost, one should be able to successfully reconstruct high purity boosted Higgs and top candidates. This is succeeded by training dedicated boosted decision trees (BDTs). The analysis strategy is constructed thereafter, which is composed of the event selection strategy and the considered resulting physics processes followed by the classification and categorization approach. The concluding signal extraction method is performed via a binned max-likelihood fit using shape templates for signal and background processes. In addition, the methods developed to model and to constrain the uncertainties of background processes are discussed.

9.2 Signal and background processes

There are several characteristics and properties of the $\mathrm{t\bar{t}H}$ signal process and the relevant backgrounds. A specific process is considered as a background to an analysis if it cannot be distinctively separated from the signal process by requiring specific thresholds, or “cuts”, on measurable observables. This definition also implies that the number of signal events remains on a reasonable level when a set of cuts is applied. The high-dimensional space defined by the domains of kinematic observables is referred to as the “phase space”.

9.2.1 Signal process

This analysis focuses on a search for a standard model Higgs boson in association with a top quark ($\mathrm{t\bar{t}H}$). Due to the high branching fraction, the Higgs boson decaying into a pair of bottom quarks ($H \to b\bar{b}$) is considered. As explained in section 3.5, the b quark arise from the hadronization of b-hadron, whose lifetime is considered long and of the order of $\tau_0 \approx 1.6 \cdot 10^{-12}$ s. Therefore, a jet can be identified as a b hadron using the state-of-art discriminators developed by CMS collaboration and described in section 3.5.1. This jet is commonly referred to as "b-tag". To study the $\mathrm{t\bar{t}H}$ process it is important to discus the decays of the top qurak. The top quark decays exclusively via weak interactions to a W boson and a bottom quark. The hadronically decays of W bosons have a branching ratio $BR_{hadr} = 0.6741 \pm 0.0023$. Depending on the momentum of the W boson and the decay angle relative to its motion, there is a probability that the decay products will be reconstructed on a large radius jet. On the other hand, the leptonic decays of the W boson are evenly distributed over the three generations with branching ratios $BR_{e\nu} = 0.1071 \pm 0.0016$ for W decaying to an electron and a neutrino and $BR_{\mu\nu} = 0.1063 \pm 0.0015$ when it decays to a muon and a neutrino. The decay of the top quark can be characterized by the decays of the subsequent W bosons, resulting into three different channels:

> **fully-hadronic (FH):** In this case both W bosons decay into quarks. Therefore, six jets are in the final state and no leptons. This final state is challenging to target as it contains large contribution on multijets (QCD) background.
>
> **dilepton (DL):** In this case, both W bosons decay leptonically, leading into two jets, two charged leptons and two neutrinos in the final state. Although this channels has a smaller branching ratio, the two opposite signed leptons give a clear signature. However, the presence of the neutrinos introduces a momentum imbalance making the event reconstruction not trivial.
>
> **single-lepton (SL):** In this case, one W boson decays into hadrons while the other one will decay into leptons. In the final state there will be four jets, one charged lepton, and one neutrino. The branching ratio is similar to the fully hadronic channel, but the presence of a charged lepton can significantly suppress multijet backgrounds.

In this analysis the decay of the $t\bar{t}$ system is considered in the fully hadronic channel. Therefore, in the final state we will have eight jets and no leptons, refereed as "all-jet-final" state. Compared to previous searches in the all jet final state that reconstruct resolved jets, this analysis focuses on fully boosted and semi-boosted topologies. The Higgs boson and top candidates can be produced with a large Lorentz boost and hence to be reconstructed in a large radius jet. Those signatures include one or more boosted jets, one of which is identified as the Higgs candidate.

In addition, to the analysis signal process, other processes of the Higgs boson decaying into particles rather than a pair of b quarks is considered. Despite their minor expected yield due to smaller branching ratios and different final-state signatures, events of those processes can potentially pass phase space selection criteria and contribute to the total number signal events.

9.2.2 Background processes

Several SM processes contribute to the background of the fully hadronic $\mathrm{t\bar{t}H}$ signal. The final state objects such as leptons and jets are used to derive experimental observables. Therefore, relevant background processes can be determined by studying their final states in simulations while varying phase space selection criteria and monitoring their impact on the yield of the signal process. The by far most dominant background comes from the QCD multijet production, as there is a finite probability that ordinary jets, from single parton radiation, will mimic the topological substructure of a top or higgs decaying jet. In addition, several processes that mainly include top quarks decays can enter in the analysis topology. In all the cases, high p_T jets can be reconstructed as a Higgs or top candidate. The background processes are described below and the corresponding feynman diagramms are illustrated in figure 9.1.

QCD multijet: This background consists of events with jets produced through strong interactions. Those events include multiple gluon radiation and have a relative large cross section. Those jets could have large p_T and thus pass the analysis cuts and be reconstructed as boosted jets. This background is exclusively modeled from the data (data-driven).

inclusive $\mathrm{t\bar{t}}$ background: The second most dominant background contribution comes from $\mathrm{t\bar{t}}$ events. Additional jets from gluon radiation can be produced, resulting in final states with large hadronic activity. The large p_T cut excludes most these soft jets, however, the contribution of this background can result from boosted jets, that pass the analysis cuts. Therefore, the final states of the SM $\mathrm{t\bar{t}}$ production is kinematically close to the signal. In addition, the cross section of this process is large, and thus we expect a large contribution in the final state. This background is modeled in simulation, but the uncertainty on the expected yield is constrained by a control region from the data.

$\mathrm{t\bar{t}}$+ Z: $\mathrm{t\bar{t}}$ production in association with a Z boson has a similar cross section to $\mathrm{t\bar{t}H}$ production. However, the branching ratio for $\mathrm{Z} \to b\bar{b}$ is lower compared to $H \to b\bar{b}$, so a lower rate is expected in the final state. This process presents a signal-like final state and therefore is an irreducible background. This background is modeled in simulation, but is treated separately from the rest of subdominant backgrounds due to its similarities with the signal process.

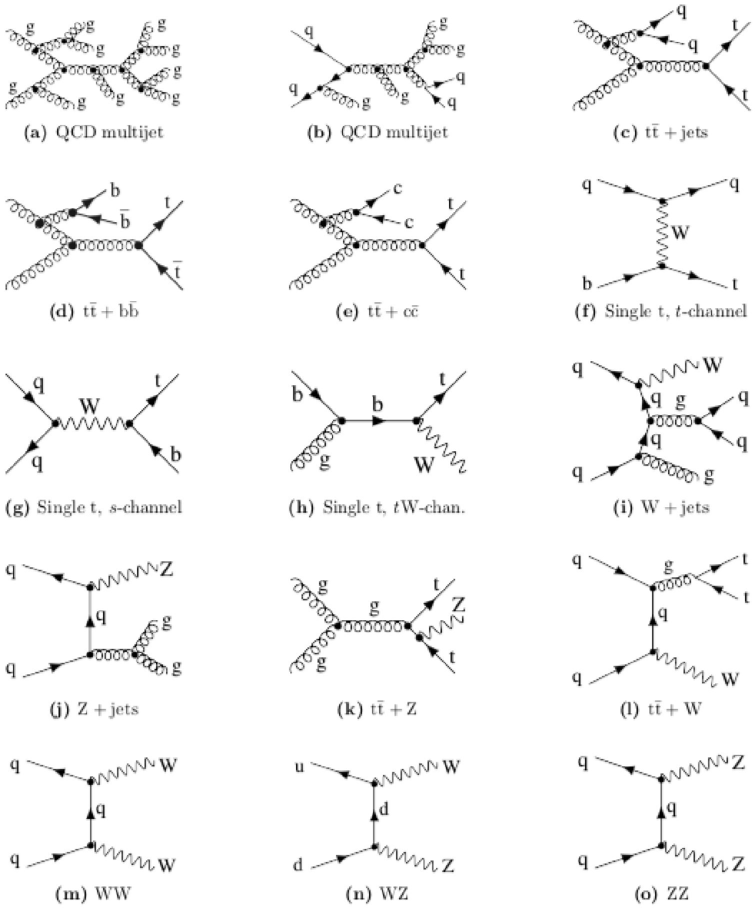
Single top quark: In this analysis there is a minor contribution from single top quark production. Although this process has a significant larger cross section than the signal, single-top events more probably will fail the selection since we require large hadronic activity in the event. The process can occur through an exchange of a W boson in the t or s-channel or in the tW-channel.

W +jets: A background contribution can arise from W+jet events. W boson production has a much larger cross section than the signal, however, to form a background it requires a significant amount of radiation and thus the final contribution is small.

Z + jets: Z boson production has a lower cross section than W boson production, but there is a small probability such events to enter in the signal region.

t$\bar{\text{t}}$ + W: t$\bar{\text{t}}$ production in association with a W boson also has a similar cross section to t$\bar{\text{t}}$H production. However, the W boson cannot decay to two b quarks and thus its contribution to the signal region is rather small.

Diboson: The production of two weak vector bosons occurs as WW, WZ or ZZ in decreasing order of cross section. The three processes have a cross section one to two orders of magnitude larger than the signal, however, this background is suppressed due to high hadronic activity required in the signal region.



(a) QCD multijet (b) QCD multijet (c) t$\bar{\text{t}}$ + jets

(d) t$\bar{\text{t}}$ + b$\bar{\text{b}}$ (e) t$\bar{\text{t}}$ + c$\bar{\text{c}}$ (f) Single t, t-channel

(g) Single t, s-channel (h) Single t, tW-chan. (i) W + jets

(j) Z + jets (k) t$\bar{\text{t}}$ + Z (l) t$\bar{\text{t}}$ + W

(m) WW (n) WZ (o) ZZ

Figure 9.1: Feynman diagrams of possible SM processes that contribute to the background of the fully hadronic t$\bar{\text{t}}$H H → b$\bar{\text{b}}$ signal.

9.3 Data and simulated samples

9.3.1 Data samples

The analysis was performed using the full data set of of center-of mass energy $\sqrt{s} = 13$ TeV pp collisions collected by CMS in 2016. The data corresponds to an integrated luminosity of $35.9\ \text{fb}^{-1}$. The certified data were collected based on different trigger types. In particular this analysis uses the JetHT dataset which contains all events selected by any of the jet and $\mathrm{H_T}$[1] based triggers. The datasets are split into different data eras that account for different conditions such as beam intensities and operating status. Those datasets are listed in table 9.1 and correspond to the reconstruction that was done on February 2017 which profits from better calibration and performance corrections. Events from these datasets are selected for further analysis if they pass the trigger requirements and the selection criteria.

2016 data sets
/JetHT/Run2016B-03Feb2017_ver2-v2/MINIAOD
/JetHT/Run2016C-03Feb2017-v1/MINIAOD
/JetHT/Run2016D-03Feb2017-v1/MINIAOD
/JetHT/Run2016E-03Feb2017-v1/MINIAOD
/JetHT/Run2016F-03Feb2017-v1/MINIAOD
/JetHT/Run2016G-03Feb2017-v1/MINIAOD
/JetHT/Run2016H-03Feb2017_ver2-v1/MINIAOD
/JetHT/Run2016H-03Feb2017_ver3-v1/MINIAOD

Table 9.1: list of data sets that used in the boosted fully hadronic search.

9.3.2 Simulated samples

In order to distinguish events coming from the targeting signal process, a deep understanding of the standard model (SM) background processes that can lead to the same final state is necessary. Any excess of events compared to the SM background expectations can be considered as signal. The $\mathrm{t\bar{t}H}$ signal and the standard model background processes are simulated with Monte Carlo (MC) event generators. The simulated samples used for the analysis, depending on the physics process, are generated as described in chapter 4, with PYTHIA, POWHEG (v1 or v2),or MADGRAPH5_aMC@NLO. In all samples, parton showering and hadronization are simulated with PYTHIA (v8.200). The parton distribution functions (PDFs) of the proton is modeled with NNPDF3.0 with α_s = 0.118 as recommended by the PDF4LHC group [102] for the second run of the LHC. More specifically, the $\mathrm{t\bar{t}H}$ signal sample is simulated to the NNLO with MADGRAPH5_aMC@NLO. For this simulation, the mass of the Higgs boson is set to $\mathrm{m_H = 125\ GeV}$ and that of the top quark is set to $\mathrm{m_t = 172.5\ GeV}$. For the background processes, the $\mathrm{t\bar{t}}$ and the single top t- and tW- channels of signal top production are modeled with POWHEG at NLO. In addition to the nominal $\mathrm{t\bar{t}}$ sample, two samples with the same generator conditions, but with a cut on an invariant mass of the two tops ($M_{t\bar{t}}$) at patron level were used. Those samples include generated events with $M_{t\bar{t}} > 700$ GeV and thus, with applying selection cuts they lead to the targeted phase space. Those samples profit with higher statistics compared to the nominal $t\bar{t}$ sample and therefore, are used for constructing $\mathrm{t\bar{t}}$ enriched control regions. The associ-

[1] H_T is defined as the scalar sum of the transverse momentum of all the jets in the event.

ated production of $\mathrm{t\bar{t}}$ with a vector boson, $\mathrm{t\bar{t}} + V$, samples are simulated at NLO with MADGRAPH5_aMC@NLO. The production of W and Z bosons with additional jets, as well as, the QCD multijet events is simulated at LO using MADGRAPH. As mentioned in the background processes description, the QCD background is estimated completely from the data. However, QCD simulated events are used to validate and to perform closure test on methods. The diboson processes, WW,WZ and ZZ are simulated at LO with PYTHIA (v.8.2). The simulated events are characterized by a set of parameters related to cut-off energy scales and energy dependence of the underlying interaction. This is referred as $tune$ [103].

To compare the simulated samples to the data, the simulated samples need to be normalized to the integrated luminosity of the data according to their predicted cross sections. The production cross section of $\mathrm{t\bar{t}H}$ signal, as well as, the Higgs boson branching ratios are calculated at NLO. On the other hand, the cross section of the $\mathrm{t\bar{t}}$ simulated events is calculated on the full next-to-next-to-leading-order (NNLO) accuracy. For this simulation the top quark mass is assumed to be $m_t = 172.5$ GeV. In addition, the $\mathrm{W} + jets$ and $\mathrm{Z} + jets$ are calculated at NNLO, and for all single top channel and dibososn background for NLO. Table 9.2 shows the simulated samples for signal and background that were used in the analysis, along with the number of generated events and the corresponding production cross sections.

9.4 Corrections to simulated events

Although the simulated generated MC samples are produced in order to describe best the data, in reality discrepancies are observed. The source of these discrepancies could reflect various effects, such as mismodelling in event simulation and reconstruction procedure, to parton showering, hadronization and to detector simulation and reconstruction algorithms. More precisely, disagreement observed in these quantities does not represent a cause, but rather a consequence of underlying mismodeling. To account on this mismodeling, corrections are derived by comparing characteristic quantities that are correlated to the effect with the data. These corrections typically result in a weight for each simulated event that is directly related to the over or underestimation of events with the particular properties of the correction. Therefore, by reweighting the simulated event the agreement between data and simulation is improved. Consequently, a thorough treatment of uncertainties is absolutely necessary. To account on this discrepancies, identification and reconstruction efficiencies are used to both simulation and data. Disagreements between efficiencies in data directly translate into deviating event rates, which must be compensated by including derived data-to-simulation ratios to event weights. This additional weight applied to simulation is called "scale factor". In this section, corrections related to the number of pile up, trigger efficiency and b tagging scale factors are discussed.

9.4.1 Number of pile up interactions

The increasing of the instantaneous luminosity of the LHC during 2016 resulted in an increase of the average rate of overlapping events over time. Pile up events can impact the object identification and reconstruction performance and therefore, any discrepancies between data and simulation need to be taken into account. In CMS,

Monte Carlo sample	Generators	Events (10^6)	cross section σ (pb)
$\mathrm{t\bar{t}H}(H \rightarrow b\bar{b})$	MG5_aMC@NLO	9.8912	0.2934
$\mathrm{t\bar{t}H}(H \nrightarrow b\bar{b})$	MG5_aMC@NLO	10.1312	0.5297
$\mathrm{t\bar{t}}$	POWHEG	77.0811	832
$\mathrm{t\bar{t}}$, $700 < M_{\mathrm{t\bar{t}}} < 1000$ GeV	POWHEG	38.4226	69.64
$\mathrm{t\bar{t}}$, $M_{\mathrm{t\bar{t}}} > 1000$ GeV	POWHEG	24.5616	16.74
$\mathrm{t\bar{t}} + \mathrm{W}$, $\mathrm{W} \rightarrow q\bar{q}'$	MG5_aMC@NLO	0.569424	0.4062
$\mathrm{t\bar{t}} + \mathrm{Z}$, $\mathrm{Z} \rightarrow q\bar{q}'$	MG5_aMC@NLO	0.396360	0.5297
$\mathrm{W} + jets$, $\mathrm{W} \rightarrow q\bar{q}'$	MG5_aMC@NLO	22.4	3539
$\mathrm{Z} + jets$, $\mathrm{Z} \rightarrow q\bar{q}'$	MG5_aMC@NLO	$32.8^{+4.3}_{-3.8}$	$30.3^{+4.0}_{-3.5}$
WW,	PYTHIA8	1.9984	51.723
ZZ,	PYTHIA8	338.456	22.29
$ST\ t$ (t-channel)	POWHEG	67.2	136.03
$ST\ \bar{t}$ (t-channel)	POWHEG	38.8	80.95
$ST\ t$W	POWHEG	6.9	35.6
$ST\ \bar{t}$W	POWHEG	6.9	35.6
QCD (H_T 300to500 GeV)	MG5_aMC@NLO	54.5	3.477e+5
QCD (H_T 500to700 GeV)	MG5_aMC@NLO	62.3	3.21e+4
QCD (H_T 700to1000 GeV)	MG5_aMC@NLO	45.4	6831
QCD (H_T 1000to1500 GeV)	MG5_aMC@NLO	15.1	1207
QCD (H_T 1500to2000 GeV)	MG5_aMC@NLO	11.8	119.9
QCD ($H_T > 2000$ GeV)	MG5_aMC@NLO	6.0	25.24

Table 9.2: Monte Carlo simulated samples used for the $\mathrm{t\bar{t}H}$ $H \rightarrow b\bar{b}$ analysis

the effect of pileup is modeled in simulated event by including the final-state particles of a number of minimum-bias events. The probability distribution of the number of pileup interactions is measured in data with a limited amount of integrated luminosity before simulation campaigns are initiated. The number of pile-up interactions for each collision depends on the instantaneous luminosity for each bunch crossing and the total inelastic cross section. The total inelastic pp cross section is measured using dedicated forward detectors and $\sigma_{inelastic} = 69.2$ mb is found to accurately describe the $\sqrt{s} =$ 13 TeV data, with an uncertainty of 4.6% [104]. The ratio between the data and simulation yields pileup correction weights, which are visualized in the bottom panel of the left part of figure 9.2, where the simulated number of primary vertex distribution is compared with the one from the data. The uncertainty on the inelastic cross section of ±4.6% is propagating to the weights description as shown in the right part of figure 9.2.

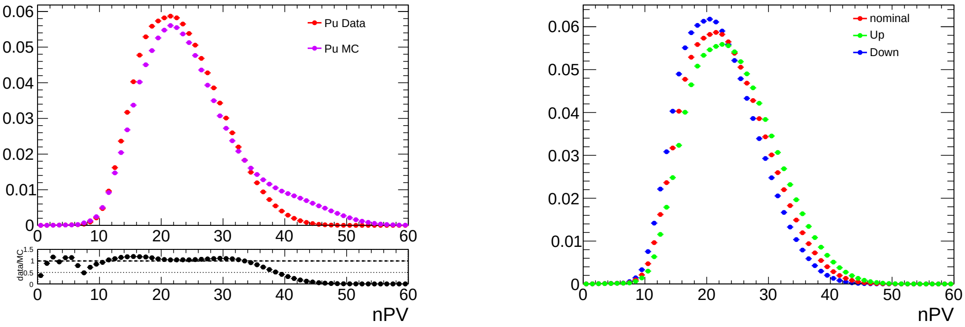


Figure 9.2: Distribution of the number of pileup (PU) events in data and simulation (left figure). The systematic uncertainty on the pileup weight is computed by varying the cross section by $\pm 4.6\%$ (right figure).

9.4.2 Trigger Efficiency and scale factors

As mentioned in the introduction of this section, the trigger performance in simulated events does not necessarily match the performance of the observed data. Therefore, the efficiencies for both data and simulation are calculated and scale factors are derived to account on that mismodeling. The trigger path employed for the collection of signal events uses L1 seeds that require the sum of the transverse momentum of the jets to be greater than $p_T > 240$ GeV. At HLT, jets are reconstructed from (online) particle flow candidates using the anti-kt algorithm with distance parameter $R = 0.8$ and with mass, after trimming of soft particles, greater than 50 GeV. Interesting events are also required to have a sum of the transverse momentum of the AK8 jet greater than 700 GeV. The aforementioned trigger path ran unprescaled for the duration of the 2016 run. In case of a prescaled trigger, the data are collected with very low-threshold seeds, resulting in reduced output rate. Since the analysis trigger was unprescaled it collected an integrated luminosity of 35.5 fb^{-1}. The L1 seeds and the signal trigger path is shown in the first two rows of table 9.3.

trigger path	purpose
L1_HTT240 OR L1_HTT255 OR L1_HTT270 OR L1_HTT280 OR L1_HTT300 OR L1_HTT320	L1 seed
HLT_AK8PFHT700_TrimR0p1PT0p03Mass50	signal path trigger
HLT_AK8PFJet200	reference trigger
HLT_Mu50	reference trigger

Table 9.3: analysis signal and reference trigger paths. The L1 seeds of the signal trigger are also mentioned.

The efficiency of a signal trigger is usually studied as a function of a quantity that reflects the analysis phase space and the trigger requirements. Therefore, such quantity for the analysis trigger could be H_T, which is defined as the scalar sum of the transverse momentum of the Ak8 jets. However, since the analysis phase space requires jets reconstructed with radius $R = 0.8$ and $R = 0.4$ and because the offline and online jet requirements are sightly different, the trigger efficiency is studied as a function of S_T, which is defined as the scalar sum of the p_T of all the jets in the event. This way, events with large hadronic activity, i.e, events with only one reconstructed Ak8 jet, but with large Ak4 multiplicity, are taken into account. The efficiency of the signal trigger is measured by monitoring rates with respect of "reference" triggers, which are trigger paths weakly or not correlated to the nominal trigger. First, the number of events that pass only the reference trigger N_{ref} and the offline analysis selection is determined. Then the fraction of events that additionally pass the signal trigger ($N_{sigTrig}$) is:

$$\epsilon = \frac{N_{ref} + N_{sigTrig} + \text{offline criteria}}{N_{ref} + \text{offline criteria}} \tag{9.1}$$

yields the efficiency of the signal trigger. The trigger performance was studied in several topologies, enriched in different events. For this purpose, two different reference triggers were used. A hadronic trigger that requires a jet with $p_T > 200$ GeV is considered for a topology enriched in multijet events. The second trigger is a muon trigger which requires the presence of a $p_T > 50$ GeV muon and is used for topologies enriched in $\mathrm{t\bar{t}}$ and $\mathrm{W} + Jet$ events. Both reference trigger paths are listed in the last two rows of table 9.3. The topologies considered for the trigger efficiency studies are described below and their selection requirements are summarized in table 9.4.

multijet topology: A topology enriched in multijets events is constructed by selecting events with no leptons. In addition, to ensure the presence of at least one Ak8 jet, the leading jet is required to have a p_T greater than 300 GeV. For this study, QCD and $\mathrm{t\bar{t}}$ simulated samples were used to model the efficiency performance in simulation.

$\mathrm{t\bar{t}}$ topology: To ensure a topology rich in $\mathrm{t\bar{t}}$ events, the presence of exactly one muon is required. In addition, as in the multijet topology, the momentum of the leading jet is required to be greater than 300 GeV. In order to suppress events coming from Drell-Yann processes, we require the missing transverse momentum of the event (E_T^{miss}) to be greater that 50 GeV. For a high purity $\mathrm{t\bar{t}}$ sample, events are required to have at least one b tagged jet. In addition to the $\mathrm{t\bar{t}}$ simulated samples, Drell-Yann simulated events were used with $H_T > 70$ GeV, to model the trigger performance in simulation.

$\mathrm{W} + jets$ topology: by inverting the b-tag requirement of the $\mathrm{t\bar{t}}$ selection (zero b tagged jets) we construct a topology enriched in $W + Jets$. For this study simulated events of leptonically decaying W+jets were used. The simulated samples were divided in bins of H_T with $H_T > 100$ GeV.

Topology	lepton	b-tagged jets	leading jet p_{T} [GeV]	p_T^{miss} [GeV]
multijet (QCD)	0	no requirement	300	no requirement
$\mathrm{t\bar{t}}$	1 muon	1bjet	300	>50
W+ jets	1 muon	0bjet	300	> 50

Table 9.4: Topologies used for trigger studies

In figure 9.3, the trigger efficiency as a function of S_T is shown for the three aforementioned topologies. In all topologies, the simulated events are compared with the data. In the same pad, the distribution of S_T is illustrated for the case with only one Ak8 jet but at least four Ak4 jets. It is clear that the trigger can be efficient withouth excluding potential signal from this type of events. For all the three topologies, the trigger is efficient above $S_T > 900$ GeV. Therefore, this cut is used to select interesting events for analysis.

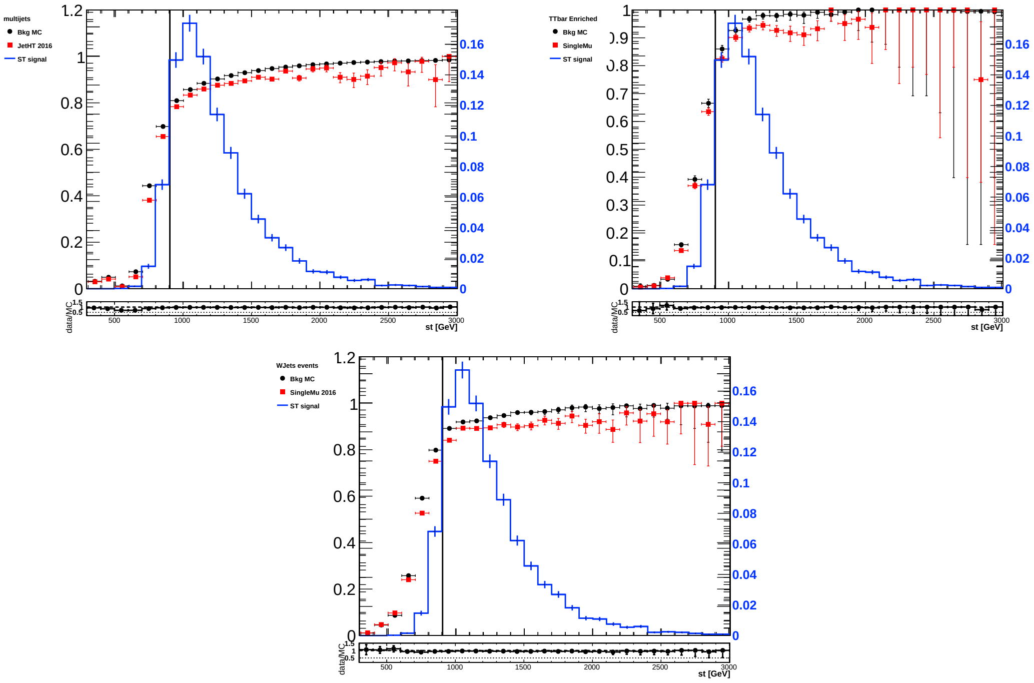


Figure 9.3: Trigger efficiency for the signal path of the analysis for the multijet (upper left), $t\bar{t}$ (upper right) and $W + jets$ (bottom) enriched regions. The distribution of S_T in case with only one Ak8 jet but at least four Ak4 jets is shown. The vertical line illustrates where the trigger becomes efficient ($S_T > 900$ GeV).

The slight differences between efficiencies in data and MC simulated events are rectified by applying scale factors, which are calculated as the ratio of the efficiency in data to that in simulation. Since the multijet topology is statistically favoured, it is used to derive a correction for the simulation to match the data. The ratio between the data and simulation is fitted with a constant, as shown in figure 9.4 and a constant scale factor of 0.97 is derived. Uncertainties on the scale factors are derived as the uncertainty coming from the fit and are treated as a systematic uncertainty.

9.4.3 Shape calibration of the btag discriminant

Another quantity that potentially shows discrepancies between data and simulation, is the performance of the chosen b-tagging algorithm, meaning that the b-jet identification efficiency and the misidentification probability can differ form the one of observed in data. In this analysis, the CSVv2 algorithm described in section 3.5.1 is used. The output of the CSVv2 algorithm, refereed as CSV discriminant, can be used in two different ways. First, a specific threshold (working point) can serve as a binary criterion

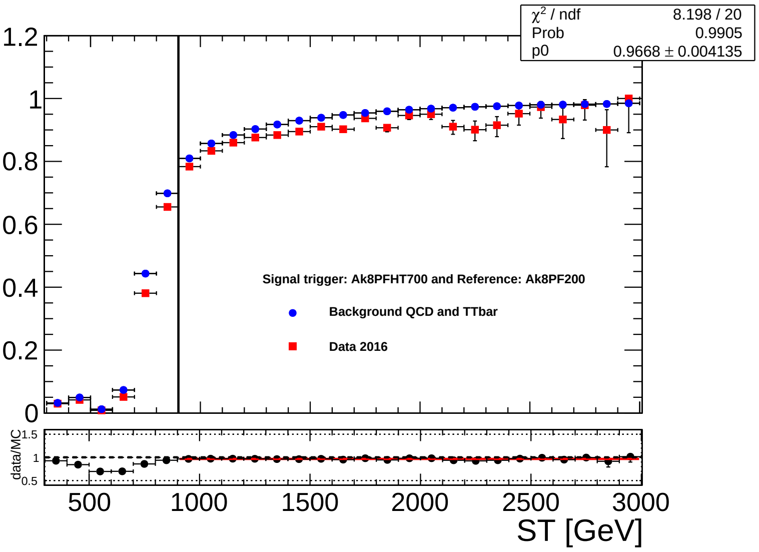


Figure 9.4: Comparison of the trigger efficiency in data and simulation for the signal path of the analysis in the multiget enriched topology. The vertical line separates the signal region, where $S_T > 900$ GeV. The constant fit in the ratio of data and simulation is used to derive a scale factor.

definition of a b-tagged jet, with specific identification and misidentification efficiencies. In addition, the shape information of the CSV discriminant, can be exploited further, as it can be used as an input to multivariate techniques for candidate selection or category classification. For all these reasons, a weight that compensates for deviating b-tagging efficiencies only, is not sufficient. Therefore, the full CSV shape needs to be calibrated. The distribution of the b-tagging discriminant in simulation is corrected by scale factors, which depend on the flavour, p_T and $|\eta|$ of the jets, in order to match with the observed distribution in data. This correction is derived separately for light-flavour and b jets from a “tag-and-probe” approach using control samples enriched in events with a Z boson and exactly two jets, and $\mathrm{t\bar{t}}$ events with no additional jets, respectively [105]. This analysis uses mixed topologies with both Ak4 and Ak8 jets and thus, the CSV discriminant is used for identifying b-Ak4 jets and $b-$Ak4 subjets. Since both jet collections are cross cleaned then the scale factor derived to account on simulation mismodelling is:

$$SF_{total} = \prod_{i}^{N_{\mathrm{jets}}} \mathrm{SF}_{\mathrm{jet_i}} = \prod_{i}^{N_{\mathrm{Ak4jets}}} \mathrm{SF}_{\mathrm{AK4jet_i}} \cdot \prod_{i}^{N_{\mathrm{Ak4subjets}}} \mathrm{SF}_{\mathrm{AK4subjet_i}} = \mathrm{SF}_{\mathrm{jet_1}} \cdot \mathrm{SF}_{\mathrm{jet_2}} \cdot .... \quad (9.2)$$

There is no calibration sample for charm jets and thus, the scale factors for them are set to 1.00 with the uncertainty derived from the calibration for b jets. The systematic uncertainties on the b tagging scale factors are discussed in section 11.5.1, where a dedicated study is performed to see the effect on the sample purity and selection efficiency.

From the multiple corrections discussed in previous sections, a global weight is derived by multiplying each separate weight. Therefore, the global weight comes from the product:

$$\text{global weight} = \prod_{i}^{\mathrm{N_{SFs}}} \mathrm{SF_i} = \mathrm{SF_{trig}} \cdot \mathrm{SF_{PU}} \cdot SF_{btag} \tag{9.3}$$

9.5 Object selection

9.5.1 Leptons

In this analysis, leptons are reconstructed for two main purposes. First and foremost, in the final selection, no leptons are required in order to ensure a high purity hadronic sample. In addition, jets could carry an electromagnetic component and thus, it is possible for a lepton to be misidentified as a jet. Since leptons are reconstructed with high efficiency, the lepton collection is reconstructed first, and then used to remove other objects, such as jets, that overlap with a reconstructed lepton within a ΔR threshold. By cross-cleaning the objects, jets are reconstructed with high efficiency and double counting events is avoided.

Muons

For the reconstruction of muons, PF candidates were used as described in section 3.4.2 and are selected based in their kinematic variables p_T and η. In particular, muons are required to have a p_T greater than 20 GeV and $|\eta| < 2.4$. Additionally, a requirement on the corrected relative muon isolation is imposed. The absolute value of the isolation variable is defined as:

$$Iso^{\mu} = \sum_{\Delta R<0.4} p_T^{h^{\pm}} + \max\Big(0, \sum_{\Delta R<0.4} E_T^{h^0} + \sum_{\Delta R<0.4} E_T^{\gamma} - \frac{1}{2} \sum_{\Delta R<0.4} p_T^{\mathrm{pu}}\Big) \tag{9.4}$$

where $p^{h^{\pm}}$, $E_T^{h^0}$, E_T^{γ} and p_T^{pu} are the transverse momentum or energy of particles identified by the PF algorithm. The indices denote the type of particle that are considered in the sums: charged hadrons from the primary vertex $(h^{\pm})$; neutral hadrons (h^0); photons (γ); and charged hadrons from other vertexes (pu). The sum is performed over all particles within a cone of $\Delta R = 0.4$ around the muon. Muons are required to pass the medium ID working point [106] and to have a relative mini-isolation less than 0.1.

Electrons

As is the case of muons, electrons are defined for the sole purpose of vetoing events containing leptons. PF reconstructed electrons (see 3.4.3) are selected based on their p_T and η, as well as, a number of isolation variables. In particular electrons with $p_T > 20$ GeV and $|\eta| < 2.4$ are selected. Electron candidates are required to pass the tight working point [107] and to have a relative mini-isolation less than 0.1.

9.5.2 Jets

The efficient reconstruction of jets is crucial for this analysis since jet properties are used in the BDT training and also the discriminating variable that is used in the final fit is the mass soft drop of the tagged Higgs jet. In this analysis two types of jets are used. Those jets are reconstructed by PF candidates as described in section 3.4.4 and are classified depending on the distance parameter R. Moreover, corrections and calibrations are applied to the jet collections in order to ensure a better description of the data. To maximize the signal significance while keeping the background low, the selection requirements on the jets are optimized by applying further requirements. In addition, both collections are cross-cleaned to avoid double counting events. In the following section, the two jet collections and the selection requirements are discussed in detail.

Large-Radius jets

The primary jet collection, used in this analysis is obtained by clustering the PF candidates with the anti-k_T algorithm using a distance parameter $R = 0.8$ and therefore, this collection is referred as "Ak8 jets". The choice of the distance parameter is such to allow containment of the full Higgs and top decay products for a certain p_T threshold. In addition, those jets have undergone charged hadron subtraction (CHS) as discussed in section 3.4.4, in order to suppress pile up. The “modified mass drop tagger” algorithm, also known as the “soft drop” (SD) algorithm (see section 10.3.2), with angular exponent $\beta = 0$, soft cutoff threshold $z_{cut} < 0.1$, and characteristic radius $R_0 = 0.8$, is applied to remove soft and wide-angle radiation from the jet. In the default configuration, the SD algorithm identifies two hard subjets with distance parameter $R = 0.4$ inside the Ak8 jet. By exploiting the kinematics of these two subjets, the 4-momentum of the Ak8 jet is calculated. A minimum transverse momentum of 200 GeV and $|\eta| < 2.4$ is required for all the large-R jets. They should also satisfy the tight PF jet identification criteria [108]. In addition, the jet energy scale and resolution corrections derived for CHS jets clustered with $R = 0.4$ are applied on the subjets. The soft drop mass, computed as the invariant mass of the two subjets, is required to be greater than 50 GeV. To avoid misidentifying leptons as jets, any large-R jets overlapping with an electron or muon with $\Delta R(\mathrm{Ak8jet}, \ell) < 0.4$ are not considered in this analysis. In addition, the CSVv2 tagging algorithm is used to identify b-subjets. More specifically, the full distribution of the CSVv2 discriminator of the two subjets is used as an input variable to the training.

Small-Radius jets

An additional jet collection is reconstructed from clustering PF candidates using the anti-k_T algorithm with distance parameter $R = 0.4$ refereed as "Ak4 jets". Charged PF candidates associated with pileup vertices are removed from the jet constituents using the charged hadron subtraction algorithm. In addition, jet energy scale and resolution are corrected to match the one from data. The Ak4 jets are required to have a $p_T > 30$ GeV and $|\eta| < 2.4$ and satisfy the tight PF jet identification criteria [108]. To avoid misidentifying leptons as jets, Ak4 jets overlapping with an electron or muon with $\Delta R(\Delta R(\mathrm{Ak4jet}, \ell) < 0.4$ are discarded. In addition there is a probability that a candidate to be reconstructed both as a Ak8 and Ak4 jet. Therefore, those two distributions are cross cleaned by requiring the $\Delta R(\mathrm{Ak8jet}, \mathrm{Ak4jet}) > 0.8$. As in Ak8 jets, the CSVv2 discriminator is used to identify b Ak4 jets. The medium working point was used which corresponds to an efficiency of around 63% to tag jets with $p_T > 20$ GeV originating from b quarks, (12%) for jets originating from c quarks, and approximately 1% misidentification probability for jets from light-flavour quarks or gluons.

9.6 Baseline selection

After preselections on the reconstructed objects, certain requirements are applied to further separate boosted Higgs and top candidates from the QCD multijet background. The events are selected if they contain at least one Ak8 jet and pass the analysis trigger, discussed in section 9.4.2. To ensure a fully efficient trigger, the S_T of the event is required to be greater than 900 GeV. In addition, since the the minimum transverse momentum of 300 GeV is required for the decay products of a Higgs boson to be fully contained within an $R = 0.8$ jet, this momentum threshold is applied to the leading jet. As discussed earlier in this section we veto on reconstructed leptons. The above baseline requirements are summarized on table 9.5.

Observable	Requirement
N_{jets}	> 0
N_{leptons}	$= 0$
$p_T^{\text{leading-jet}}$	> 300 GeV
m_{SD}^{jets}	> 50 GeV
S_T	> 900 GeV

Table 9.5: Baseline selection requirements.

Chapter 10

Identification of Highly Lorentz-Boosted Hadronically Decaying Higgs and Top candidates

10.1 Introduction

The efficient identification ("tagging") of highly Lorentz-boosted hadronically decaying massive particles such as top quarks and W, Z and Higgs bosons have become necessary at the LHC, since it provides useful handles for both searches for new physics and probes of the SM in the high-momentum regime. In addition, as luminosity leveling will be used extensively in HL-LHC, one should be able to fully exploit the high-momentum phase space.

At the LHC, jets are ubiquitously produced, as a result from the hadronisation of quarks. The vast majority of the jets comes from a single guark or gluon, and therefore are not particularly interesting for physics analysis. However, a high-momentum electroweak scale particle, such as the top quark and the W, Z, and Higgs boson, with a subsequent hadronic decay, results in a triplet or pair of highly collimated quarks, which then can lead to a single large-radius jet, instead of several well-separated jets that correspond to each individual quark. The large-radius jets that initiated by highly Lorentz-boosted hadronically decaying massive particles serve several characteristics. Therefore, they can be distinguished from the ubiquitous jets initiated by a single quark or gluon. This is succeeded by studying the internal structure, or "substructure", of the jets.

There is a high interest for the study of jet substructure, both from theoretical and experimental point of view. In collider physics, jet substructure techniques are now commonly used in searches for new heavy particles that subsequently decay to highly boosted top quarks or W, Z and Higgs bosons [109, 110] and to SM processes [111, 112] in the high-momentum regime. On the theoretical side, there is a variety of new substructure observables and techniques developed to address the need of better discrimination power and robustness of the jet substructure–based tagging algorithms. Those techniques are motivated by the need of deeper understanding of Quantum Chromodynamics (QCD) that could shed light on strong interaction. In addition, advance machine learning techniques have brought further progress and deeper insights into jet substructure. In this chapter, the jet substructure techniques used by CMS are presented and the machine learning techniques used in this analysis are discussed.

10.2 Jet Substructure

To identify boosted massive particles, Ak8 jets, introduced in section 3.4.4 are used. The Ak8 jets come form the clustering of PF candidates using the anti-k_T algorithm with a distance parameter $\mathrm{R} = 0.8$. In analyses looking for highly energetic particles, the opening angle between the decay products of a Lorentz boosted particle becomes so small that the highly boosted particle appears as a single large jet instead of two well-separated smaller jets. The distance between the two quarks [113], in the case of a hadronic decay, depends on the mass (m) of the particle and its p_T as

$$\Delta R \sim \frac{2\mathrm{m}}{p_T} \tag{10.1}$$

From the above relation, it is clear that for jets clustered with $R = 0.8$, they can contain all decay products from W and Z bosons with $p_T \geq 200$ GeV, Higgs boson with $p_T \geq 300$ GeV and top quarks with $p_T \geq 400$ GeV.

A sketch of the two different situations is shown in figure 10.1. If the W, Z, Higgs boson and top quark p_T is well below the aforementioned momentum thresholds, their decay products are two well-defined jets (right). However, once their transverse momenta is approximately as the given threshold, both the quarks are completely contained within a single jet (left), referred to as a boosted W, Z, Higgs or top jet respectively.

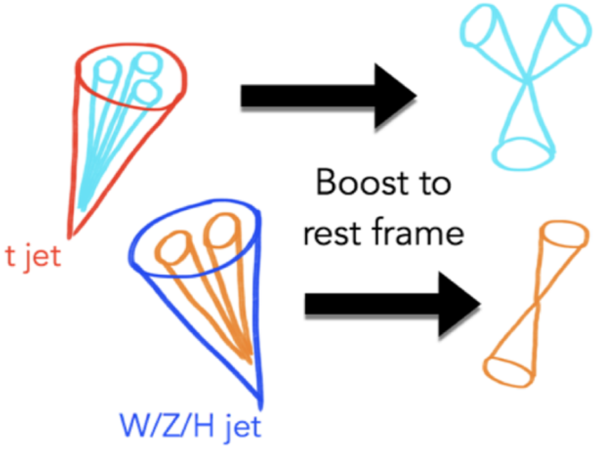


Figure 10.1: Sketch showing two different scenarios. In the left part, the p_T of each candidate is enough to be reconstructed as a boosted jet, whereas, in the right part the p_T is lower than the boosted p_T threshold, resulting into separately reconstructed decaying products.

At leading order in the perturbation theory, the quark and gluon jets should have a mass closer to zero, while the mass of the signal jets should be much higher, close to the intrinsic mass of the top quark or the W, Z or Higgs boson. Therefore, it is clear that the jet mass would in principle be a good discriminant to distinguish jets from hadronically decaying boosted particles from those of quarks or gluons produced by QCD. At very high transverse momenta, however, the width and therefore the mass of QCD jets, may become equally large. In addition, diffuse radiation caused by the underlying event and pileup, give rise to a significant number of additional particles in the event contributing to the total jet mass. Therefore, being able to accurately and efficiently separate highly boosted QCD jets from highly boosted particles requires other methods. In order to remove the underlying event and pileup, algorithms like PUPPI or CSH introduced in section 3.4.4, can be used. In order to improve the mass resolution further, dedicated grooming algorithms must be applied.

10.3 Jet Grooming

Jet grooming is an additional 'post-processing' treatment of large radius jets in order to remove unwanted soft radiation and to allow the underlying hard substructure associated with a two-prong (e.g. Higgs boson) or three-prong (e.g. top quark) decay to be identified more efficiently. In CMS, the grooming methods considered are: trimming, pruning, modified mass drop tagger and soft drop [114]. They all involve the identification of subjets within an original jet, and share the characteristic that they attempt to remove subjets carrying less than some (small) fraction of the original jet's momentum.

10.3.1 Trimming and Pruning

The trimming algorithm [115] is a grooming algorithm mostly used at trigger level in CMS. The first step of the algorithm is to recluster a large anti-k_T or C/A jet, in order to create subjets of some size. It then proceeds to check whether each subjet has a momentum fraction above a certain threshold,

$$p_{T,i}/p_{T,jet} > p_{T,frac} \tag{10.2}$$

The procedure continues as follows, if the subjet fails this requirement, it is removed. The remaining subjets are then assembled into a new "trimmed" jet.

In addition to removing soft particles, the pruning algorithm has an additional requirement on the distance between any recombination that is at wide angle. In particular, it proceeds by reclustering the jet with the C/A algorithm, requiring at each step that

$$z_{ij} = \frac{min(p_{T,i}, p_{T,j})}{p_{T,P}} > z_{cut} \qquad \text{and} \qquad \Delta R_{i,j} < D_{cut} = \frac{2 r_{cut} m_j}{p_T} \tag{10.3}$$

where m_J and p_T are the mass and transverse momentum of the originally-clustered jet, and z_{cut} and r_{cut} are parameters of the algorithm. If these two requirements are not satisfied, i and j are not merged and instead the softer of the two clusters is removed. The z_{ij} requirement ensures that soft particles are discarded, and the ΔR requirement ensures that wide-angle particles are discarded. The resulting jet is referred to as the "pruned" jet.

10.3.2 Modified Mass Drop Tagger and Soft Drop

Like any grooming method, the soft drop tagger removes wide-angle soft radiation from a jet in order to mitigate the effects of contamination from initial state radiation (ISR), underlying event, and pileup. The mass drop tagger (MDT) [116] is based on the assumption that a highly boosted jet is formed by two quark subjets and therefore, the mass of each subjet is much smaller than their combined mass. On the contrary, a

QCD jet is formed by continuous soft radiation, meaning that its heaviest subjet is close to the mass of the jet itself. The MDT tagger therefore, starts from a jet j clustered with the C/A algorithm and then declusters it again, defining the subjets $s1$ and $s2$, where $m_{s1} > m_{s2}$. If a significant mass drop occurred during declustering, $m_{s1} < \mu m_j$, where μ is the mass soft drop parameter [117] and m_j is the mass of the jet j, and the splitting is not too asymmetric, $min(p_{T,s1}^2, p_{T,s2}^2)\Delta R(s1, s2)/m_j^2 > y_{cut}$, then the jet j is selected as the tagged jet. Otherwise j is set equal to s_1 and the procedure starts over. The modified mass drop tagger (mMDT) [118] is a modification of MDT where the subjet with the largest transverse mass, $m^2 + p_T^2$, instead of bare mass, m, is followed. The benefit of that decision is that in cases where the mass drop and asymmetry conditions are not satisfied, the more energetic rather than the heavier branch is followed. In addition, the mMDT algorithm by default uses z_{cut} criteria, as for pruning, rather than y_{cut} criteria. The modified mass drop condition is generalized through the soft drop declustering method [119], simply called Soft Drop, which allows for different types of angular requirements to enter the condition. The Soft Drop condition is the following,

$$\frac{min(p_{T,1}, p_{T,2})}{p_{T,1} + p_{T,2}} > z_{cut} \frac{\Delta R_{12}^{\beta}}{R_0} \tag{10.4}$$

where the asymmetry condition now is defined directly through the transverse momentum fractions of the subjets, rather than through a k_T distance to the jet mass. If the splitting is not too asymmetric, the condition is met and the full jet is considered the softdrop jet. Otherwise only the highest-pT subjet is kept and the declustering continues. If the jet can not be declustered any further, it can either be removed from consideration, so-called “tagging”-mode, or deemed the final soft-dropped jet, “grooming”-mode. The values used in CMS are $z_{cut} = 0.1$ and $\beta = 0$, providing the best signal/background discrimination while maintaining an excellent signal mass resolution. In figure 10.2 the soft drop mass is showed for boosted jets that are matched at parton level with a Higgs (orange) and a top (purple) parton. The transverse momentum of each reconstructed jet is such as all the decay products to be merged in a large radius jet. Those distributions are compared with the M_{SD} distribution of QCD jets. It is clear that for distinguishing Higgs candidates from QCD jets, only the M_{SD} is not sufficient. Therefore, for the purpose of the analysis described in this thesis, methods that exploit machine learning techniques were developed. Those are described in section 10.5.

10.4 N-subjettiness

To further separate signal from background jets, the energy distribution inside the jet is taken into account. For a boosted massive particle, the partons from the decay typically all carry a sizeable fraction of the initial particle’s momentum. Therefore, the resulting jet tends to have multiple hard cores (“prongs”). In contrast, quark and gluon jets, which are dominated by the radiation of soft gluons, typically have only one hard core in each jet. This is illustrated in figure 10.3 showing the decay jets of a Higgs boson, a top quark and a QCD jet.

This characteristic is quantified through the N-subjettiness variable, τ_N, defined as

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} min(\Delta R_{1,k}, \Delta R_{2,k}, ..., \Delta R_{N,k}), \tag{10.5}$$

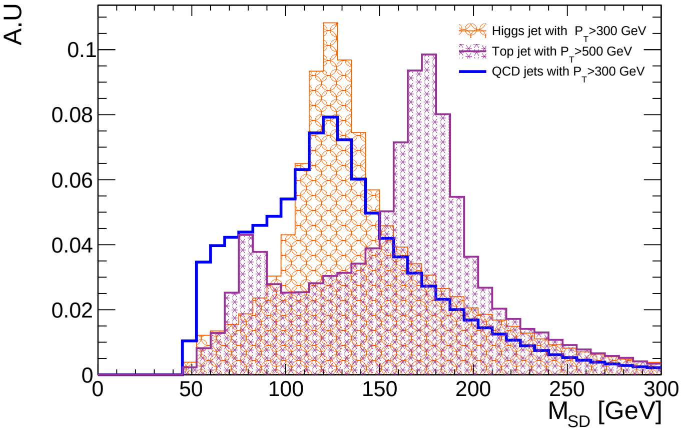


Figure 10.2: M_{SD} distribution for a Higgs (orange) a top (purple) reconstructed jet that is matched with a Higgs and top parton respectively. Those distributions are compared with the M_{SD} distribution of QCD jets.

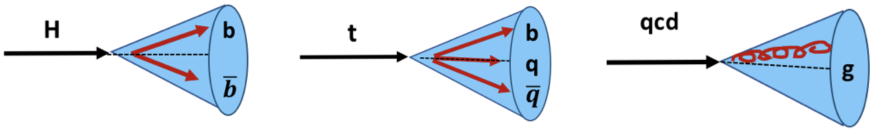


Figure 10.3: A jet stemming from the decay of a Higgs boson will usually have two well-separated high-p_T subjets and a jet stemming from the decay of a top quark will have three well-separated high-p_T subjets, while a jet with a single-prong origin consists of one hard core inside the jet.

where k runs over all the jet constituents, $p_{T,k}$ is the constituent transverse momentum, and $\Delta R_{i,k}$ is the distance between the constituent and candidate subjet axes. The quantity d_0 is a normalization constant defined as:

$$d_0 = \sum_k p_{T,k} R_0, \tag{10.6}$$

where R_0 corresponds to the cone size of the initial jet. The “subjets” are found by performing the exclusive k_T algorithm on the jet constituents before the application of any grooming techniques. With this definition, jets with $\tau_N = 0$ have most of their constituents aligned along the subjet axes. However, if $\tau_N \gg 0$, a large fraction of the energy is radiated away from the subjet directions and the jet is more likely to have more than N subjets. Better discrimination power can be achieved by the ratio of different τ_N variables [120]. This is due to the fact that, while signal jets are expected to have a large τ_1, quark/gluon can similarly have large τ_1 because of the diffuse radiation present. On the contrary, QCD jets with a large τ_1 tend to have an equally large τ_2, while signal jets do not. Therefore, the ratio τ_2/τ_1 can be used to discriminate the two prong W,Z and Higgs jets from QCD jets, while the ratio τ_3/τ_1 can be used for tagging the 3-prong top quark jets.

10.5 Machine Learning

Over the last few decades the use of Machine Learning (ML) methods gained popularity due to recent technology developments and increased computational power. In High–Energy Physics (HEP) in particular, those methods are widely used as they can provide a powerful tool for physics analysis. The most fundamental idea behind ML is the concept of training an algorithm on a preferably large data-set from which it can learn the desired patterns. The training process can either be *supervised*, meaning that the example training dataset comes with a set of true labels or outputs which are known a priori. This is the most common type of ML algorithms. However, there are also *unsupervised* learning techniques in which the algorithm does not have access to the true desired outcome during its training phase. Moreover, there is a third category called *reinforced* learning in which the algorithm does not have access to the true desired output, but instead its predicted outcome is either penalized or rewarded after which is can update its predictions for a next training iteration. The three main approaches and some of their applications are illustrated in figure 10.4.

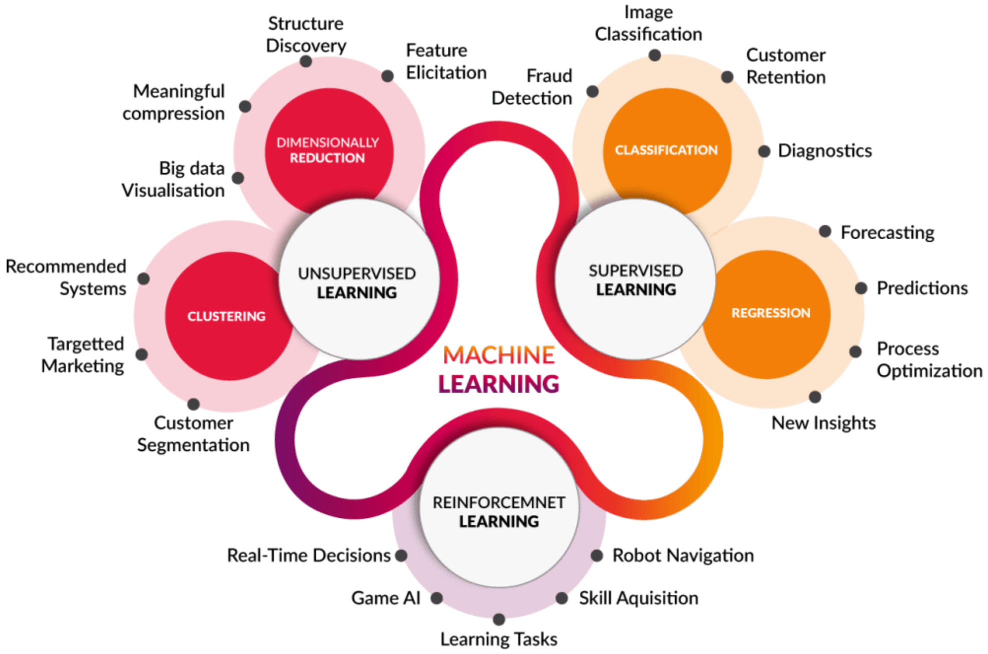


Figure 10.4: Three main approaches in machine learning and examples of the corresponding applications.

In order to make accurate predictions on unseen examples over a large range of the considered phase space, ML algorithms rely heavily on the availability of very large training datasets. The production of large–scale simulations with truth–level information, for HEP analyses provides the perfect environment to train very complex algorithms. Additionally, the training of complex algorithms with thousands or millions of internal parameters on a set of millions of training data requires efficient and fast computing resources. This issue is addressed by the development of Graphics Processing Units (GPU) that allow powerful and resourceful algorithms to be trained at reasonable time scale.

10.5.1 Boosted decision trees (BDTs)

The boosted decision trees are a special application of the principle of $gradient\ boosting$. A boosting algorithm uses many weak-base-classifiers to construct a strong classifier while avoiding over-training. However, instead of training many weak learners at the same time, boosting is an iterative procedure. The first step of the algorithm consists by training the first base–classifier which yields a certain accuracy of the classification performance. Thereafter, a weight is assigned to the account on its accuracy. In addition, the training events are also reweighted by assigning higher weights to misclassified events. This is done to ensure that the misclassified events become more important in the training of the next base–classifier. The algorithm proceeds iteratively, updating the event–weights such that the events which are hardest to classify correctly are given a higher and higher importance throughout the training. The first iterations are able to detect the rough features that separate signal from background, whereas the last iterations should focus on more delicate differences between the features of the different categories. The final ensemble output is a weighed average of the outputs of each base–classifier, with the weights determined by the accuracy of each individual weak learner. Within the context of ML, jet tagging is an ideal task for classification algorithms. For instance, it is very common in particle physics to train BDTs to distinguish between signal and background events as it provides an observable with excellent discriminating power. Therefore, for the analysis presented in this thesis three dedicated BDTs were trained. This will be discussed in detail in the next section.

10.6 BDT-Based Higgs Boson and Top Quark Tagging Algorithms

To distinguish jets coming from Higgs and top decays from QCD jets, all the techniques described in this chapter were exploited in order to improve the performance for identifying boosted candidates with high purity. Therefore, three independent boosted decision tress were trained. All of them used jet-level observables as input variables with sufficient discriminating power between signal and background. Since the analysis strategy is to perform a max-likelihood fit on the groomed soft drop mass (M_{SD}), any potential correlation with the training variables must be avoided. In case of strong correlation the undesired effect of making the QCD background more peak-like rather than smoothly falling, is consequently making QCD more difficult to distinguish from the signal. Therefore, the M_{SD}, although it provides excellent discrimination, is not used in the training. The same holds for the transverse momentum of the boosted jets, since kinematically, it has strong correlation with the groomed mass. The variables that finally used in the training are summarized below:

”N-subjettiness” τ_1,τ_2,τ_3: these variables show an estimate of the energy distribution inside the jet. The number of each variable reflects the number of hard radiation centers. Therefore, it provides strong discrimination between two-prong (Higgs), three-prong (Top) and one-prong (QCD) jets. The "N-subjettiness” τ_1,τ_2,τ_3 are shown in figure 10.5 where distributions from different jets matched at parton level are compared.

mass of the leading and subleading subjet: Signal and background jets show differences in the kinematics of the subjets, and thus they can be exploited to provide extra information to the training. A very discriminating variable of that kind is the mass of the leading and subleading subjet. For instance, when the top jet has a p_T high enough to merge all its decay products in one fat-jet, then a peak on around the mass of the W is expected in the subjet mass distribution. This is not the case for a QCD or a Higgs jet, where a smoothing falling distribution is expected. This is illustrated in the upper pad is figure 10.6 where diiferent jet's leading and subheading mass distributions are compared.

Heavy flavor tagging discriminant: The identification of top quarks and Higgs bosons can benefit greatly from tagging the b quark resulting from their decays. In addition, the distribution of the algorithm used for b-tagging is a powerful variable. This is illustrated in the lower part of figure 10.6. Therefore, the full distribution of the CSVv2 b-tagging discriminant of the subjets is used as an input to the BDTs.

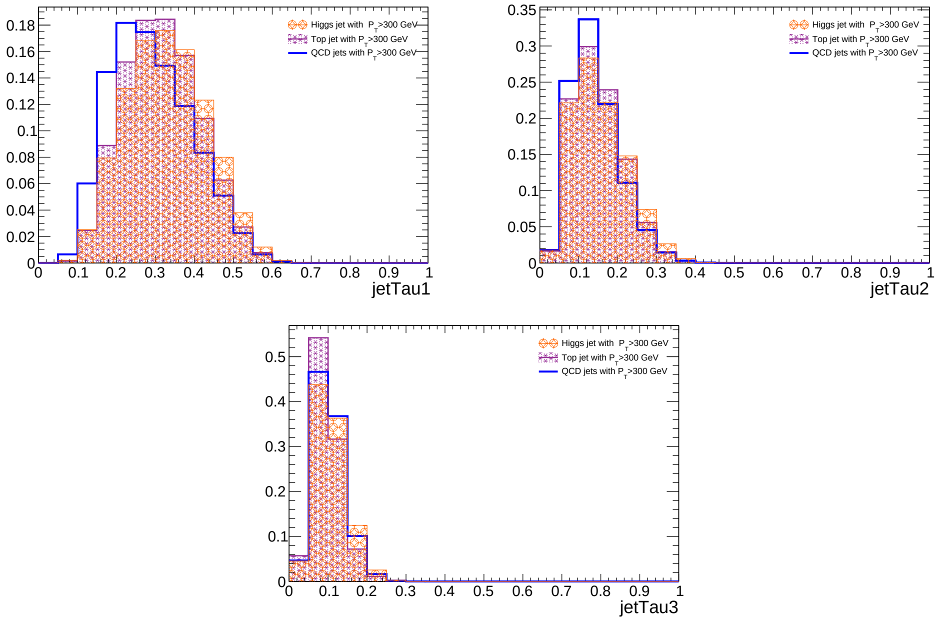


Figure 10.5: τ_1,τ_2,τ_3, variables for a parton level matched Higgs (orange), a top (purple) and a QCD jet (blue). For all distributions the transverse momentum of the jet is greater than 300 GeV.

The top quark and Higgs boson tagging BDTs are trained with simulated jets using the Toolkit for Multivariate Data Analysis (TMVA)[121]. Events that pass the analysis trigger with $p_T > 300\ \mathrm{GeV}$ are selected. In addition, to have a pure hadronic sample, a lepton veto is applied. For the top quark tagging BDT, the signal jets are those matched to top quarks at parton level, whereas, for the Higgs quark tagging BDT, the signal jets are those matched to the Higgs boson parton. The matching requires that a signal Ak8 jet is close to a top quark or a Higgs boson parton, with $\Delta R(\mathrm{jet}, \mathrm{parton}) < 0.3$. In addition to the other selection cuts, we exclude events with $S_T < 900\ \mathrm{GeV}$, where the trigger is not fully efficient. Monte Carlo $\mathrm{t\bar{t}H}$ and $\mathrm{t\bar{t}}$ simulated samples were used to construct signal samples for Higgs tagging BDT and top tagging BDT respectively. In order to have enough statistics for the QCD background sample we did not consider the unmatched jets from the previous samples as background jets. Instead, QCD simu-

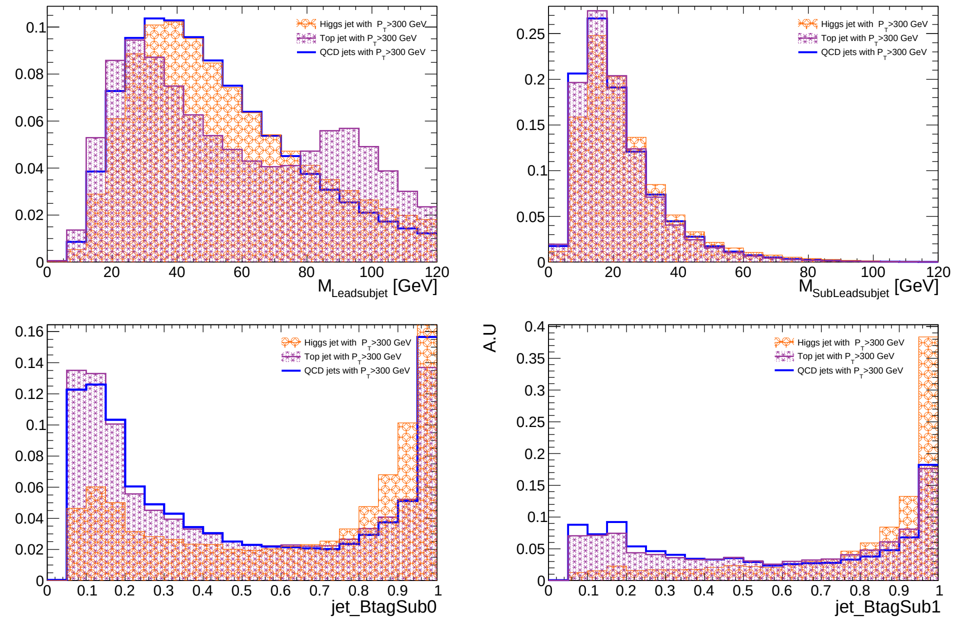


Figure 10.6: The mass (upper pad) and the CSVv2 (bottom pad) distribution of the leading (left) and the subleading (right) for a matched at parton level Higgs (orange), top (purple) and a QCD jet (blue). For all distributions the transverse momentum of the jet is greater than 300 GeV.

lated samples were used, where background jets are those that pass the basic selection criteria.

The two signal samples are trained independently against the same QCD background sample, resulting in two BDT discriminants. Those BDTs are composed of 500 trees and trained with the Gradient Boost method with shrinkage parameter equal to 0.1. The BDT response of the two training is shown in the upper part of figure 10.7. The training and test samples for signal and background are also illustrated. For both BDTs no overtraining is observed. In addition to the two BDTs trained to identify boosted Higgs and Top candidates, a third BDT with the same parameters is trained to distinguish between the two. The response of the training is shown in the lower pad of figure 10.7. The option of multitraining, i.e, training simultaneously for all signal and background to derive one discriminant, was also considered. However, the combination of three BDTs along with other selection cuts, outperforms. This way, by performing optimized cuts on the three BDT discriminants combining with selection requirements, we are able to identify boosted Higgs jets and therefore to reconstruct the M_{SD} of the Higgs candidate, where the max likelihood fit is going to be performed.

10.7 Higgs Boson and Top quark reconstruction

In analyses performed so far that include a Higgs boson decaying to a pair of quarks, the Higgs boson is reconstructed with the "resolved" approach, meaning that the two well separated Ak4 jets from its decays are used for its reconstruction. However, this analysis used a novel approach that considers mainly Lorentz-boosted jets and therefore, the Higgs boson is reconstructed with a single, large-R jet (“boosted jet”). The exploitation of the boosted regime, provides several advantages compared to the tradi-

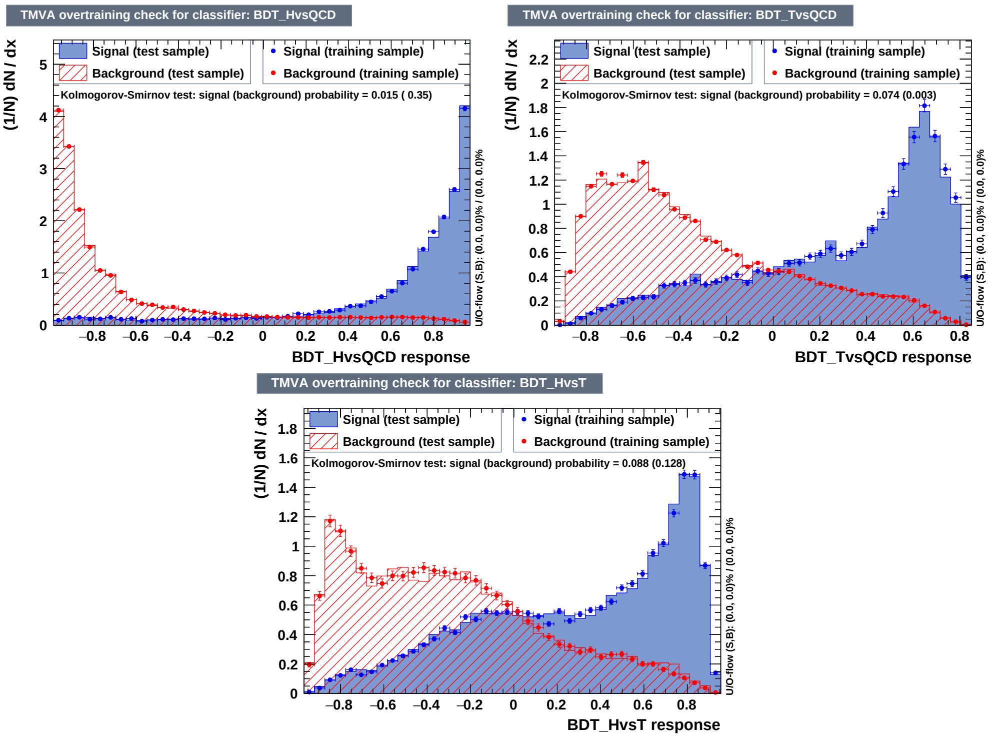


Figure 10.7: BDT response of the three BDTs, taken from the TMVA GUI. In all responses the signal and background test and training samples are shown. No overtraining is observed. The upper left plots illustrates the HvsQ BDT, and the upper right the TvsQ. In the bottom pad the HvsT is shown.

tional searches. Firstly, the boosted regime has better signal purity, though at the cost of a significantly lower signal acceptance. However, the signal acceptance is going to be improved during HL-LHC where luminosity levelling will favour analysis performed in the high momentum regime. In addition, both quarks from the Higgs boson decay can be captured in one jet, which allows for better exploitation of the correlation between the two quarks, and helps to avoid combinatorial backgrounds that can arise in the resolved-jet approach. The same arguments holds when considering boosted top quark candidates. In this analysis the identification of boosted top candidates is done for the purpose of classification the events into several categories based on the number of Ak8 jets. In this section, the identification and reconstruction of boosted top and Higgs candidates, as well as, the performance and validation of the taggers are going to be discussed.

10.7.1 Tagging strategy for Boosted Higgs and Top Candidates

One of the biggest challenges for this analysis is to efficiently reconstruct and identify the boosted Higgs bosons and top quarks, while rejecting background jets arising from multijet production. For this purpose, three BTDs based on substructure observables were trained as described in section 10.6. The output of these BDTs provides three individual scores that separate boosted Higgs jets form QCD jets (HvsQ), boosted top jets from QCD jets (TvsQ) and boosted Higgs jets from boosted top jets (HvsT). Although, the individual trainings provide a good discrimination between signal and background, further requirements are necessary in order to identify boosted objects with high purity.

In this analysis the Higgs candidate is always reconstructed as a boosted jet with a distance parameter of $R = 0.8$. Therefore, the tagging strategy begins with identifying the boosted Higgs boson. An Ak8 jet that pass the baseline selection is considered a boosted Higgs candidate if it has the highest sum of the HvsQ and HvsT scores. In addition, to further suppress the background, the score of HvsQ should be greater than 0.8 and HvsT is required to be greater than 0.1. In order to capture all the decay products of the Higgs boson inside the $R = 0.8$ radius jet, the transverse momentum of the candidate is required to be greater than $300\ \mathrm{GeV}$. In addition, the mass soft drop of the Higgs candidate should be greater than $70\ \mathrm{GeV}$.

After identifying the Higgs boson candidate, we try to identify one or two boosted top quarks. The boosted top quark candidate would be the Ak8 jet with the highest TvsQ score. To further suppress the background from multijet events the TvsQ score must be greater than 0.5. Moreover, the mass soft drop is required to be inside the top mass window ($130 < M_{SD} < 220\ \mathrm{GeV}$). As in the boosted Higgs tagging the transverse momentum of the boosted top candidate is required to be greater than $300\ \mathrm{GeV}$. The selection requirements for identifying boosted Higgs boson and top quark candidates are summarized in table 10.1.

Boosted Higgs Candidate	Boosted Top Candidate
jet with the highest $\mathrm{HvsQ} + \mathrm{HvsT}$	jet with the highest TvsQ score
$\mathrm{HvsQ} > 0.8$ and $\mathrm{HvsT} > 0.1$	$\mathrm{TvsQ} > 0.5$
$p_T^{\text{jet}} > 300\ \mathrm{GeV}$	$p_T^{\text{jet}} > 300\ \mathrm{GeV}$
$m_{SD}^{\text{jet}} > 70\ \mathrm{GeV}$	$130 < m_{SD}^{\text{jet}} < 220\ \mathrm{GeV}$

Table 10.1: Boosted Higgs and Top tagging requirements.

Figure 10.8 shows the composition of the reconstructed Higgs candidate (left) and top candidate (right). The reconstructed Higgs candidate which is matched with a Higgs parton with minimum $\Delta R(\mathrm{jet}, \mathrm{parton}) < 0.3$ (light green distribution) correspond to 56% of the overall reconstructed Higgs candidates. In addition, there is a certain amount of reconstructed Higgs candidates that are matched with a top parton (purple distribution) and a small fraction of reconstructed Higgs candidates that are not matched with either a Higgs or a top parton. Moreover, the purity of top reconstructed candidates is 88%, while there is a small fraction of reconstructed tops that are matched with a Higgs parton and an even smaller fraction that remains unmatched. In addition to the purity, the efficiency of identifying matched Higgs bosons and top quarks as a function of the parton p_T are shown in figure 10.9.

10.7.2 Validation of BDT-taggers in data

The validity of a BDT is typically tested by comparing predicted distributions obtained for simulated and measured data events. A source for potential discrepancies emerges from differences in the modeling of utilized input variables. The selection of well

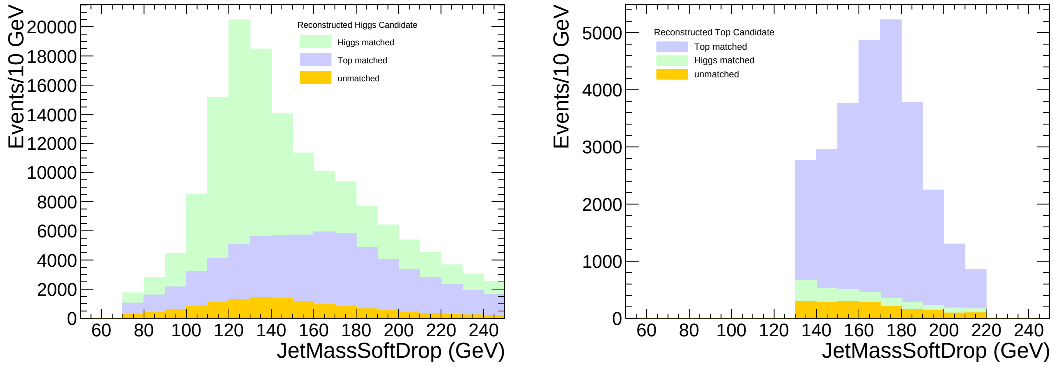


Figure 10.8: Composition of the reconstructed Higgs (left), top (right) candidates taken from simulation. All candidates are matched to a Higgs or top parton with minimum $\Delta R(\mathrm{jet}, \mathrm{parton}) < 0.3$. The purity of the Higgs candidate samples is 56%, whereas, for the top candidate the purity is 88%. There is a certain amount of reconstructed candidates that are matched to the other candidate and a small fraction of them that remains unmatched.

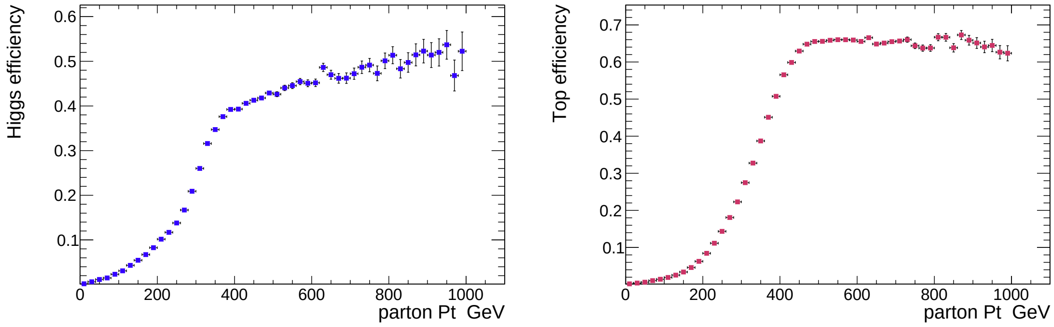


Figure 10.9: efficiency of identifying matched Higgs bosons (left) and top quarks (right) as a function of the parton p_T. The reconstructed candidates have a $p_T > 300$ GeV.

modeled variables is often based on comparisons of distributions between data and simulation. In order to check the robustness of the three BDT tagging algorithms, their output distributions as well as, the variables used in the training are validated in data using a sample enriched in $\mathrm{t\bar{t}}$ events. The $\mathrm{t\bar{t}}$ validation sample and the analysis signal sample consists of mutually excluded events, and therefore it is unbiased. A region enriched in hadronically decaying $\mathrm{t\bar{t}}$ events was constructed by requiring events that pass the basic selection and consists of two Ak8 jets with $p_T > 300$ GeV and $M_{SD} > 70$ GeV. In addition, both Ak8 jets have at least one b tagged subjet. To further suppress the QCD background a tighter cut on the $\mathrm{TvsQ} > 0.7$ was applied. In order to be orthogonal to the signal region and also to enhance hadronically decaying $\mathrm{t\bar{t}}$ events, the HvsT cut is inverted. The selection requirements are summarized in table 10.2. The simulated distributions on the variables used in the training is compared to the one of the data. The comparisons of the N-subjettiness τ_1, τ_2 and τ_3 are shown in figure 10.10. The rest of the variables used in the training that concern the two subjets are illustrated in figure 10.11. In addition, the three BDT distributions are compared with the data, as showed in figure 10.12. The expected yield of the simulated $\mathrm{t\bar{t}}$ events is corrected by a factor derived in section 11.4, to account on simulation mismodelling. Overall, the shapes in data are compatible with the expectation from simulation within uncertainties.

Observable	Requirement
N_{jets}	> 1
N_{leptons}	$= 0$
$Nb_{\mathrm{leading\,subJet}}$	> 0
$Nb_{\mathrm{subleading\,subJet}}$	> 0
$p_T^{\mathrm{leading-jet}}$	> 300 GeV
m_{SD}^{jets}	> 70 GeV
TvsQ	> 0.7
HvsT	< 0.1

Table 10.2: selection requirements for $\mathrm{t\bar{t}}$ enriched validation region

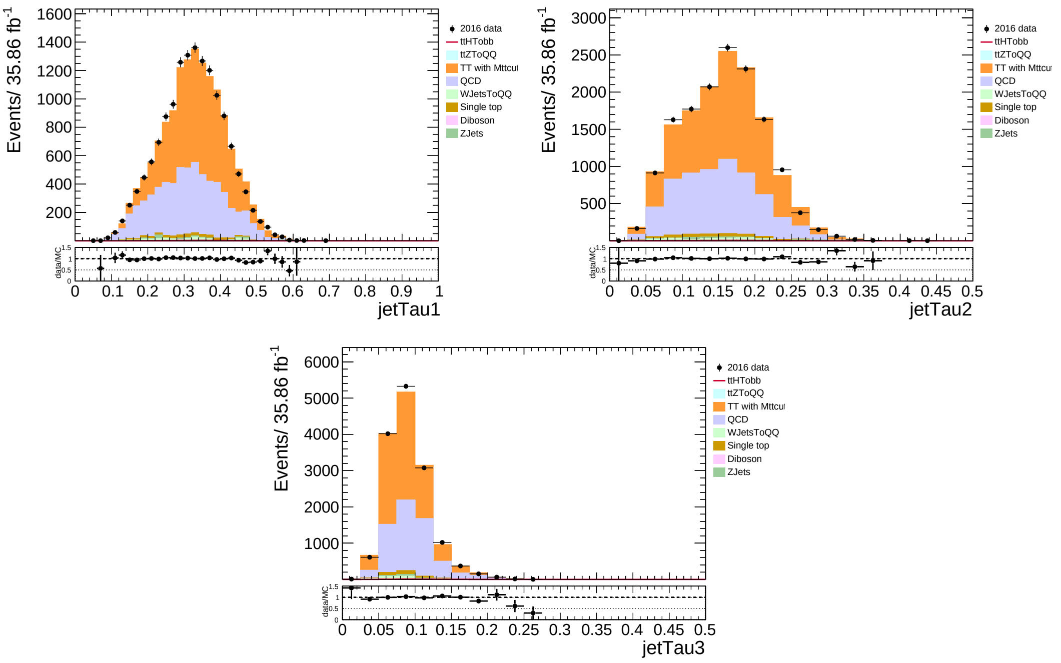


Figure 10.10: N-subjettiness variables τ_1, τ_2 and τ_3 in the $\mathrm{t\bar{t}}$ enriched region. The simulated distributions are compared with the one from the data. The lower pad shows the data to simulation ratio.

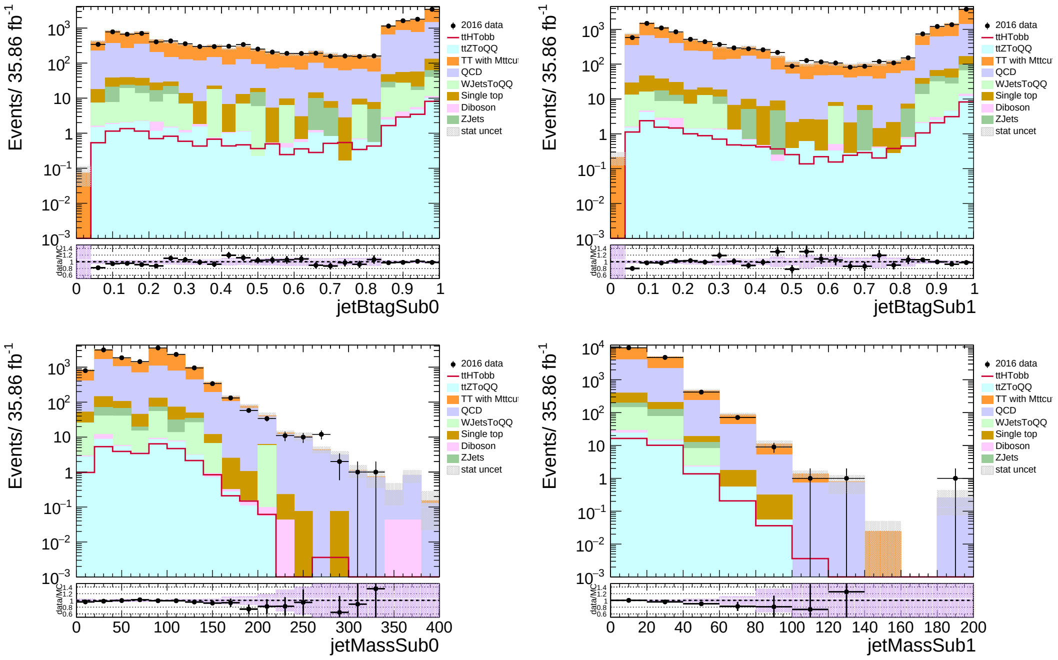


Figure 10.11: The distributions of the CSVv2 discriminant and the mass of the leading (right) and the subleading (left) jet in the $t\bar{t}$ enriched region. The simulated distributions are compared with the one from the data. The lower pad shows the data to simulation ratio.

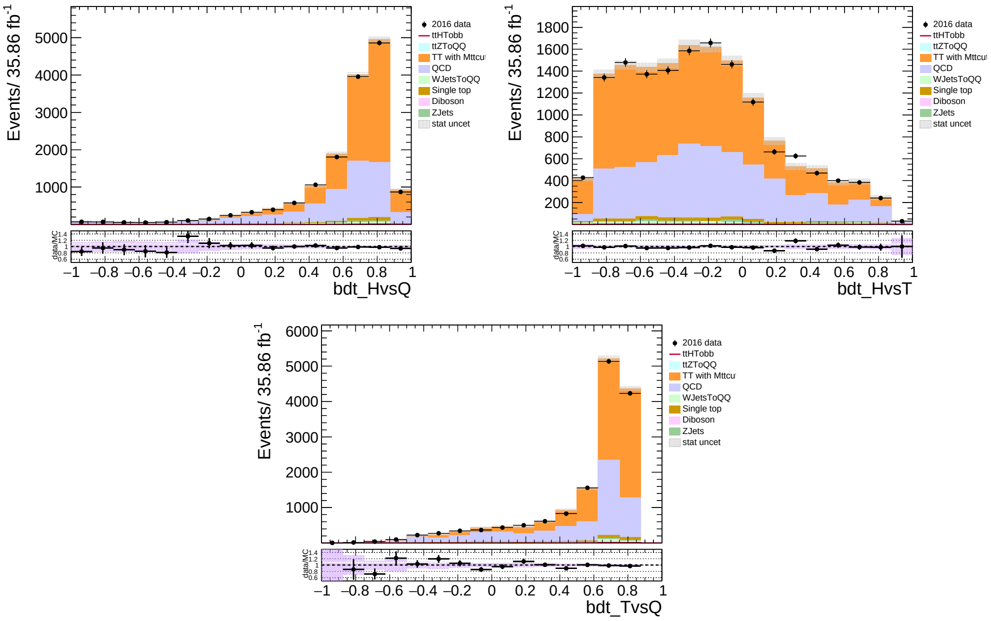


Figure 10.12: The distributions of the BDTs trained for tagging boosted Higgs and top candidates. The HvsQ is illustrated on the upper right, the HvsT in the upper left and the TvsQ on the bottom part. The simulated distributions are compared with the one from the data. The lower pad shows the data to simulation ratio.

10.7.3 Comparison of data and simulation

This analysis is blinded in a jetMassSoftDrop (m_{SD}) window that is close to the mass of the Higgs Candidate for all the signal categories. Therefore, all the selection cuts are optimized in simulation, while no decisions are made by "looking" at the data. However, in order to ensure the robustness of the BDTs, as well as, there are no big discrepancies between data and simulation, simulated distributions of interest are compared with the corresponding ones from the data. In particular, the BDT distributions and the variables used in the training are compared for events that pass the baseline section described in table 9.5 and also have one Ak8 jet reconstructed as the Higgs candidate. Figure 10.13 shows the shapes of the three simulated BDTs distributions. The shaded grey distribution corresponds to the total simulated background which is compared to the distribution of the data. Overall, the agreement between data and simulation is reasonable. For comparison, the BDT discriminants of simulated $\mathrm{t\bar{t}H}$ events are also illustrated on the same figures. It is clear, that the BDTs are able to discriminate potential signal events.

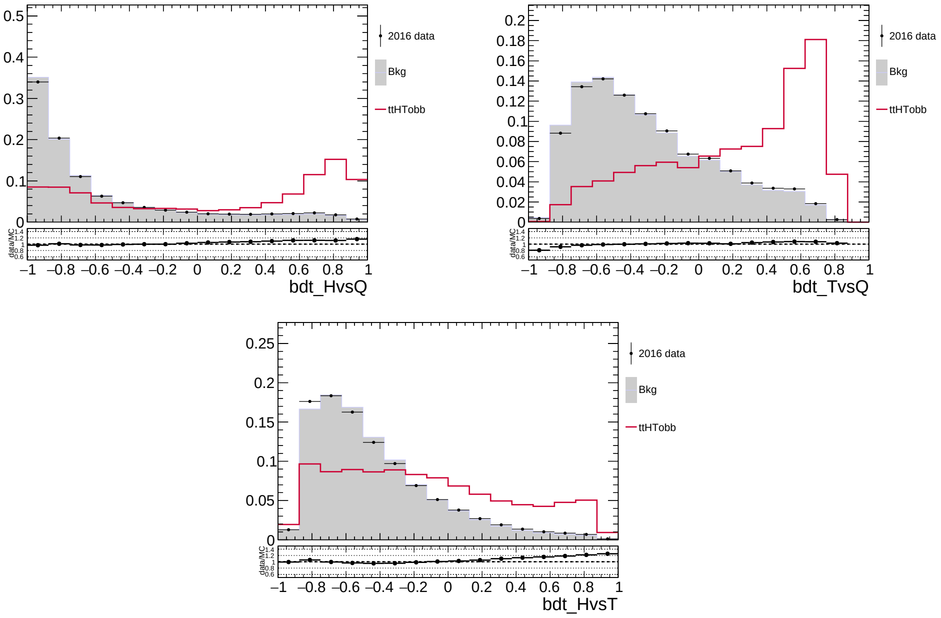


Figure 10.13: Comparison of HvsQ (upper left), TvsQ (upper right), HvsT (bottom) simulated distributions with the data. The distributions are normalized so that the total integral is equal to one and thus, the small discrepancies are due to differences in the shape. On the lower pad, the ratio between data and simulation is shown. For comparison purposes, the BDT distribution of the simulated signal sample is illustrated (red line) in the same plot.

In addition, the N-subjettiness variables τ_1, τ_2 and τ_3, (figure 10.14), and the mass and CSVv2 distributions of the leading and subleading jets, (figure 10.15), are compared for data and simulation. Moreover, two simulated $\mathrm{t\bar{t}H}$ signals, one for a Higgs boson decaying into a pair of quarks and one that considers all the other decays, are illustrated on the same plots. All the simulated distributions are scaled by their corresponding cross section. The QCD multijet background has been estimated from MC simulation, but due to its large cross section uncertainty, it is scaled to fill the gap in yield between other simulated backgrounds and the data. This factor is between 0.9 and 1.1. In general there is a good agreement between data and simulation.

After validating the BDTs both in the signal region and in the orthogonal region enriched in $\mathrm{t\bar{t}}$ events, then it is important to check the consistency between data and simulation of the reconstructed M_{SD} distribution of the Higgs candidate. The check

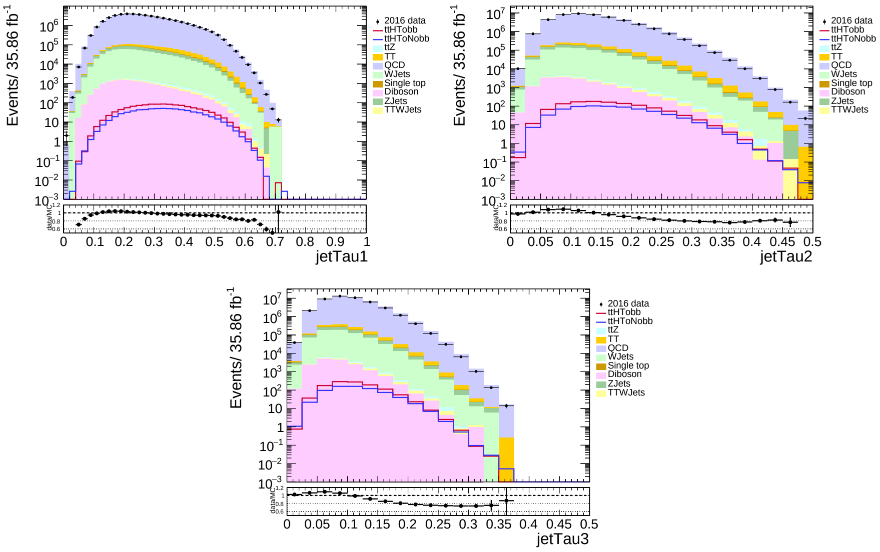


Figure 10.14: The N-subjettiness variables τ_1, τ_2 and τ_3 for data and simulation. Each background contribution is illustrated with a different color and is compared with the data (black dots). On the lower pad, the ratio of data and simulation is shown. For comparison, two simulated $\mathrm{t\bar{t}H}$ signals, one for a Higgs boson decaying into a pair of quarks and one that considers all the other decays, are illustrated on the same plots. The simulated backgrounds are first scaled to the luminosity of the data, and then the simulated QCD multijet background is rescaled to match the yield in data.

is performed before applying additional requirements to classify events in analysis categories. Therefore, we can evaluate the reconstructed Higgs candidate and spot any discrepancies while remaining unblinded. The reconstructed Higgs candidate is shown on the upper part of figure 10.16. In addition, the reconstructed top candidate is illustrated on the bottom part of figure 10.16. All in all, there is a good agreement between data and simulation.

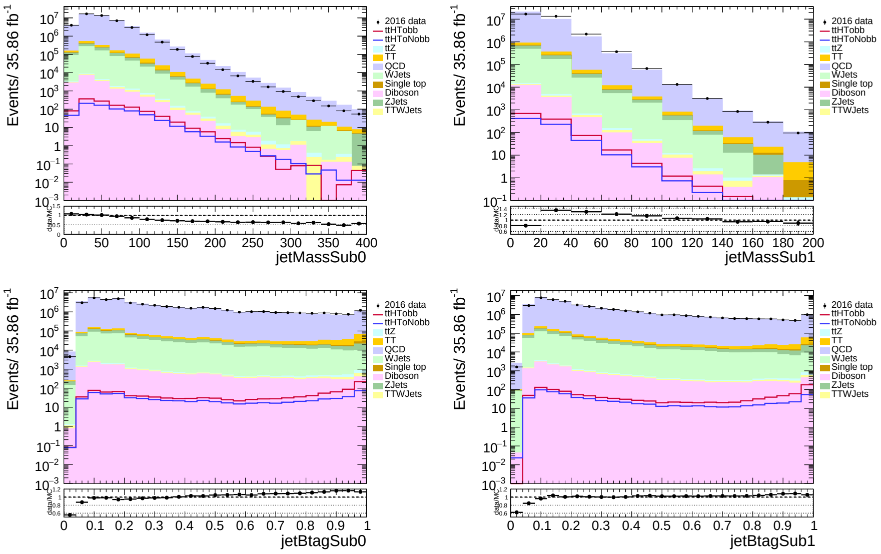


Figure 10.15: The distributions of the CSVv2 discriminant and the mass of the leading (right) and the subleading (left) jet are compared for data and simulation. Each background contribution is illustrated with a different color and is compared with the data (black dots). On the lower pad, the ratio of data and simulation is shown. For comparison, two simulated $t\bar{t}H$ signals, one for a Higgs boson decaying into a pair of quarks and one that considers all the other decays, are illustrated on the same plots. The simulated backgrounds are first scaled to the luminosity of the data, and then the simulated QCD multijet background is rescaled to match the yield in data.

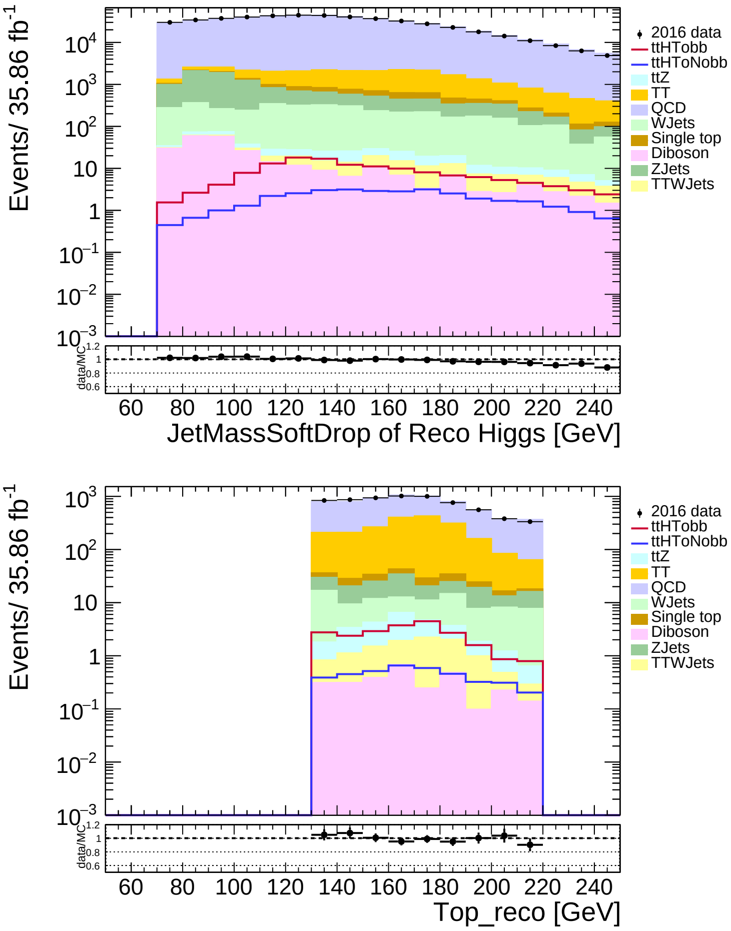


Figure 10.16: Comparison between data and simulation of the reconstructed Higgs (up) and top (down) candidates before entering in the analysis categories. Each background contribution is illustrated with a different color and is compared with the data (black dots). On the lower pad, the ratio of data and simulation is shown. For comparison, two simulated $t\bar{t}H$ signals, one for a Higgs boson decaying into a pair of quarks and one that considers all the other decays, are illustrated on the same plots. The simulated backgrounds are first scaled to the luminosity of the data, and then the simulated QCD multijet background is rescaled to match the yield in data.

Chapter 11

Analysis methods

11.1 Event Categories

The definition of the analysis phase space is obtained by the selection procedure. As described in section 9.6, events that do not pass the baseline selection are discarded. In addition, a set of dedicated selection cuts discussed in section 10.7.1, are optimized in order to reconstruct efficiently boosted Higgs and top candidates. A typical $\mathrm{t\bar{t}H}$ signal in the high-p_T regime has a clear signature of a Lorentz-boosted Higgs boson with transverse momentum greater than $300~\mathrm{GeV}$. Depending on the transverse momentum of the top quark ($p_T > 400~\mathrm{GeV}$), its decay products can be reconstructed in a large radius jet and therefore an Ak8 jet has a probability to be tagged as top. However, even though an event could have large Ak8 jet multiplicity, there is a probability the Ak8 jets to fail the top candidate tagging requirements. Moreover, in cases when the transverse momentum is below the given threshold, then the top quark decay products can be reconstructed with the resolved approach. In this case, the events will include a large number of Ak4 jets some of which will be b-tagged.

Therefore, to further enhance the analysis sensitivity, we split the phase space into orthogonal categories based on the Ak8 multiplicity, that account on all the aforementioned cases. Events with a boosted Higgs candidate are selected and then, cases with three, two or one Ak8 jets are considered. The categories are further divided by the presence or not of top tagged Ak8 jets, the Ak4 jet multiplicity and the presence or not of b tagged Ak4 jets, leaving in total nine mutual excluded categories. The nine signal categories along with their requirements are summarized on figure 11.1 and are described below.

categories with three Ak8 jets

– *category* 0: In this category one of the three Ak8 jets is tagged as the Higgs candidate and at least one of the remaining two Ak8 jets are tagged as a top candidate. Since we expect all the products of the $\mathrm{t\bar{t}H}$ decays to be merged in each of the three Ak8 jets, events with no b-tagged resolved jets are selected. However, to enhance the signal acceptance, only events with loose b-tagged resolved jets are discarded. The loose working point corresponds to an efficiency of tagging a real b-jet of 81% with misidentification probability for jets from light-flavour quarks or gluons of 9%. The composition of the reconstructed Higgs candidate of category 0 is illustrated in figure 11.2a. This category has the largest purity of reconstructed candidates which is almost 73%. The purity of the top candidate is illustrated in figure 11.3d which reflects in a purity of 87%.

- $category$ 1: As mentioned before, there is a finite probability of a real high-momentum top quark to fail to be reconstructed as a top candidate. Therefore, this category consists of three Ak8 jets, one of which is tagged as the Higgs boson and none of the other two is tagged as a top quark. The same argument concerning resolved b-jet multiplicity holds here as well and so events with loose b-tagged resolved jets are discarded. The purity of the reconstructed Higgs candidate for category 1 is illustrated in figure 11.2b where 51% of all the reconstructed Higgs candidates are matched with a Higgs parton within $\Delta R < 0.3$.

categories with two Ak8 jets

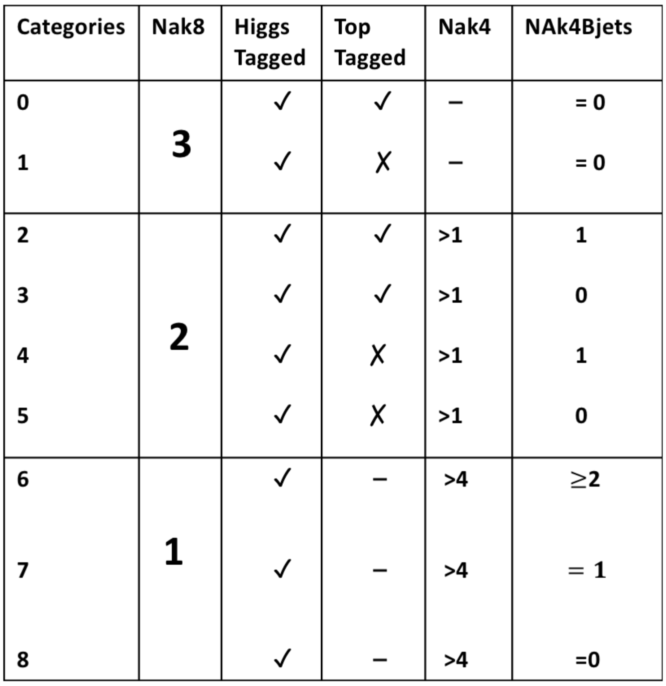
- $category$ 2: This category includes events with two Ak8 jets, one tagged as Higgs and the other tagged as top. The decay products of the other top quark can be reconstructed with the resolved approach and thus, events with at least two resolved jets, one of which is b-tagged, are considered. The purity of the reconstructed Higgs candidate for category 2 is 53% and is illustrated in figure 11.2c. The purity for tagging the top candidate is 88% and is illustrated in figure 11.3e.
- $category$ 3: The only difference between category 2 and category 3 is that in the latter, no medium b-tagged resolved jet is found. For the medium working point, the CSVv2 probability of tagging a real b-tag jet is 62% and therefore, there is a finite probability real b jets from low-momentum top quark decays to fail to be tagged. The purity of the reconstructed Higgs candidate for category 3 is 62% and is illustrated in figure 11.2d. The purity for tagging the top candidate is 87% and is illustrated in figure 11.3f.
- $category$ 4: The only difference between category 2 and category 4, is that in the latter, neither of the jets is tagged as a top. This can happen if the second jet fails for instance the TvsQ cut. The purity of the reconstructed Higgs candidate for category 4 is 52% and is illustrated in figure 11.2e.
- $category$ 5: Category 5 consists of events with two Ak8 jets, one of which is tagged as the Higgs boson while the other one is not tagged as a top candidate. In addition, no b-tagged resolved jet is found. The purity of the reconstructed Higgs candidate for category 5 is 60% and is illustrated in figure 11.2f.

categories with one Ak8 jet

- $category$ 6: In this category the only Ak8 jet is tagged as a Higgs candidate. The decay products of the two top quarks of the $\mathrm{t\bar{t}H}$ signal are reconstructed in the resolved regime and thus, events with at least 5 resolved jets, at least two of which b-tagged, are selected. The purity of the reconstructed Higgs candidate for category 6 is 38% and is illustrated in figure 11.3a. This category has the lowest purity compared to the other categories. However, combining with the rest signal categories contributes in the overall signal sensitivity.

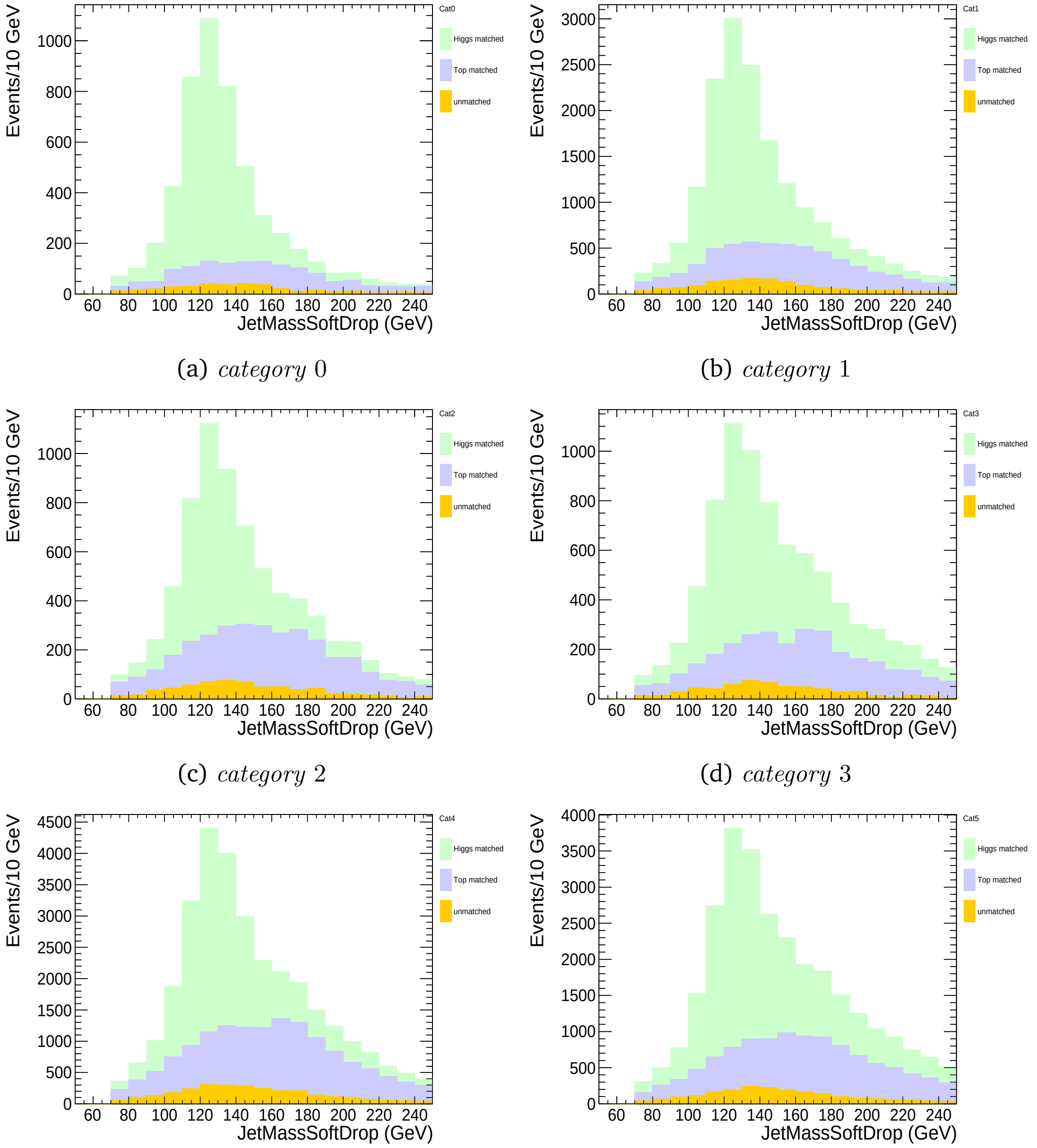
- *category* 7: This category consists of events that have only one Ak8 jet tagged as the Higgs candidate and at least 5 resolved jets. However, compared to category 6, cases with exactly one b-tagged jets are considered. The purity of the reconstructed Higgs candidate for category 7 is 51% and is illustrated in figure 11.3b.

- *category* 8: This category consists of events that have only one Ak8 jet tagged as the Higgs candidate and at least 5 resolved jets. However, compared to category 6, cases with no b-tagged jets are considered. The purity of the reconstructed Higgs candidate for category 8 is 57% and is illustrated in figure 11.3c.



Categories	Nak8	Higgs Tagged	Top Tagged	Nak4	NAk4Bjets
0	3	✓	✓	–	= 0
1		✓	✗	–	= 0
2	2	✓	✓	>1	1
3		✓	✓	>1	0
4		✓	✗	>1	1
5		✓	✗	>1	0
6	1	✓	–	>4	≥2
7		✓	–	>4	= 1
8		✓	–	>4	=0

Figure 11.1: Analysis categories. Events are selected if a boosted Higgs candidate is identified and then are further divided into 9 mutual excluded categories based on Ak8 jet multiplicity. The categories are further divided by the presence or not of top tagged boosted jets, the Ak4 jet multiplicity and the presence or not of b tagged Ak4 jets.



(a) *category* 0

(b) *category* 1

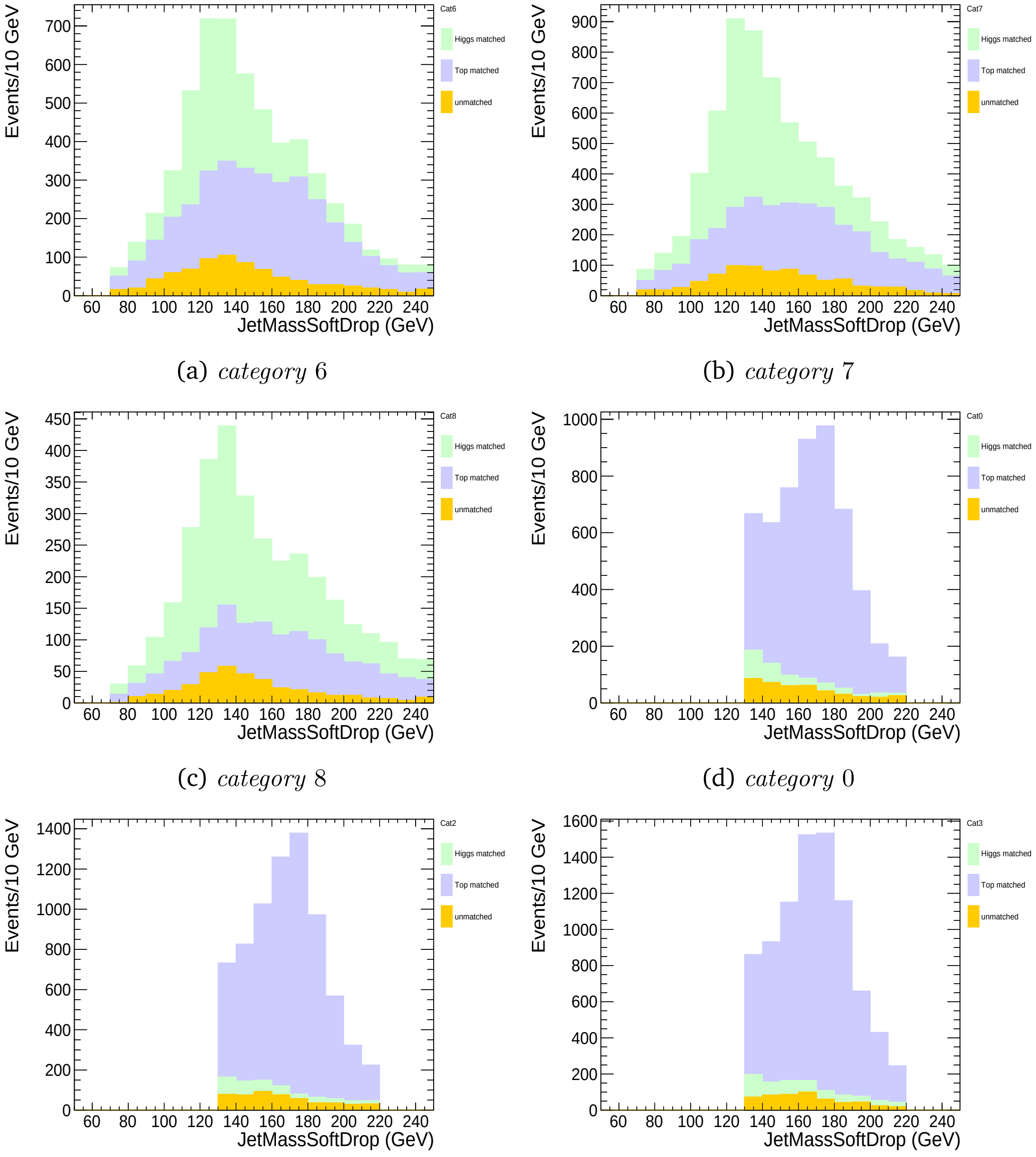
(c) *category* 2

(d) *category* 3

(e) *category* 4

(f) *category* 5

Figure 11.2: Composition of the reconstructed Higgs candidates for categories with three (a,b) and two (c,d,e,f) Ak8 jets. For all categories, the Higgs candidates are matched to a Higgs parton with minimum $\Delta R(\mathrm{jet}, \mathrm{parton}) < 0.3$. In addition, there is a small probability of the reconstructed Higgs candidate to be matched with a top parton and a smaller to remain unmatched.



(a) *category* 6

(b) *category* 7

(c) *category* 8

(d) *category* 0

(e) *category* 2

(f) *category* 3

Figure 11.3: Composition of the reconstructed Higgs candidates for categories with one (a,b,c) Ak8 jet. For all categories, the Higgs candidates are matched to a Higgs parton with minimum $\Delta R(\mathrm{jet}, \mathrm{parton}) < 0.3$. In addition, there is a small probability of the reconstructed Higgs candidate to be matched with a top parton and a smaller to remain unmatched. Composition of the reconstructed top candidate for categories that require top tagging (d,e,f). For all categories, the top candidates are matched to a top parton with minimum $\Delta R(\mathrm{jet}, \mathrm{parton}) < 0.3$. In addition, there is a small probability of the reconstructed top candidate to be matched with a Higgs parton and a smaller to remain unmatched.

After the application of all selection cuts and classification of events in the analysis categories, a certain number of generated events remain for further analysis. Those are presented in the upper part of figure 11.4 which shows the expected yields for signal and background processes for each analysis category. The QCD multijet background has been estimated from MC simulation, but due to its large cross section uncertainty, it is scaled to fill the gap in yield between other simulated backgrounds and the data. In the lower part of figure 11.4 the background composition is shown for each category. It is clear that for all the categories the dominant background comes from multijet processes. In addition, categories that require a top candidate such as category 1 have a larger contribution from $t\bar{t}$ events compared to categories with zero tagged top quarks. There is also a small contribution of subdominant backgrounds in all categories.

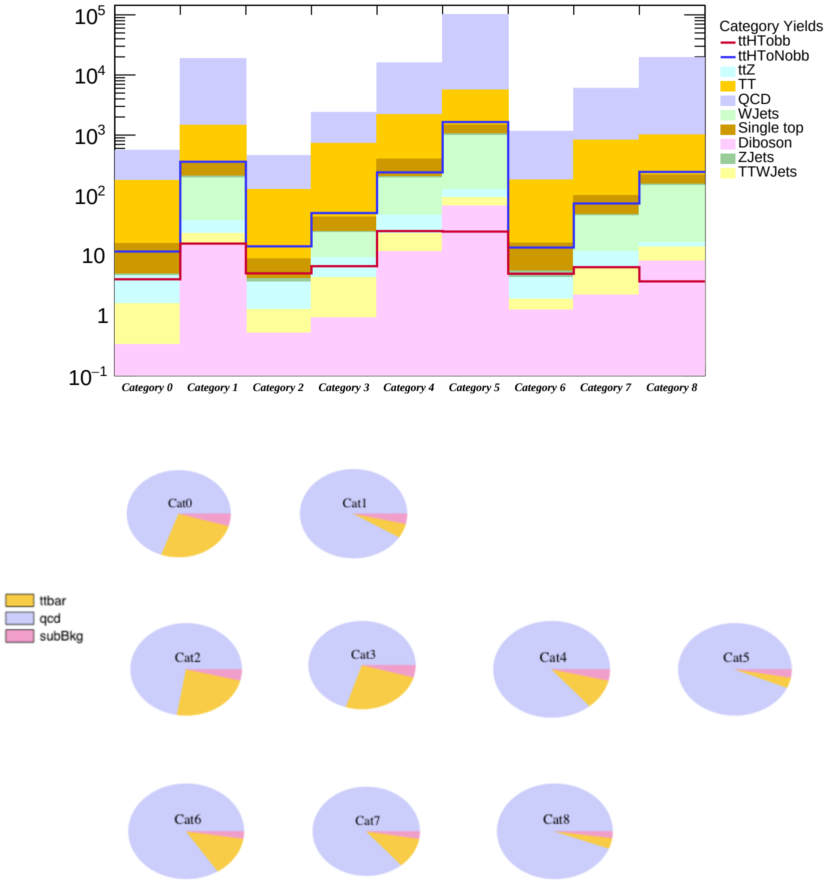


Figure 11.4: In the upper part, the expected signal and background yields for all categories are shown. The simulated backgrounds are first scaled to the luminosity of the data, and then the simulated QCD multijet background is rescaled to match the yield in data. In the lower part, the background composition of each category is illustrated for each category. The dominant background comes from mutijet processes. Categories that require a top candidate have a larger contribution from $t\bar{t}$ events compared to the ones who do not. There is also a small contribution of subdominant backgrounds in all categories.

11.2 Comparison between data and simulation

To ensure a good understanding of the background processes contributing to the selected events before entering in the analysis categories, the distributions of various quantities that are used to classify the events in the analysis categories have been compared in data and simulation. More specifically, the agreement for the multiplicity of Ak8 jets and Ak4 jets for events that pass the baseline selection and have a reconstructed Higgs candidate is shown in the upper part of figure 11.5. In addition, the number of b-tagged Ak4 jets under the condition that in the event there are at least five resolved jets, is illustrated in the lower part of figure 11.5. This requirement holds for categories with one Ak8 jet. The agreement between the data and simulation is good and therefore the classification of events in the categories is well-modeled.

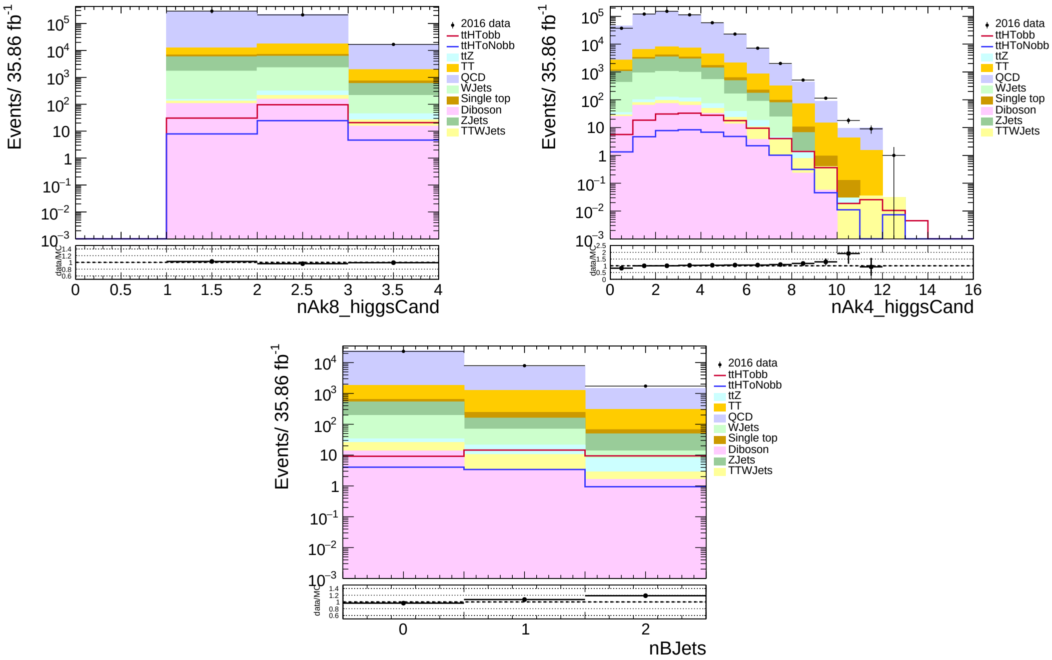


Figure 11.5: Ak8 (left) and Ak4 (right) jet multiplicity for events that pass the baseline selection. In addition, the number of b-tagged resolved jets (bottom) is shown under the condition that in the event there are at least five resolved jets. The last bin includes the overflow bin. On the lower pad, the ratio of data and simulation is shown. For comparison, two simulated $t\bar{t}H$ signals, one for a Higgs boson decaying into a pair of quarks and one that considers all the other decays, are illustrated on the same plots. The simulated backgrounds are first scaled to the luminosity of the data, and then the simulated QCD multijet background is rescaled to match the yield in data.

11.3 Estimation of the QCD background

As mentioned in section 9.2.2, the by far most dominant background consists of events from QCD multijet production, as there is a finite probability that ordinary jets, from single parton radiation, will mimic the topological substructure of a top or higgs decaying jet. Modelling this background from simulation will introduce large uncertainties due to the large uncertainties on the QCD cross sections. In addition, it is unfeasible to calculate and generate simulated events at a scale that is needed. Therefore, it is difficult to obtain reliable descriptions of the QCD background. In final states involving many jets, data-driven methods are used to estimate the backgrounds to the processes of interest. A data-driven method is any method that makes use of the data in the “background” dominated region to estimate background contribution in the “signal” region, where interesting events are expected. For the QCD multijet background modelling, two independent data-driven methods were developed. The first one is used to model the QCD shape and is derived from a control region from the data. The second method gives a prediction on the expected QCD yield for each category.

11.3.1 Estimation of the QCD shape

The dedicated BDTs trained for Higgs and Top tagging manage to suppress significantly the contribution of the QCD background. The remaining QCD background is taken directly from the data using a control region which is orthogonal to the signal regions, but with similar kinematic properties, and thus capable to describe the expected QCD shape. The control region in the data is constructed by selecting 2 Ak8 jets with transverse momentum greater than $300\ \mathrm{GeV}$ and $M_{SD} > 70\ \mathrm{GeV}$. To ensure orthogonality with the signal region the BDT scores are reversed. However, in order to be as signal-like as possible lower limits are given to the BDTs. In particular, only events with $0.1 < \mathrm{HvsQ} < 0.8$ and $0.1 < \mathrm{TvsQ} < 0.5$ scores are selected. Those requirements are summarized on table 11.1. The BDT side-bands were chosen in order to maintain enough statistics in the control region and on the same time to preserve similar kinematic characteristics with the signal.

In the QCD control region no Ak8 jet is tagged as the Higgs candidate. Therefore, the observable used to model the QCD shape is the m_{SD} of the subleading jet. The m_{SD} of the subleading jet is compared for data and simulation as shown in figure 11.6. It is clear that this region is enriched in multijets events, where the contribution from other backgrounds is rather small. In the lower pad, the ratio for data and simulation is shown.

In order to validate the method and to ensure that there is a good description of the expected QCD shape in the signal region, a study was performed in simulation. In particular, QCD simulated events were used to construct the control region and the nine signal regions. Then, the shape of m_{SD} of the subleading jet of the control region is compared with the shape of the m_{SD} of the reconstructed Higgs candidate for QCD events that end up in the signal region. The choice of the m_{SD} of the subleading jet is motivated by the fact that for categories that require a top tagged Ak8 jet, the Higgs candidate most of the time corresponds to the second leading jet. In addition, the second leading jet distribution describes well the rest of the categories. This is shown in figure 11.7 for categories with two or more Ak8 jets and in figure 11.8 for categories with only one Ak8 jet.

The consistency check in simulation is illustrated on the lower pad of each comparison plot of figure 11.7 and 11.8 and the agreement is observed within the statistical uncertainties. However, in order to correct the observed mismodelling, the ratio is fitted with a linear function and a transfer factor is assigned to each analysis category. The corrected QCD shape from the data is compared with the corresponding simulated shape from the signal region in figure 11.9 for categories with two and three Ak8 jets and figure 11.10 for categories with only one Ak8 jet. The corrected distributions show a better closure and are used to estimate the QCD shape for each analysis categories.

Observable	Requirement
N_{jets}	$= 2$
N_{leptons}	$= 0$
$p_T^{\text{jet0,jet1}}$	> 300 GeV
$m_{SD}^{\text{jet0,jet1}}$	> 50 GeV
$\text{HvsQ}_{\text{jet0,jet1}}$	$0.1 < \text{HvsQ} < 0.8$
$\text{TvsQ}_{\text{jet0,jet1}}$	$0.1 < \text{TvsQ} < 0.5$
$\text{HvsT}_{\text{jet0,jet1}}$	$\text{HvsT} < 0.1$

Table 11.1: Selection for QCD control region.

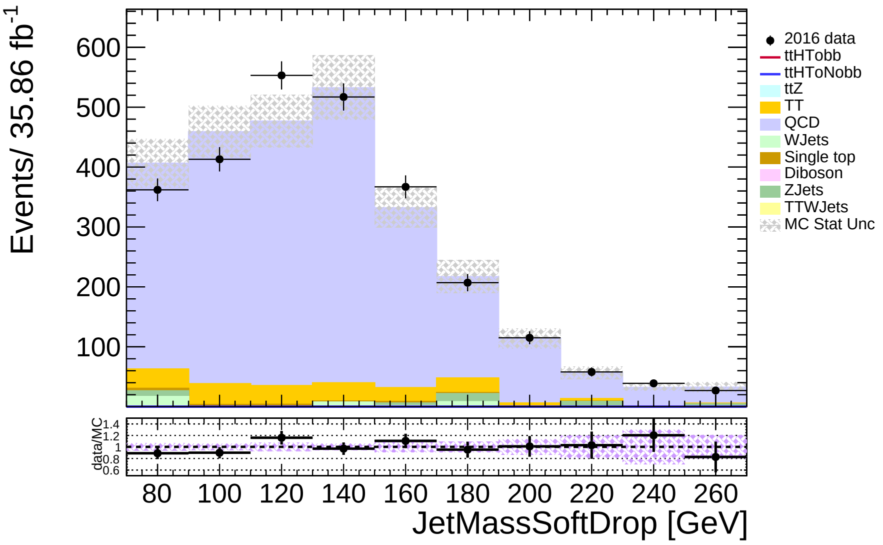
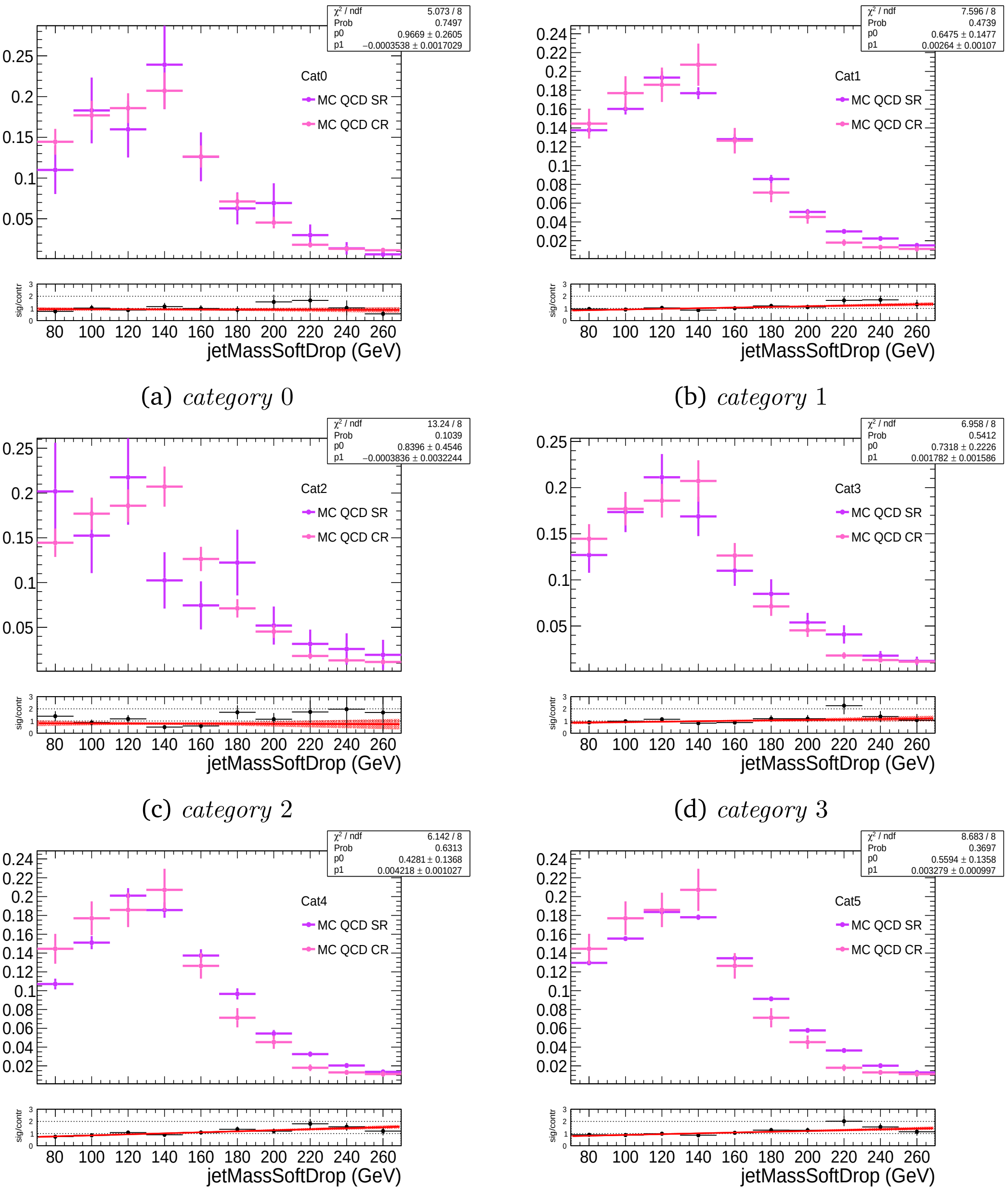


Figure 11.6: M_{SD} distribution of the QCD control region for data and simulation. Each background contribution is illustrated with a different color and is compared with the data (black dots). On the lower pad, the ratio of data and simulation is shown. The simulated backgrounds are first scaled to the luminosity of the data, and then the simulated QCD multijet background is rescaled to match the yield in data.



(a) *category* 0

(b) *category* 1

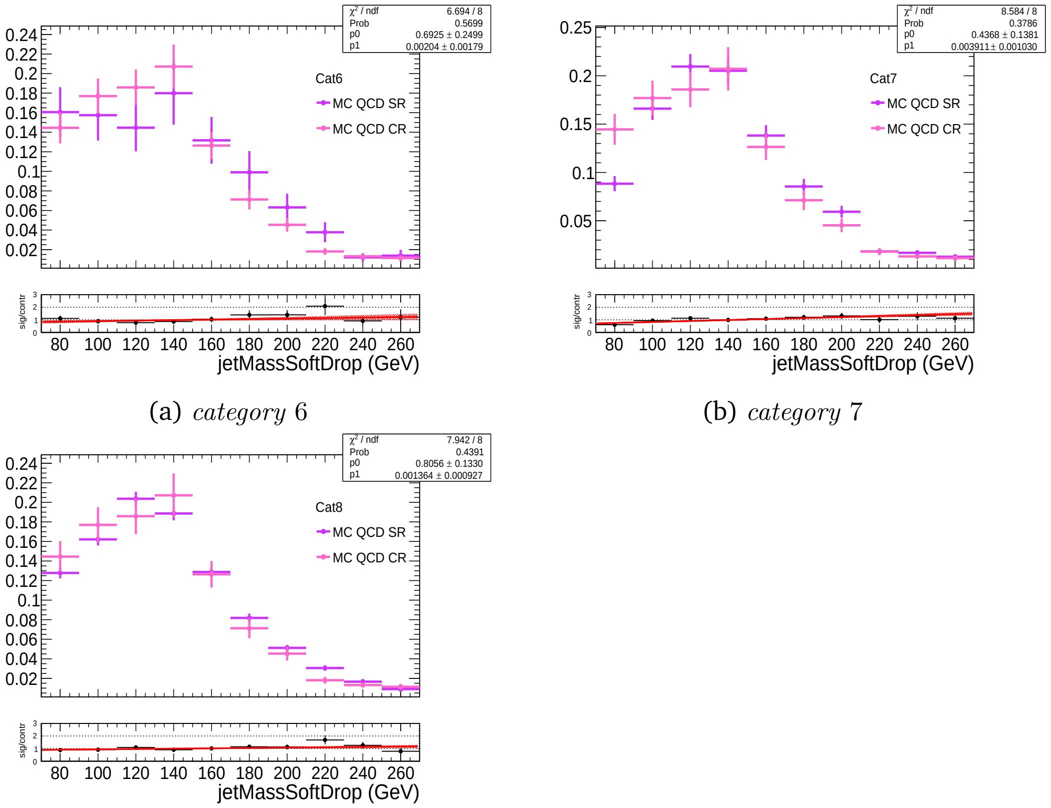
(c) *category* 2

(d) *category* 3

(e) *category* 4

(f) *category* 5

Figure 11.7: Closure test on the estimation of the QCD shape for categories with three Ak8 jets (cat0, cat1) and two Ak8 jet (cat2, cat3, cat4, cat5). For each category, QCD simulated events were used to construct the distributions of m_{SD} in the signal region and in the control region. The distributions are normalized so that the total integral is equal to one and thus, the small discrepancies are due to differences in the shape. The ratio of the two distributions is fitted with a linear function and a transfer factor in obtained for each category.

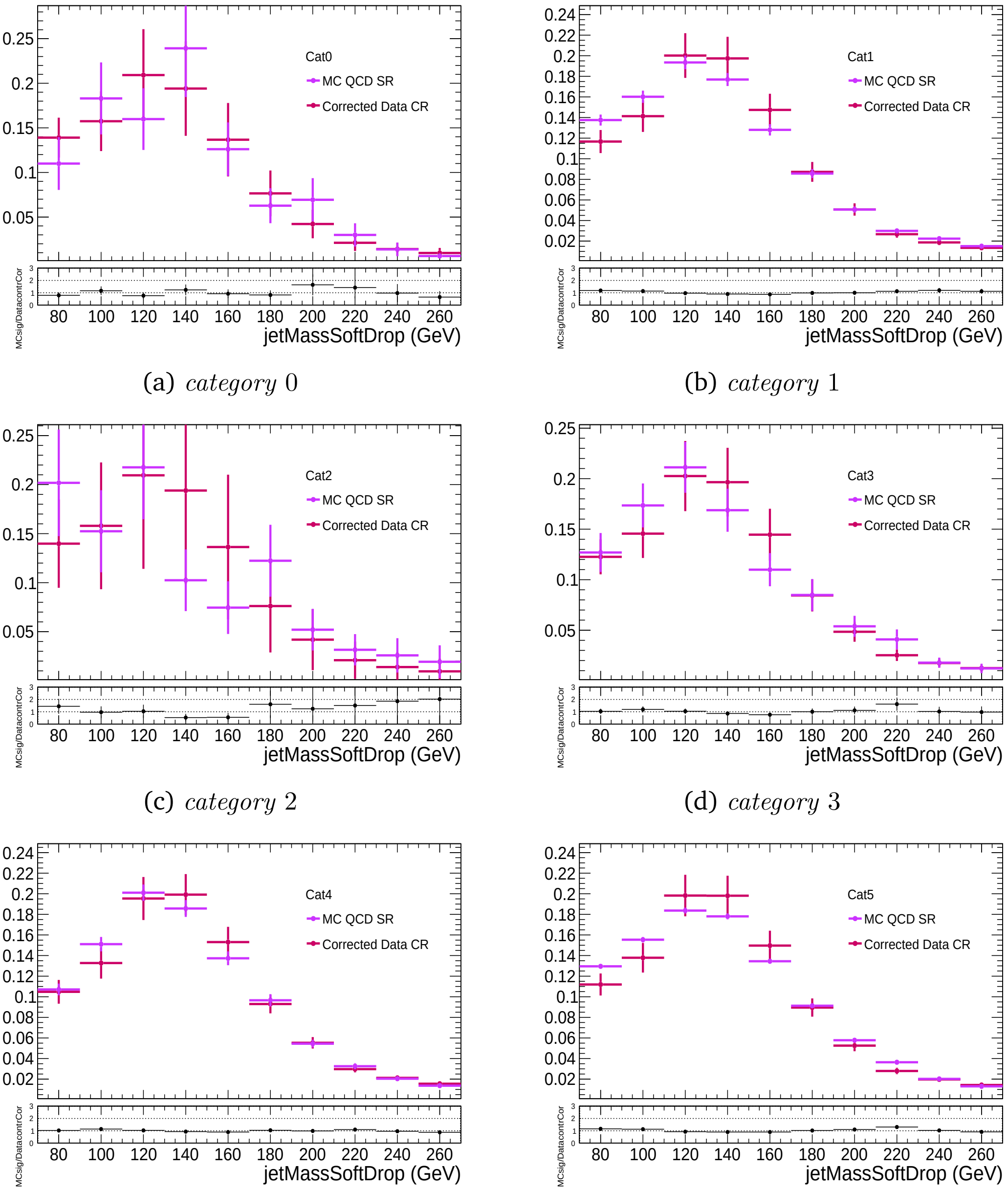


(a) *category* 6

(b) *category* 7

(c) *category* 8

Figure 11.8: Closure test on the estimation of the QCD shape for categories with one Ak8 jet (cat6, cat7, cat8). For each category, QCD simulated events were used to construct the distributions of m_{SD} in the signal region and in the control region. The distributions are normalized so that the total integral is equal to one and thus, the small discrepancies are due to differences in the shape. The ratio of the two distributions is fitted with a linear function and a transfer factor in obtained for each category.



(a) *category* 0

(b) *category* 1

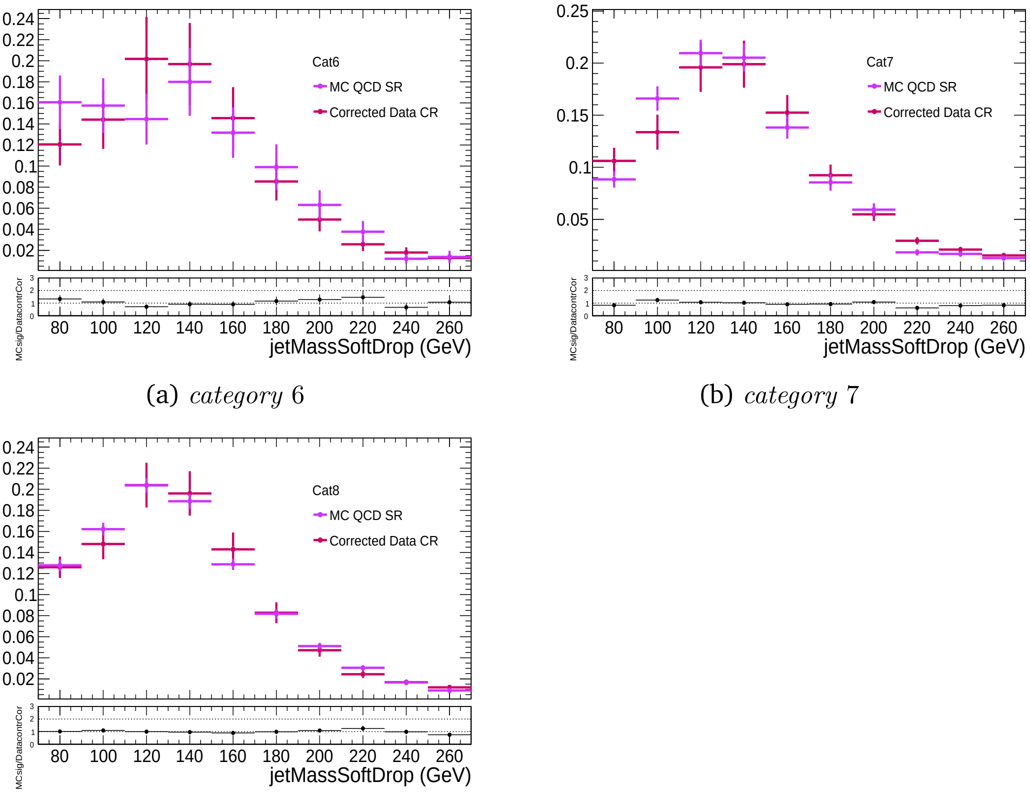
(c) *category* 2

(d) *category* 3

(e) *category* 4

(f) *category* 5

Figure 11.9: The distribution of the corrected QCD shape is taken from the data (red dots) for categories with three Ak8 jets (cat0, cat1) and two Ak8 jets (cat2, cat3, cat4, cat5). The distributions are normalized so that the total integral is equal to one and thus, the small discrepancies are due to differences in the shape. This distribution is compared with the simulated distribution in the signal region. On the lower pad the ratio of the two distributions is illustrated.



(a) *category* 6

(b) *category* 7

(c) *category* 8

Figure 11.10: The distribution of the corrected QCD shape is taken from the data (red dots) for categories with only one Ak8 jet (cat2, cat3, cat4, cat5). This distribution is compared with the simulated distribution in the signal region. The distributions are normalized so that the total integral is equal to one and thus, the small discrepancies are due to differences in the shape. On the lower pad the ratio of the two distributions is illustrated.

11.3.2 Estimation of the expected QCD yield

The method developed in the previous section provides an estimation only on the overall QCD shape. For many processes, calculations to next-to-leading order (NLO) in strong interactions are accessible through modern Monte Carlo generators. However, the NLO calculations introduce large uncertainties. In addition, as the number of final-state hadronic jets increase, the accuracy becomes steadily worse. For the above reasons, simulation is not able to provide an accurate prediction of the expected QCD events. Instead, a commonly used data-driven method called "ABCD" method is used to estimate the expected yield of the background. This method basically uses interpolations from measurements performed on the signal side bands. Assuming that there are two uncorrelated variables $\mathrm{v1}$ and $\mathrm{v2}$, then in the absence of any other information, the minimal assumption is that the background would have a smooth distribution. Therefore, the information from three regions can be used to approximate the background in the fourth. As illustrated in figure 11.11, two selections are required that form part of the definition of the signal region, $\mathrm{region\ A}$, which can be inverted in order to define three further regions, $\mathrm{region\ B}$, C, and D. Then the expected number of events in the signal $\mathrm{region\ A}$ is given by:

$$N_A = \frac{N_C^{bkg} \cdot N_B^{bkg}}{N_D^{bkg}} \tag{11.1}$$

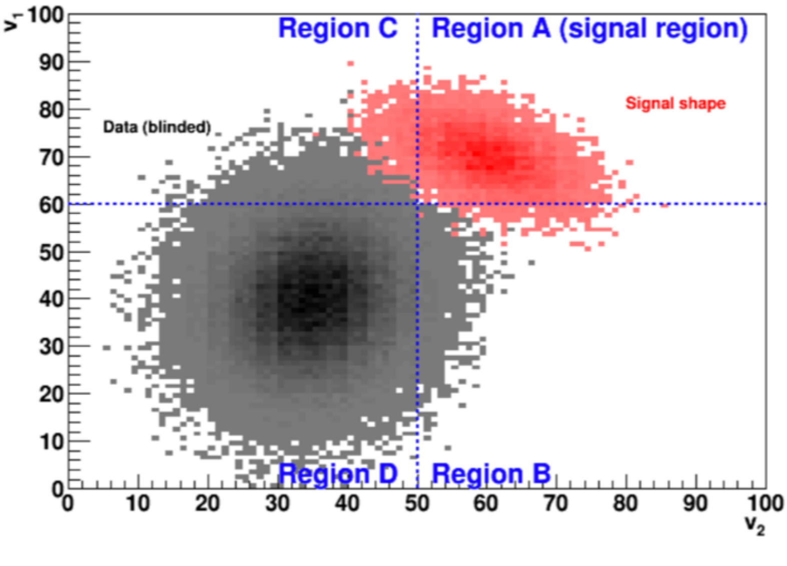


Figure 11.11: Illustration of an example of an ABCD method. Two uncorrelated variables $\mathrm{v1}$ and $\mathrm{v2}$ are used and the scatter plot is divided in four regions. The three regions are used to estimate the expected number of events in the signal region.

For the application of the ABCD method in high energy physics analysis, two main points need to be taken into consideration. First and foremost, in order for the equation 11.1 to be valid, the two individual variables need to be uncorrelated. In addition, the control regions need to be as pure as possible, meaning that contributions from signal or other background processes need to be negligible.

This analysis has a complex phase space of multiple categories with different requirements and therefore, it is not feasible to derive individual predictions for each category. Instead of performing the method in each category's phase space, an extended phase space is used. More specifically, the extended phase space is divided based on the Ak8 multiplicity but with less requirements compared to the analysis categories. Therefore, the expected number of QCD events is derived in events for three, two and only one Ak8 jets.

Since no selection was made on the training samples based on the Ak4 jet multiplicity, two variables that show minimum correlation are the HvsQ discriminant and the multiplicity of Ak4 jets. In addition, the analysis categories require events with a certain number of Ak4 jets with a specific HvsQ threshold ($\mathrm{HvsQ} > 0.8$). Therefore, it is easy to define an extended signal region and three control regions. A sketch of the regions used for the ABCD method is shown in figure 11.12 for three cases of Ak8 multiplicity. In each case the signal region is a subset of the $\mathrm{region\ A}$, which is estimated from the other three regions. Figure 11.13 shows the control and the extended signal regions in simulated QCD events. More specifically, figure 11.13a shows the two discriminating variables HvsQ and the number of b-tagged Ak4 jets ($\mathrm{nAk4BJets}$) in the 2 dimensional plane for categories with three Ak8 jets. The extended signal region consists of events with $\mathrm{HvsQ} > 0.8$ and zero b-tagged Ak4 jets. Reversing those cuts we construct the three other regions. Figure 11.13b and figure 11.13c show the two dimensional distributions of HvsQ and the number of Ak4 jets ($\mathrm{nAk4Jets}$). The side-bands used for categories with two and one Ak8 jets show minimum correlation between the variables. By requiring the $\mathrm{HvsQ} > 0.8$ and the $\mathrm{nAk4Jets} > 1$ ($\mathrm{nAk4Jets} > 4$) we get the extended signal region for two (one) Ak8 jet categories.

QCD prediction in simulation- MC closure test

In order to validate the ABCD method and to ensure a robust prediction of the expected yield, a Monte-Carlo closure test is performed. The extended signal and control regions were constructed from QCD simulated events as shown on figure 11.13 for three (11.13a), two (11.13b) and one (11.13b) Ak8 jets. Then, the regions B, C and D are used to predict the expected number of QCD events in the extended region A from the relation 11.1. The prediction from simulation is illustrated on the third column of table 11.2 and it is compared with the prediction form the ABCD method illustrated on the second column. The uncertainties on the prediction from simulation includes the statistical uncertainties. The number of events in regions B, C and D can be taken as independent, and so the error of the ABCD prediction can be easily calculated using linear error propagation. As shown from table 11.2 the predictions are within uncertainties and therefore, the method shows good closure.

$N_{\mathrm{Ak8\,jets}}$	ABCD prediction	MC prediction
3 Ak8 jets	17634 ± 317	17466 ± 283
2 Ak8 jets	220331 ± 2479	216853 ± 590
1 Ak8 jets	186976 ± 1375	187086 ± 958

Table 11.2: ABCD MC closure test. The errors of the predictions reflect the statistical uncertainties.

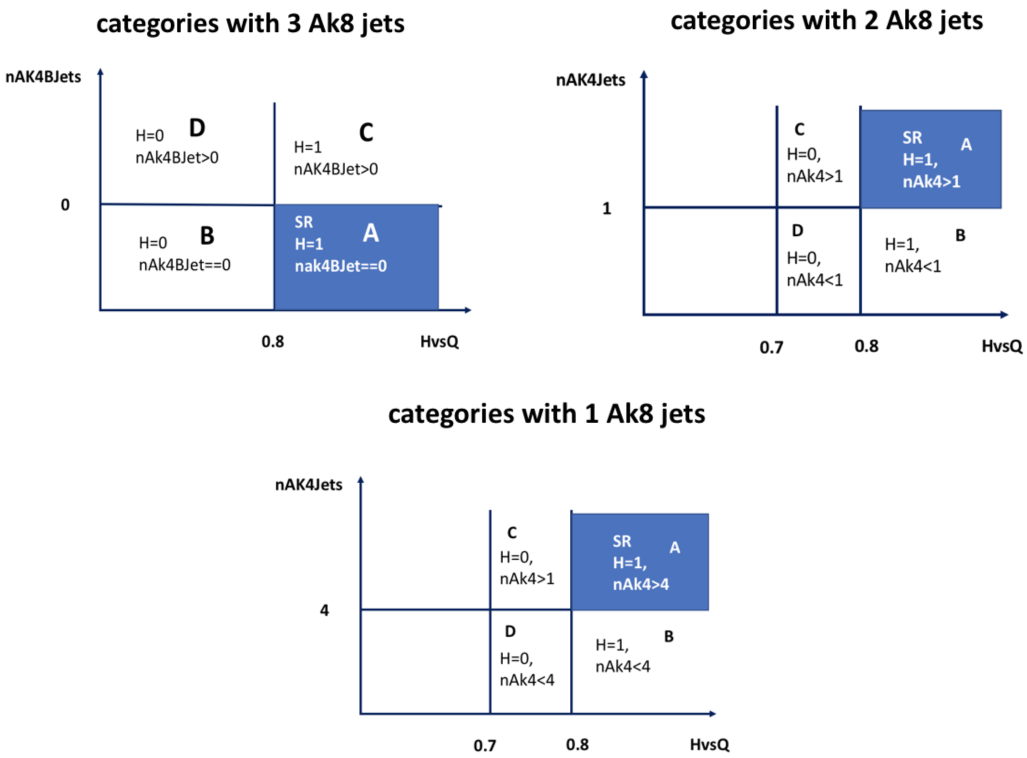
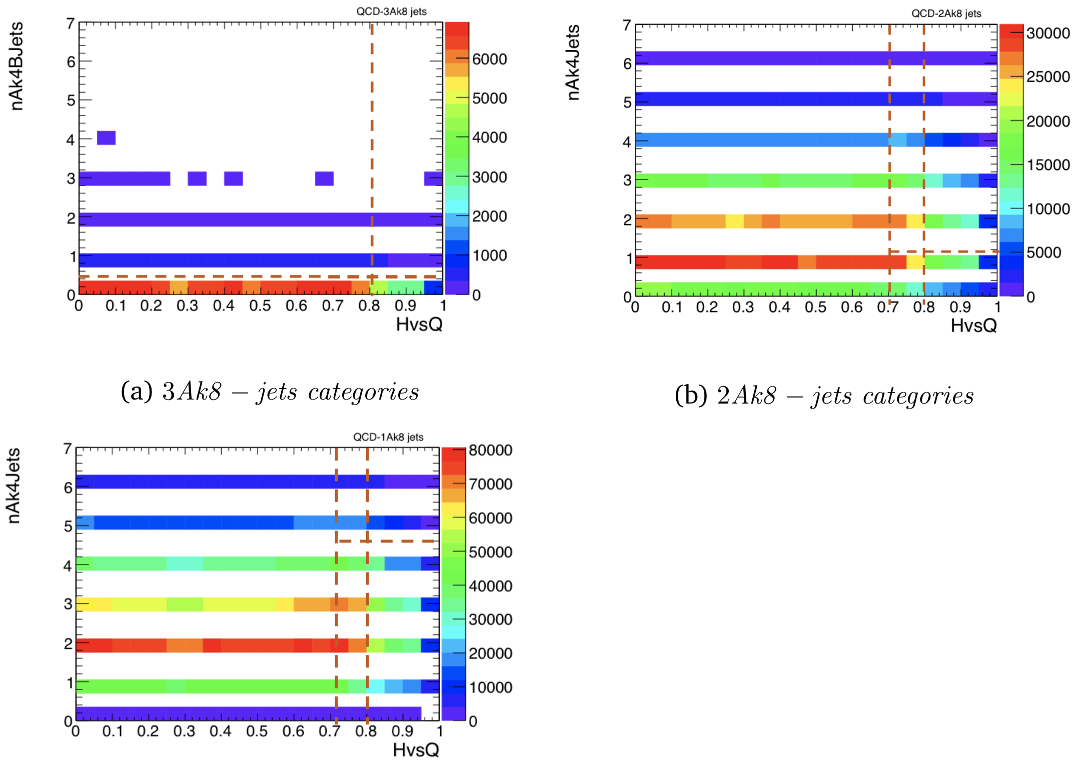


Figure 11.12: A sketch of the regions used for the ABCD method for events with three, two and one Ak8 jet. The regions are divided by the HvsQ threshold and the number of Ak4 jets. In case of three Ak8 jets a requirement of b-tagged Ak4 jets is made. The side-bands in each category reflect to regions that show minimum correlation between the variables.

QCD Prediction in the data

After performing a closure test on the ABCD method in simulation, the method is applied to the data. The control and the extended signal regions were constructed for events that pass the baseline selection. Although the regions in data are enriched in mutijet events, there is a finite probability events from $t\bar{t}$ decays to pass the selection cuts of each category. There is also a minor contribution from the rest of the background processes. Therefore, since those background contributions are non-negligible, they are subtracted from the data using simulated events. The procedure starts by calculating the expected number of events for all the background processes in the control regions for each Ak8 multiplicity scenario. Then, the background contribution is subtracted from the corresponding control region from the data. After the background contributions are subtracted from regions B, C and D, then the expected number of QCD events in the extended region is derived from relation 11.1. Table 11.3 shows the predictions derived from applying the ABCD method on the data for the three extended regions. Since we are not blinded in each of the extended regions, the actual number of data on those regions are compared with the predictions. The errors presented on table 11.3 are the statistical uncertainties. By comparing the two columns, we can conclude that the method performed on the data is robust.



(a) $3Ak8 - jets\ categories$

(b) $2Ak8 - jets\ categories$

(c) $1Ak8 - jet\ categories$

Figure 11.13: ABCD method performed on simulated QCD events. The regions are divided by the HvsQ threshold and the number of Ak4 jets. In case of three Ak8 jets a requirement of b-tagged Ak4 jets is made. The side-bands in each category reflect to regions that show minimum correlation between the variables.

$N_{\text{Ak8 jets}}$	ABCD prediction	Extended Signal region
3 Ak8 jets	13580 ± 151	13420 ± 136
2 Ak8 jets	171144 ± 1224	171993 ± 249
1 Ak8 jets	157820 ± 597	157858 ± 408

Table 11.3: ABCD method performed in the data. The errors of the predictions reflect the statistical uncertainties.

11.3.3 Final estimation of the QCD background

Two independent data-driven methods for estimating the QCD background were developed and tested in simulation. The estimation of the QCD shape is derived for each category. However, the ABCD method described on section 11.3.2 gives an estimation on the expected QCD events in an extended phase space. In order to extrapolate from the extended phase space to the analysis phase space, the fraction of events that are classified in each category is needed. A schematic view of the extended and the analysis phase space is illustrated on figure 11.14. As showed in section 11.2 the variables that are used to classify the events in analysis categories, such as the Ak8 and Ak4 multiplicity and the number of btagged Ak4 jets under the condition that there are at least five resolved jets in the event, show good agreement between data and simulation. Since those variables are well-modeled, we can argue that the simulation gives an accurate description of the fraction of events classified in each analysis category. Therefore, the QCD yield for each category is the product of the prediction from the ABCD method preformed on the data and the fraction that accounts on the number of events that are selected for each category. Finally, the QCD estimation ($\mathrm{QCD_{cat_i}}$) is derived by scaling the normalized QCD shape distribution $\mathrm{D(QCD_{cati})}$ for each category i to the corresponding yield as:

$$\mathrm{QCD_{cat_i}} = \mathrm{NAk8_{ext}} \cdot \mathrm{frac_{cati}} \cdot \mathrm{D(QCD_{cati})} \tag{11.2}$$

where $\mathrm{NAk8_{ext}}$ is the estimated yield of the extended region from ABCD method and $\mathrm{frac_{cati}}$ is the fraction that accounts on the percentage of $\mathrm{NAk8_{ext}}$ events that will be classified in category i.

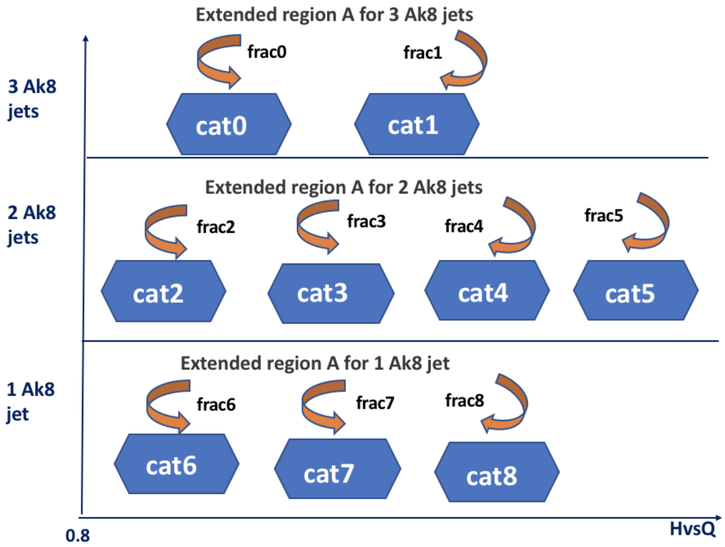


Figure 11.14: Schematic view of the extended and the analysis phase space. To extrapolate from the extended phase space to the analysis phase space the fraction of events classified in each category is needed. The fraction of events is estimated from simulation, and then multiplied to the prediction taken from the data to derive the expected events for each category.

11.4 $\mathrm{t\bar{t}}$ signal extraction

The second background contribution comes from $\mathrm{t\bar{t}}$ events. This background is modelled completely from simulation. Although the $\mathrm{t\bar{t}}$ process is simulated quite accurately, there are some differences observed between different MC generators and the data. In particular, the p_T distribution of the top quark is a variable that shows discrepancies between data and simulation. In the low-p_T spectrum the disagreement is recovered by reweighting the p_T distribution of the generated $\mathrm{t\bar{t}}$ events [122]. However, this is not possible in the high-momentum regime targeted in this analysis. To account on that mismodelling, a specific correction for this analysis phase-space (i.e large hadronic activity with high-momentum jets) is derived. Therefore, the $\mathrm{t\bar{t}}$ signal is extracted via a binned max likelihood fit on the m_{SD} of the top candidate, which is considered the leading jet, from the $\mathrm{t\bar{t}}$ enriched region described in section 10.7.2. Applying the selection requirements summarized in table 10.2, a pure sample of $\mathrm{t\bar{t}}$ events is obtained. Figure 11.15 illustrates the m_{SD} of the leading jet which is considered as the top candidate. For this selection, 98.4% of the reconstructed top candidates are matched with a top parton with minimum $\Delta R < 0.3$.

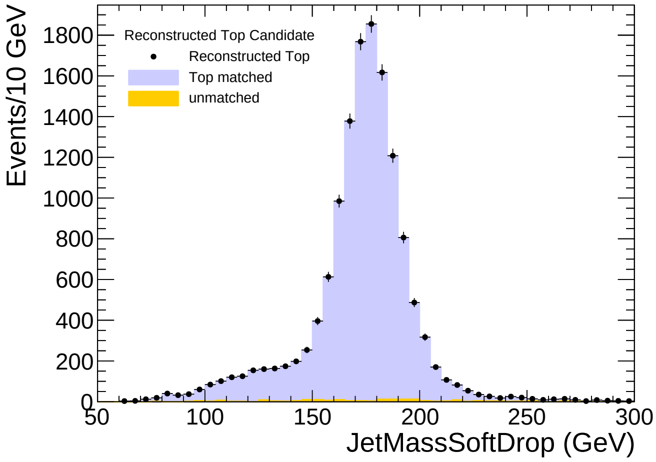


Figure 11.15: M_{SD} of the reconstructed top candidate (leading jet) in the $\mathrm{t\bar{t}}$ enriched region which corresponds to a sample with purity of 98.4%.

For the $\mathrm{t\bar{t}}$ signal extraction, simulated $\mathrm{t\bar{t}}$ events with a cut at parton level in the invariant mass of the $\mathrm{t\bar{t}}$ system ($M_{\mathrm{t\bar{t}}} > 700$ GeV) were used. This choice is motivated from the fact that those samples profit from better statistics in the particular phase space of the $\mathrm{t\bar{t}}$ enriched region. This is illustrated in figure 11.16, where the invariant mass of the reconstructed $\mathrm{t\bar{t}}$ system is compared between the nominal-all phase space and the $M_{\mathrm{t\bar{t}}}$-cut simulated sample.

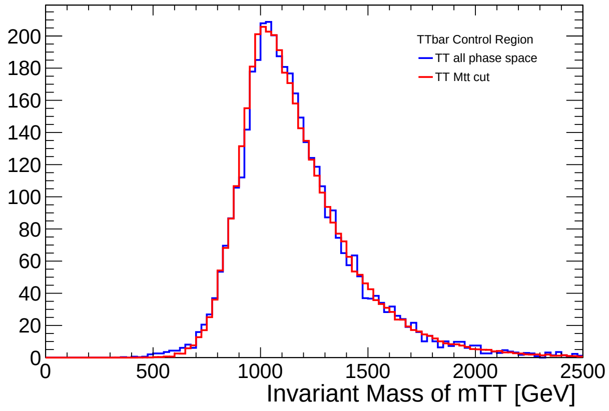


Figure 11.16: Reconstructed invariant mass of the $\mathrm{t\bar{t}}$ system for the nominal and the $M_{\mathrm{t\bar{t}}}$ cut sample for the phase space of the $\mathrm{t\bar{t}}$ enriched region.

The fit was performed on the m_{SD} of the top candidate twice. Both fits used the same tool, the Toolkit for Data Modeling with ROOT (RooFit). Template histograms for signal and background are fitted to the data as shown in figure 11.17. The second fit was done using the Higgs PAG combine tool [123] which provides a command line interface to many different statistical techniques available inside RooFit/RooStats and is used widely inside CMS. Both fits give the same value within the uncertainties arising from the fit. The fitted $\mathrm{t\bar{t}}$ yield ($N_{\mathrm{t\bar{t}}}$) can be converted to a fiducial cross section (σ) by:

$$\sigma = N_{t\bar{t}}/L \tag{11.3}$$

The measured cross section is found to be a factor $r = 0.746$ lower than the theoretical prediction (Powheg+Pythia8), which is compatible with the measurement of the top-anti-top differential production cross section of high transverse momentum top quarks in the all-hadronic final state performed with the same data-set [124]. Therefore, the expected yield of the $\mathrm{t\bar{t}}$ events is corrected by this factor. The distribution of the m_{SD} and p_T of the top candidate is compared for $\mathrm{t\bar{t}}$ simulated events and the data in figure 11.18 and figure 11.19 respectively. Background contributions from multijet production and other processes are also illustrated. It is clear that after the correction there is a good agreement between data and simulation.

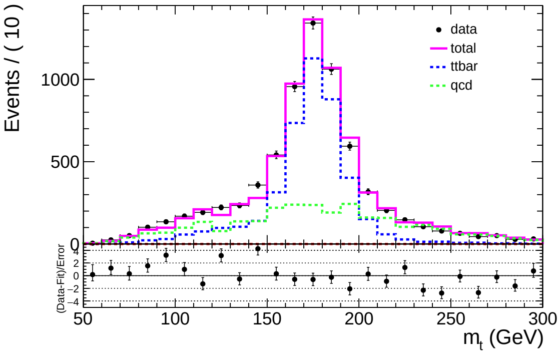


Figure 11.17: The m_{SD} distributions for data and simulation for events that pass the selection requirements of the $\mathrm{t\bar{t}}$ enriched region. The lower panel shows the ratio of the observed data to the post expectation from simulation.

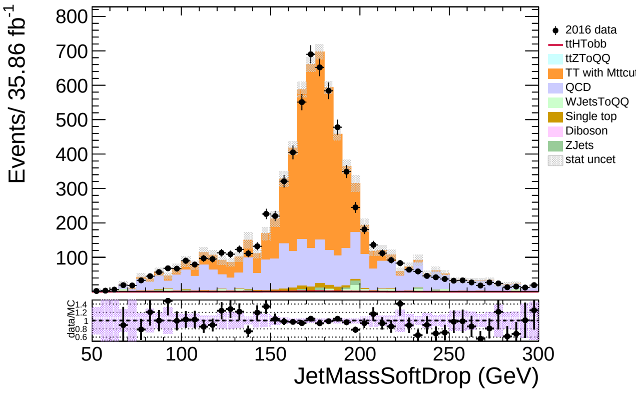


Figure 11.18: M_{SD} of the reconstructed leading jet that is "tagged" as the top candidate. The simulated events are compared to the data after the correction on the cross section applied. On the lower pad, the ratio of data and simulation is shown. The simulated backgrounds are first scaled to the luminosity of the data, and then the simulated QCD multijet background is rescaled to match the yield in data.

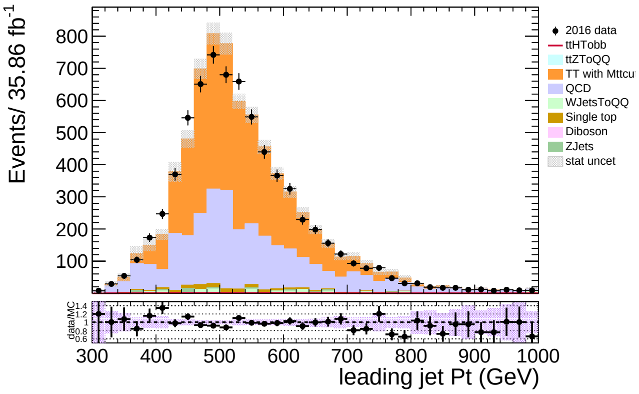


Figure 11.19: p_T of the reconstructed leading jet that is "tagged" as the top candidate. The simulated events are compared to the data after the correction on the cross section applied. On the lower pad, the ratio of data and simulation is shown. The simulated backgrounds are first scaled to the luminosity of the data, and then the simulated QCD multijet background is rescaled to match the yield in data.

11.5 Systematic Uncertainties

This section summarizes the systematic uncertainties that are relevant to the analysis of the thesis. The signal extraction procedure is based on the likelihood $L(\mu, \theta|x)$ of μ and the nuisance parameters θ given measured data x as described in detail in section 5.1. A nuisance can only affect the rate of a process. However, rate-changing nuisances can alter the composition of processes in a specific category and therefore lead to effective changes in the shape of the distribution. Rate-changing nuisances introduced in the following sections are modeled through log-normal distributions (ℓnN) whose widths are based on a-priori knowledge from previous measurements or on theoretical reasoning. The shape changing effect is modeled from alternative shape distributions that result from variations of the respective uncertainty.

11.5.1 Experimental Uncertainties

The uncertainties that rise from experimental sources are due to limited measurement accuracies. Such uncertainties could relate to the measurement of the integrated luminosity, the calibration of the jet energy scale and resolution, the corrections applied to event weights, as well as the limited amount of simulated events that affect the precision of predicted distributions. Those uncertainties are fully correlated among all processes.

Integrated Luminosity

The efficient determination of luminosity is essential for ensuring high data quality over a large amount of events under measurement conditions. The measure of the online, real-time collision rate is used for monitoring beam and detector performance while the integrated luminosity over time, describing the absolute amount of recorded events for high-level physics analyses, is determined offline. This is done through the pixel detector which yields the best precision and a good pileup linearity up to 150 interactions per bunch crossing [125]. The overall uncertainty in the measured integrated luminosity recorded by the CMS experiment is estimated to be 2.5%. This source of uncertainty is treated as a single, correlated between the categories, nuisance parameter, since it affects equally the expected rates of all the physical processes.

Pile-up

As described in section 9.4.1 the uncertainty in the distribution in the number of pileup interactions is evaluated by changing the minimum-bias cross section by 4.6% relative to its nominal value. Since the value of a weight is a function of the number of pileup events, the associated nuisance parameter has a shape-changing effect, which is demonstrated in the right part of figure 9.2. In addition, the rate of predicted contributions can change due to a slightly different pileup profile in the phase space of each category. This nuisance is treated as fully correlated among all processes.

Trigger Efficiency

As described in section 9.4.2 in order to recover any discrepancies of the performance of the employed trigger path, a scale factor was derived by fitting the efficiency ratio of data and simulation with a linear function. The correction factor is 0.97 ± 0.004. In order to study the effect of the trigger, up and down systematic variations are derived by assigning event weights of $\mathrm{sf} \pm \mathrm{err}_{\mathrm{sf}}$. Those variations have a small impact both on the yield and shape and thus they are already covered from the normalization and shape uncertainties assigned to each process.

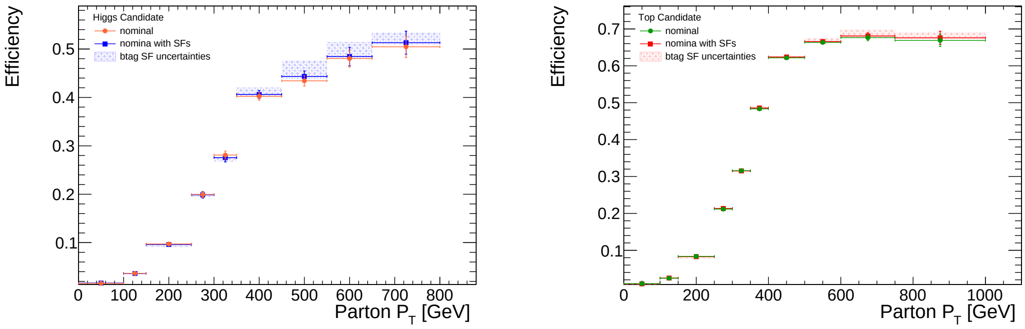
Shape calibration of the b-tagging discriminant

Scale factors are used to correct any mismodeling of the shape of the b-tagging discriminant (CSVv2), as described in detail in section 9.4.3. Those corrections are applied to each simulated jet and they depend on the transverse momentum, pseudorapidity, hadron flavor and the actual value of the CSV discriminant. The correction, applied as an event weight, is obtained from equation 9.2 by multiplying the scale factors of all selected jets of an event. The systematic uncertainties arise from those scale factors have three main sources: JES, purity of heavy- or light-flavour jets in the control sample used to obtain the scale factors, and the statistical uncertainty of the event sample used in their extraction. In addition, a large uncertainty is assigned to charm-flavour jets due to the lack of a reliable data-based calibration. The b-tag shape calibration is not intended to change the expected yield of events with the same jet multiplicity. However, after applying b-tag selection criteria, the average correction weight can be different from unity. Therefore, employed nuisances have both a shape- and a rate-changing effect. Each component of these systematic b-tagging uncertainties is considered uncorrelated from the others, resulting in nine separate nuisance parameters in the final fit.

Effect of scale factors and btag uncertainties in analysis

The corrections concerning the trigger efficiency, pile up and b-tag reshaping were applied to account on any potential mis-modelling between data and simulation. Each correction, has dedicated sources of uncertainties that were treated separately resulting in a rate-plus-shape effect on the final templates. Due to this analysis particular phase space, considering mixed-topologies of both Ak8 and Ak4 jets, the largest impact is expected from the b-tag reshaping corrections. As described in section 10.6, the CSVv2 output discriminant is used as an input variable to the BDT training and thus, the choice of the reconstructed Higgs candidate is subsequently affected. In addition, the resolved b jet multiplicity is used to classify events in several categories and thus, the application of those weights can result in event mitigation through categories. Therefore, it is important to ensure that there is not any inconsistencies concerning the construction of the analysis strategy. For the above reasons the efficiency of the reconstructed Higgs (figure 11.20a) and top (figure 11.20b) candidate is calculated as a function of parton's transverse momentum before and after applying the scale factors. The shaded region shows all the CSVv2 reshaping systematic variations added in quadrature. The efficiency for both the reconstructed Higgs and top candidate does not show any peculiar trend after applying the scale factors. The shaded regions corresponds to the nine independent sources of CVSv2 uncertainty.

Moreover, the effect of the scale factors and the corresponding b-tag uncertainties is shown for the reconstructed Higgs candidate on figure 11.21. The simulated $t\bar{t}H$ events are scaled to the luminosity of the data so differences between before and after the scale factor application reflect on differences in expected yields and shapes. As described previously, the btag scale factors affect the classification of events in analysis categories. This is illustrated in figures 11.22 and 11.23 where the reconstructed Higgs candidate per analysis category is shown before and after the application of scale factors. In general, the effect of the expected yield of the analysis categories varies between $0.1 - 10\%$.



(a) Higgs candidate

(b) Top candidate

Figure 11.20: Reconstructed Higgs (left) and top (right) efficiency as a function of the parton p_T before and after applying the scale factors. The shaded regions corresponds to the nine independent sources of CVSv2 uncertainty.

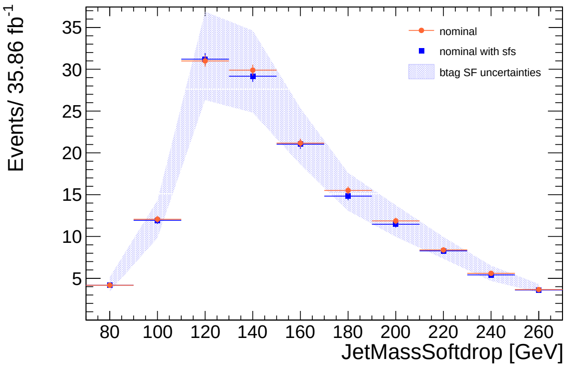


Figure 11.21: Reconstructed Higgs candidate before (orange) and after (blue) applying the scale factors. The shaded regions corresponds to the nine independent sources of CVSv2 uncertainty. The simulated $\mathrm{t\bar{t}H}$ events are scaled to the luminosity of the data so differences between before and after the scale factor application reflect on differences in expected yields and shape.

Jet energy scale and resolution

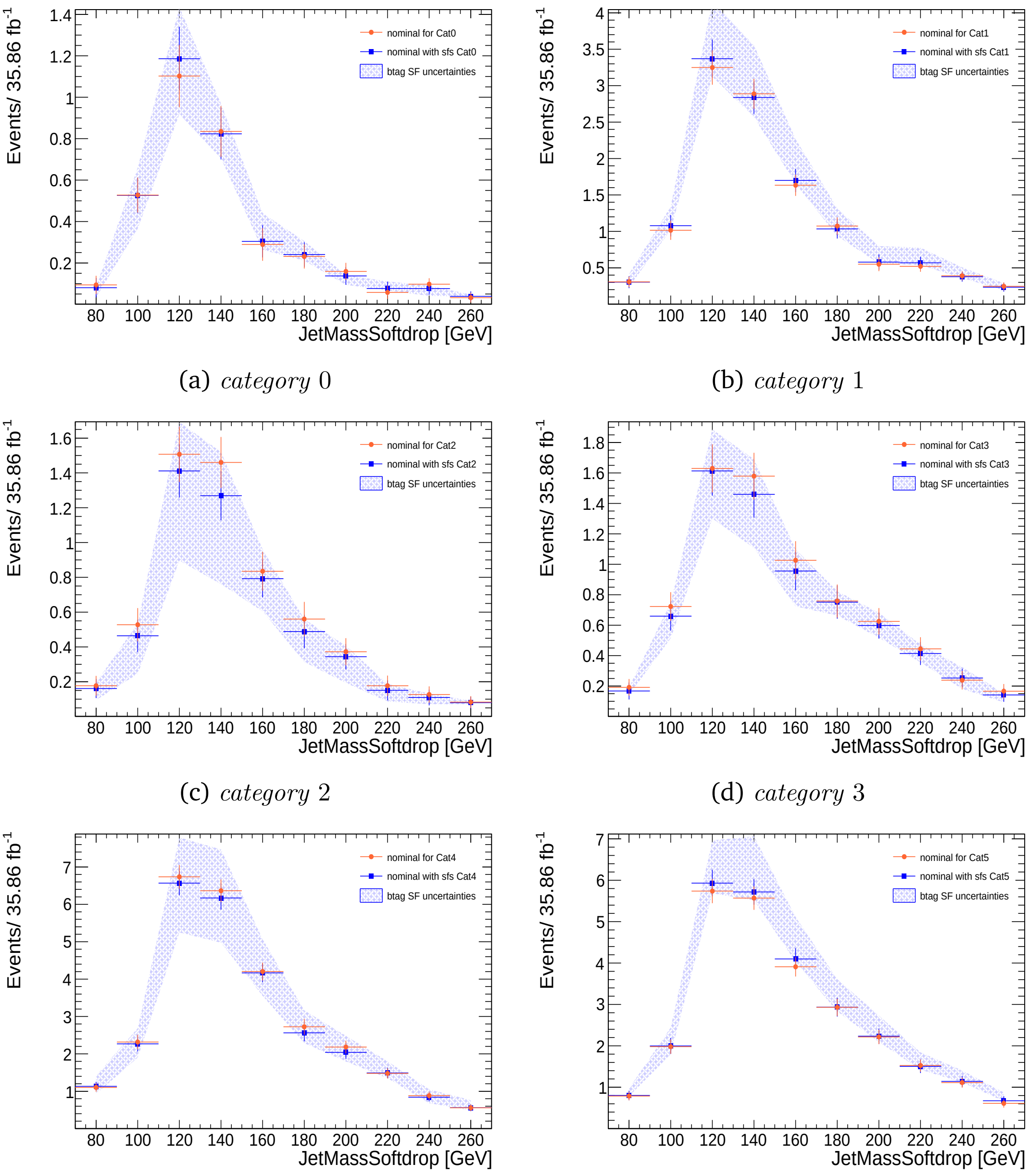
As mentioned in section 3.4.4, corrections are applied to the jets to account on differences observed in data. Analysis that have final states with high jet multiplicity can be sensitive on the JES corrections since those can result in migration of events between analysis categories. The impact of the uncertainty on the JES correction is evaluated for each jet of the simulated events, by shifting the nominal correction by 1σ [126]. For each variation a new jet collection is created and the event interpretation is repeated. The effect of the JES uncertainty results not only in variations of the p_T scale itself, but may also lead to different Higgs and top candidates.
The impact on the measurement due to the jet energy resolution (JER) is determined by smearing the jets according to the JER uncertainty. The JER uncertainty is evaluated by increasing and decreasing the difference between the reconstructed-level and particle level jet energy, according to the standard CMS prescription. The effect is estimated by recalculating all the kinematic quantities.

11.5.2 Theoretical and background modelling Uncertainties

In addition to uncertainties that rise from experimental sources, one should also consider uncertainties related to theoretical predictions and modeling of simulated events. In particular, this analysis considers uncertainties related to variations of parton distribution functions, renormalization and factorization scales, and modeling uncertainties that can affect the shape and the rate of specific backgrounds. Depending on the type of the uncertainty, the nuisance parameters can affect either a specific process or a group of processes.

Parton Distribution function

A probabilistic description of the parton momentum fractions is provided by the Parton distribution functions (PDFs). PDFs mostly depend on the momentum transfer factor Q^2 of the interaction and on the flavor of the involved parton, as described in section 4.1.1. PDFs are used for cross section calculations, as well as for the simu-



(a) *category* 0

(b) *category* 1

(c) *category* 2

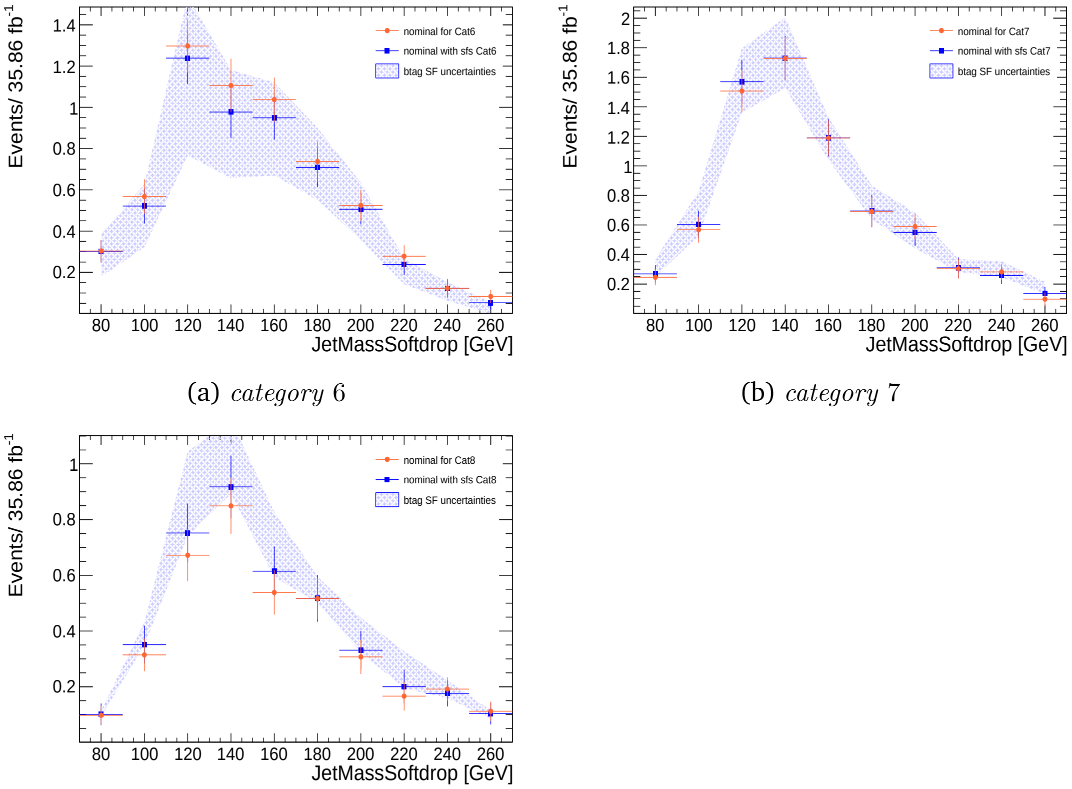
(d) *category* 3

(e) *category* 4

(f) *category* 5

Figure 11.22: Reconstructed Higgs candidate for categories with three and two Ak8 jets, before (orange) and after (blue) the scale factor application. The shaded regions corresponds to the nine independent sources of CVSv2 uncertainty. The simulated $t\bar{t}H$ events are scaled to the luminosity of the data so differences between before and after the scale factor application reflect on differences in expected yields and shape.

lation of proton-proton collisions in event generators. To estimate systematic uncertainties related to the choice of a specific PDF, different variations are used. Those variations result in a rate-plus shape effect and are treated as separate nuisance parameters. Changes of simulated distribution shapes are modeled through a reweighting of events that considers variations of 100 replicas of the nominal NNPDF3.0 PDF set [76] as recommended by the PDF4LHC group for the second run of the LHC [102]. This results in two variations that reflect the up $(+1\sigma)$ and down (-1σ) variations of the



(a) *category* 6

(b) *category* 7

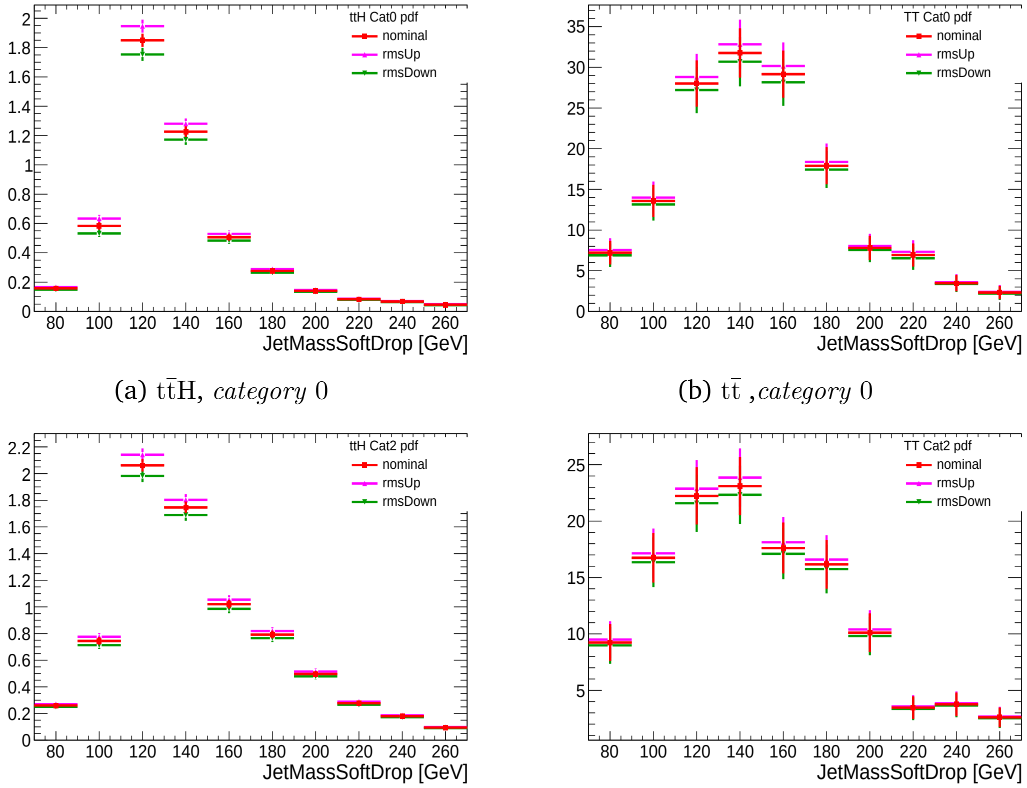
(c) *category* 8

Figure 11.23: Reconstructed Higgs candidate for categories with only one Ak8 jet, before (orange) and after (blue) the scale factor application. The shaded regions corresponds to the nine independent sources of CVSv2 uncertainty. The simulated $\mathrm{t\bar{t}H}$ events are scaled to the luminosity of the data so differences between before and after the scale factor application reflect on differences in expected yields and shape.

nominal PDF uncertainty. Contributions from signal and subdominant $\mathrm{t\bar{t}}$ background were considered. On the contrary, PDF uncertainties are not applied on QCD multijet background, since this contribution is estimated completely from the data. Those variations result to changes of the rate of each process. This effect is showed in figure 11.24 for signal, $\mathrm{t\bar{t}H}$, and subdominant $\mathrm{t\bar{t}}$ background process for *category* 0 and *category* 2 .

Renormalization (μ_R) and factorization (μ_F) scales-Strong coupling constant (α_S) uncertainty.

Renormalization and factorization scales define the value of the cutoff scale for absorbing infrared and ultraviolet divergences in the calculation of scattering amplitudes. Therefore, the choice of the renormalization and factorization scales may have an impact on the kinematical distributions of the final–state objects. To estimate the systematic uncertainties related to those scales, the μ_R and μ_F are varied by a factor of $1/2$ and 2. The unphysical anticorrelated variations are discarded, yielding a total of 7 combinations of the renormalization and factorization scales. Similarly to the PDF uncertainties, up ($+1\sigma$) and down (-1σ) variations are derived and included as event weights in the simulated samples. Those variations result to a slight rate plus shape effect.



(a) $t\bar{t}H$, *category* 0

(b) $t\bar{t}$,*category* 0

(c) $t\bar{t}H$ *category* 2

(d) $t\bar{t}$, *category* 2

Figure 11.24: Up (pink) and down (green) PDF uncertainty variations of the signal $t\bar{t}H$ and $t\bar{t}$ background processes for *category* 0 and *category* 2. Those shapes are compared to the nominal (red) distribution of each process.

The uncertainty associated with the (α_S) is estimated by applying event weights corresponding to higher and lower values of (α_S) for the matrix element using the variations of the NNPDF set.

Uncertainties on the QCD prediction

One of the main challenges to address is the accurate estimation of the overwhelming background coming from multijet processes. Therefore, two data-driven methods were developed that estimate the expected yield and shape of the QCD multijet background (section 11.3). This way, many uncertainties related to MC simulation are avoided. However, there are some remaining uncertainties to consider. Those uncertainties are due to the limited statistics of the data and MC simulation of each control region. In addition, systematic uncertainties are considered that reflect on the data-driven methods used for modelling the multijet background. Therefore, a throughout treatment of the systematic uncertainties is performed, that account on the QCD normalization and the QCD shape of each category.

QCD shape uncertainty

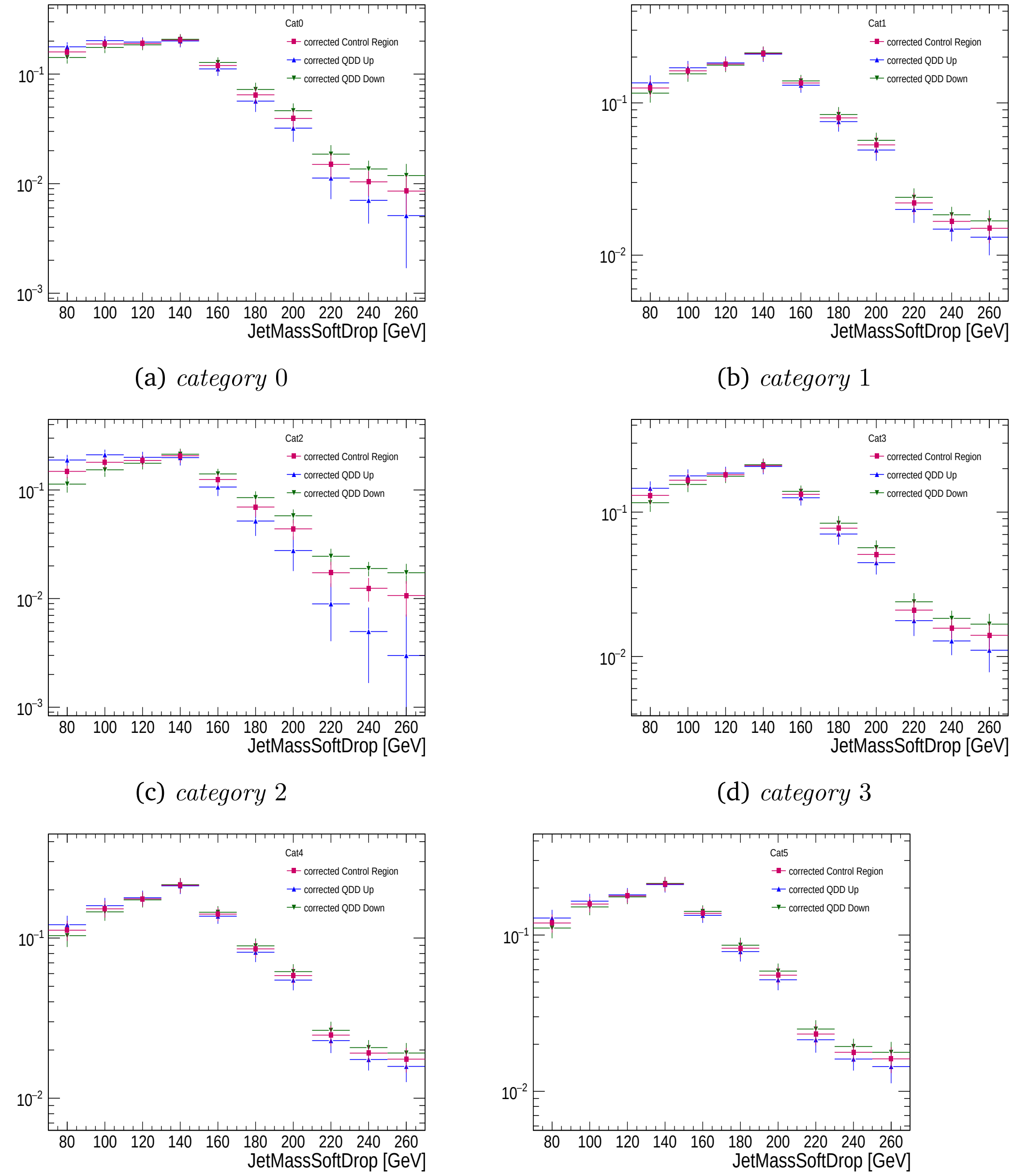
To assign a systematic uncertainty on the overall QCD shape of each category, up and down variations are used. More specifically, two alternative QCD shapes are derived by shifting up and down the nominal QCD shape of each category, using the transfer factor derived in section 11.3.1. This transfer factor results from a linear fit of the ratio of the simulated QCD shape in the signal and in the control region. The linear fit results into two parameters where their correlation is given by the covariance matrix D. In order to de-correlate them, the covariance matrix is diagonilized using an orthogonal transformation as:

$$\mathrm{A} = \mathrm{PDP}^{-1} \tag{11.4}$$

where, P is the orthogonal matrix that has the eigenvectors of D as columns. The eigenvalues of the diagonal covariance matrix A reflect to the variance of the fit parameters. The up and down variations are obtained by shifting the nominal distribution by 1σ up and down respectively. The varied up and down simulated shapes are shown on figure 11.25 for each category. The nominal QCD shape is also shown for each category for comparison.

QCD multijet normalization

The uncertainties of the multijet background normalisation have the largest impact on the analysis sensitivity, and therefore a data-driven method is developed to constrain this uncertainty. The ABCD method gives a prediction of the expected yield of the multijet events in an extended phase space based on Ak8 multiplicity. Therefore, three nuisance parameters are used in the final fit to account on the uncertainty of the ABCD prediction in the data, as showed in table 11.3. Those three independent nuisances are treated as correlated between categories with the same Ak8 multiplicity. However, the final QCD prediction comes from equation 11.2 that considers the fraction of events classified in each category. Therefore, a 10% systematic uncertainty is assigned for the nine analysis categories. This uncertainty is treated as nine uncorrelated nuisance parameters. By predicting the expected contribution of the multijet background from the data, the normalization in the fit is constrained, which results in a 30% lower expected limit.



(a) *category* 0

(b) *category* 1

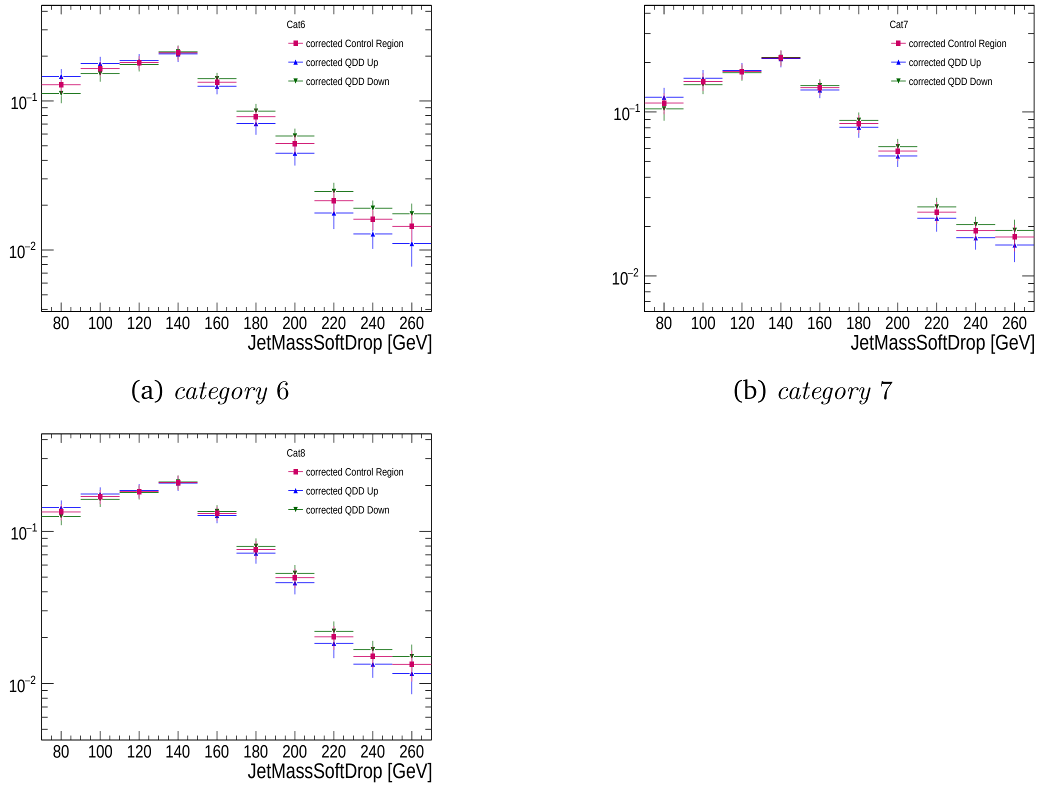
(c) *category* 2

(d) *category* 3

(e) *category* 4

(f) *category* 5

Figure 11.25: QCD corrected shape of the control region in simulation and up and down variations for categories all the analysis categories.



(a) *category* 6

(b) *category* 7

(c) *category* 8

Figure 11.26: QCD corrected shape of the control region in simulation and up and down variations for categories all the analysis categories.

Uncertainties on the $\mathrm{t\bar{t}}$ prediction

In this analysis the subdominant background comes from $\mathrm{t\bar{t}}$ events and thus, the accurate modeling of simulated variable distributions, as well as the description of the systematic uncertainties of those processes is crucial. The $\mathrm{t\bar{t}}$ normalization uncertainty is constrained by a control region enriched in $\mathrm{t\bar{t}}$ events, whereas, the $\mathrm{t\bar{t}}$ shape uncertainty is estimated from alternative $\mathrm{t\bar{t}}$ simulated samples with varied generator parameters. Those parameters are treated as independent nuisances and are considered correlated between the analysis categories. The $\mathrm{t\bar{t}}$ shape uncertainties are summarized below:

> **Final state radiation (FSR):** This uncertainty is estimated from alternative MC samples with reduced and increased value for the strong coupling constant used by Pythia8 to generate final state radiation (α_{FSR} by factors $\sqrt{2}$ and $1/\sqrt{2}$).
>
> **Initial state radiation (ISR):** Similarly to FSR uncertenty, this uncertenty is estimated by two alternative samples with varied nominal scales by factors of 2 and $1/\sqrt{2}$.
>
> **Matrix element – parton shower matching (ME-PS):** In the POWHEG matrix element to parton shower (ME-PS) matching scheme, the re-summation damping factor hdamp is used to regulate high-p_T radiation. Uncertainties in hdamp are parameterized by considering alternative simulated samples with hdamp varied by $\mathrm{hdamp} = \mathrm{m_t}$ and $\mathrm{hdamp} = 2.24\,m_t$.
>
> **Underlying event tune:** Uncertainties related to parton showering and the modeling of global hadron production, are referred as the "underlying event" (UE). In general, the corresponding tuning parameters are summarized by the CUETP8M2T4 tune [127]. This uncertainty is estimated from alternative Monte Carlo samples with the tune CUETP8M2T4 parameters varied by $\pm 1\sigma$.

Separate $\mathrm{t\bar{t}}$ simulated samples were generated for ISR, FSR, ME-PS, and UE variations, and applied to identical event selection and classification procedures. Since the normalization of the process will be determined via the simultaneous fit of the signal and control regions, we keep only the shape-changing effect of those uncertainties, while the rate-changing effect is factored out by rescaling the $\mathrm{t\bar{t}}$ yields in the up and down variations to preserve the overall yield of each category.

To constrain the $\mathrm{t\bar{t}}$ normalization, the M_{SD} of the leading jet of a region enriched in $\mathrm{t\bar{t}}$ events is fitted simultaneously with the M_{SD} of the Higgs candidate of all the analysis categories. The $\mathrm{t\bar{t}}$ enriched region is constructed from events that pass the requirements summarized in table 11.4. This region consists of events with two Ak8 jets with at least one medium b-tagged subjet. In order to enhance the $\mathrm{t\bar{t}}$ event selection, lower thresholds are given to TvsQ and HvsT scores. In addition, since both tops are reconstructed as high-momentum Ak8 jets, no b-tagged Ak4 jets are required in the event. This selection leads to a sample with purity 99%.

In order to model the contribution of the QCD multijet background in the $\mathrm{t\bar{t}}$ control region, the same data-driven techniques developed for the analysis regions are used. More specifically, the shape of QCD is derived from the data using an orthogonal control region. This region has the same requirements as $\mathrm{t\bar{t}}$ control region but with reversed b-tagging requirements on the subjets of the leading and subleading jet. To preserve

Observable	Requirement
N_{Ak8jets}	$= 2$
N_{leptons}	$= 0$
$Nb_{\text{leading Jet}}$	> 0
$Nb_{\text{subleading Jet}}$	> 0
$p_T^{\text{leading}-\text{jet}}$	> 300 GeV
m_{SD}^{jets}	> 70 GeV
TvsQ	> 0.7
HvsT	< 0.1
$N_{\text{Ak4b4jets}}$	$= 0$

Table 11.4: selection requirements for new $t\bar{t}$ enriched validation region

similar kinematic characteristics with the signal region, lower and upper thresholds are set to TvsQ. Those requirements are summarized in table 11.5. Figure 11.27 shows the data-simulation comparison of the QCD control region constructed to model this contribution in $t\bar{t}$ region. It is clear that this region is enriched in multijet events where the signal ($t\bar{t}$) is only 3%. A Monte Carlo closure test is performed to check the agreement between the two regions. This is shown in figure 11.28 where a linear fit is performed on the ratio of the QCD shape in the signal and in the control region. To set a systematic uncertainty on the overall QCD shape, up and down variations are considered as described in section 11.5.2.

Observable	Requirement
N_{Ak8jets}	$= 2$
N_{leptons}	$= 0$
$Nb_{\text{leading Jet}}$	$= 0$
$Nb_{\text{subleading Jet}}$	$= 0$
$p_T^{\text{leading}-\text{jet}}$	> 300 GeV
m_{SD}^{jets}	> 70 GeV
TvsQ	$0.5 <$ TvsQ < 0.7
HvsT	< 0.1
$N_{\text{Ak4b4jets}}$	$= 0$

Table 11.5: selection requirements of the QCD CR to model QCD contribution for $t\bar{t}$ enriched validation region.

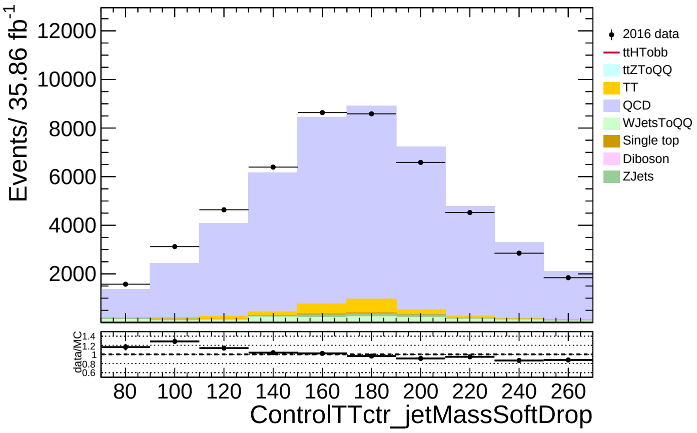


Figure 11.27: M_{SD} distribution of a multijet enriched region constructed to model the shape of QCD events in the ttbar control region. The distributions of data and simulation are shown. Each background contribution is illustrated with a different color and is compared with the data (black dots). On the lower pad, the ratio of data and simulation is shown. The simulated backgrounds are first scaled to the luminosity of the data, and then the simulated QCD multijet background is rescaled to match the yield in data.

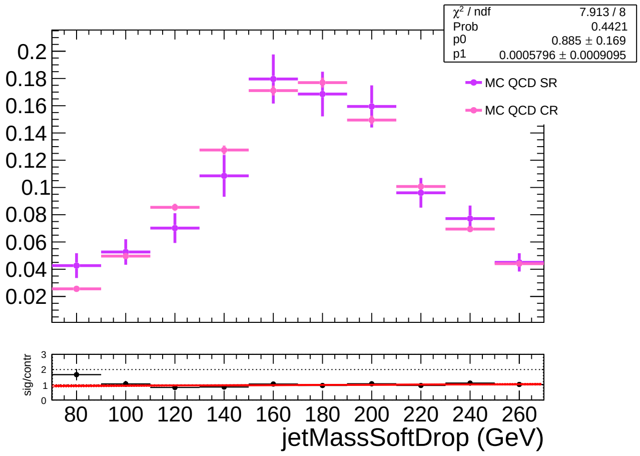


Figure 11.28: Monte Carlo closure test performed for the QCD region constructed to model the shape of QCD events in the ttbar control region. QCD simulated events were used to construct the distributions of m_{SD} of the leading in the signal region and in the control region. The distributions are normalized so that the total integral is equal to one and thus, the small discrepancies are due to differences in the shape. The ratio of the two distributions is fitted with a linear function and a transfer factor in obtained.

The expected QCD events is derived by performing the ABCD method described in section 11.3.2. The BDT responses and the Ak4 multiplicity are two variables that show minimum correlation, and therefore they can be used to perform the ABCD method. Three orthogonal control regions are constructed by inverting the TvsQ and $\mathrm{nAk4BJet}$ requirements. A schematic view of the regions used is shown in figure 11.29. The signal region A consists of events after the selection summarized in table 11.5. The expected number of events is given by equation 11.1. The ABCD method is tested in simulated $t\bar{t}$ events as shown in figure 11.30. The prediction of the ABCD method as well as, the expected simulated $t\bar{t}$ events are illustrated in table 11.6. It is clear that the method shows good closure.

The ABCD method is performed in data and the prediction is illustrated in the first column of table 11.7. The signal region is dominated by $t\bar{t}$ events and therefore the expected number of multijet events is derived by subtracting the $t\bar{t}$ contribution from table 11.7. Therefore the final prediction of events from multijet processes in the region enriched in $t\bar{t}$ events is 2683 ± 91.

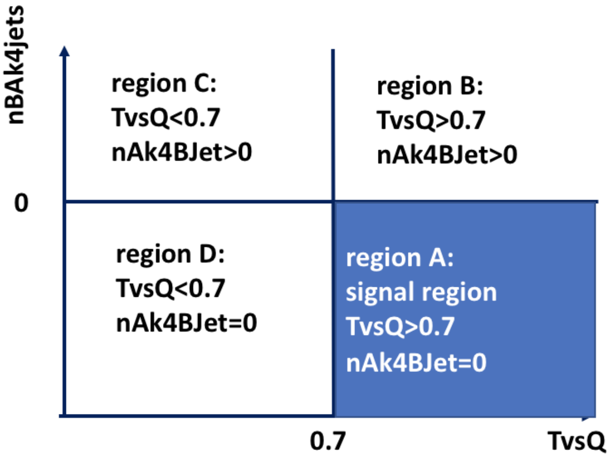


Figure 11.29: A sketch of the regions used for the ABCD method in $t\bar{t}$ events. The regions are divided by the TvsQ threshold and the number of Ak4 b-tagged jets.

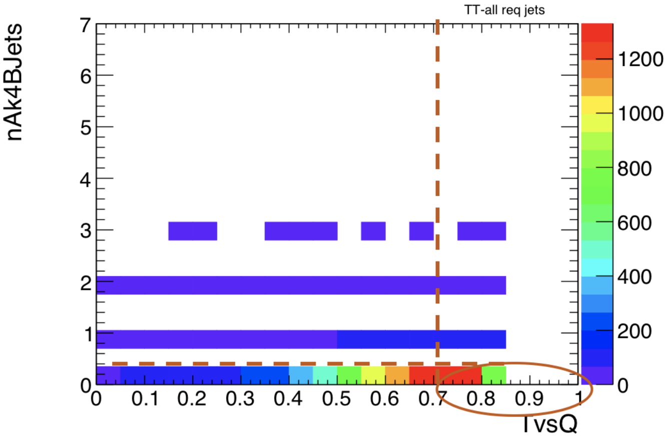


Figure 11.30: MC closure test

	ABCD prediction	MC prediction
$t\bar{t}$	3393 ± 32	3317 ± 43

Table 11.6: ABCD MC closure test performed in simulated $t\bar{t}$ events. The errors of the predictions reflect the statistical uncertainties.

	ABCD prediction	Signal Region
Data	6076 ± 85	6119 ± 78

Table 11.7: ABCD method in data. The errors of the predictions reflect the statistical uncertainties.

Subdominant Background modelling

The rest of the background contributions are modelled completely from simulation. Due to similarities between $t\bar{t}Z$ and $t\bar{t}H$ signal, the $t\bar{t}Z$ background is treated separately from the rest of the background processes. To take into account any potential mismodelling on the $t\bar{t}Z$ background, a 20% systematic uncertainty is assigned. The rest of the subdominant background contributions are added together in a single template for each category and a 50% uncertainty is assigned.

Uncertainty due to limited amount of simulated events

Template distributions from Monte-Carlo generators are subject to statistical fluctuations due to finite number of events in samples. The influence of these fluctuations can be expected to be significant in regions of low amounts of Monte-Carlo events. For incorporating such uncertainties into likelihood function the Barlow and Beeston method [128] was used. In this method, instead of requiring separate parameters per process, a single nuisance parameter is assigned to scale the sum of the process yields to each bin. The advantage of this method is that it minimizes the number of parameters required in the maximum-likelihood fit.

Chapter 12

Results

12.1 Results

The metric used to assess the analysis performance is the expected exclusion limit at 95% CL on the signal strength modifier, as defined in a fully frequentist approach. The strategy to extract exclusion limits in the presented analysis relies on a binned maximum likelihood fit to data of the m_{SD} distribution in the signal and control regions simultaneously in all the categories and channels. Therefore, the $\mathrm{t\bar{t}H}(b\bar{b})$ signal is extracted via a binned maximum likelihood fit on the mass of the Higgs candidate, with the signal and the control regions in all the analysis categories included to the fit simultaneously. The benefit of such approach is that not only the expected event yields for the signals and backgrounds, but also the distribution of those events are taken into account. Furthermore, this method is convenient when the background cannot be predicted reliably a-priori, resulting in a better discrimination between a signal-like and background-like excess. Another advantage is that it provides an in-situ normalization of the background. The observed distributions of the events in data and the expectations from the signal and all backgrounds are provided as histograms all with the same binning. A total of 10 bins are used in the fit for each region, with a bin width of 20 GeV corresponding roughly to the m_{SD} resolution. The signal extraction is performed with a simultaneous fit of the $\mathrm{t\bar{t}}$ background enriched control region and the signal region divided into nine orthogonal categories. In the signal region, the fitted distribution corresponds to the m_{SD} of the reconstructed Higgs candidate whereas in the $\mathrm{t\bar{t}}$ region the m_{SD} of the leading jet is used.

Figure 12.1 shows the expected m_{SD} distributions of the $\mathrm{t\bar{t}H}(H \rightarrow b\bar{b})$ signal and the background processes in the signal region after the fit, as well as the observed m_{SD} distribution in data. Overall, good agreement is observed between the predicted background and the observed data.

The signal strength modifier of the $\mathrm{t\bar{t}H}(H \rightarrow b\bar{b})$, which is defined as the ratio of the measured signal yield to the SM prediction, is determined $\hat{r} = \mu_{\mathrm{t\bar{t}H}(H \rightarrow b\bar{b})} = -1.5^{+5.3}_{-3.5}$ which is consistent with the SM expectation. The systematic uncertainties discussed in section 11.5 are taken into account in the fit as nuisance parameters, which allow for variations in the shape and normalization of the m_{SD} distributions during the fit. Figures 12.2, 12.3 and 12.2 show the post fit values of the nuisance parameters and their impacts on the signal strength modifier. The vast majority of the nuisances is pulled below 1σ. There are however, few nuisances pulled up to 2σ. Generally, these nuisances have a small effect on the signal strength modifier. Overall, the impacts show

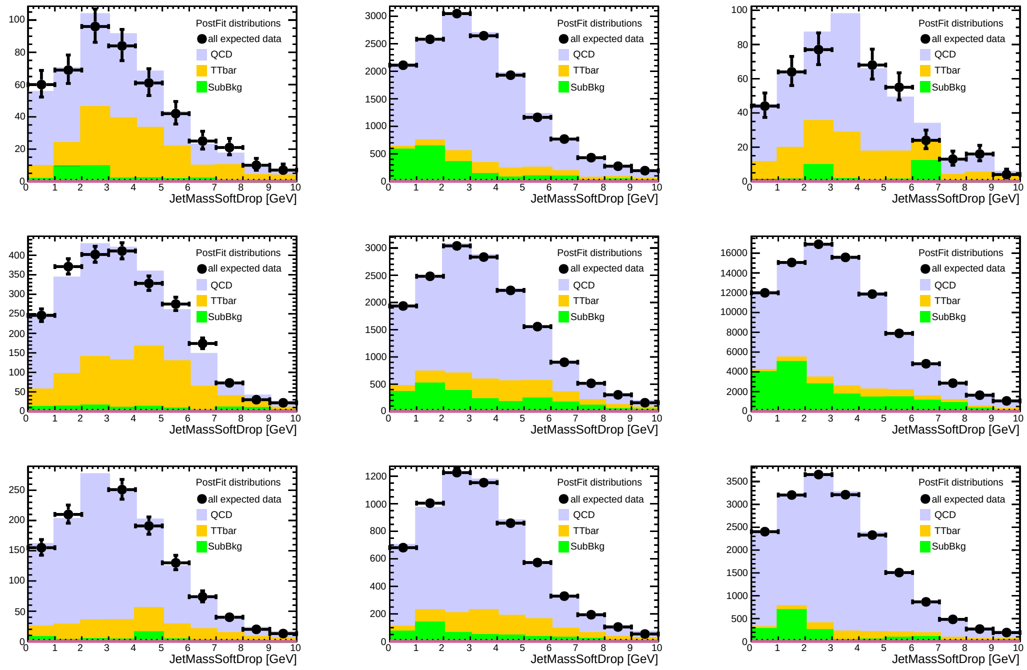


Figure 12.1: Post fit distributions of data and simulation for each analysis category and $t\bar{t}$ control region. Starting from the top left with category 0 to category 8. The data corresponds to an integrated luminosity of 35.6 fb^{-1}.

a fair/expected distribution of the pulls.

In addition, the impact of various systematics has been evaluated and those who turned out to have the larger effect are the ones related to btagging, $t\bar{t}$ shape, theoretical uncertainties and pile up. To study and to quantify the effect of each nuisance several tests were performed where one (or more) categories were ruled out and then a fit was performed using an Asimov dataset ($\mu = 1$). The conclusions are described in detail below:

- **$t\bar{t}$ shape uncertainty:** The $t\bar{t}$ shape uncertainty is correlated between all the categories and the $t\bar{t}$ CR. The $t\bar{t}$ CR in particular, is able to constrain the nuisances related to the $t\bar{t}$ shape. In addition, the pre-fit estimation for those uncertainties is more conservative as it can also be seen by analysis using a similar phase space.

- **uncertainties related to b-tagging:** The b-jet multiplicity is a key variable to classify events in each category. Therefore, the change on the b-tag rate is affecting significantly the ability of each nuisance to be constrained. Moreover, a larger constrain is expected by correlating the b-tag uncertainties between the categories. To see that effect, we performed a fit with an Asimov dataset only for categories with 3Ak8 jets (cat0 and cat1) and then for categories with 3Ak8 and 2 Ak8 jets with at least one b-tag jet (Categories 0-5). It looks like that adding categories that require bjet multiplicity (cat0-cat8) the uncertainties on b tagging are significantly constrained. In particular, for the Cat0-Cat1 fit all the btagging uncertainties are less that (50%) constrained whereas for the Cat0-Cat5 fit only (5-6%).

- **pile up uncertainties:** As for the bjet multiplicity, the analysis is sensitive in the multiplicity of jets and therefore, the pile up. A larger constraint is achieved by correlating the pile-up uncertainty between categories and processes.

- **theoretical uncertainties:** The theoretical uncertainties that affect the modelling of the signal and ttbar, are fitted as correlated between all categories and processes. A large constraint is expected from the $t\bar{t}$ CR where the effect is 5-6%.

- **background modelling uncertainties:** For the uncertainties related to our estimation of the backgrounds, dedicated methods were developed in order to constrain them. In some of the cases the estimation was on the conservative side (pre-fit).

Since the fitted value is also compatible with the background-only hypothesis, we set an upper limit on the signal strength parameter. The observed upper limit obtained is found to be 9.4 times the standard model expectations. The observed value is compatible with the expected value which is 10.4 with 1σ side bands 7.6 and 14.3.

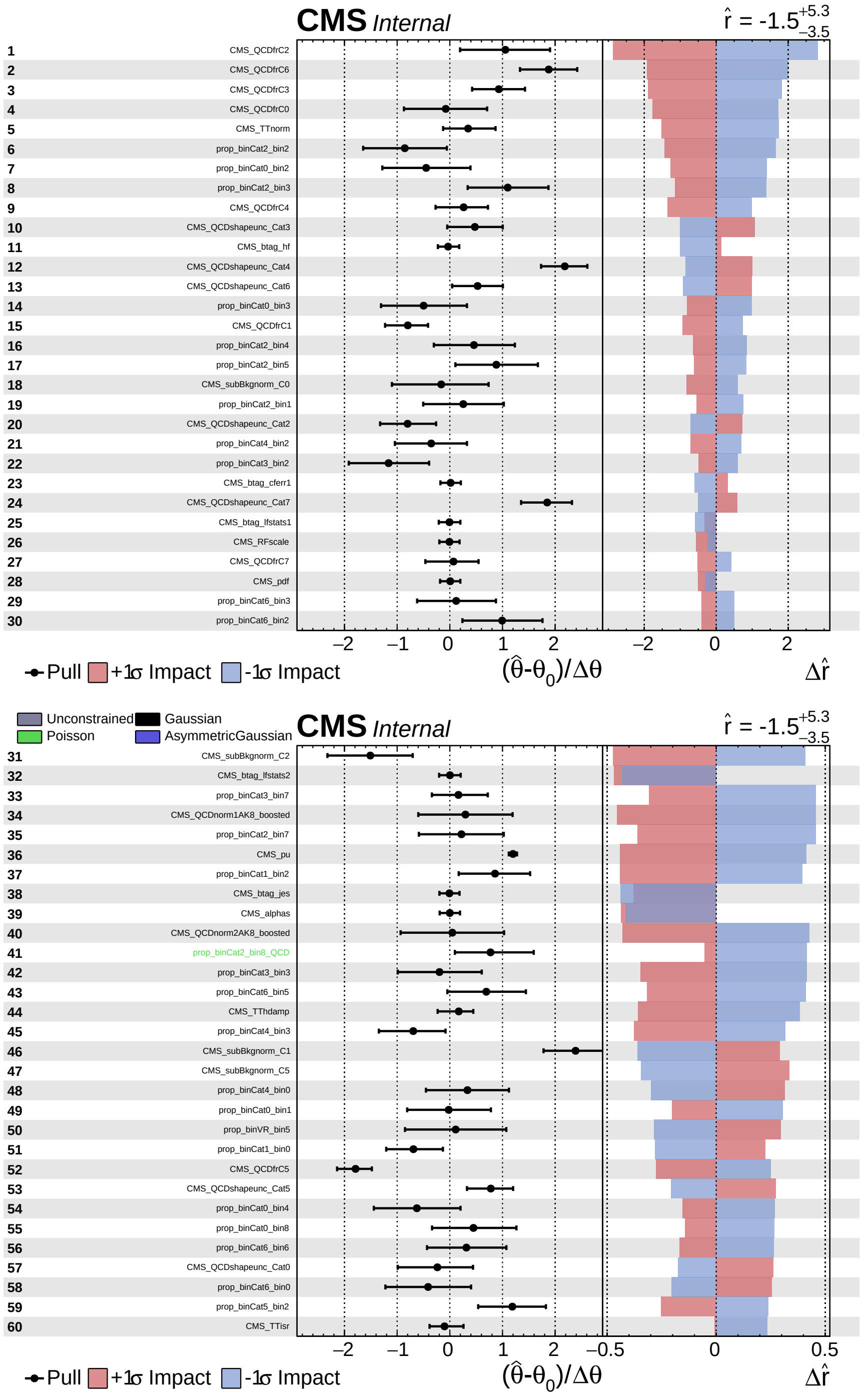


Figure 12.2: Post-fit impacts of the various systematics uncertainties on the signal strength.

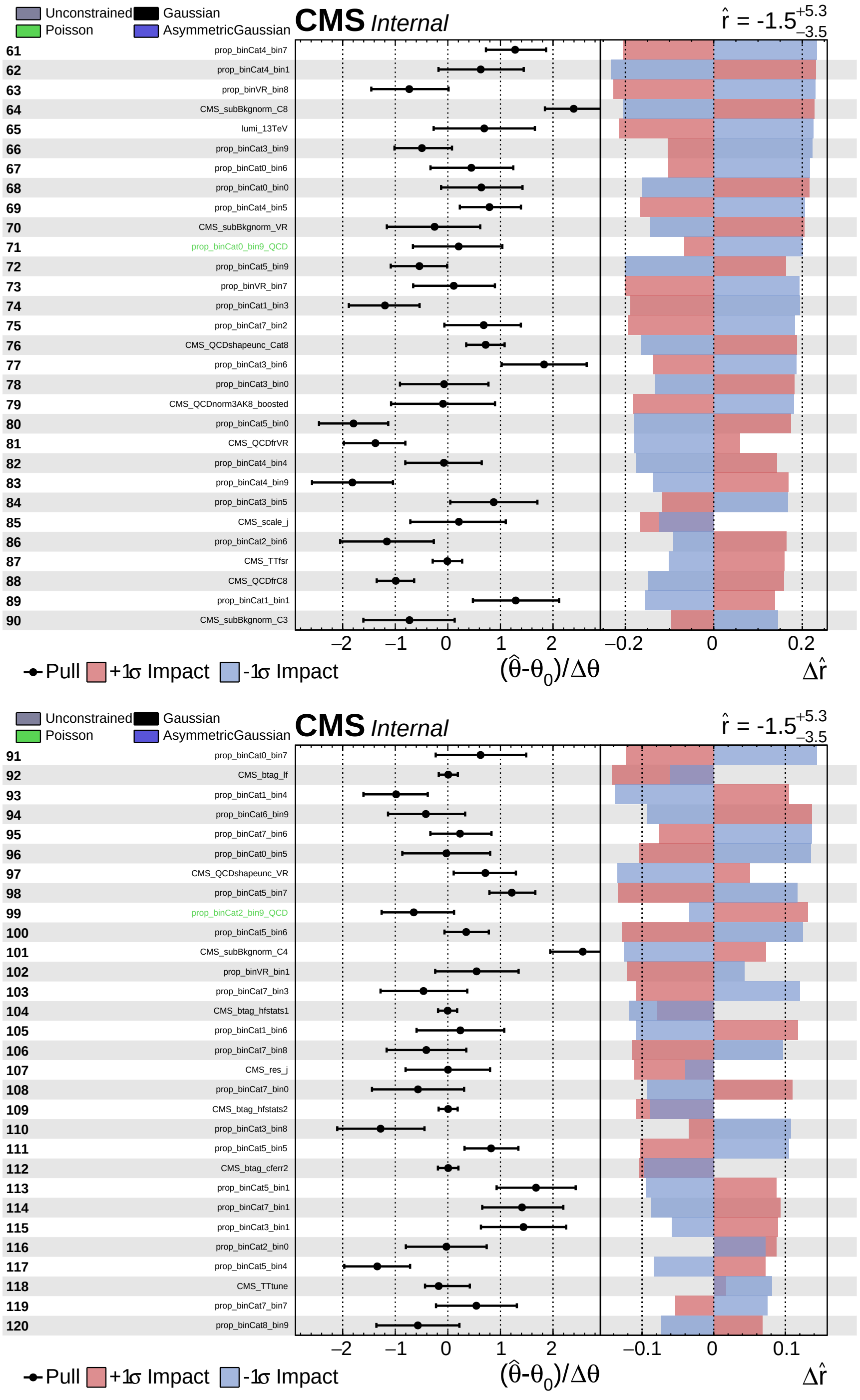
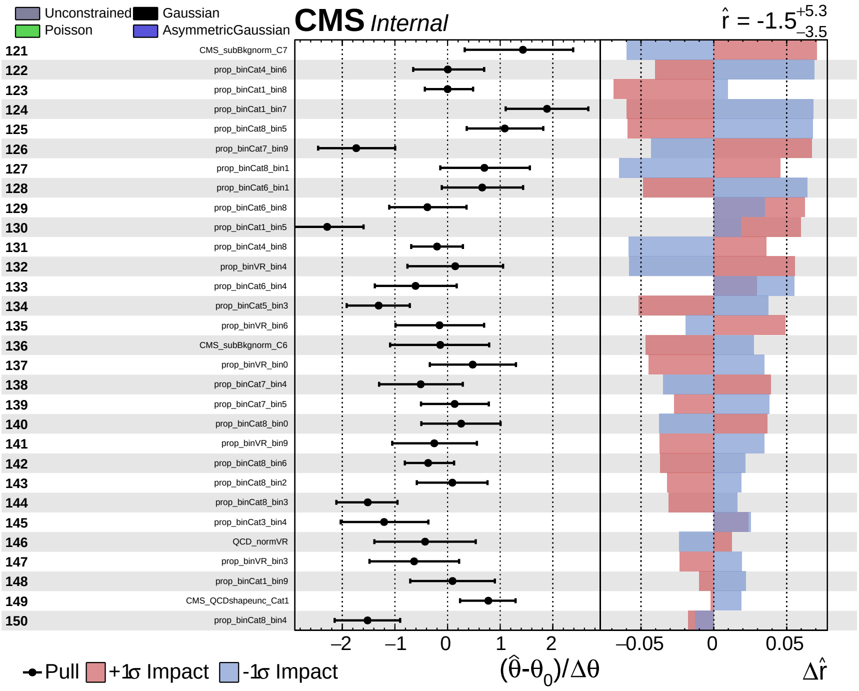


Figure 12.3: Post-fit impacts of the various systematics uncertainties on the signal strength.



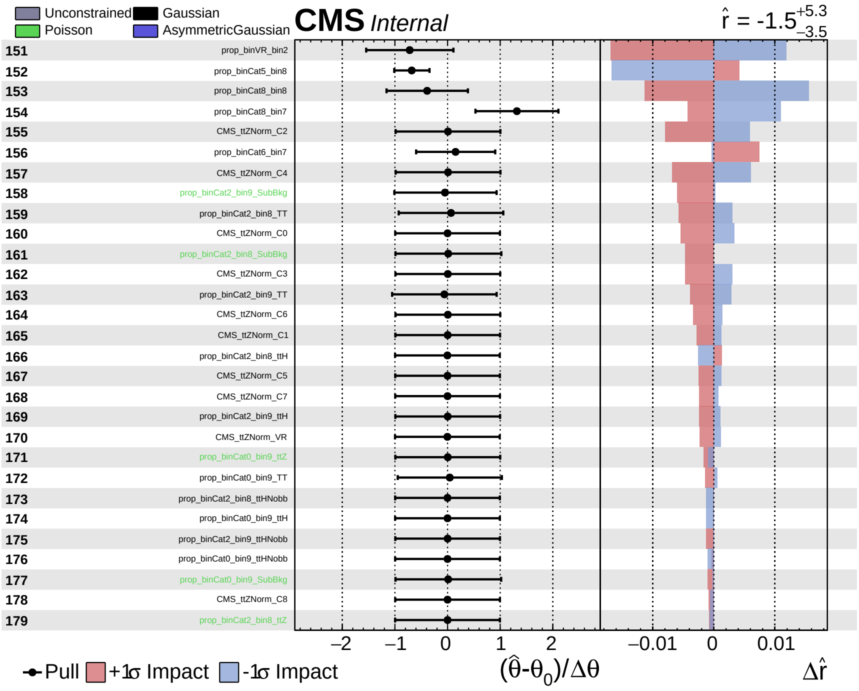


Figure 12.4: Post-fit impacts of the various systematics uncertainties on the signal strength.

12.2 Summary

A search is presented for the standard model Higgs boson produced in association with a top quark pair ($\mathrm{t\bar{t}H}$) in the all-jet final state using large-radius jets with data collected with the CMS detector in pp collisions at a centre-of-mass energy of 13 TeV in 2016, corresponding to an integrated luminosity of 35.9 fb^{-1}. The fitted (expected) 95$\%$ CL exclusion limit on the signal strength at 125 GeV is found to be 9.4 $(7.6 < 10.4 < 14.3)$ times the $\sigma\mathrm{BR}(\mathrm{t\bar{t}H} \to b\bar{b})$ predicted by the Standard Model. The analysis described in this note reports the first search of the production of a standard model Higgs boson in association with a top quark pair ($\mathrm{t\bar{t}H}$) in the fully-hadronic, boosted regime. A significant improvement is expected using the full Run2 dataset (2016+2017+2018). However, the larger improvement is expected during the HL-LHC where luminosity leveling will favor analyses using high-p_T^{miss} jets.

Part IV

Ελληνική Περίληψη

Κεφάλαιο 13

Θεωρητική Εισαγωγή

13.1 Το καθιερωμένο πρότυπο

Ένας από τους κυριότερους ερευνητικούς στόχους της φυσικής στοιχειωδών σωματιδίων είναι η ενοποίηση των θεμελιωδών αλληλεπιδράσεων της φύσης (ηλεκτρομαγνητική, ισχυρή, ασθενής, και βαρύτητα). Η προσπάθεια αυτή οδήγησε στη δημιουργία μιας θεωρίας που περιγράφει όλες τις θεμελιώδεις δυνάμεις εκτός από την βαρύτητα. Αυτή η θεωρία περιγράφει τις τρεις αλληλεπιδράσεις με την βοήθεια της τοπικά αναλλοίωτης θεωρίας βαθμίδας η οποία περιλαμβάνει την κβαντική ηλεκτροδυναμική, την ηλεκτρασθενή αλληλεπίδραση και την κβαντική χρωμοδυναμική και είναι γνωστή ως το Καθιερωμένο Πρότυπο (ΚΠ). Το Καθιερωμένο πρότυπο (ΚΠ) -Standard Model - της σωματιδιακής φυσικής έχει προκύψει από διαχρονικές επιτυχίες της πειραματικής και θεωρητικής φυσικής στην περιγραφή της πολυπλοκότητας που μας περιβάλλει, χρησιμοποιώντας θεμελιώδη σωματίδια και αλληλεπιδράσεις. Αποτελεί τη συμπαγή διατύπωση μιας σειράς θεωριών οι οποίες ερμηνεύουν και περιγράφουν τη συμπεριφορά της ύλης σε επίπεδο στοιχειωδών σωματιδίων. Εξακολουθεί να παραμένει το πιο ολοκληρωμένο μοντέλο που συμβάλλει στην κατανόηση του σύμπαντος μας. Όπως αναφέρθηκε προηγουμένως, περιγράφει τις τρεις από τις τέσσερις γνωστές θεμελιώδης αλληλεπιδράσεις μεταξύ των στοιχειωδών σωματιδίων στα πλαίσια μιας θεωρίας κβαντικών πεδίων της οποίας η θεμελιώδης ποσότητα (Lagrangian) είναι αναλλοίωτη κάτω από μια κατηγορία συνεχών τοπικών μετασχηματισμών βαθμίδας (Gauge Theory).

Η ηλεκτρασθενής θεωρία και η κβαντική χρωμοδυναμική ενοποιούνται σε μια θεωρία αναλλοίωτη κάτω από την ομάδα μετασχηματισμών βαθμίδας.

$$\mathrm{G_{SM}} = \mathrm{SU(3)_C} \otimes \mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y} \tag{13.1}$$

Η θεωρία βαθμίδας (gauge group theory) του ΚΠ, απαγορεύει την ύπαρξη μάζας σε όλα τα σωματίδια. Έτσι γεννάται το ερώτημα, γιατί τα στοιχειώδη σωματίδια φέρουν μάζα. Η εισαγωγή του μηχανισμού Higgs, δίνει μάζα στα διανυσματικά μποζόνια (Vector Bosons: W, Z) καθώς επίσης και σε ολόκληρο το φάσμα των στοιχειωδών σωματιδίων, ενώ παράλληλα αφήνει το φωτόνιο χωρίς μάζα.

Τον Ιούλιο του 2012 τα πειράματα CMS και ATLAS ανακοίνωσαν ταυτόχρονα την ανακάλυψη ενός νέου σωματιδίου μάζας περίπου 125 GeV, με ιδιότητες συμβατές με αυτές του μποζονίου Higgs.

Το ΚΠ απέχει από το να χαρακτηριστεί ως μια πλήρης θεωρία των θεμελιωδών αλληλεπιδράσεων, διότι δεν περιλαμβάνει τη φυσική της σκοτεινής ύλης και ενέργειας και αδυνατεί να εξηγήσει τη πλήρη θεωρία της βαρύτητας όπως περιγράφεται από τη γενική σχετικότητα. Εντούτοις, το ΚΠ είναι πολύ σημαντικό εξίσου για τη θεωρητική όσο και για τη πειραματική σωματιδιακή φυσική. Οι θεωρητικοί το χρησιμοποιούν ως βάση για τον σχεδιασμό εξωτικών μοντέλων που εμπεριέχουν υποθετικά σωματίδια, επιπλέον διαστάσεις και πραγματεύονται συμμετρίες. Αντίστοιχα οι πειραματικοί φυσικοί έχουν ενσωματώσει το ΚΠ σε προσομοιωτές ώστε να ερευνήσουν τη φυσική πέρα από το καθιερωμένο πρότυπο. Παράλληλα το ΚΠ μελετάται εξονυχιστικά από τους πειραματικούς σε διάφορες ενέργειες και πειράματα.

13.2 Περιγραφή του Καθιερωμένου Προτύπου

Το ΚΠ της σωματιδιακής φυσικής περιγράφει σχεδόν όλες τις θεμελιώδεις αλληλεπιδράσεις των σωματιδίων. Μπορεί να χωριστεί σε τρία μέρη: Το πρώτο περιλαμβάνει τα βασικά σωματίδια της ύλης, τα φερμιόνια με spin-1/2. Τα φερμιόνια μπορούν να χωριστούν σε δύο είδη, τα κουάρκ (Πίνακας 1) και τα λεπτόνια (Πίνακας 2) τα οποία κατατάσσονται σε τρεις γενιές με αυξανόμενη μάζα. Τα σωματίδια υψηλότερης γενιάς διασπώνται μέσω της ασθενούς αλληλεπίδρασης σε σωματίδια της πρώτης γενιάς. Υπάρχουν επίσης έξι γεύσεις (flavor) των κουάρκς. Το up (u), charm (c), top(t), τα οποία φέρουν +2/3 ηλεκτρικό φορτίο, και το down (d), strange (s), bottom (b) τα οποία φέρουν φορτίο -1/3. Επίσης τα λεπτόνια έχουν έξι γεύσεις, το ηλεκτρόνιο (e), το μιόνιο (μ) και το ταυ (τ) και τα αντίστοιχα νετρίνο τους. Στα πλαίσια του καθιερωμένου προτύπου τα νετρίνο έχουν μηδενικό φορτίο και αρχικά θεωρούνταν άμαζα. Όμως πρόσφατες μελέτες από πειράματα ταλαντώσεων νετρίνων έδειξαν ότι τα νετρίνο δεν έχουν αμελητέα μάζα. Σε κάθε ένα από αυτά τα δώδεκα σωματίδια αντιστοιχεί ένα αντισωματίδιο το οποίο έχει αντίθετο φορτίο.

Πίνακας 1.1: Κουάρκς

Πρώτη γεννιά	up (u)	down (d)
Δεύτερη γεννιά	charm (c)	strange (s)
Τρίτη γεννιά	top (t)	bottom (b)

Πίνακας 1.2: Λεπτόνια

Πρώτη γεννιά	electron (e)	electron neutrino (ν_e)
Δεύτερη γεννιά	muon (μ)	muon netrino (ν_μ)
Τρίτη γεννιά	tau (τ)	tau neutrino (ν_τ)

Πίνακας 1.3: Θεμελιώδεις αλληλεπιδράσεις

Interaction	Strength	Theory	Mediator boon	Charge
Strong	1	Chromodynamics	8 colored gluons	0
Electromagnetic	10^{-2}	Electrodynamics	γ	0
Weak	10^{-7}	Flavordynamics	W^{+}, W^{-}, Z	+1,-1,0
Gravitational	10^{-39}	General Relativity	Graviton	0

Το δεύτερο μέρος περιλαμβάνει τις θεμελιώδεις αλληλεπιδράσεις (Πίνακας 1.3) στα οποία αντιστοιχούν τα μποζόνια με spin-1. Οι δυνάμεις αυτές είναι διαφορετικής εμβέλειας και δύναμης. Οι φορείς τους είναι: το γλουόνιο για την ισχυρή αλληλεπίδραση, το φωτόνιο για την ηλεκτρομαγνητική, τα δυο W και Z για την ασθενή και το υποθετικό γκραβιτόνιο για τη βαρύτητα.

Το τρίτο μέρος περιλαμβάνει τον μηχανισμό Higgs που εισήχθει από τον Englert-Brout-Higgs (1964) και επιτρέπει στα σωματίδια να αποκτούν μάζα ανάλογα με την επίδρασή τους με το πεδίο Higgs. Οι μάζες καθορίζονται ανάλογα με το πόσο ισχυρή είναι η σύζευξη (coupling) με το πεδίο Higgs. Το SM Higgs έχει μάζα 125 GeV και είναι αποτέλεσμα του αυθόρμητου σπασίματος της ηλεκτρασθενούς συμμετρίας.

13.2.1 Μηχανισμός **Higgs**

Το καθιερωμένο πρότυπο συμπληρώνεται με τον μηχανισμό Higgs σύμφωνα με τον οποίο μέσω του αυθόρμητου σπάσιμου της ηλεκτρασθενούς συμμετρίας τα μποζόνια βαθμίδας $W^{\pm}$ και Z, φορείς της ασθενούς αλληλεπίδρασης αποκτούν μάζα, ενώ το φωτόνιο παραμένει άμαζο.

Εισάγοντας μια SU_2 διπλέττα βαθμωτών μιγαδικών πεδίω,ν

$$\Phi = \begin{pmatrix} \phi^{+} \\ \phi^{0} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi 3 + i\phi_4 \end{pmatrix} \tag{13.2}$$

η απλούστερη επανακανονικοποιησιμη μορφή δυναμικού είναι:

$$\mathrm{V}(\Phi) = \mu^2 \Phi^{\dagger}\Phi + \lambda(\Phi^{\dagger}\Phi)^2 \tag{13.3}$$

όπου μ μια πραγματική παράμετρος και λ μια μιγαδική παράμετρος.

Η λανγραντζιανή της ηλεκτρασθενούς θεωρίας είναι αναλλοίωτη κάτω από τον μετασχηματισμό $SU(2)_L \times U(1)_Y$ και με την συνεισφορά του δυναμικού της σχέσης 13.3 παίρνει την μορφή:

$$\mathcal{L} = T - V = (D_{\mu}\Phi)^{\dagger}(D^{\mu}\Phi) - (\mu^2\Phi^{\dagger}\Phi + (\lambda\Phi^{\dagger}\Phi)^2) \tag{13.4}$$

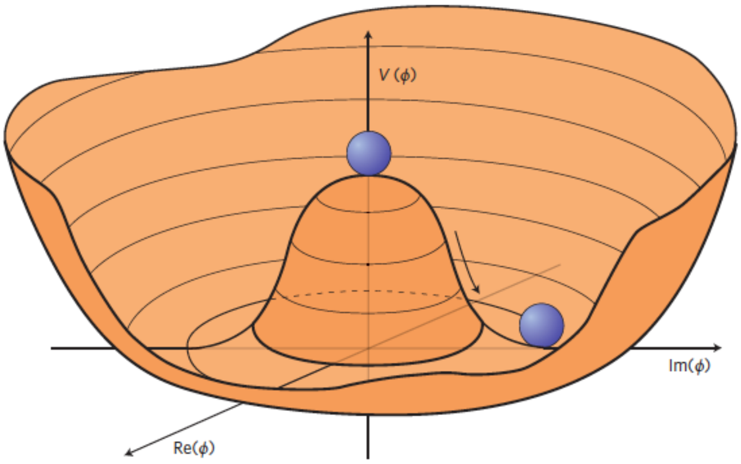
όπου $D_{\mu} = \partial_{\mu} + \frac{ig}{2}\sigma W_{\mu} + \frac{ig'}{2} Y B_{\mu}$ είναι η συναλλοίωτη παράγωγος του πεδίου Φ. Το W αντιστοιχεί στο $SU(2)_L$ και το B στο $U(1)$ πεδίο βαθμίδας της ηλεκτρασθενούς θεωρίας.

Ελαχιστοποιώντας το δυναμικό παίρνουμε την αναμενόμενη τιμή του κενού v. Στην περίπτωση που το $\mu^2 > 0$ η αναμενόμενη τιμή του κενού είναι 0 και έτσι η συμμετρία διατηρείται. Στην περίπτωση όμως που το $\mu^2 < 0$, τότε το δυναμικό ελαχιστοποιείται όταν:

$$|\Phi^2| = \Phi^{\dagger}\Phi = \frac{-\mu^2}{2\lambda} \equiv \frac{v^2}{2} \tag{13.5}$$

Επιλέγοντας λοιπόν την κατάσταση ελάχιστης ενέργειας καταλήγουμε στη σχέση:

$$\langle\phi\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v\end{pmatrix}, \quad \text{όπου} \quad v = \sqrt{-\frac{\mu^2}{\lambda}} \tag{13.6}$$



Σχήμα 13.1: Δυναμικό με μορφή 'Μεξικάνικου καπέλου' από το οποίο προέρχεται το αυθόρμητο σπάσιμο της συμμετρίας.

Η επιλογή ενός συγκεκριμένου κενού οδηγεί σε σπάσιμο της συμμετρίας. Το βαθμωτό πεδίο μπορεί να αναπτυχθεί γύρω από την αναμενόμενη τιμή του κενού v (σχήμα 13.1)

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_1 + \mathrm{i}\phi_2\\ v + h + \mathrm{i}\alpha^0\end{pmatrix} \tag{13.7}$$

Εισάγοντας την σχέση 13.7 στην Λανγρατζιανή του καθιερωμένου προτύπου, προκύπτουν τρία άμαζα **Goldstone** μποζόνια που αντιστοιχούν στα πεδία ϕ_1, ϕ_2 και α^0. Με αυτόν τον τρόπο τα μποζόνια Z και W αποκτούν μάζα

$$M_W^2 = \frac{g^2u^2}{4}, M_Z^2 = \frac{(g'^2 + g^2u^2)}{4}. \tag{13.8}$$

Το βαθμωτό δυναμικό πεδίο h που προκύπτει είναι το μποζονιο **Higgs** με μάζα

$$m_H = \sqrt{-2\mu^2} = \sqrt{2\lambda}v \tag{13.9}$$

Συνοψίζοντας, μέσω του αυθόρμητου σπάσιμου της ηλεκτρασυενούς συμμετρίας αποκτούν μάζα τα μποζόνια βαθμίδας $W^{\pm}$ Z^0 . Ο διαδότης της νέας αλληλεπίδρασης είναι το ουδέτερο μποζόνιο **Higgs** μη μηδενικής μάζας.

13.3 Επεκτάσεις του καθιερωμένου προτύπου

13.3.1 Κίνητρα για την Υπερσυμμετρία

Το καθιερωμένο πρότυπο της σωματιδιακής φυσικής αποτελούμενο από το ηλεκτραθενές μοντέλο και τη κβαντική χρωμοδυναμική είναι μια κβαντική θεωρία πεδίου εξαιρετικά επιτυχής στις προβλέψεις της. Ωστόσο υπάρχουν αρκετά αναπάντητα ερωτήματα πρακτικού αλλά κυρίως φιλοσοφικού περιεχομένου τα οποία το ΚΠ αποτυγχάνει να αντιμετωπίσει. Το μοντέλο περιέχει 19 ελεύθερες παραμέτρους τις οποίες η θεωρία αδυνατεί να προβλέψει και μετριούνται από το πείραμα. Δεν παρέχει όρους μάζας για τα νετρίνα ούτε ερμηνεία της ασυμμετρίας ύλης-αντιύλης του σύμπαντος. Επίσης αδυνατεί να παρέχει κατάλληλο μαθηματικό περιβάλλον ενοποίησης με τη θεωρία βαρύτητας της γενικής σχετικότητας. Δεν παρέχει υποψήφια σωματίδια τα οποία θα μπορούσαν να ερμηνεύσουν την παρατηρούμενη σκοτεινή ύλη και σκοτεινή ενέργεια του σύμπαντος. Τέλος εμπεριέχει το πρόβλημα ιεραρχίας, το οποίο συνοψήζεται στην απαίτηση μη φυσικών τεράστιων διορθώσεων, όταν γίνεται υπολογισμός της μάζας του Higgs σε μεγαλύτερες τάξεις της θεωρίας διαταραχών.

Η θεωρία της Υπερσυμμετρίας είναι μία από τις πολλά υποσχόμενες θεωρίες, η οποία μελετάται από τα πειράματα του μεγάλου αδρονικού επιταχυντή στο CERN, στη Γενεύη της Ελβετίας. Η Υπερσυμμετρία καταφέρνει να δώσει λύση στο πρόβλημα της ιεραρχίας αφού παρέχει μια πιο μεγάλη συμμετρία στο μοντέλο και δίνει τη δυνατότητα ενοποίησης των δυνάμεων σε πιο μεγάλη ενεργειακή κλίμακα για συγκεκριμένα μοντέλα. Παρέχει επίσης νέα σωματίδια υποψήφια για την ερμηνεία της σκοτεινής ύλης. Η ιδέα της υπερσυμμετρίας πηγάζει από τη παρατήρηση ότι η κβαντική ηλεκτροδυναμική δεν παρουσιάζει κανένα πρόβλημα σε υπολογισμούς μεγαλύτερων τάξεων, γιατί ακριβώς προστατεύεται από συγκεκριμένες συμμετρίες. Επεκτείνοντας τις ήδη υπάρχουσες συμμετρίες του μοντέλου σε ανταλλαγή μποζονίων και φερμιονίων, εμφανίζονται νέοι όροι αλληλεπίδρασης, οι οποίοι εξαφανίζουν το πρόβλημα της ιεραρχίας.

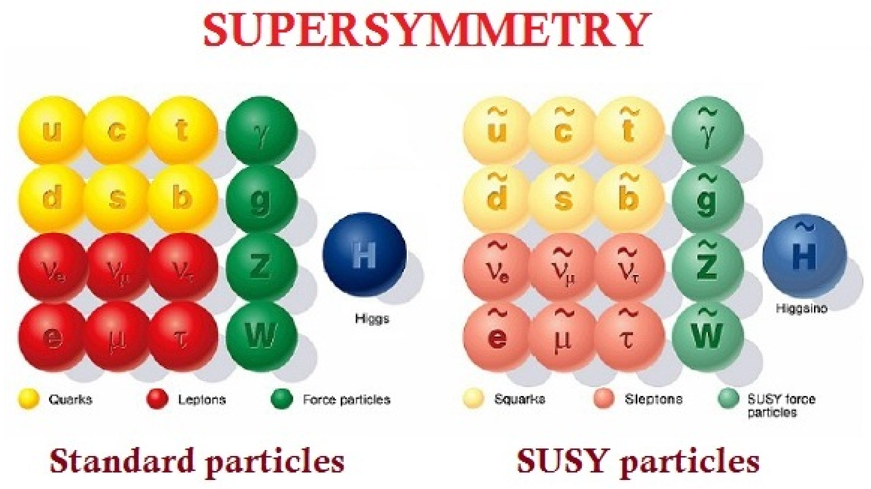
13.3.2 Υπερσυμμετρία (**Supersymmetry)**

Όπως προαναφέρθηκε η υπερσυμμετρία (SUSY) ειναι μια δημοφιλής επέκταση του ΚΠ η οποία προβλέπει την ύπαρξη ενός υπερσωματιδίου για κάθε σωματίδιο του ΚΠ. Αυτά τα υπερσωματίδια έχουν τους ίδιους κβαντικούς αριθμούς με τους αντίστοιχους των σωματιδίων του ΚΠ αλλά διαφέρουν κατά μισή μονάδα στο spin. Έτσι το κάθε σωματίδιο αποκτά ένα υπερσυμμετρικό "σύντροφο". Τα νέα σωματίδια επιπλέον θα πρέπει να είναι αρκετά βαριά ώστε να μην έχουν εως σήμερα παρατηρηθεί -η υπερσυμμετρία θα πρέπει να είναι μια σπασμένη συμμετρία- αλλά όχι πολύ πάνω από την ηλεκτρασθενή κλίμακα. Στο σχήμα 13.2 παρουσιάζονται τα σωματίδια του καθιερωμένου προτύπου με τους αντίστοιχους υπερσυμμετρικούς συντρόφους.

13.3.3 Το ελάχιστο Υπερσυμμετρικό καθιερωμένο πρότυπο **(MSSM)**

Το ελάχιστο Υπερσυμμετρικό καθιερωμένο πρότυπο (MSSM) είναι μια φαινομενολογική προσέγγιση μοντέλων υπερσυμμετρίας σε σχετικά χαμηλή ενέργεια. Είναι η πιο απλή επέκταση του Καθιερωμένου Προτύπου η οποία προτάθηκε το 1981 και παρέχει λύση στο πρόβλημα της ιεραρχίας. Στα πλαίσια του MSSM ο αριθμός των προβλεπόμενων σωματιδίων είναι ο μικρότερος δυνατός, έχουμε διατήρηση της R-parity και η θεωρία παραμένει αναλοίωτη κάτω από Gauge και Poincare μετασχηματισμούς. Όπως σε κάθε θεωρία Υπερσυμμετρίας έτσι και εδώ υπάρχει ένας υπερσυμμετρικός σύντροφος για κάθε σωματίδιο του ΚΠ με ίδιους κβαντικούς αριθμούς και με διαφορά στο spin 1/2. Επίσης η θεωρία πρέπει να είναι συμβατή με τις ιδιότητες του καθιερωμένου προτύπου, όπως η ύπαρξη χεραλικών φερμιονίων και η παραβίαση της ομοτιμίας. Η Λαγκρανζιανή του MSSM είναι της μορφής

$$\mathcal{L}_{MSSM} = \mathcal{L}_{SUSY} + \mathcal{L}_{Breaking} \tag{13.10}$$



Σχήμα 13.2: Τα σωματίδια του ΚΠ μαζί με τους αντίστοιχους υπερσυμμετρικούς συντρόφους τους.

όπου η $\mathcal{L}_{SUSY}$ είναι η Λαγκρανζιανή της Υπερσυμμετρίας και περιλαμβάνει τις αλληλεπιδράσεις Yukawa και βαθμίδας διατηρώντας την υπερσυμμετρική ανναλοιώτητα τους καθώς και η $\mathcal{L}_{Breaking}$ η οποία περιγράφει το σπάσιμο της υπερσυμμετρίας. Οι παράμετροι του MSSM μπορούν να περιγραφούν ξεχωριστά για το διατηρήσιμο και μη κομμάτι.

Έτσι το MSSM μπορεί να ενσωματώσει την υπερσυμμετρία στο καθιερωμένο πρότυπο κάνωντας ελάχιστες προσθέσεις (minimal) οι οποίες είναι απαραίτητες για τη μετάβαση από το ΚΠ στη θεωρία της Υπερσυμμετρίας. Οι βασικές του υποθέσεις είναι:

- θεωρεί υπερυμετρικούς συντρόφους στα μποζόνια βαθμίδας (gauginos)
- θεωρεί υπερσυμετρικούς συντόφους στα φερμιόνια (sparticles)
- θεωρεί συνήθως πάνω από ένα υπερσυμετρικο συντόφο για το πεδίο του Higgs (higginos)
- προσθέτει όρους "soft symmetry breaking"
- προσθέτει μια δεύτερη διπλέτα του Higgs.

Οι Πίνακες 13.1 και 13.2 παρουσιάζουν τις ιδιοκαταστάσεις του MSSM. Είναι δυνατή η μίξη μεταξύ των gauginos και higgsinos του ίδιου φορτίου καθώς και διαφόρων υπερσυμμετρικών φερμιονίων (sfermions) του ίδιου φορτίου. Μόνο στο γκλουίνο δεν επιτρέπεται τέτοια μίξη αφού το φορτίο χρώματος (color charge) εμποδίζει τη μίξη τους με άλλα σωματίδια. Τα ουδέτερα higgsinos ($\widetilde{H}_u^0$ και $\widetilde{H}_d^0$) συνδυάζονται με τα ουδέτερα gauginos ($\widetilde{B}^0$ και $\widetilde{W}^0$) για να δημιουργήσουν ιδιοκαταστάσεις μάζας γνωστές ως νετραλίνο (*neutrlinos*) ($\widetilde{\chi}_i^0$, i=1,2,3,4). Τα φορτισμένα higgsinos ($\widetilde{H}_u^+$ και $\widetilde{H}_d^-$) συνδυάζονται σε δυο καταστάσεις μάζας γνωστές ως *charginos* ($\widetilde{\chi}_i^\pm$, i=1,2).

13.3.4 Gauge Mediated Supersymmetry Breaking scenario

Το σενάριο για τη Gauge mediated Supersymmetry Breaking (GMSB) είναι από τα απλούστερα και παλαιότερα σενάρια. Έχει πολύ θεωρητικό ενδιαφέρον για τη νέα φυσική αφού όχι μόνο σταθεροποιεί τη μάζα του Higgs του ΚΠ αλλά αποφεύγει τα ουδέτερα ρεύματα αλλαγής γεύσης που υφίστανται σε άλλα σενάρια σπασίματος της υπερσυμμετρίας. Η βασική

Particles		spin-0	spin-1/2	$SU(3)_C$	$SU(2)_L$	$U(1)_Y$
squarks quarks,	Q	$(\tilde{u}_L, \tilde{d}_L)$	(u_L, d_L)	3	2	$\frac{1}{6}$
3 families	$\bar{u}$	$\tilde{u}_R^*$	$u_R^\dagger$	$\bar{3}$	1	$-\frac{2}{3}$
	$\bar{d}$	$\tilde{d}_R^*$	$d_R^\dagger$	$\bar{3}$	1	$\frac{1}{3}$
sleptons, leptons	L	$(\tilde{\nu}\ \tilde{e}_L)$	νe_L	1	2	$-\frac{1}{2}$
3 families	$\bar{e}$	$(\tilde{e}_R^*)$	$e_R^\dagger$	1	1	1
Higgs, higgsinos	H_u	(H_u^+, H_u^0)	$(\tilde{H}_u^+, \tilde{H}_u^0)$	1	2	$\frac{1}{2}$
Higgs, higgsinos	H_d	(H_d^0, H_d^-)	$(\tilde{H}_d^0, \tilde{H}_d^-)$	1	2	$-\frac{1}{2}$

Table 13.1: Οι χεραλικές υπερδιπλέτες του MSSM

Particles	spin-1/2	spin-1	$SU(3)_C$	$SU(2)_L$	$U(1)_Y$
gluino, gluon	$\tilde{g}$	g	8	1	0
winos, W bosons	$\tilde{W}^\pm$, $\tilde{W}^0$	$W^\pm$, W^0	1	3	0
bino, B boson	$\tilde{B}^0$	B^0	1	1	0

Table 13.2: Οι βαθμωτές υπερδιπλέτες του MSSM

αρχή αυτής της θεωρίας είναι πως οι υπερσυμετρικοί σύντροφοι των φερμιονίων του καθιερωμένου προτύπου αποκτούν μάζα μέσω αλληλεπιδράσεων βαθμίδας. Το μοντέλο αυτό αναπτύχθηκε την δεκαετία του 80 και περιέχει όλες τις παραμέτρους από του MSSM καθώς και το 'μαλακό' σπάσιμο της Υπερσυμμετρίας. Στο GMSB σενάριο το ελαφρύτερο υπερσυμμετρικό σωματίδιο είναι το gravitino ($\widetilde{G}$) ενώ το αμέσως ελαφρύτερο υπερσυμμετρικό σωματίδιο είναι το νετραλίνο ($\widetilde{\chi}_1^0$).

Κεφάλαιο 14

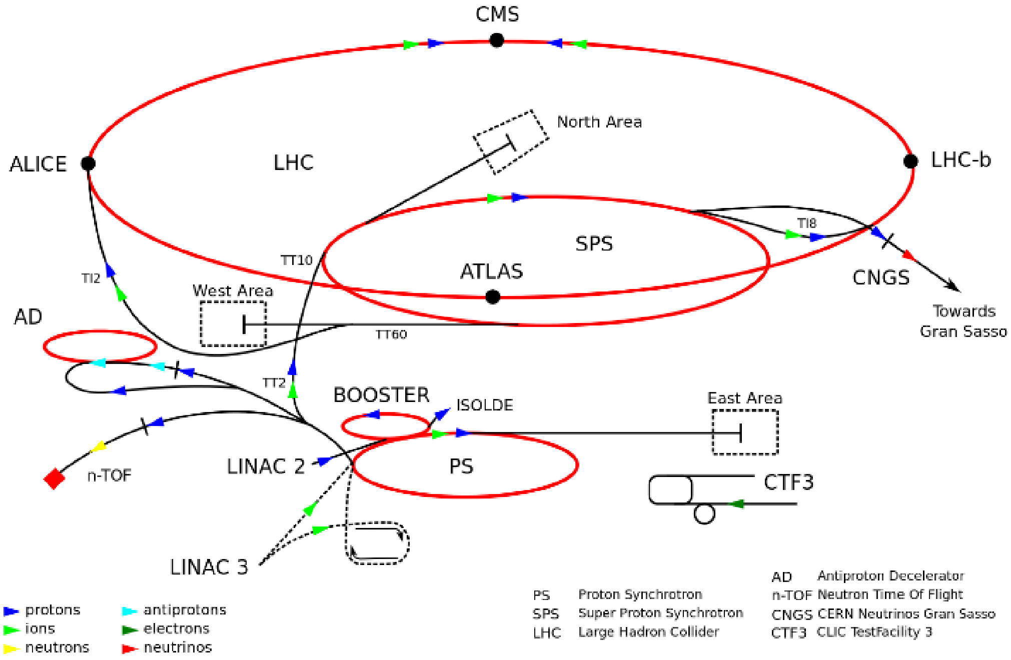
Το Πείραμα **CMS**

14.1 Το Ευρωπαικό Κέντρο Πυρηνικών Ερευνών

Το Ευρωπαϊκό Κέντρο Πυρηνικών Ερευνών CERN είναι ένα από τα μεγαλύτερα κέντρα πυρηνικής και σωματιδιακής φυσικής στον κόσμο. Το 1954 δώδεκα Δυτικές Ευρωπαϊκές χώρες, μεταξύ αυτών και η Ελλάδα συμφώνησαν στην ίδρυση του οργανισμού. Σήμερα το CERN αριθμεί 22 χώρες-μέλη. Στο CERN πραγματοποιούνται πολλά πειράματα με διεθνείς συνεργασίες. Ο βασικός τους στόχος είναι να παρέχει τα απαραίτητα εργαλεία (επιταχυντές σωματιδίων, ανιχνευτές) για τη μελέτη της βασικής έρευνας στη φυσική. Ένα από τα μεγαλύτερα επιτεύγματα του CERN πέρα από τη βασική έρευνα είναι η δημιουργία του Παγκόσμιου Ιστού (World Wide Web).

14.2 Ο Μεγάλος Αδρονικός Επιταχυντής **(LHC)**

Ο Μεγάλος Αδρονικός Επιταχυντής LHC είναι ένας επιταχυντής αδρονίων με περιφέρεια περίπου 26.7 km. Η εγκατάσταση του LHC έγινε στη ήδη υπάρχουσα υπόγεια κυκλική σήραγγα του επιταχυντή ηλεκτρονίου-ποζιτρονίου LEP. Η σήραγγα αυτή διασχίζει τα σύνορα Γαλλίας-Ελβετίας σε βάθος μεταξύ 45 και 175 m. Η εισαγωγή των πρωτονίων στον δακτύλιο επιτάχυνσης του LHC γίνεται με χρήση μιας σειράς άλλων επιταχυντών που λειτουργούν στις εγκαταστάσεις του CERN. Τα πρωτόνια που δημιουργούνται μέσω ιονισμού υδρογόνου ξεκινούν από τετράπολα ραδιοσυχνότητας (Radio Frequency Quadrupoles) έχοντας ενέργεια 750 keV. Ακολούθως επιταχύνονται αποκτώντας ενέργεια 50 MeV με τη βοήθεια του γραμμικού επιταχυντή LINAC και κατευθύνονται στον προωθητή (Booster) όπου αναπτύσσουν ενέργεια ίση με 1.4 GeV. Στη συνέχεια οδηγούνται στο σύγχροτο πρωτονίων (Proton Synchrotron) όπου επιταχύνονται μέχρι την ενέργεια των 25 GeV. και κατόπιν στο μεγάλο σύνγχροτο πρωτονίων (Super Proton Synchrotron) όπου καταλήγουν στην ενέργεια των 40 GeV. Από εκεί κατευθύνονται προς τους δύο ομόκεντρους δακτυλίους του LHC, με αντίθετες κατευθύνσεις μέχρι να αποκτήσουν την επιθυμητή ενέργεια. Κατόπιν συγκρούονται στα σημεία που είναι εγκατεστημένοι οι ανιχνευτές των τεσσάρων πειραμάτων ATLAS, CMS, ALICE, και LHCb. Όλα αυτά συνοψίζονται στην παρακάτω εικόνα.



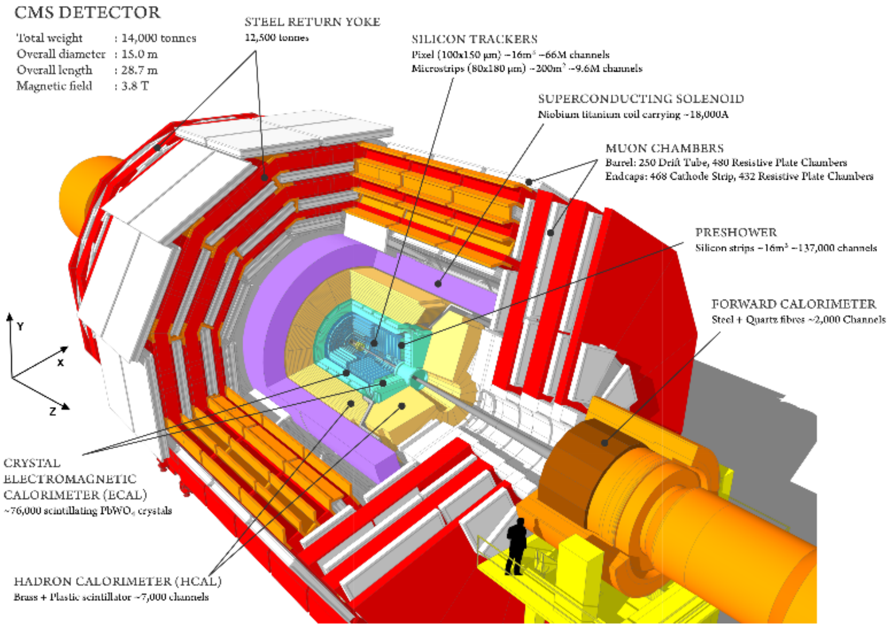
Σχήμα 14.1: Ολόκληρο το σύμπλεγμα του CERN μαζί με τα σημεία σύγκρουσης και τα αντίτοιχα πειράματα

14.3 Το **CMS**

Το Compact Muon Solenoid (CMS) είναι ένας γενικού σκοπού ανιχνευτής τοποθετημένος στο Σημείο 5 του δακτυλίου του LHC κοντά στο γαλλικό χωριό Cessy στην Γαλλία. Κατασκευάστηκε για να δώσει απαντήσεις σε πληθώρα ερωτημάτων, όπως ο ακριβής μηχανισμός του σπασίματος της ηλεκτραθενούς συμμετρίας, ψάχνοντας για το μποζόνιο του Higgs καθώς επίσης και για αποκλίσεις από το Καθιερωμένο Πρότυπο ψάχνοντας για ενδείξεις νέας φυσικής όπως η υπερυμμετρία, η σκοτεινή ύλη και οι έξτρα διαστάσεις. Στην εικόνα 14.2 παρουσιάζεται ο ανιχνευτής CMS.

Ο ανιχνευτής περιλαμβάνει ένα υπεραγώγιμο σωληνοειδή μαγνήτη με πολλαπλούς υποανιχνευτές μέσα σε αυτόν. Το σωληνοειδές παρέχει ένα μαγνητικό πεδίο 3.8 T κατά μήκος του ανιχνευτή έτσι ώστε να μπορεί να στρέψει τις τροχιές των φορτισμένων σωματιδίων. Σχεδιάστηκε με στόχο την μελέτη των προϊόντων σύγκρουσης πρωτονίων-πρωτονίων και ως εκ τούτου δύναται να ανιχνεύσει φωτόνια, ηλεκτρόνια, αδρόνια και μιόνια μετρώντας με ακρίβεια την ενέργειά τους. Παράλληλα καταγράφει έμμεσα τα παράγωγα που δεν αλληλεπιδρούν με τα υλικά του ανιχνευτή, μετρώντας το έλλειμμά στην εγκάρσια συνιστώσα της ορμής ή αλλιώς μετρώντας την ελλείπουσα εγκάρσια ενέργεια. Το σύστημα συντεταγμένων που έχει υιοθετηθεί από το πείραμα CMS, έχει την αρχή του (0,0) στο κέντρο του ανιχνευτή, στο σημείο σύγκρουσης. Ο άξονας y δείχνει κάθετα προς τα πάνω, ενώ ο άξονας χ δείχνει προς το κέντρο του LHC . Η αζιμουθιανή γωνία ϕ μετράται από τον άξονα χ στο επίπεδο χy ενώ η πολική γωνία θ μετράται τον άξονα z . Επιπροσθέτως, το έλλειμμά της ενέργειας όπως υπολογίζεται το εγκάρσιο επίπεδο, συμβολίζεται ως p_T^{miss}.

Οι διαστάσεις του ανιχνευτή CMS είναι 21.6 m μήκος, 14.6 m διάμετρος και συνολι-


CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
Y
X
Z
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Σχήμα 14.2: Ο ανιχνευτής CMS . Φαίνονται τα βασικά του χαρακτηριστικά κάθε κομματιού του ανιχνευτή.

κό βάρος 12500 t. Οι τροχιές των φορτισμένων σωματιδίων μετρώνται από το εσωτερικό σύστημα τροχιάς (inner tracking system), το οποίο αποτελείται από ένα ανιχνευτή pixel (pixel detector) και ένα ανιχνευτή τροχιάς πυριτίου (silicon strip tracker) καλύπτοντας $0 < \phi < 2\pi$ και $|\eta| < 2.5$ όπου η η ψευδοωκύτητα. Το σύστημα τροχιάς περικλείεται από το ηλεκτρομαγνητικό θερμιδόμετρο (ECAL) και το αδρονικό θερμιδόμετρο (HCAL). Το ECAL αποτελείται από 61200 κρυστάλλους βολφραμικού μολύβδου ($PbWO_4$) οι οποίοι είναι εγκατεστημένοι στο κεντρικό μέρος του βαρελιού (ECAL Barrel, EB) και 7324 που βρίσκονται τοποθετημένες στις άκρες (ECAL Endcap, EE). Επιπλέον, ένας ανιχνευτής προκαταιγισμού (preshower detector) έχει τοποθετηθεί μπροστά από τους κρυστάλλους στο EE, με σκοπό την αναγνώριση των ουδέτερων πιονίων την περιοχή $1.653 < |\eta| < 2.6$. Συντελεί επίσης στο διαχωρισμό των ηλεκτρονίων από τα σωματίδια γνωστά ως minimum bias ionizing και επιτρέπει τον προσδιορισμό της θέσης των ηλεκτρονίων και των φωτονίων. Το σύστημα σκανδαλιμού (trigger system) είναι οργανωμένο σε δυο επίπεδα και έχει ως στόχο την επιλογή των πιο ενδιαφέρον γεγονότων όπως προκύπτουν μετά από συγκρούσεις μεταξύ πρωτονίων.

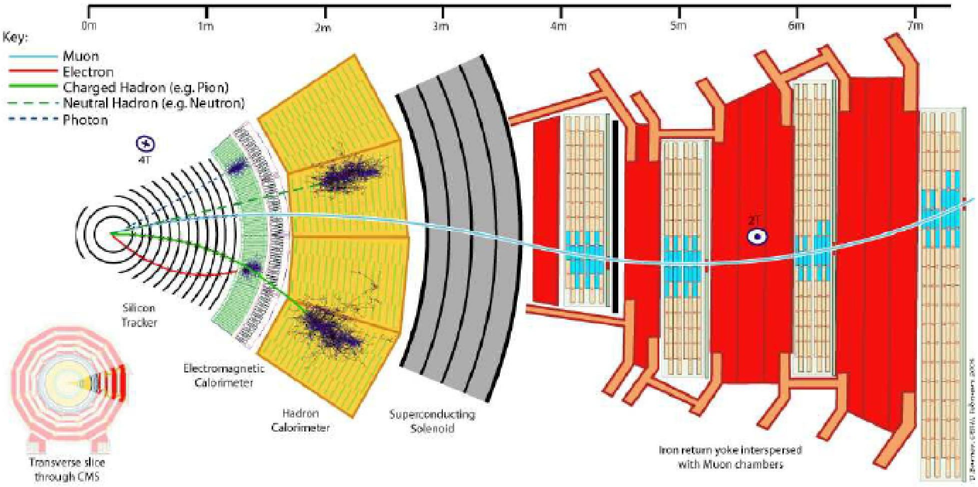
Κεφάλαιο 15

Ανακατακευή Γεγονότων και Προσομοιώσεις

Η ανακατασκευή των γεγονότων και των αντικειμένων πραγματοποιείται μέσω του αλγορίθμου Particle Flow (PF) ο οποίος έχει σχεδιαστεί ώστε να μπορεί να ταυτοποιήσει τα σταθερά σωματίδια που προέρχονται από τις συγκρούσεις πρωτονίων συνδυάζοντας πληροφορίες από επιμέρους υπό-ανιχνευτές του CMS. Αυτό είναι δυνατόν χάρη στην εξαιρετική διακριτική ικανότητα του ανιχνευτή τροχιών πυριτίου και του ηλεκτρομαγνητικού θερμιδομέτρου. Για την ταυτοποίηση τροχιών χρησιμοποιούνται αλγόριθμοι επανάληψης υψηλής αποδοτικότητας (interative tracking algorithm) κατάλληλα σχεδιασμένοι στο να ταυτοποιήσουν πίδακες (jets) δίνοντας μια μικρή πιθανότητα να ανακατασκευάσουν λάθος την τροχιά. Αυτή η πιθανότητα υπολογίζεται πως είναι μικρότερη του 1% ακόμα και για τις πιο περίπλοκες περιπτώσεις χαμηλής ορμής με τροχιές που έχουν προέλθει από τον άξονα της δέσμης. Για τη ενέργεια που εναποτίθεται στα θερμδόμετρα υπάρχει ένας αλγόριθμος ομαδοποίησης (clustering algorithm) ικανός να ξεχωρίζει κοντινές εναποθέσεις διατηρώντας υψηλή απόδοση ακόμα και σε χαμηλές ενέργειες. Ο αλγόριθμος PF αναγνωρίζει μιόνια συγκρίνοντας τροχιές από τον ανιχνευτή τροχιών πιριτίου και τροχιές από τα μιονικά συστήματα ανίχνευσης του CMS. Τα ηλεκτρόνια ανακατασκευάζονται χρησιμοποιώντας μεταβλητές από το ηλεκτρομαγνητικό θερμιδόμετρο και τον ανιχνευτή τροχιών. Όλα τα υπόλοιπα σωματίδια μπορούν να ταυτοποιηθούν ως αδρόνια, ουδέτερα αδρόνια ή φωτόνια συγκρίνοντας την ορμή των τροχιών με την εναπόθεση που άφησαν η μη στο θερμιδόμετρο. Στα σωματίδια που ανακατασκευάστηκαν εφαρμόζονται αυστηρότερα κριτήρια και χρησιμοποιούνται ως βασική παράμετρος σε άλλους αλγορίθμους όπως την ανακατασκευή των πιδάκων ή στον υπολογισμό της ελλείπουσας εγκάρσιας ορμής. Στο σχήμα 15.1 φαίνεται η γενική εικόνα ανακατασκευής των αντικειμένων που χρησιμοποιούνται στην ανάλυση.

15.1 Τροχιές και Κορυφές

Οι τροχιές των στοιχειωδών σωματιδίων αναγνωρίζονται χρησιμοποιώντας την Kalman Filter μέθοδο στην οποία πραγματοποιείται μια προσαρμογή (fit) λαμβάνοντας υπόψιν τις πολλαπλές σκεδάσεις με αποτέλεσμα τη παράμετρο κρούσης και την αρχική ορμή. Οι Βασικές Τροχιές (Primary Vertices) ανακατασκευάζονται από την ομαδοποίηση τροχιών οι οποίες είναι συμβατές με τη περιοχή της βασικής αλληλεπίδρασης. Η θέση της κάθε κορυφής προσαρμόζεται από την αντίστοιχη κορυφή χρησιμοποιώντας το adaptive vertex fitting.


0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS

Σχήμα 15.1: Μια τομή του ανιχνευτή CMS μαζί με τι αλληλεπιδράσεις από διάφορα σωματίδια.

15.2 Φωτόνια

Ένα φωτόνιο που παράγεται στο σημείο της σύγκρουσης, πρώτα περνάει από τον ανιχνευτή τροχιών και έπειτα εισέρχεται στο ηλεκτρομαγνητικό θερμιδόμετρο ECAL όπου χάνει όλη του την ενέργεια μέσω του ηλεκτρομαγνητικού καταιγισμού. Υπάρχουν δύο περιπτώσεις. Στην πρώτη περίπτωση το φωτόνιο περνάει από τον ανιχνευτή τροχιών χωρίς να αλληλεπιδράσει και αφήνει σχεδόν όλη του την ενέργεια στους κρυστάλλους του ECAL. Ένα τέτοιο φωτόνιο ονομάζεται unconverted. Στην δεύτερη περίπτωση, το φωτόνιο μετατρέπεται σε ζεύγος ηλεκτρονίου-ποζιτρονίου πριν εισέλθει στο ηλεκτρομαγνητικό θερμιδόμετρο. Το ηλεκτρόνιο και το ποζιτρόνιο που δημιουργήθηκε επηρεάζονται από το μαγνητικό πεδίο και αφήνουν την ενέργεια στο θερμιδόμετρο σε μεγαλύτερες γωνίες ϕ. Αυτά τα φωτόνια ονομάζονται converted. Για να συμπεριλάβουν όλη την εναπόθεση ενέργειας, τα φωτόνια ανακατασκευάζονται χρησιμοποιώντας "superclusters". Η ενέργεια των φωτονίων συλλέγεται αθροίζοντας την ενέργεια που έχει εναποτεθεί στους κρυστάλλους και την ενέργεια που έχει συλλεχθεί στις άκρες από τον ανιχνευτή προ-καταιγισμού("pre shower detector").

15.3 Ηλεκτρόνια

Τα ηλεκτρόνια ανακατασκευάζονται συνδυάζοντας τα supercluster του θερμιδιμέτρου με μια τροχιά. Η υποψήφια τροχιά προκύπτει κάνοντας προσαρμογή των κτυπημάτων του ανιχνευτή τροχιών με τον αλγόριθμο Gaussian-sum filter (GSF) στην οποία μοντελοποιείται η απώλεια ενέργειας λόγω ακτινοβολίας πέδησης με ένα άθροισμα από γκοαυσιανές συγκεκριμένου βάρους.

15.4 Μιόνια

Τα μιόνια ανακατασκευάζονται χρησιμοποιώντας τη μέθοδο global muon reconstruction η οποία συνδυάζει την πληροφορία από τον ανιχνευτή μιονίων και τον ανιχνευτή τροχιών.

15.5 Ανακατασκευή Αδρονικών πιδάκων

Μια από τις μεγαλύτερες προκλήσεις της φυσικής των υψηλών ενεργειών είναι η επιτυχής ταυτοποίησή και ανακατασκευή σωματιδίων των οποίων οι τελικές καταστάσεις περιέχουν κουαρκς. Τα κουαρκ που παράγονται από ανελαστική σκέδαση δεν μπορούν να υπάρξουν ως ελευθέρα σωματίδια λόγω περιορισμών από την QCD. Αυτό που μπορεί να μετρηθεί σε έναν ανιχνευτή είναι το αποτέλεσμα από μια διαδικασία μετατροπής coloured partons σε ένα καταιγισμό από colourless partons. Το αποτύπωμα των κουαρκ και των γλουωνίων ονομάζεται jet. Ένα jet περιέχει κυρίως σωματίδια όπως πιόνια, καόνια ακόμα και πρωτόνια και νετρόνια.

Η ανακατασκευή των jets ξεκινά από δεδομένα που προέρχονται από τα θερμιδόμετρα (ECAL και HCAL). Έπειτα χρησιμοποιείται ο αλγόριθμος anti-k_T ο οποίος λαμβάνει υπ' όψιν του το σωματίδιο με τη μεγαλύτερη ορμή και ομαδοποιεί τα γειτονικά σωματίδια σε ένα πίδακα με βάση την ενέργεια και την ορμή. Τέλος, στα ανακατασκευασμένα jets πραγματοποιείται μια ενεργειακή βαθμονόμηση καθώς και επιπλέον διορθώσεις έτσι ώστε η ταυτοποίηση τους να έχει μεγάλη ακρίβεια.

15.6 Ταυτοποίηση του **b quark**

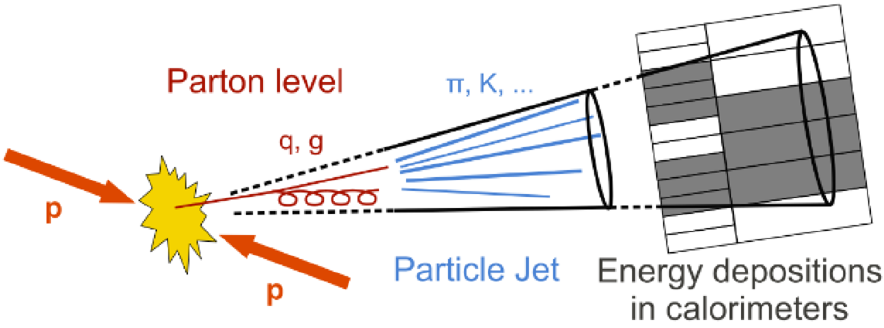
Η επιτυχής ταυτοποίηση του b quark είναι απαραίτητη σε πολλές αναλύσεις της πειραματικής φυσικής υψηλών ενεργειών καθώς αυτά βρίσκονται σε πολλές τελικές καταστάσεις. Για αυτό τον λόγο αυτό στο CMS έχουν αναπτυχθεί ειδικοί αλγόριθμοι που μπορούν να ταυτοποιήσουν ένα b jet με μεγάλη ακρίβεια. Αυτοί βασίζονται στις ιδιότητες του b quark. Για παράδειγμα ο αρκετά υψηλός χρόνο ζωής πριν αυτό διασπαστεί οδηγεί στην δημιουργία δευτερευουσών κορυφών (secondary vertices) οι οποίες αντιστοιχούν στα σημεία διασπασής του. Ο αλγόριθμος CSV (Combined Secondary Vertex) χρησιμοποιεί τις παραπάνω ιδιότητες για να ταυτοποιήσει ένα b jet και είναι αυτός που χρησιμοποιήθηκε για την ανάλυση των δεδομένων του 2016.

15.7 Ελλείπουσα εγκάρσια ορμή

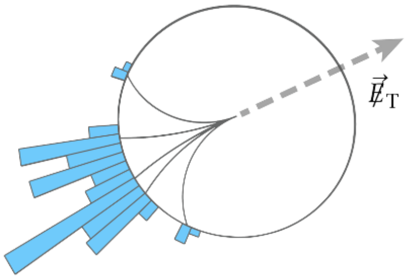
Η ελλείπουσα εγκάρσια ορμή p_T^{miss} ανακατασκευάζεται από τα υποψήφια ηλεκτρόνια, μιόνια, φωτόνια και από τα φορτισμένα αδρόνια που προκύπτουν από τον particle flow αλγόριθμο. Όπως έχει αναφερθεί τα σωματίδια εμφανίζονται ως τροχιές και εναποθέσεις ενέργειας. Ο PF αλγόριθμος αποσκοπεί στο να αναγνωρίσει όλα τα σταθερά σωματίδια συνδυάζοντας πληροφορία από κάθε επιμέρους ανιχνευτή. Η συνολική εγκάρσια ορμή ορίζεται ως το διανυσματικό άθροισμα όλων των ανακατασκευασμένων σωματιδίων του γεγονότος. Η εγκάρσια ελλείπουσα ορμή ορίζεται ως το αρνητικό διανυσματικό άθροισμα όλων των ανακατακευασμένων σωματιδίων του γεγονότος:

$$\vec{p_T}^{miss} = -\sum_i \vec{p_i}, \qquad i = PF \;\; candidates \tag{15.1}$$

Η p_T^{miss} είναι το μέτρο αυτού του διανύματος. Η p_T^{miss} μπορεί να προέρχεται από πολλές πηγές του πειράματος. Μπορεί να υπάρχει "αληθινή" εγκάρσια ελλείπουσα ορμή από διεργασίες που περιέχουν νετραλίνο και δεν ανιχνεύονται από τον ανιχνευτή, ή από υπογραφές υπερσυμμετρίας όπου ο ελαφρύτερος υπερσυμμετρικός σύντροφος συντελεί στην



Σχήμα 15.2: Μια αναπαράσταση της δημιουργίας ενός πίδακα από τις συγκρούσεις πρωτονίων στο CMS και οι αντίστοιχες εναποθέσεις ενέργειας στα θερμιδόμετρα



Σχήμα 15.3: Ελλείπουσα εγκάρσια ορμή ορίζεται ως το αρνητικό άθροισμα των εγκάρσιων ορμών όλων των PF υποψηφίων.

εμφάνιση υψηλής p_T^{miss}. Όμως η p_T^{miss} μπορεί να προέρχεται από κακή μέτρηση της εγκάρσιας ορμής κατά τη διάρκεια ανακατασκευής μέσω των μεθόδων που περιγράψαμε. Ο ακριβής υπολογισμός της είναι πολύ σημαντικός στο πείραμα CMS και έχουν δημιουργηθεί πολλές τεχνικές ώστε η μέτρηση της να είναι ακριβής.

15.8 Προσωμοιώσεις **Monte Calro**

Στη φυσική υψηλών ενεργειών χρησιμοποιούνται προσομοιώσεις του ανιχνευτή και των διαφόρων φυσικών διεργασιών. Αυτές οι μοντελοποιήσεις οδηγούν στο να βελτιώθουν οι τεχνικές ανάλυσης και να κατανοηθεί η επίδοση πολλών και περίπλοκων υποανιχνευτών όπως αυτών που περιέχονται στο CMS. Οι προσομοιώσεις πραγματοποιούνται με τη μέθοδο Monte Carlo στην οποία χρησιμοποιούνται τυχαία δείγματα στα θεωρητικά μοντέλα ώστε να προβλεφθεί η αναμενόμενη συμπεριφορά τους κάτω από ρεαλιστικές συνθήκες. Βασίζεται σε υπολογιστικές προσομοιώσεις και μπορεί να δώσει σωστές απαντήσεις εκεί που δεν μπορεί να δωθεί ντετερμινιστική λύση. Παραδείγματα στην φυσική υψηλών ενεργειών περιλαμβάνουν προσομοιώσεις γεγονότων όπου τα σωματίδια παράγονται σε τυχαία κατεύθυνση και θέση υπακούγοντας σε θεωρητικούς περιορισμούς καθώς και σε προσωμοιώσεις του ανιχνευτή όπου λαμβάνονται υπόψιν διάφοροι παράμετροι όπως ο ηλεκτρονικός θόρυβος κ.α. Τα βασικά προγράμματα προσομοίωησης για τη δημιουργία γεγονότων είναι η

PYTHIA , POWHEG, MADGRAPH και TAUOLA. Η PYTHIA είναι ένας γεννήτορας γεγονότων γενικού σκοπού. Περιλαμβάνει βασικές διεργασίες όπως της κβαντικής χρωμοδυναμικής QCD και της Υπερσυμμετρίας, της παραγωγής του Higgs αλλά και εξωτικής φυσικής. Η PYTHIA χρησιμοποιείται σε συνδυασμό και με άλλους γεννήτορες όπως τον MADGRAPH και POWHEG .Η μέθοδος POWHEG είναι μια βελτιωμένη εκδοχή της PYTHIA αφού χρησιμοποιεί Next to leading order (NLO) υπολογισμούς λαμβάνοντας υπόψιν και τον καταιγισμό των παρτονίων. Η PYTHIA είναι πολύ αποτελεσματική στη περιγραφή απλών 2 → 2 διεργασιών. Παρόλα αυτά τις περισσότερες φορές στις τελικές καταστάσεις έχουμε περισσότερα σωματίδια. Ο γεννήτορας MADGRAPH μπορεί να δώσει καλύτερη περιγραφή τέτοιων περίπλοκων τελικών καταστάσεων.

Η περιπλοκότητα του ανιχνευτή CMS απαιτεί μια πολύ πολύπλοκη προσομοίωση ώστε να αναπαραχθεί η συμπεριφορά του ανιχνευτή στην παρουσία σωματιδίων από τις συγκρούσεις των πρωτονίων. Αυτό συμβαίνει χρησιμοποιώντας το πακέτο GEANT4 το οποίο περιγράφει τον ανιχνευτή λαμβάνοντας υπόψιν την ακριβή του γεωμετρία, το υλικό κατασκευής κ.α.

Κεφάλαιο 16

Έρευνα για υπερσυμμετρία με τελικές καταστάσεις φωτονίων και εγκάρσιας ελλείπουσας ορμής

Αυτό το κεφάλαιο περιλαμβάνει τα αποτελέσματα της αναζήτησης νέας φυσικής σε τελικές καταστάσεις φωτονίων και εγκάρσιας ελλείπουσας ορμής. Τα δεδομένα που χρησιμοποιήθηκαν συλλέχτηκα από τον ανιχνευτή CMS το 2016 και αντιστοιχούν σε ολοκληρωμένη φωτεινότητά 35.9 fb^{-1}.

16.1 Φαινομενολογία τελικών καταστάσεων

Σε πολλά διαδεδομένα σενάρια νέας φυσικής πέραν του Καθιερωμένου Προτύπου και κυρίως σε σενάρια Υπερσυμμετρίας (SUSY), περιέχονται μοντέλα με τελικές υπογραφές φωτονίων και υψηλής εγκάρσιας ελλείπουσας ορμής (p_T^{miss}). Σε αυτά τα μοντέλα συνήθως το σπάσιμο της υπερσυμμετρίας μεταφέρεται σε χαμηλότερες ενεργειακές κλίμακες μέσω διαμεσολαβητών βαθμίδας (gauge mediated Supersymmetry breaking - GMSB). Αυτά τα μοντέλα έχουν ένα σύνολο κοινών χαρακτηριστικών και περιλαμβάνουν ένα σταθερό, ασθενώς αλληλεπιδρών ελαφρότερο υπερσυμμετρικό σύντροφο (Lightest Supersymmetric Partner-LSP). Ως εγκάρσια ελλείπουσα ορμή ορίζουμε το μέτρο του αντίθετου διανυσματικού αθροίσματος των εγκάρσιων ορμών όλων των «ορατών» σωματιδίων του γεγονότος και μπορεί να προκύπτει από μη ανιχνεύσιμα σωματίδια, όπως τα νετρίνα και τα υποθετικά υπερσυμμετρικά σωματίδια. Στα GGM (General Gauge Mediation) μοντέλα ως LSP θεωρείται το σχεδόν άμαζο gravitino ($\widetilde{G}$) με το επόμενο ελαφρότερο σωματίδιο (Next to the Lightest Supersymmetric Particle (NLSP)) να είναι ένα ουδέτερο νετραλίνο ($\widetilde{\chi}$). Η παρούσα έρευνα αφορά στην αναζήτηση νέας φυσικής συμβατής με το μοντέλο GGM SUSY σε τελικές καταστάσεις οι οποίες αποτελούνται από δυο φωτόνια και υψηλές τιμές εγκάρσιας ελλείπουσας ορμής. Παρόμοιες αναλύσεις έχουν πραγματοποιηθεί στο παρελθόν με δεδομένα που συλλέχθηκαν από τον ανιχνευτή CMS μικρότερης ολοκληρωμένης φωτεινότητας. Τα δεδομένα που χρησιμοποιήθηκαν στην συγκεκριμένη έρευνα συλλέχθηκαν με ενέργεια κέντρου μάζας $\sqrt{s} = 13$ TeV γεγονός που συνέβαλε θετικά στην ευαισθησία των μετρήσεων. Για την ερμηνεία των αποτελεσμάτων, δύο μοντέλα χρησιμοποιήθηκαν που ανήκουν στα λεγόμενα απλοποιημένα μοντέλα φυσικής (Simplified Models) τα οποία βασίζονται σε μικρό αριθμό υποθέσεων και παραμέτρων έτσι ώστε οι τελικές καταστάσεις που προβλέπουν να μπορούν να ταυτοποιηθούν από πειράματα επιταχυντών. Το πρώτο από αυτά είναι το μοντέλο T5gg κατά το οποίο πραγματοποιείται ταυτόχρονη παραγωγή δυο γλουονίων (gluon ($\widetilde{g}$)) και το δεύτερο, το T6gg, όπου πραγματοποιείται ταυτόχρονη παραγωγή δυο squark. Και στα δύο μοντέλα το LSP είναι το σχεδόν άμαζο gravitino ($\widetilde{G}$) και το αμέσως ελαφρύτερο υπερσυμμετρικό σωματίδιο είναι το νετραλίνο ($\widetilde{\chi}$).

Σε αυτά τα μοντέλα υπάρχει διατήρηση της R-ομοτιμίας γεγονός που διασφαλίζει πως το LSP είναι άμαζο και δεν αλληλεπιδρά με τον ανιχνευτή. Για αυτό τον λόγο στις συγκρούσεις πρωτονίων όπου μπορούν να παραχθούν υπερσυμμετρικά σωματίδια αναμένεται σημαντική ελλείπουσα εγκάρσια ορμή. Επίσης στα μοντέλα αυτά, όλα τα νετραλίνο θα διασπαστούν σε ένα gravitino ($\widetilde{G}$) και ένα φωτόνιο με αποτέλεσμα χαρακτηριστικές καταστάσεις γεγονότων με δύο φωτόνια και υψηλή ελλείπουσα εγκάρσια ορμή. Οι παραπάνω διαδιακίες περιγράφονται στα σχήματα της εικόνας 16.1.

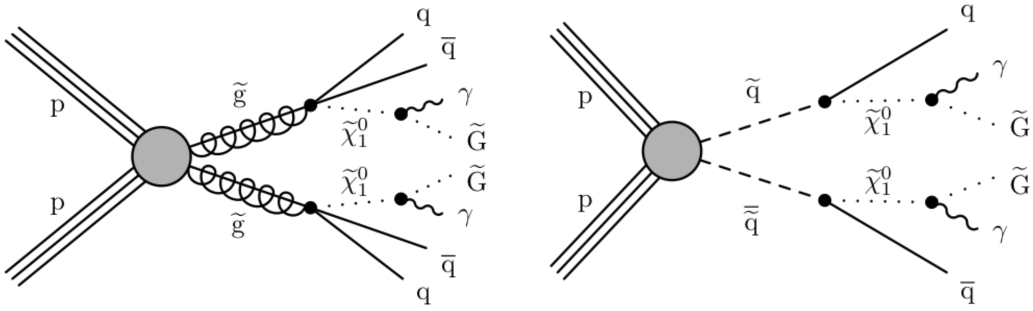


Figure 16.1: Διαγράμματα που δείχνουν τις διαδικασίες T5gg (αριστερά)και T6gg (δεξιά).

Γεγονότα αποτελούμενα από δυο φωτόνια και p_T^{miss} είναι δυνατόν να παραχθούν από διάφορες γνωστές διαδικασίες στα πλαίσια του Καθιερωμένου Προτύπου, οι οποίες περιλαμβάνουν την απευθείας παραγωγή δυο φωτονίων μέσω ακτινοβολίας αρχικής κατάστασης (Initial State Radiation- ISR), καθώς και πολυαδρονικά γεγονότα (multijet events) με πιθανή παράλληλη παραγωγή φωτονίων. Το κοινό χαρακτηριστικό αυτών των διαδικασιών είναι πως ενώ περιέχουν φωτόνια στην τελική κατάσταση η ελλείπουσα εγκάρσια ορμή του γεγονότος δεν είναι εγγενής αφού η ύπαρξη της οφείλεται σε εσφαλμένη μέτρηση. Επιπλέον, υπάρχει η πιθανότητα εσφαλμένης μέτρησης της αδρονικής δραστηριότητάς του γεγονότος στην περίπτωση που αδρονικοί πίδακες (jets) με πλούσιο ηλεκτρομαγνητικό φορτίο ταυτοποιηθούν εσφαλμένα ως φωτόνια. Μικρότερης κλίμακας υπόβαθρα μπορούν να προέλθουν από γεγονότα με ύπαρξη εγγενούς ελλείπουσας ορμής, η οποία προέρχεται κυρίως από το νετρίνο που δεν αλληλεπιδρά με τον ανιχνευτή. Τέτοιες διαδικασίες περιλαμβάνουν γεγονότα κυρίως από διαδικασίες όπως $\mathrm{W}\gamma$ και $\mathrm{W}+\mathrm{jets}$ όπου το W μποζόνιο διασπάται λεπτονικά και το ηλεκτρόνιο ανακατασκευάζεται εσφαλμένα ως φωτόνιο. Το νετρίνο του γεγονότος παραμένει μη ανιχνεύσιμο και οδηγεί σε εγγενή ελλείπουσα εγκάρσια ορμή. Τέλος υπάρχει μια μικρή συνεισφορά γεγονότων υποβάθρου από $Z\gamma\gamma \to \nu\nu\gamma\gamma$.

16.2 Δεδομένα και προσημειώσεις **Monte Carlo**

Τα πειραματικά δεδομένα συλλέχθηκαν με τον ανιχνευτή CMS ο οποίος περιγράφεται στη παράγραφο 14.3 και αντιστοιχούν σε ολοκληρωμένη φωτεινότητα 35.9 fb^{-1}.

Προσομοιώσεις Monte Carlo των διαδικασιών του σήματος και του υποβάθρου χρησιμοποιήθηκαν για τον καθορισμό της αποδοτικότητάς του σήματος καθώς και για τον προσδιορισμό ορισμένων από τα μικρότερα υπόβαθρα. Η γεννήτρια γεγονότων πρώτης τάξης MadGraph5_$aMC@NLO$ χρησιμοποιήθηκε για την προσομοίωση του σήματος, το οποίο δημιουργήθηκε με δύο gluinos ή δυο squarks και μέχρι δυο επιπλέον παρτόνια (partons) στην μήτρα υπολογισμού των στοιχείων (ΜΕ= Matrix Element). Ο καταιγισμός παρτονίων, η αδρονοποίηση, οι αλληλεπιδράσεις πολλαπλών παρτονίων και το βασικό γεγονός περιγράφηκαν από την γεννήτρια γεγονότων PYTHIA. Οι συναρτήσεις κατανομής παρτονίων λήφθηκαν από το πακέτο NNNPDF3.5. Για τις διεργασίες υποβάθρου, η απόκριση του ανιχνευτή προσομοιώθηκε με τη χρήση του λογισμικού GEANT4, ενώ η γρήγορη προσομοίωση (FastSim) για το CMS χρησιμοποιήθηκε για την παραγωγή των γεγονότων του σήματος.

16.3 Ανακατασκευή γεγονότων

Τα δεδομένα που χρησιμοποιήθηκαν σε αυτή την ανάλυση επιλέχθηκαν μέσα από ένα σύστημα σκανδαλισμού δυο φωτονίων (diphoton trigger), το οποίο απαιτεί το φωτόνιο μεγαλύτερης ορμής (κύριο) να υπερβαίνει τα 30 GeV ενώ το δευτερεύον φωτόνιο να έχει ορμή μεγαλύτερη από 18 GeV. Επίσης, τα δύο αυτά φωτόνια πρέπει να έχουν αναλοίωτη μάζα μεγαλύτερη των 95 GeV ($\mathrm{M}_{\gamma\gamma} > 95\,\mathrm{GeV}$). Τα φωτόνια που επιλέγονται οφείλουν να πληρούν συγκεκριμένα κριτήρια απομόνωσης καθώς και μορφολογικά κριτήρια. Επίσης συλλέχθηκαν και γεγονότα ηλεκτρονίων εφαρμόζοντάς τα ίδια κριτήρια με αυτά των φωτονίων αλλά με την απαίτηση να υπάρχουν τουλάχιστον δυο εν-αποθέσεις ενέργειας στον ανιχνευτή ψηφίδων. Επιπρόσθετα στις παραπάνω συλλογές δεδομένων, μια τρίτη, ορθογώνια συλλογή κατασκευάστηκε η οποία αποτελείται κυρίως από εσφαλμένα κατασκευασμένα φωτόνια (fake photons). Για την ανακατασκευή αυτών των αντικειμένων χρησιμοποιήθηκαν τα ίδια κινηματικά κριτήρια με τη συλλογή των φωτονίων όμως με τη διαφορά ότι η συλλογή των fakes δεν πληροί τα ίδια μορφολογικά κριτήρια με αυτή των φωτονίων. Με αυτό τον τρόπο καταλήγουμε σε μια συλλογή που μοιάζει αρκετά κινηματικά με το σήμα και ταυτόχρονα είναι αμερόληπτη, δίνοντας έτσι την δυνατότητα αυτή η συλλογή να χρησιμοποιηθεί για την εκτίμηση του υποβάθρου.

Όλες οι παραπάνω συλλογές περιέχουν αποκλειστικά αποκλειόμενα γεγονότα. Επιπλέον γεγονότα που περιέχουν λεπτόνια δεν λαμβάνονται υπόψιν αφού τα μοντέλα ενδιαφέροντος δεν περιέχουν λεπτόνια στην τελική κατάσταση.

16.4 Περιοχές σήματος και έλεγχου

Τα γεγονότα από τις τρεις αμοιβαία αποκλειόμενες συλλογές ταξινομήθηκαν ανάλογα με το είδος των υψηλότερων ενεργειακά ηλεκτρομαγνητικών αντικειμένων με βάση την εγκάρσια ορμή τους. Έτσι καταλήγουμε σε κατηγορίες με δυο φωτόνια ($\gamma\gamma$), με δυο fakes (ff) και με κατηγορίες που έχουν ένα ανακατασκευασμένο φωτόνιο και ένα ηλεκτρόνιο ($e\gamma$). Επιπλέον, η αναλλοίωτη μάζα των δυο ηλεκτρομαγνητικών αντικειμένων επιλέχθηκε να είναι μεγαλύτερη των 110 GeV. Η περιοχή του σήματος ορίζεται από γεγονότα της κατηγορίας $\gamma\gamma$ με $p_T^{\mathrm{miss}} \geq 100\,\mathrm{GeV}$ και χωρίζεται σε έξι τμήματα: $100 \leq p_T^{\mathrm{miss}} < 115\,\mathrm{GeV}$, $115 \leq p_T^{\mathrm{miss}} < 130\,\mathrm{GeV}$, $130 \leq p_T^{\mathrm{miss}} < 150\,\mathrm{GeV}$, $150 \leq p_T^{\mathrm{miss}} < 185\,\mathrm{GeV}$, $185 \leq p_T^{\mathrm{miss}} < 250\,\mathrm{GeV}$, και $p_T^{\mathrm{miss}} \geq 250\,\mathrm{GeV}$. Τα παραπάνω τμήματα της εγκάρσιας ελλείπουσας ορμής επιλέχθηκαν με τέτοιο τρόπο ώστε να υπάρχει ένας επαρκής αριθμός δεδομένων σε κάθε ένα από αυτά. Η περιοχή με δυο φωτόνια και $p_T^{\mathrm{miss}} < 100\,\mathrm{GeV}$ χρησιμοποιείτε ως περιοχή ελέγχου και είναι παράλληλα και ορθογώνια με τη περιοχή του σήματος που μόλις ορίστηκε. Επίσης, οι περιοχές $e\gamma$ και ff χρησιμοποιούνται σαν περιοχές ελέγχου εκτίμησης του υποβάθρου.

16.5 Εκτίμηση του υποβάθρου

Το σημαντικότερο υπόβαθρο για την συγκεκριμένη ανάλυση προέρχεται από παραγωγή πολλαπλών πιδάκων αδρονίων (multijet events) τα οποία προέρχονται από διαδικασίες κβαντικής χρωμοδυναμικής (QCD). Το κύριο χαρακτηριστικό αυτού του υποβάθρου είναι η απουσία εγγενούς p_T^{miss} η οποία προκύπτει από τις εσφαλμένες μετρήσεις της αδρονικής δραστηριότητας του γεγονότος. Αυτή η συνεισφορά μοντελοποιήθηκε πλήρως από τα δεδομένα (data driven) μέσω της χρήσης του δείγματος ελέγχου fakes (ff). Το δείγμα αυτό αποτελείται κυρίως από γεγονότα χωρίς εγγενή p_T^{miss} και συνεπώς είναι το πλέον κατάλληλο για να μοντελοποιήσει την παρατηρούμενη p_T^{miss} που προέρχεται από την QCD.

Η μέθοδος που χρησιμοποιήθηκε για την εκτίμηση του υποβάθρου είναι η λεγόμενη "ratio method" και βασίζεται στην παρατήρηση ότι ο λόγος των δύο φωτονίων $\gamma\gamma$ και των δυο fakes ff δεν έχει σημαντική εξάρτηση από την p_T^{miss} και για αυτό τον λόγο αυτό μπορεί να μοντελοποιηθεί με μια απλή συνάρτηση. Η μοντελοποίηση γίνεται αρχικά στην περιοχή ελέγχου ($p_T^{\text{miss}} < 100$ GeV) και έπειτα η συνάρτηση παρεκτείνεται στην περιοχή σήματος ($p_T^{\text{miss}} \geq 100$ GeV) παίρνοντας έτσι την εκτίμηση του υποβάθρου για κάθε τμήμα σήματος. Στην περίπτωση που το ff δείγμα ελέγχου είχε την ίδια σύσταση με το υποψήφιο δείγμα ελέγχου $\gamma\gamma$ δεν αναμένεται μεγάλη εξάρτηση από την p_T^{miss} και έτσι ο λόγος των δυο δειγμάτων συναρτήσει της p_T^{miss} αναμένεται να είναι σταθερός. Όμως, το δείγμα ff έχει χαμηλότερη καθαρότητα σε σχέση με το δείγμα ελέγχου $\gamma\gamma$ και επομένως η συνάρτηση που περιγράφει αυτή την εξάρτηση είναι μια εκθετική συνάρτηση της μορφής $p_0 e^{-p_1 x}$. Έτσι στην παραπάνω συνάρτηση γίνεται μια προσαρμογή (fit) στο λόγο των δυο κατανομών στην περιοχή ελέγχου. Τελικά, μετά την παρέκταση της προσαρμογής στην περιοχή σήματος, ο προβλεπόμενος αριθμός του υποβάθρου από γεγονότα QCD δίνεται από την σχέση:

$$\mathrm{N}^{\mathrm{i}}_{\mathrm{QCD}} = \mathrm{g}^{\mathrm{i}}_{\mathrm{ave}} \mathrm{N}^{\mathrm{i}}_{\mathrm{ff}} \tag{16.1}$$

όπου $\mathrm{N}^{\mathrm{i}}_{\mathrm{ff}}$ είναι ο αριθμός των παρατηρούμενων ff γεγονότων και το $\mathrm{g}^{\mathrm{i}}_{\mathrm{ave}}$ είναι η μέση τιμή της προσαρμογής του i τμήματος.

Για να ελεγχθεί η εγκυρότητα της μεθόδου, έγινε ένας τεστ αξιοπιστίας της με μια άλλη μέθοδο "data driven". Σε αυτή την μέθοδο το δείγμα ελέγχου ff διορθώνεται ώστε να έχει την ίδια σύσταση με το $\gamma\gamma$ δείγμα. Έπειτα ο λόγος των δυο κατανομών σχεδιάζεται και πραγματοποιείται μια γραμμική προσαρμογή. Για την εκτίμηση του υποβάθρου χρησιμοποιήθηκε η σχέση 16.1. Οι εκτιμήσεις σε κάθε τμήμα του σήματος συμφωνούν μέσα στα πλαίσια της αβεβαιότητας της κάθε μεθόδου. Αυτή η μέθοδος χρησιμοποιήθηκε για να θέσει μια συστηματική αβεβαιότητα στην κύρια μέθοδο.

Η άλλη συνεισφορά του υποβάθρου προέρχεται από το ηλεκτρασθενές υπόβαθρο (EWK) που περιλαμβάνει κυρίως $W\gamma$ γεγονότα όπου το W διασπάται κυρίως σε ένα ηλεκτρόνιο και ένα νετρίνο με το ηλεκτρόνιο να ανακατασκευάζεται λαθεμένα σαν φωτόνιο. Λόγω της παρουσίας του νετρίνο υπάρχει εγγενής p_T^{miss} στην τελική κατάσταση. Για να προσδιορίσουμε την εκτίμηση αυτού του υποβάθρου στην περιοχή του σήματος, πρέπει να εκτιμηθεί ο βαθμός κατά τον ένα ηλεκτρόνιο ανακατασκευάζεται λαθεμένα ως φωτόνιο ($\mathrm{f}_{\mathrm{e}\to\gamma}$). Αυτός προσδιορίζεται συγκρίνοντας της κορυφή της μάζας από μια συλλογή δυο ηλεκτρονίων (ee) με την κορυφή της μάζας από την $e\gamma$ συλλογή. Έπειτα εφαρμόζεται μια προσαρμογή μέγιστης πιθανοφάνειας στην κορυφή της μάζας των δειγμάτων για την υπόθεση ύπαρξης σήματος και υποβάθρου. Ο βαθμός λαθεμένης ανακατασκευής δίνεται από την σχέση $\mathrm{f}_{\mathrm{e}\to\gamma} = \mathrm{N}_{\mathrm{e}\gamma}/(2\mathrm{N}_{\mathrm{ee}} + \mathrm{N}_{\mathrm{e}\gamma})$, όπου $N_{e\gamma}$ και N_{ee} είναι ο αριθμός των γεγονότων των δυο δειγμάτων που λαμβάνεται από την προσαρμογή. Η τελική συνεισφορά του τελικού EWK υποβάθρου προσδιορίζεται διορθώνοντας των αριθμό των γεγονότων $e\gamma$ του δείγματος ελέγχου με ένα παράγοντα $\mathrm{f}_{\mathrm{e}\gamma\to\gamma\gamma} = \mathrm{f}_{\mathrm{e}\to\gamma}/(1 - \mathrm{f}_{\mathrm{e}\to\gamma}) = (2.63 \pm 0.79)\%$.

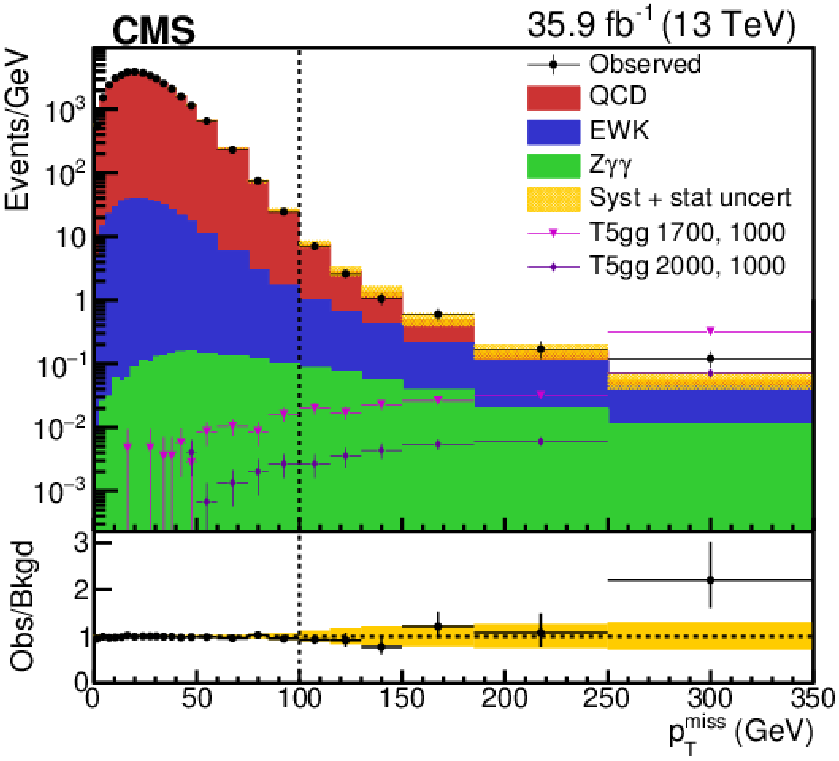
Επιπλέον υπάρχει μια μικρή συνεισφορά γεγονότων από το $Z\gamma\gamma$ διεργασίες. Αυτές εκτιμώνται εξ ολοκλήρου από προσομοιώσεις προσθέτοντας μια αβεβαιότητα 50% για να καλύψει τυχόν σφάλματα μοντελοποίησης.

16.6 Πηγές συστηματικής αβεβαιότητας

Συστηματικές αβεβαιότητες που αφορούν την ανάλυση προκύπτουν από τις εκτιμήσεις του εκάστοτε υποβάθρου. Άλλες πηγές συστηματικής αβεβαιότητας προκύπτουν από τον προσδιορισμό της αποδοτικότητας του σήματος καθώς και από τον προσδιορισμό της ολοκληρωμένης φωτεινότητας (2.5%). Η μεγαλύτερη συνεισφορά προκύπτει από το QCD υπόβαθρο το οποίο περιλαμβάνει την συνεισφορά της στατιστικής αβεβαιότητας του δείγματος ελέγχου (ff) (7-79%) και την αβεβαιότητα της μεθόδου εκτίμησης. Η τελευταία περιλαμβάνει την αβεβαιότητα από την ίδια την προσαρμογή (2-5%) και την αβεβαιότητα από την μέθοδο αξιολόγησης (10-83%) που περιγράφτηκε στην παράγραφο 16.5. Η συστηματική αβεβαιότητα που αντιστοιχεί στο EWK υπόβαθρο περιλαμβάνει την στατιστική αβεβαιότητα του δείγματος ελέγχου $e\gamma$ και μια 30% αβεβαιότητα που αντιστοιχεί στην μέθοδο εκτίμησης του υποβάθρου. Επίσης, λαμβάνονται υπόψιν μικρότερης κλίμακας αβεβαιότητες που αφορούν το μέγεθος των προσομοιώσεων του σήματος (2-45%) καθώς και αβεβαιότητες που αφορούν τις συναρτήσεων κατανομής παρτονίων (19-35%), την γνώση της ενεργειακής κλίμακας των jets (1-30%) καθώς και αβεβαιότητες που αφορούν την ταυτοποίηση των φωτονίων και τις αποδόσεις κατασκευής τους (2.5%).

16.7 Αποτελέσματα

Για την έρευνα για υπερσυμμετρία σε τελικές καταστάσεις δυο φωτονίων και ελλείπουσας εγκάρσιας ορμής, αναπτύχθηκαν μέθοδοι που εκμεταλλεύονται πλήρως τα δεδομένα για να εκτιμήσουν τις δυο βασικές συνεισφορές του υποβάθρου. Οι μέθοδοι της ανάλυσης βελτιστοποιήθηκαν στην περιοχή ελέγχου ($p_T^{\text{miss}} < 100$ GeV) και αφού πραγματοποιήθηκαν όλοι οι απαιτούμενοι έλεγχοι της αξιοπιστίας των μεθόδων, οι εκτιμώμενες τιμές των τμημάτων της περιοχή σήματος συγκρίθηκαν με τις παρατηρούμενες τιμές. Αυτό φαίνεται στο σχήμα 16.2 όπου παρουσιάζεται η κατανομή της p_T^{miss} για την περιοχή ελέγχου και σήματος. Επίσης στο ίδιο διάγραμμα παρατηρούμε δυο περιπτώσεις σήματος με μάζες gluino 1700 και 2000 GeV αντίστοιχα. Όπως φαίνεται στο τελευταίο τμήμα της κατανομής υπάρχει περίσσεια πειραματικών γεγονότων σε σχέση με τα προβλεπόμενα. Αυτή η περίσσεια γεγονότων έχει σημαντικότητα 2.4σ λαμβάνοντας υπόψιν όλα τα τμήματα του σήματος. Στα επόμενα τμήματα ο αριθμός των γεγονότων στην περιοχή σήματος συμφωνεί με την εκτίμηση του υποβάθρου μέσα στις εκτιμώμενες αβεβαιότητες.

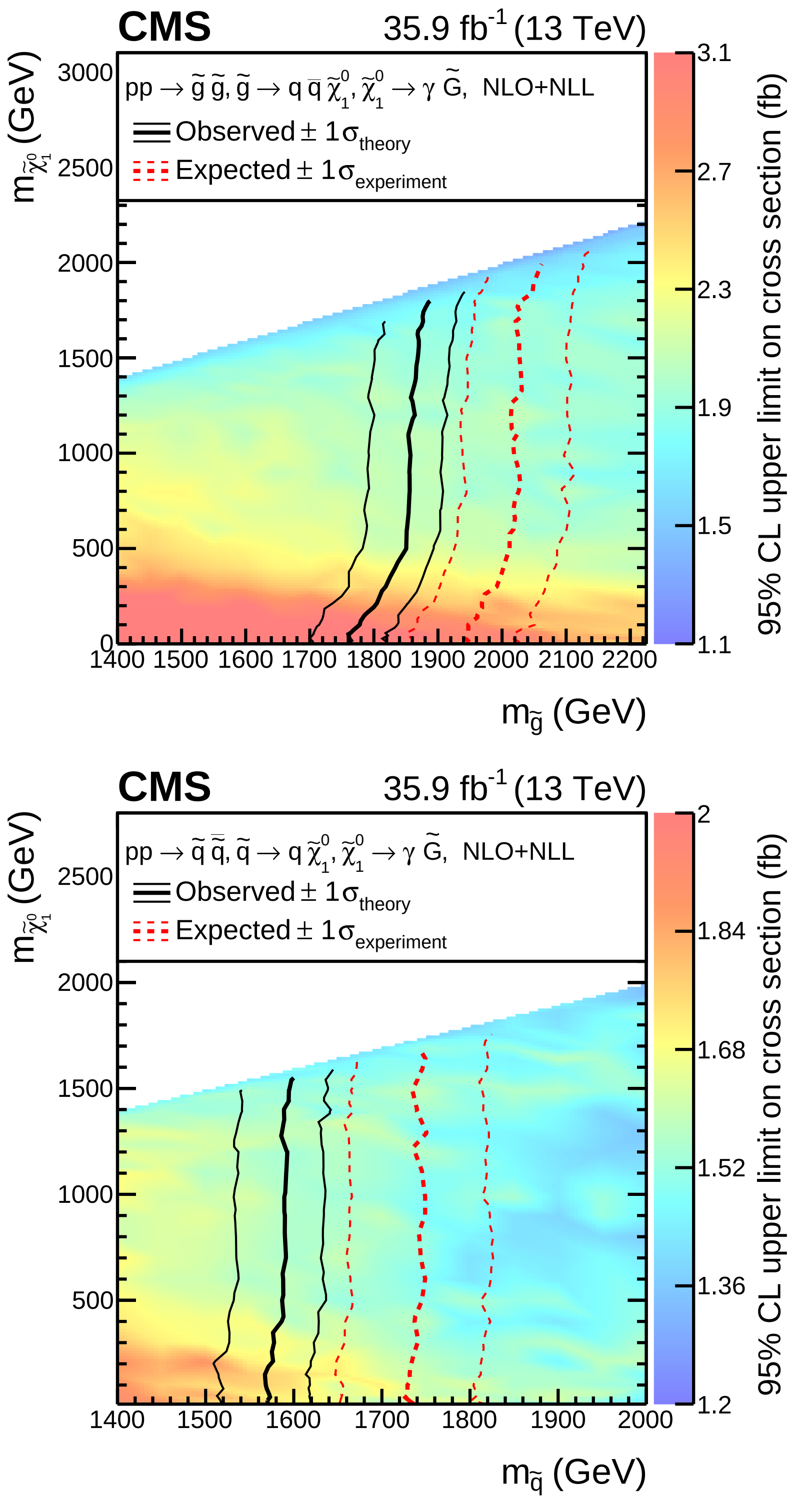


Σχήμα 16.2: Ολική εκτίμηση υποβάθρου στην περιοχή ελέγχου και στην υποψήφια περιοχή συναρτήσει της εγκάρσιας ελλείπουσας ενέργειας, μαζί με τα παρατηρηθέντα πειραματικά δεδομένα. Παράλληλα φαίνονται τα αναμενόμενα γεγονότα για δυο διαφορετικές περιπτώσεις σήματος.

16.8 Ανώτατα όρια και ερμηνείες

Τα αποτελέσματα χρησιμοποιήθηκαν για να υπολογιστούν τα 95% επίπεδα εμπιστοσύνης (CL) για τα ανώτερα όρια των ενεργών διατομών παραγωγής ενός ζεύγους gluino και squark. Χρησιμοποιήθηκε η μέθοδος frequentist CLs η οποία βασίζεται σε ένα λογαριθμικής πιθανοφάνειας στατιστικό έλεγχο, που συγκρίνει την πιθανότητα της ύπαρξης μόνο του Καθιερωμένου Προτύπου με την πιθανότητα της επιπλέον παρουσίας ενός σήματος στην συνεισφορά των διαδικασιών του Καθιερωμένου Προτύπου. Η συνάρτηση πιθανοφάνειας κατασκευάζεται από τις κατανομές εγκάρσιας ελλείπουσας ορμής του υποβάθρου και του σήματος στις έξι περιοχές ανίχνευσης p_T^{miss}. Οι συστηματικές αβεβαιότητες που περιγράφτηκαν στην παράγραφο 16.7 περιλαμβάνονται στο στατιστικό έλεγχο ως ελεύθερες παράμετροι με λογαριθμική-κανονική κατανομή πιθανότητας.

Τα απλουστευμένα μοντέλα που χρησιμοποιήθηκαν στην ανάλυση περιγράφτηκαν στην παράγραφο 16.1. Στην εικόνα 16.3 φαίνονται τα αναμενόμενα και τα παρατηρούμενα όρια απόρριψης ως προς τις μάζες των gluino και των squark . Για τυπικές μάζες νετραλίνο, αναμένουμε να αποκλείσουμε μάζες gluino από 2.02 TeV και για μάζες squark αποκλείονται μάζες πάνω από 1.74 TeV. Τα παρατηρούμενα όρια είναι 1.86 TeV για μάζες gluino και 1.59 TeV για μάζες μάζες squark. . Τα συγκεκριμένα όρια αυξάνουν την ευαισθησία της μέτρησης κατά 300 GeV σε σύγκριση με έρευνες που πραγματοποιήθηκαν στο CMS στο παρελθόν με δεδομένα χαμηλότερης ολοκληρωμένης φωτεινότητας.



Σχήμα 16.3: κακκ

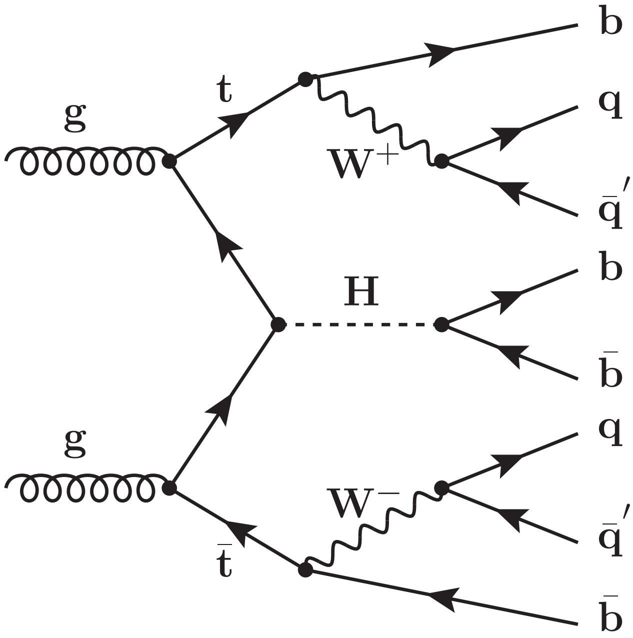
16.9 Σύνοψη

Παρουσιάστηκαν τα αποτελέσματα μιας έρευνας για υπερσυμμετρία που βασίζεται σε μια θεωρεία διαμεσολαβητών βαθμίδας. Χρησιμοποιήθηκαν δεδομένα από συγκρούσεις πρωτονίων που συλλέχθηκαν από το ανιχνευτή CMS το 2016 και αντιστοιχούν σε ολοκληρωμένη φωτεινότητα 35.9 fb^{-1} κέντρου μάζας $\sqrt{s} = 13$ TeV. Για την ανάλυση λήφθηκαν υπόψιν τελικές καταστάσεις δυο φωτονίων και μεγάλης ελλείπουσας εγκάρσιας ορμής. Μια περίσσεια γεγονότων σημαντικότητας 2.4 σ παρατηρήθηκε στα δεδομένα και προσδιορίστηκαν τα όρια στις υπερσυμμετρικές μάζες δυο απλουστευμένων μοντέλων χρησιμοποιώντας τεχνικές data driven για την εκτίμηση του υποβάθρου. Τα παρατηρούμενα όρια αυξάνουν την ευαισθησία της ανάλυσης κατά 210 GeV ενώ τα αναμενόμενα όρια κατά 300 GeV σε σχέση με προηγούμενες έρευνες.

Κεφάλαιο 17

Έρευνα για την ταυτόχρονη παραγωγή ενός συμβατού με το καθιερωμένο πρότυπο μποζονίου **Higgs** με ένα ζεύγος **top-antitop quark** και την μετέπειτα διάσπασή του σε ένα ζεύγος **b quark** στην πλήρως αδρονική τελική κατάσταση χρησιμοποιώντας πίδακες μεγάλης ακτίνας

Σε αυτό το κεφάλαιο παρουσιάζεται μια έρευνα με δεδομένα που συλλέχθηκαν το 2016 από το πείραμα CMS και αντιστοιχούν σε ολοκληρωμένη φωτεινότητα 35.9 fb^{-1}. Η ανάλυση αφορά στην ταυτόχρονη παραγωγή ενός μποζονίου Higgs με ένα ζεύγος top-antitop quark ($\mathrm{t\bar{t}H}$ production) και την μετέπειτα διάσπαση του σε ένα ζεύγος b-quark ($\mathrm{t\bar{t}H}(H \to (b\bar{b}))$. Στην συγκεκριμένη ανάλυση τα W μποζόνια από τις διασπάσεις των δυο top quarks διασπώνται σε light quarks, καταλήγοντας σε τελικές καταστάσεις με τουλάχιστον οκτώ quarks, τέσσερα από τα οποία είναι b quark. Αυτή η τελική κατάσταση ονομάζεται πλήρως αδρονική με πειραματική υπογραφή jets τα οποία παράγονται σε μεγάλες ακτίνες σε σχέση με τη δέσμη. Αυτά τα jets αναμένεται να έχουν σχετικά μεγάλη εγκάρσια ορμή. Η διαδικασία αυτή περιγράφεται από το διάγραμμα του σχήματος 17.1. Όταν η ορμή υπερβαίνει ένα ενεργειακό κατώφλι, τότε τα προϊόντα της αντίδρασης παράγονται αρκετά ευθυγραμμισμένα με αποτέλεσμα να ανακατασκευάζονται ως μεγάλης-ακτίνας jets ("boosted jets"). Συλλέχθηκαν λοιπόν δεδομένα που περιλαμβάνουν τέτοια jets με τουλάχιστον ένα από αυτά να είναι ανακατασκευασμένο ως υποψήφιο Higgs jet.Τα μεγάλης-ακτίνας jets περιέχουν όλη την πληροφορία της διάσπασης του μποζονίου Higgs και των top quark η οποία μπορεί να χρησιμοποιηθεί για την ανακατασκευή υποψήφιων boosted Higgs και boosted top jets. Για τον λόγο αυτό αναπτύχθηκαν τεχνικές πολλών μεταβλητών (MultiVariate Techniques (MVA)) ικανές να ταυτοποιήσουν Higgs και top jets με μεγάλη ακρίβεια. Επιπλέον οι τεχνικές που αναπτύχθηκαν για την εκτίμηση του υποβάθρου παρουσιάζονται στην συνέχεια.



Σχήμα 17.1: διάγραμμα της διαδικασίας $ttH(H \to bb)$ στην πλήρως αδρονική διάσπαση

17.1 Καταστάσεις Υπόβαθρου

Πολλές διεργασίες του καθιερωμένου προτύπου συνεισφέρουν στην τελική κατάσταση του πλήρους αδρονικού $\mathrm{t\bar{t}H}$ σήματος. Το κυρίαρχο υπόβαθρο αυτών των καταστάσεων προέρχεται από την παραγωγή πολλών jets (QCD multijet production), όπου είναι πιθανόν jets από τέτοιες διεργασίες να μιμούνται την τοπολογική σύσταση ενός Higgs και top boosted jet.Για την εκτίμηση αυτού του υποβάθρου αναπτύχθηκαν τεχνικές που βασίζονται μόνο στα δεδομένα (data driven techniques).

Το δεύτερο σε σημαντικότητα υπόβαθρο προέρχεται από γεγονότα των διαδικασιών της διάσπασης δυο ζευγών top quark ($\mathrm{t\bar{t}}$), όπου τέτοια γεγονότα μοιάζουν κινηματικά με το σήμα. Αυτό το υπόβαθρο εκτιμήθηκε με την βοήθεια προσομοιώσεων, όμως η αβεβαιότητα της εκτιμώμενης συνεισφοράς περιορίστηκε από μια περιοχή ελέγχου στα δεδομένα.

Επίσης υπάρχουν διάφορες διεργασίες του καθιερωμένου προτύπου που συνεισφέρουν στο σήμα. Αυτές περιλαμβάνουν διεργασίες όπως την παραγωγή ενός top quark (Single top production), γεγονότα $W+$ jets, $Z+$ jets, $\mathrm{t\bar{t}}$ + jets, και γεγονότα παραγωγής δυο ασθενών μποζονίων WW, WZ και ZZ. Η συνεισφορά τους στο σήμα είναι μικρή και για αυτό αυτές οι διεργασίες έχουν εκτιμηθεί με προσομοιώσεις.

Τέλος υπάρχει μια συνεισφορά γεγονότων προερχόμενα από την ταυτόχρονη παραγωγή ενός ζεύγους top-antitop ($\mathrm{t\bar{t}}$) σε συνδυασμό με ένα μποζόνιο Z ($\mathrm{t\bar{t}}Z$). Αυτή η διεργασία έχει ακριβώς την ίδια τελική υπογραφή με το σήμα και για αυτό τον λόγο είναι ένα υπόβαθρό που δεν μπορεί να μειωθεί. Αυτό το υπόβαθρο εκτιμάται εξολοκλήρου από την προσομοίωση.

17.2 Δεδομένα και προσομοιώσεις **Monte Carlo**

Τα πειραματικά δεδομένα συλλέχθηκαν με τον ανιχνευτή CMS ο οποίος περιγράφεται στη παράγραφο 14.3 και αντιστοιχούν σε ολοκληρωμένη φωτεινότητα 35.9 fb^{-1}.

Προσομοιώσεις Monte Carlo των διαδικασιών του σήματος και του υποβάθρου χρησιμοποιήθηκαν για τον καθορισμό της αποδοτικότητας του σήματος, τον ποιοτικό έλεγχο των μεθόδων που αναπτύχθηκαν καθώς και για τον προσδιορισμό ορισμένων από τα μικρότερα υπόβαθρα. Ανάλογα με την φυσική διαδικασία διάφοροι γεννήτορες δεδομένων χρησιμοποιήθηκαν όπως η PYTHIA , POWHEG, MADGRAPH5_*aMC@NLO*. Οι συναρτήσεις κατανομής παρτονίων λήφθηκαν από το πακέτο NNNPDF3.5. Για τις διεργασίες υποβάθρου, η απόκριση του ανιχνευτή προσομοιώθηκε με τη χρήση του λογισμικού GEANT4, ενώ η γρήγορη προσομοίωση (FastSim) για το CMS χρησιμοποιήθηκε για την παραγωγή των γεγονότων του σήματος. Συγκεκριμένα το σήμα προσομοιώθηκε χρησιμοποιώντας το MADGRAPH5_*aMC@NLO* όπου το μποζόνιο Higgs έχει μάζα $\mathrm{m_H} = 125$ GeV και το top quark $\mathrm{m_t} = 175$ GeV.

17.3 Ανακατασκευή γεγονότων

Τα δεδομένα που χρησιμοποιήθηκαν σε αυτή την ανάλυση επιλέχθηκαν μέσα από ένα σύστημα σκανδαλισμού ο οποίος απαιτεί την παρουσία boosted jets ακτίνας $\mathrm{R} = 0.8$ και μάζας μεγαλύτερης των 50 GeV. Επίσης τα γεγονότα που συλλέχτηκα απαιτούν το άθροισμα της εγκάρσιας ορμής των ανακατασκευασμένων jets (H_T) να είναι μεγαλύτερο των 700 GeV. Όμως σε αυτή την ανάλυση υπάρχει συνεισφορά γεγονότων όχι μόνο από jets μεγάλης ακτίνας R αλλά και από jets μικρότερης ακτίνας ($\mathrm{R} = 0.4$). Έτσι η αποδοτικότητα του σκανδαλιστή μελετήθηκε συναρτήσει του αθροίσματος των εγκάρσιων ορμών όλων των ανακατασκευασμένων jets του γεγονότος (S_T). Συγκεκριμένα βρέθηκε πως για γεγονότα με $S_\mathrm{T} > 900$ GeV ο σκανδαλιστής είναι πλήρως αποδοτικός χωρίς να αποκλείονται ενδιαφέροντα γεγονότα σήματος.

Σε αυτή την ανάλυση κατασκευάζονται δυο συλλογές λεπτονίων (ηλεκτρονίων και μιονίων) για δυο κύριους λόγους. Αρχικά, στα δεδομένα που επιλέγονται για ανάλυση αποκλείονται τα λεπτόνια ώστε να διασφαλιστεί ένα δείγμα δεδομένων υψηλής καθαρότητας. Επιπλέον, τα jets μπορούν να περιέχουν ηλεκτρομαγνητική σύσταση και έτσι υπάρχει η πιθανότητα ένα λεπτόνιο να ανακατασκευάζεται λαθεμένα ως jet. Η συλλογή αυτή λοιπόν χρησιμοποιήθηκε και για να απορρίψουμε τέτοια jets που μπορεί να περάσουν τα κριτήρια επιλογής.

Η κύρια συλλογή αυτής της ανάλυσης είναι η συλλογή των jets. Ανακατασκευάστηκαν λοιπόν, δυο ειδών jets τα οποία χρησιμοποιούν αντικείμενα προερχόμενα από τον particle flow αλγόριθμο (PF candidates) και κατατάσσονται ανάλογα με τη ακτίνα απόστασης R. Πιο συγκεκριμένα, η κύρια συλλογή αποτελείται από PF jets ακτίνας $\mathrm{R} = 0.8$ και για αυτό τον λόγο ονομάζεται Ak8 jet συλλογή. Η επιλογή της ακτίνας απόστασης είναι τέτοια ώστε να περιέχει όλα τα προϊόντα της διάσπασης του μποζονίου Higgs και του top quark. Για να μειωθούν τα jets που έχουν προέρθει από δευτερεύουσες συγκρούσεις εφαρμόζονται συγκεκριμένες τεχνικές τόσο στα jets όσο και στα μικρότερα jets (subjets) που βρίσκονται μέσα τους. Επιπλέον, τα επιλεγμένα jets πρέπει να έχουν εγκάρσια ορμή μεγαλύτερη των 200 GeV και να βρίσκονται μέσα στην περιοχή ευκρίνειας του tracker. Επίσης, η αναλλοίωτη μάζα των δυο subjets, γνωστή ως $\mathrm{m_{SD}}$ είναι μεγαλύτερη από 50 GeV.

Τα μικρής ακτίνας jets χρησιμοποιούν PF jets που έχουν ανακατασκευαστεί με τον αλγόριθμο anti$-k_T$ με παράμετρο απόστασης $\mathrm{R} = 0.4$. Η συλλογή των Ak4 jets απαιτεί τα jets να έχουν εγκάρσια ορμή μεγαλύτερη από 30 GeV και να βρίσκονται μέσα στην περιοχή ανίχνευσης του tracker.

Για την ταυτοποίηση των b jets και των b subjets χρησιμοποιήθηκε ο αλγόριθμος CSVv2. Επίσης στην συλλογή των jets εφαρμόστηκαν όλες οι απαραίτητες ενεργειακές και βαθμονομικές διορθώσεις. Οι δυο συλλογές των jets έχουν κατασκευαστεί έτσι ώστε να αποτελούνται από αμοιβαία αποκλειόμενα γεγονότα.

Η τελική επιλογή γεγονότων για ανάλυση περιλαμβάνει τουλάχιστον ένα ανακατασκευασμένο Ak8-jet με εγκάρσια ορμή μεγαλύτερη των 300 GeV και μάζας $\mathrm{m_{SD}} > 50$ GeV, κανένα λεπτόνιο στην τελική κατάσταση και την συνολική S_{T} του γεγονότος να είναι μεγαλύτερη από 900 GeV.

17.4 Εκπαίδευση πολλών μεταβλητών για την ταυτοποίηση **boosted Higgs** και **boosted Top** υποψηφίων.

Μια από τις σημαντικότερες προκλήσεις αυτής της ανάλυσης είναι η επιτυχής ταυτοποίηση των υποψηφίων Higgs και top που έχουν παραχθεί με υψηλό Lorentz-boost. Για την επιτυχή ανίχνευση τους χρησιμοποιούνται τεχνικές που βασίζονται στην σύσταση των boosted jets (jet-substructure techniques). Επιπλέον ο συνεχής αναπτυσσόμενος χώρος των τεχνικών πολλαπλών μεταβλητών σε συνδυασμό με την υπολογιστική ισχύ μπορούν να συντελέσουν στην βαθύτερη κατανόηση της σύστασης των jets και κατ΄ επέκταση μπορούμε να ξεχωρίσουμε boosted jets που προέρχονται από το υπόβαθρο της QCD από boosted jets του σήματος. Για τους παραπάνω λόγους, εκπαιδεύτηκαν τρία διαφορετικά ενισχυμένα δέντρα απόφασης (Boosted Desicion Trees (BDTs)) που χρησιμοποιούν μεταβλητές της σύστασης των jets που δείχνουν μεγάλη διακριτική ικανότητα μεταξύ σήματος και υποβάθρου. Αυτές οι μεταβλητές είναι:

1. "N-subjettines": τ_1,τ_2,τ_3.
   Αυτές οι μεταβλητές δείχνουν την ενεργειακή κατανομή μέσα στο boosted jet και έχουν διακριτική ισχύ μεταξύ boosted jets με τρεις (Higgs), δυο (top) και ένα (QCD) ενεργειακούς πυρήνες.

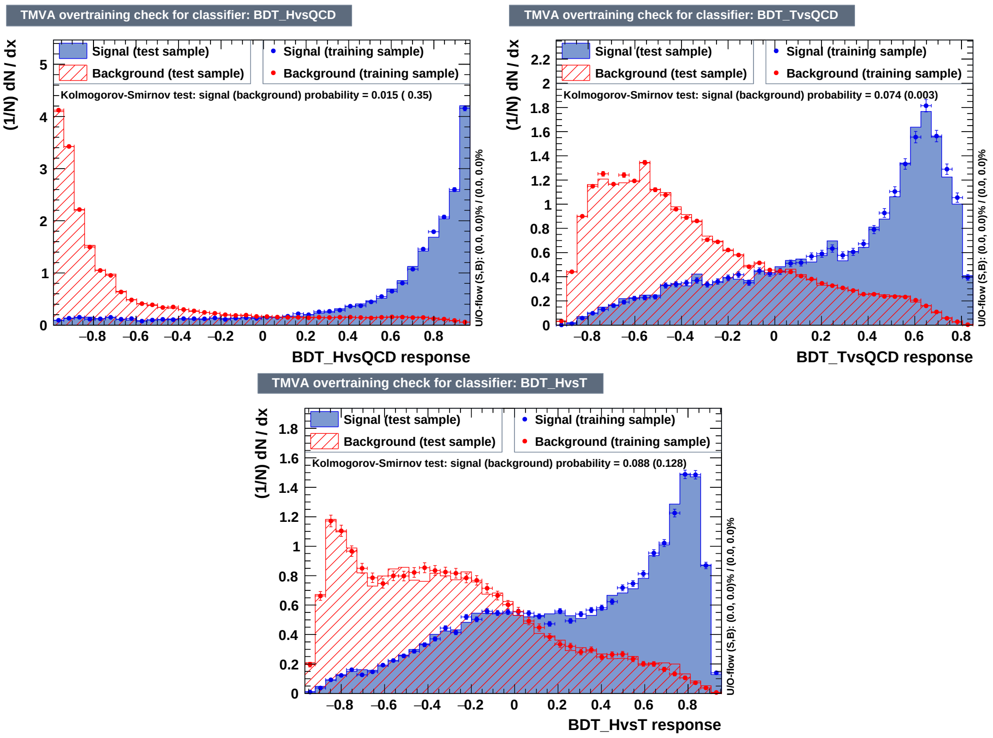
2. μάζα των δυο subjets.
   βασίζεται στην διαφορετική κινηματική συμπεριφορά των subjets

3. CSVv2 κατανομές των δυο subjets
   Η ίδια η κατανομή του αλγορίθμου των δυο subjets μπορεί να αποτελέσει μεταβλητή μεγάλης διακριτικής ικανότητας

Για την εκπαίδευση των ενισχυμένων δέντρων απόφασης χρησιμοποιήθηκε το πακέτο Toolkit for Multivariate Data Analysis (TMVA). Χρησιμοποιήθηκαν προσομειωμένα jets που περνούν την βασική επιλογή με εγκάρσια ορμή μεγαλύτερη των 300 GeV. Επιπλέον, γεγονότα με λεπτόνια αποκλείονται. Τα δείγματα σήματος χωρίστηκαν έτσι ώστε τα jets να έχουν ταυτοποιηθεί με ένα Higgs παρτόνιο ή με ένα top παρτόνιο. Επιπλέον, χρησιμοποιήθηκαν jets από προσομοιώσεις Monte Carlo του υποβάθρου QCD. Τα δύο δείγματα σήματος εκπαιδεύτηκαν ανεξάρτητα ενάντια στο ίδιο υπόβαθρο QCD και έπειτα και μεταξύ τους. Το αποτέλεσμα της εκπαίδευσης είναι τρεις διαφορετικές αποκρίσεις (HvsQ,TvsQ, HvsT) οι οποίες χρησιμοποιούνται μαζί και με άλλες απαιτήσεις για να ταυτοποιήσουμε υποψήφια Higgs και top jets. Η απόκριση για καθένα από τα BDT φαίνεται στο σχήμα 17.2.

17.5 Ταυτοποίηση **boosted Higgs** και **boosted Top** υποψηφίων.

Στην ανάλυση αυτή απαιτείται τουλάχιστον ένα ανακατασκευασμένο boosted jet το οποίο να έχει ταυτοποιηθεί ως Higgs candidate. Για αυτό τον λόγο η στρατηγική της ανάλυσης



Σχήμα 17.2: Απόκριση του BDT για την εκπαίδευσή HvsQ, TvsQ και HvsT

ξεκινά με την ταυτοποίηση των jets. Ο Higgs candidate θα είναι το jet με το μεγαλύτερο άθροισμα των σκορ των HvsQ και HvsT. Επιπλέον, απαιτείται το HvsQ $>$ 0.8 και HvsT $>$ 0.1. Τα υποψήφια jets πρέπει να έχουν εγκάρσια ορμή μεγαλύτερη από 300 GeV και μάζα, m_{SD}, μεγαλύτερη από 70 GeV. Αυτή η επιλογή αντιστοιχεί σε μια συλλογή καθαρότητας 54%. Μετά την ταυτοποίηση του Higgs candidate το jet με το μεγαλύτερο TvsQ σκορ θεωρείται ως το υποψήφιο top jet. Επιπλέον, το σκορ TvsQ πρέπει να είναι μεγαλύτερο από 0.5 έτσι ώστε να αποκλειστούν περισσότερα jets υποβάθρου. Τέλος η μάζα του top πρέπει να είναι μεταξύ 130-220 GeV. Αυτή η επιλογή αντιστοιχεί σε μια συλλογή καθαρότητας 88%.

17.6 Κατηγορίες Γεγονότων

Για να ενισχυθεί η ευαισθησία της ανάλυσης η περιοχή σήματος χωρίστηκε σε εννιά αμοιβαία αποκλειώμενες περιοχές (κατηγορίες) όπου πάντα υπάρχει ένας ταυτοποιημένος Higgs candidate.Αυτές χωρίζονται με βάση τον αριθμό των ανακατασκευασμένων boosted jets, τον αριθμό των top jets και σε κατηγορίες με λιγότερα από τρία boosted jets χωρίζονται και με βάση τον αριθμό των Ak4 jets και Ak4 b jets. Επίσης ελέγχθηκε η συμφωνία των κατανομών των μεταβλητών που χρησιμοποιήθηκαν για την κατηγοριοποίηση των γεγονότων σε δεδομένα και προσομοίωση. Έτσι πιστοποιήθηκε πως η προσομοίωση περιγράφει σωστά τα γεγονότα.

17.7 Εκτίμηση του **QCD** υποβάθρου

Το υπόβαθρο με την μεγαλύτερη συνεισφορά στην ανάλυση αποτελείται από γεγονότα με παραγωγή πολλαπλών jets, αφού υπάρχει πεπερασμένη πιθανότητα συνηθισμένα jets από την ακτινοβολία ενός παρτονίου να μιμούνται την εσωτερική τοπολογία ενός boosted jet. Η μοντελοποίηση αυτού του υποβάθρου από την προσομοίωση εισαγάγει μεγάλες αβεβαιότητες και για αυτό το λόγο μοντελοποιήθηκε πλήρως από τα δεδομένα. Συγκεκριμένα αναπτύχθηκαν data driven τεχνικές για να εκτιμήσουν τόσο την αναμενόμενη μορφολογία της QCD (QCD shape), όσο και τον αριθμό των αναμενόμενων γεγονότων (QCD rate).

Το QCD shape μοντελοποιήθηκε από μια περιοχή ελέγχου στα δεδομένα αντίστοιχης κινηματικής αλλά με διαφορετικές απαιτήσεις στα scores των BDTs. Η εγκυρότητα της μεθόδου ελέγχθηκε στην προσομοίωση και μια διόρθωση εφαρμόστηκε στην κατανομή των δεδομένων η οποία χρησιμοποιείται για τη περιγραφή της κατανομής της QCD στην περιοχή σήματος.

Για την εκτίμηση του αριθμού της QCD χρησιμοποιήθηκε μια μέθοδος "ABCD". Αυτή η μέθοδος δίνει την εκτίμηση των γεγονότων σε μια περιοχή σήματος ("A") από μετρήσεις που πραγματοποιούνται σε περιοχές ελέγχου στα όρια της περιοχής σήματος. Πιο συγκεκριμένα, για δυο ασυσχέτιστες μεταβλητές, ο αριθμός των γεγονότων στην περιοχή ("A") θα δίνεται από τον λόγο:

$$N_A = \frac{N_C^{bkg} \cdot N_B^{bkg}}{N_D^{bkg}} \tag{17.1}$$

Η μέθοδος αυτή εφαρμόστηκε σε τρεις εκτεταμένες περιοχές του σήματος με βάση τον αριθμό των ανακατασκευασμένων jets του γεγονότος χρησιμοποιώντας ως ασυσχέτιστες μεταβλητές το score των BDTs και τον αριθμό των Ak4 jets. Αυτές οι περιοχές αποτελούνται από γεγονότα που περνάνε την βασική επιλογή αλλά σε αυτές εφαρμόζονται λιγότερα κριτήρια. Έπειτα, το ποσοστό κατά το οποίο τα γεγονότα κατατάσσονται στις διάφορες κατηγορίες προσδιορίζεται από την προσομοίωση και χρησιμοποιείτε στην τελική εκτίμηση του υποβάθρου της QCD. Η τελική εκτίμηση της QCD θα δίνεται από την σχέση:

$$\mathrm{QCD}_{\mathrm{cat_i}} = \mathrm{NAk8}_{\mathrm{ext}} \cdot \mathrm{frac}_{\mathrm{cati}} \cdot \mathrm{D}(\mathrm{QCD}_{\mathrm{cati}}) \tag{17.2}$$

όπου το $\mathrm{NAk8}_{\mathrm{ext}}$ αντιστοιχεί στα εκτιμώμενα γεγονότα της εκτεταμένης περιοχής, το $\mathrm{frac}_{\mathrm{cati}}$, στο ποσοστό των γεγονότων κάθε κατηγορίας και το $\mathrm{D}(\mathrm{QCD}_{\mathrm{cati}})$ στην κατανομή της QCD.

17.8 Εκτίμηση του υποβάθρου από ταυτόχρονη παραγωγή δυο **top quark**

Το αμέσως επόμενο σε σημαντικότητα υπόβαθρο αποτελείται από γεγονότα $\mathrm{t\bar{t}}$. Αυτό το υπόβαθρο μοντελοποιείται αποκλειστικά από την προσημείωση. Όμως, υπάρχουν διαφορές μεταξύ διαφορετικών Monte Carlo γεννητόρων και δεδομένων. Για αυτό τον λόγο, υπολογίστηκε μια διόρθωση από μια περιοχή ελέγχου στα δεδομένα με περίσσεια από γεγονότα $\mathrm{t\bar{t}}$. Αυτή η περιοχή ελέγχου χρησιμοποιήθηκε επίσης και για να περιορίσει την αβεβαιότητα του αριθμού των εκτιμώμενων γεγονότων του υποβάθρου $\mathrm{t\bar{t}}$.

17.9 Πηγές συστηματικής αβεβαιότητας

Σε αυτή την ενότητα περιγράφονται οι πηγές της συστηματικής αβεβαιότητας που αφορούν την ανάλυση. Αυτές ερμηνεύονται σαν παράμετροι ενόχλησης (nuisance parameters) στη τελική προσαρμογή για την ανίχνευση σήματος.

Οι πειραματικές αβεβαιότητες αφορούν στη μοντελοποίηση των υποβάθρων, καθώς αναπτύχθηκαν τεχνικές και χρησιμοποιήθηκαν προσομοιώσεις για την εκτίμηση τους. Άλλες αβεβαιότητες προκύπτουν από τον προσδιορισμό της ολοκληρωμένης φωτεινότητας, την γνώση της ενεργειακής κλίμακας των jets και του αριθμού των αληθινών αλληλεπιδράσεων ανά bunch crossing. Επίσης λήφθηκαν υπόψιν αβεβαιότητες που αφορούν στην εφαρμογή διορθώσεων για την καλύτερη συμφωνία δεδομένων και προσομοίωσης (scale factors).

Οι θεωρητικές αβεβαιότητες αφορούν την προσημείωση αυτή καθεαυτή. Αυτές περιλαμβάνουν την αβεβαιότητα που προκύπτει από τις συναρτήσεις κατανομής παρτονίων καθώς και αβεβαιότητες που αφορούν στην επανακανονικοποίηση και τη παραγοντοποίηση της κλίμακας που αφορά τον υπολογισμό του πλάτους σκέδασης της κάθε διεργασίας.

17.10 Αποτελέσματα

Η μετρική που χρησιμοποιήθηκε για τον προσδιορισμό της ευαισθησίας της ανάλυσης είναι το εκτιμώμενο αποκλειόμενο όριο (expected exclusion limit) με διάστημα εμπιστοσύνης 95%. Η στρατηγική για τον προσδιορισμό του ορίου περιλαμβάνει μια προσαρμογή μέγιστης πιθανοφάνειας στην κατανομή του ανακατασκευασμένου Higgs στα δεδομένα ταυτόχρονα σε όλες τις κατηγορίες σήματος και υποβάθρου και στις κατηγορίες ελέγχου. Επιπλέον, όλες οι αβεβαιότητες λήφθηκαν υπόψιν ως nuisances. Το παρατηρούμενο (αναμενόμενο) όριο για την ανάλυση σε διάστημα εμπιστοσύνης 95% είναι 9.4 ($7.6 < 10.4 < 14.3$) φορές από την εκτίμηση του καθιερωμένου προτύπου.

17.11 Συμπεράσματα

Παρουσιάστηκε μια έρευνα για την ταυτόχρονη παραγωγή ενός συμβατού με το καθιερωμένο πρότυπο μποζονίου Higgs με ένα ζεύγος top-antitop quark και την μετέπειτα διάσπασή του σε ένα ζεύγος b quark στην πλήρως αδρονική κατάσταση χρησιμοποιώντας πίδακες μεγάλης ακτίνας. Τα δεδομένα αντιστοιχούν σε ολοκληρωμένη φωτηνότητα 31.5 fb^{-1} και συλλέχθηκαν από τον ανιχνευτή CMS το 2016. Η συγκεκριμένη ανάλυση παρουσιάζει μια καινούργια προσέγγιση, αφού θεωρεί jets μεγάλης εγκάρσιας ορμής. Για την εκτίμηση των υποβάθρων αναπτύχθηκαν τεχνικές που βασίζονται στα δεδομένα και επίσης εκτιμήθηκαν και συνυπολογίστηκαν όλες οι παράγοντες αβεβαιότητας. Η ευαισθησία της ανάλυσης μελετήθηκε ως προ το αναμενόμενο όριο (9.4 ($7.6 < 10.4 < 14.3$) φορές των εκτιμήσεων του ΚΠ).

Αυτό είναι το πρώτο αποτέλεσμα σε φασικό χώρο με υψηλές εγκάρσιες ορμές. Αναμένεται μεγάλη βελτίωση της ευαισθησίας της ανάλυσης στην HL-LHC όπου λόγω της αυξημένης φωτεινότητας η συγκεκριμένη προσέγγιση θα είναι ευνοημένη.

Bibliography

[1] Andrew Purcell. Go on a particle quest at the first CERN webfest. Le premier webfest du CERN se lance à la conquête des particules. (BUL-NA-2012-269. 35/2012):10, Aug 2012.

[2] John Ellis, Mary K Gaillard, and Dimitri V Nanopoulos. A Historical Profile of the Higgs Boson. An Updated Historical Profile of the Higgs Boson. (arXiv:1504.07217. KCL-PH-TH-2015-20. LCTS-2015-10. CERN-PH-TH-2015-098):22 p, Apr 2015.

[3] LHC Higgs Cross Section Working Group. https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG, 2019.

[4] Joshua T. Ruderman and David Shih. General neutralino nlsps at the early lhc. *Journal of High Energy Physics*, 2012(8):159, Aug 2012.

[5] CMS Supersymmetry Physics Results. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSUS, 2019.

[6] Christiane Lefèvre. The CERN accelerator complex. Complexe des accélérateurs du CERN. Dec 2008.

[7] CMS luminosity Public Results. https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults, 2019.

[8] CMS detector design. http://cms.web.cern.ch/news/cms-detector-design, 2011.

[9] The CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *Journal of Instrumentation*, 9(10):P10009–P10009, oct 2014.

[10] S. Chatrchyan et al. The CMS Experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[11] Giovanni Abbiendi. The CMS muon system in Run2: preparation, status and first results. *PoS*, EPS-HEP2015:237, 2015.

[12] James Brooke, David Cussans, Ricky Frazier, GP Heath, D Newbold, Sameera Galagedera, Abid Shah, and Sadaf Madani. Design and implementation of the global calorimeter trigger for cms. 10 2019.

[13] Performance of b tagging algorithms in proton-proton collisions at 13 TeV with Phase 1 CMS detector. https://twiki.cern.ch/twiki/bin/view/CMSPublic/BTV13TeV2017FIRST2018, 2018.

[14] Stefan Höche. Introduction to parton-shower event generators. In *Proceedings, Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders (TASI 2014): Boulder, Colorado, June 2-27, 2014*, pages 235–295, 2015.

[15] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt. Parton distributions for the lhc. *The European Physical Journal C*, 63(2):189–285, Jul 2009.

[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 2. (CERN-TH-412):80 p, Feb 1964.

[18] Steven Weinberg. A model of leptons. *Phys. Rev. Lett.*, 19:1264–1266, Nov 1967.

[19] Abdus Salam. Weak and Electromagnetic Interactions. *Conf. Proc.*, C680519:367–377, 1968.

[20] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964.

[21] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13:508–509, Oct 1964.

[22] Jeffrey Goldstone, Abdus Salam, and Steven Weinberg. Broken symmetries. *Phys. Rev.*, 127:965–970, Aug 1962.

[23] M Tanabashi, PD Grp, K Hagiwara, K Hikasa, Katsumasa Nakamura, Y Sumino, F Takahashi, J Tanaka, K Agashe, G Aielli, Claude Amsler, Mario Antonelli, DM Asner, Howard Baer, S Banerjee, RM Barnett, T Basaglia, Christian Bauer, and J. Beatty. Review of particle physics: Particle data group. *Physical Review D*, 98, 08 2018.

[24] ATLAS Collaboration. Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector. *Phys. Lett. B*, 784(arXiv:1806.00425):173–191. 19 p, Jun 2018.

[25] CMS Collaboration. Observation of $\mathrm{t\bar{t}H}$ production. *Phys. Rev. Lett.*, 120(CMS-HIG-17-035. CMS-HIG-17-035-003):231801. 17 p, Apr 2018.

[26] CMS collaboration. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716(1):30 – 61, 2012.

[27] ATLAS collaboration. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1 – 29, 2012.

[28] Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s}$ =7 and 8 TeV.

[29] "CMS and ATLAS Collaborations". Combined measurement of the higgs boson mass in pp collisions at $\sqrt{s} = 7$ and 8 tev with the atlas and cms experiments.

[30] Combined measurements of the Higgs boson's couplings at $\sqrt{s} = 13$ TeV. Technical Report CMS-PAS-HIG-17-031, CERN, Geneva, 2018.

[31] CMS Collaboration. Measurements of properties of the higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$tev. *Journal of High Energy Physics*, 2017(11):47, Nov 2017.

[32] CMS Collaboration. Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 11(CMS-HIG-16-040. CMS-HIG-16-040-003):185. 56 p, Apr 2018.

[33] Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s} = 13$ TeV. Technical Report CMS-PAS-HIG-16-042, CERN, Geneva, 2018.

[34] CMS collaboration. Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector. *Phys. Lett. B*, 779(CMS-HIG-16-043. CMS-HIG-16-043-003):283–316. 34 p, Aug 2017.

[35] ATLAS Collaboration. Observation of $H \rightarrow b\bar{b}$ decays and VH production with the ATLAS detector. *Phys. Lett. B*, 786(arXiv:1808.08238):59–86. 28 p, Aug 2018.

[36] CMS collaboration. Observation of Higgs boson decay to bottom quarks. *Phys. Rev. Lett.*, 121(arXiv:1808.08242. CMS-HIG-18-016-003. 12):121801. 20 p, Aug 2018.

[37] Search for the standard model higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the atlas detector. *Phys. Rev. D*, 97:072016, Apr 2018.

[38] Measurement of $\mathrm{t\bar{t}H}$ production in the $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$ decay channel in $41.5\,\mathrm{fb}^{-1}$ of proton-proton collision data at $\sqrt{s} = 13\,\mathrm{TeV}$. Technical Report CMS-PAS-HIG-18-030, CERN, Geneva, 2019.

[39] CMS collaboration. Search for $\mathrm{t\bar{t}H}$ production in the all-jet final state in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 06(CMS-HIG-17-022. CMS-HIG-17-022-003):101. 46 p, Mar 2018.

[40] ATLAS collaboration. Search for the standard model higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in pp collisions at $\sqrt{s} = 13$tev with the atlas detector. *Journal of High Energy Physics*, 2016(5):160, May 2016.

[41] Stefan P. Martin. *Prespectives of Supersymmetry 2*. World Scientific.

[42] Günther Christian. An introduction to quantum field theory, by michael e. pestin and daniel v. schroeder. *Journal of Applied Mathematics and Stochastic Analysis*, 10, 01 1997.

[43] Julius Wess and Jonathan Bagger. *Supersymmetry and Supergravity*. Princeton Series in Physics.

[44] Ulrich Ellwanger, Cyril Hugonie, and Ana M. Teixeira. The next-to-minimal supersymmetric standard model. *Physics Reports*, 496(1-2):1–77, Nov 2010.

[45] Gian Francesco Giudice and Riccardo Rattazzi. Theories with Gauge-Mediated Supersymmetry Breaking. *Phys. Rep.*, 322(hep-ph/9801271. CERN-TH-97-380):419–499. 103 p, Jan 1998.

[46] Simon Knapen and Diego Redigolo. Gauge mediation at the lhc: status and prospects. *Journal of High Energy Physics*, 2017(1):135, Jan 2017.

[47] Johan Alwall, Philip C. Schuster, and Natalia Toro. Simplified models for a first characterization of new physics at the lhc. *Phys. Rev. D*, 79:075020, Apr 2009.

[48] Simplified models for LHC new physics searches. *Journal of Physics G: Nuclear and Particle Physics*, 39(10):105005, sep 2012.

[49] Interpretation of searches for supersymmetry with simplified models. *Phys. Rev. D*, 88:052017, Sep 2013.

[50] Lyndon Evans and Philip Bryant. LHC machine. *Journal of Instrumentation*, 3(08):S08001–S08001, aug 2008.

[51] *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical Design Report CMS. CERN, Geneva, 2006. There is an error on cover due to a technical problem for some items.

[52] The CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *Journal of Instrumentation*, 9(10):P10009–P10009, oct 2014.

[53] *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.

[54] Addendum to the CMS ECAL technical design report: Changes to the CMS ECAL electronics. 2002.

[55] The performance of the CMS muon detector in proton-proton collisions at $\sqrt{s} =$ 7 TeV at the LHC. *JINST*, 8(CMS-MUO-11-001. CMS-MUO-11-001. CERN-PH-EP-2013-072):P11002. 101 p, Jun 2013. Comments: Submitted to JINST.

[56] The CMS trigger system. The CMS trigger system. *JINST*, 12(CMS-TRG-12-001. 01):P01020. 122 p, Sep 2016. Replaced with the published version. Added the journal reference and DOI. All the figures and tables can be found at http://cms-results.web.cern.ch/cms-results/public-results/publications/TRG-12-001/index.html.

[57] CMS Technical Design Report for the Pixel Detector Upgrade. Technical Report CERN-LHCC-2012-016. CMS-TDR-11, Sep 2012. Additional contacts: Jeffrey Spalding, Fermilab, Jeffrey.Spalding@cern.ch Didier Contardo, Universite Claude Bernard-Lyon I, didier.claude.contardo@cern.ch.

[58] D Contardo, M Klute, J Mans, L Silvestris, and J Butler. Technical Proposal for the Phase-II Upgrade of the CMS Detector. Technical Report CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02, Geneva, Jun 2015. Upgrade Project Leader Deputies: Lucia Silvestris (INFN-Bari), Jeremy Mans (University of Minnesota) Additional contacts: Lucia.Silvestris@cern.ch, Jeremy.Mans@cern.ch.

[59] CMS collaboration. Particle-flow reconstruction and global event description with the CMS detector. *Journal of Instrumentation*, 12(10):P10003–P10003, oct 2017.

[60] Description and performance of track and primary-vertex reconstruction with the CMS tracker. *JINST*, 9(CMS-TRK-11-001. CERN-PH-EP-2014-070. CMS-TRK-11-001):P10009. 80 p, May 2014. Comments: Replaced with published version. Added journal reference and DOI.

[61] W. Adam, B. Mangano, T. Speer, and T. Todorov. Track reconstruction in the CMS tracker. 2005.

[62] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-ktjet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, Apr 2008.

[63] Particle-flow reconstruction and global event description with the CMS detector. Particle-flow reconstruction and global event description with the CMS detector. *JINST*, 12(CMS-PRF-14-001. CMS-PRF-14-001-004. 10):P10003. 82 p, Jun 2017. Replaced with the published version. Added the journal reference and DOI. All the figures and tables can be found at http://cms-results.web.cern.ch/cms-results/public-results/publications/PRF-14-001 (CMS Public Pages).

[64] W Adam, R FrÃŒhwirth, A Strandlie, and T Todorov. Reconstruction of electrons with the gaussian-sum filter in the CMS tracker at the LHC. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9–N20, jul 2005.

[65] *Journal of Instrumentation*, 10(08):P08010–P08010, Aug 2015.

[66] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. Fastjet user manual. *The European Physical Journal C*, 72(3), Mar 2012.

[67] V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al. Jet energy scale and resolution in the cms experiment in pp collisions at 8 tev. *Journal of Instrumentation*, 12(02):P02014–P02014, Feb 2017.

[68] Daniele Bertolini, Philip Harris, Matthew Low, and Nhan Tran. Pileup per particle identification. *Journal of High Energy Physics*, 2014(10), Oct 2014.

[69] Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. *Journal of Instrumentation*, 12(02):P02014–P02014, feb 2017.

[70] Matteo Cacciari and Gavin P. Salam. Pileup subtraction using jet areas. *Physics Letters B*, 659(1-2):119–126, Jan 2008.

[71] A.M. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al. Identification of heavy-flavour jets with the cms detector in pp collisions at 13 tev. *Journal of Instrumentation*, 13(05):P05011–P05011, May 2018.

[72] The CMS collaboration. Identification of b-quark jets with the CMS experiment. *Journal of Instrumentation*, 8(04):P04013–P04013, apr 2013.

[73] Performance of missing transverse momentum in pp collisions at sqrt(s)=13 TeV using the CMS detector. Technical Report CMS-PAS-JME-17-001, CERN, Geneva, 2018.

[74] Stefan Weinzierl. Introduction to monte carlo methods, 2000.

[75] Stefan Höche. Introduction to parton-shower event generators, 2014.

[76] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Christopher S. Deans, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Nathan P. Hartland, José I. Latorre, and et al. Parton distributions for the lhc run ii. *Journal of High Energy Physics*, 2015(4), Apr 2015.

[77] Bo Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand. Parton Fragmentation and String Dynamics. *Phys. Rept.*, 97:31–145, 1983.

[78] Bo Andersson. The Lund model. *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.*, 7:1–471, 1997.

[79] V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, M. Friedl, and et al. Charged particle multiplicities in pp interactions at $\sqrt{s} = 0.9$, 2.36, and 7 tev. *Journal of High Energy Physics*, 2011(1), Jan 2011.

[80] S. Agostinelli et al. GEANT4: A Simulation toolkit. *Nucl. Instrum. Meth.*, A506:250–303, 2003.

[81] Sezen Sekmen. Recent developments in cms fast simulation, 2017.

[82] J. S. Conway. Incorporating nuisance parameters in likelihoods for multisource spectra, 2011.

[83] F. James. *Statistical Methods in Experimental Physics*. World Scientific Publishing, Singapore.

[84] A.L Read. Linear interpolation of histograms. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 425(1):357 – 360, 1999.

[85] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[86] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2), Feb 2011.

[87] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):426–482, 1943.

[88] A L Read. Presentation of search results: theCLstechnique. *Journal of Physics G: Nuclear and Particle Physics*, 28(10):2693–2704, sep 2002.

[89] G Petrucciani, A Rizzi, and C Vuosalo. Mini-AOD: A new analysis data format for CMS. *Journal of Physics: Conference Series*, 664(7):072052, dec 2015.

[90] V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, and et al. Event generator tunes obtained from underlying event and multiparton scattering measurements. *The European Physical Journal C*, 76(3), Mar 2016.

[91] A. Kulesza and L. Motyka. Soft gluon resummation for the production of gluino-gluino and squark-antisquark pairs at the lhc. *Physical Review D*, 80(9), Nov 2009.

[92] Sezen Sekmen. Recent developments in cms fast simulation, 2017.

[93] EgammaIDRecipesRun2 CMS Collaboration. https://twiki.cern.ch/twiki/bin/viewauth/CMS/EgammaIDRecipesRun2, 2019.

[94] V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al. Search for supersymmetry in events with photons and missing transverse energy in pp collisions at 13 tev. *Physics Letters B*, 769:391–412, Jun 2017.

[95] CMS Collaboration. Search for supersymmetry in final states with photons and missing transverse momentum in proton-proton collisions at 13 tev, 2019.

[96] ATLAS Collaboration. Search for supersymmetry in a final state containing two photons and missing transverse momentum in $\sqrt{s} = 13$ tev pp collisions at the lhc using the atlas detector, 2016.

[97] ATLAS Collaboration. Search for photonic signatures of gauge-mediated supersymmetry in 13 tev pp collisions with the atlas detector, 2018.

[98] CMS Collaboration. Combined search for supersymmetry with photons in proton-proton collisions at $\sqrt{s} = 13$ tev, 2019.

[99] CMS Collaboration. Search for supersymmetry in events with a photon, a lepton, and missing transverse momentum in proton-proton collisions at $\sqrt{s} = 13$ tev, 2018.

[100] A. M. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, and et al. Search for supersymmetry in events with at least one photon, missing transverse momentum, and large transverse event activity in proton-proton collisions at s = 13

$$\sqrt{s} = 13$$

tev. *Journal of High Energy Physics*, 2017(12), Dec 2017.

[101] CMS Collaboration. Search for gauge-mediated supersymmetry in events with at least one photon and missing transverse momentum in pp collisions at $\sqrt{s} =$ 13 tev, 2017.

[102] Juan Rojo, Alberto Accardi, Richard D Ball, Amanda Cooper-Sarkar, Albert de Roeck, Stephen Farry, James Ferrando, Stefano Forte, Jun Gao, Lucian Harland-Lang, and et al. The pdf4lhc report on pdfs and lhc data: results from run i

and preparation for run ii. *Journal of Physics G: Nuclear and Particle Physics*, 42(10):103103, Sep 2015.

[103] P. Skands, S. Carrazza, and J. Rojo. Tuning pythia 8.1: the monash 2013 tune. *The European Physical Journal C*, 74(8), Aug 2014.

[104] Utilities for Accessing Pileup Information for Data. https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupJSONFileforData, 2019.

[105] B tagging discriminant shape calibration using event weights with a tag-and-probe method. https://twiki.cern.ch/twiki/bin/view/CMS/BTagShapeCalibration, 2019.

[106] Muon Identification and Isolation efficiency on full 2016 dataset. https://cds.cern.ch/record/2257968/files/DP2017_007.pdf, 2019.

[107] Egamma RunII Recommendations. https://twiki.cern.ch/twiki/bin/view/CMS/EgammaRunIIRecommendations#Spring16_mva_Summer16_cut_based, 2019.

[108] Jet algorithms performance in 13 TeV data. Technical Report CMS-PAS-JME-16-003, CERN, Geneva, 2017.

[109] CMS Collaboration. Search for direct production of supersymmetric partners of the top quark in the all-jets final state in proton-proton collisions at sqrt(s) = 13 tev, 2017.

[110] CMS Collaboration. Search for resonant $\mathrm{t\bar{t}}$ production in proton-proton collisions at $\sqrt{s} = 13$ tev, 2018.

[111] A.M. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al. Inclusive search for a highly boosted higgs boson decaying to a bottom quark-antiquark pair. *Physical Review Letters*, 120(7), Feb 2018.

[112] CMS Collaboration. Measurement of the integrated and differential t-tbar production cross sections for high-pt top quarks in pp collisions at sqrt(s) = 8 tev, 2016.

[113] Andrew J. Larkoski, Ian Moult, and Benjamin Nachman. Jet substructure at the large hadron collider: A review of recent advances in theory and machine learning. *Physics Reports*, 841:1–63, Jan 2020.

[114] Mrinal Dasgupta, Alessandro Fregoso, Simone Marzani, and Gavin P. Salam. Towards an understanding of jet substructure. *Journal of High Energy Physics*, 2013(9), Sep 2013.

[115] David Krohn, Jesse Thaler, and Lian-Tao Wang. Jet trimming. *Journal of High Energy Physics*, 2010(2), Feb 2010.

[116] Jonathan M. Butterworth, Adam R. Davison, Mathieu Rubin, and Gavin P. Salam. Jet substructure as a new higgs search channel at the lhc, 2008.

[117] Mrinal Dasgupta, Alessandro Fregoso, Simone Marzani, and Gavin P. Salam. Towards an understanding of jet substructure. *Journal of High Energy Physics*, 76(9), 2013.

[118] Mrinal Dasgupta, Alessandro Fregoso, Simone Marzani, and Gavin P. Salam. Towards an understanding of jet substructure. *Journal of High Energy Physics*, 2013(9), Sep 2013.

[119] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, and Jesse Thaler. Soft drop. *Journal of High Energy Physics*, 2014(5), May 2014.

[120] Jesse Thaler and Ken Van Tilburg. Identifying boosted objects with n-subjettiness. *Journal of High Energy Physics*, 2011(3), Mar 2011.

[121] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, A. Christov, D. Dannheim, K. Danielowski, S. Henrot-Versille, M. Jachowski, K. Kraszewski, A. Krasznahorkay Jr., M. Kruk, Y. Mahalalel, R. Ospanov, X. Prudent, A. Robert, D. Schouten, F. Tegenfeldt, A. Voigt, K. Voss, M. Wolter, and A. Zemla. Tmva - toolkit for multivariate data analysis, 2007.

[122] pt(top-quark) based reweighting of ttbar MC. https://twiki.cern.ch/twiki/bin/viewauth/CMS/TopPtReweighting, 2019.

[123] Documentation of Higgs PAG combine. http://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/, 2019.

[124] CMS Collaboration. Measurement of differential $\mathrm{t\bar{t}}$ production cross sections using top quarks at large transverse momenta in pp collisions at $\sqrt{s} = 13$ tev, 2020.

[125] CMS Collaboration. CMS Luminosity Measurements for the 2016 Data Taking Period. 2017.

[126] V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al. Jet energy scale and resolution in the cms experiment in pp collisions at 8 tev. *Journal of Instrumentation*, 12(02):P02014–P02014, Feb 2017.

[127] V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, and et al. Event generator tunes obtained from underlying event and multiparton scattering measurements. *The European Physical Journal C*, 76(3), Mar 2016.

[128] Roger Barlow and Christine Beeston. Fitting using finite monte carlo samples. *Computer Physics Communications*, 77(2):219 – 228, 1993.