



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ
ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ
ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΣΧΟΛΗ ΜΗΧΑΝΟΛΟΓΩΝ
ΜΗΧΑΝΙΚΩΝ

ΕΚΕΦΕ «ΔΗΜΟΚΡΙΤΟΣ»
ΙΝΣΤΙΤΟΥΤΟ ΝΑΝΟΕΠΙΣΤΗΜΗΣ
ΚΑΙ ΝΑΝΟΤΕΧΝΟΛΟΓΙΑΣ
ΙΝΣΤΙΤΟΥΤΟ ΠΥΡΗΝΙΚΗΣ ΚΑΙ
ΣΩΜΑΤΙΔΙΑΚΗΣ ΦΥΣΙΚΗΣ



Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών

«Φυσική και Τεχνολογικές Εφαρμογές»

**Search for the exotic decay of the Higgs boson
to four bottom-quarks $H \rightarrow 4b$ in the associated production
of the Higgs with a Z boson at the LHC**

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της Άρτεμις Ταξείδη

Επιβλέπουσα:

Γεωργία Καραποστόλη

Αθήνα, Μάρτιος, 2025

Abstract

In this thesis, we present a search for non-standard decay of the Standard Model-like Higgs boson to a pair of pseudoscalar lighter bosons $H \rightarrow aa$ with $12 \leq m_a \leq 60$ GeV, as predicted in models with extended Higgs sectors, like the two Higgs doublet model with a scalar singlet field (2HDM+s). Monte Carlo simulated samples of the RUN2 2017 data taking period of CMS were used. The integrated luminosity of this period was $L_{int} = 41.5 fb^{-1}$ and the center of mass energy of pp collisions was $\sqrt{s} = 13$ TeV.

In this analysis, the ZH associated production is considered where the Z boson decays either leptonically (2-lepton channel) or invisibly (0-lepton channel).

As "a" boson mass decreases, the b-quarks of its decay tend to be collimated. Thus, in each channel we define two regimes, the resolved and the merged one. The resolved regime, is defined for "a" boson masses in the range 20-60 GeV and the standard jet reconstruction and b-tagging methods of CMS are used to identify the final state b-quarks. Meanwhile, the merged regime is defined for lower "a" masses in the range 15-40 GeV where we introduce large-radius jets, named fat jets and a double b-tag discriminator to reconstruct two high-Lorentz boosted b-quarks into a single fat jet.

Then, variables with high discriminative power are chosen for multi-variable analysis. The boosted decision trees (BDT) classification algorithm was chosen to distinguish between signal and background.

The output distribution of BDT was used to conduct binned shape analysis. We used RooFit to generate MC pseudo-data and perform binned maximum likelihood fit.

Finally, Bayesian method was chosen to compute credible intervals of the expected signal yield, using RooStats Markov Chain Monte Carlo (MCMC) algorithm implementation. From these intervals, we derive the 95% confidence level upper limits on the branching ratio $Br(h \rightarrow aa \rightarrow 4b)$, for all "a" masses in each regime of each channel.

Keywords: particle physics, beyond the standard model(BSM), exotic Higgs decay, multivariable analysis, boosted decision tree (BDT), significance, maximum likelihood fit, RooFit, Roostats, Bayesian inference, Markov chain Monte Carlo method, credible interval, upper limits.

Ευχαριστίες

Καταρχάς θα ήθελα να ευχαριστήσω ιδιαίτερος την οικογένειά μου που με στήριξε όλο αυτό το διάστημα που ασχολήθηκα με τη συγκεκριμένη εργασία και ήταν δίπλα μου στις όμορφες αλλά και στις δύσκολες στιγμές.

Θα ήθελα επίσης να ευχαριστήσω τους συναδέλφους μου Πέννυ Ρουσελιωτάκη και Κωνσταντίνο Πάσχο, με τους οποίους συνεργαζόμαστε στενά και υπήρξαν εκεί σε κάθε μου βήμα κατά το μεγαλύτερο διάστημα αυτής της πορείας.

Ένα πολύ μεγάλο ευχαριστώ στην επιβλέπουσα μου Καραποστόλη Γεωργία. Καταρχάς για την μεγάλη ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο με πολλές προοπτικές και κατά δεύτερον που με καθοδήγησε με μεγάλη ζέση και υπομονή. Κατά τη διαδικασία της εκπόνησης αυτής της εργασίας, σαφώς μου έμαθε πάρα πολλά τόσο σε επιστημονικό επίπεδο, αλλά και με ωρίμασε σαν άνθρωπο και με έμαθε να προσπαθώ να βελτιώνομαι συνεχώς και να έχω θάρρος και επιμονή. Ευχαριστώ επίσης τα άλλα δύο μέλη της τριμελούς επιτροπής, τους κυρίους Κουσουρή και Τσιπολίτη, που κατά την περίοδο των πρώτων μου βημάτων, μου έδωσαν σημαντικό feedback στα εβδομαδιαία status report meetings που διεξάγονταν.

Τέλος θα ήθελα να ευχαριστήσω τον κύριο Ευάγγελο Γαζή που στα προπτυχιακά μου χρόνια στη ΣΗΜΜΥ μου ενέπνευσε την αγάπη για την φυσική, η οποία βρισκόταν σε λανθάνουσα κατάσταση, και τελικά επέλεξα να στραφώ προς αυτό το αντικείμενο μεταπτυχιακών σπουδών και μετέπειτα πορείας.

Εκτεταμένη Περίληψη

Το καθιερωμένο Πρότυπο

Η σωματιδιακή φυσική είναι ο κλάδος που μελετά τις δομικές συνιστώσες της ύλης, δηλαδή τα στοιχειώδη σωματίδια και τις μεταξύ τους αλληλεπιδράσεις, που αποτελούν τις θεμελιώδεις δυνάμεις. Στόχος της είναι ο προσδιορισμός της δομής του μικρόκοσμου και η καλύτερη κατανόηση των νόμων του Σύμπαντος, ώστε να απαντηθούν καίρια ερωτήματα που αφορούν την εξέλιξή του. Το Καθιερωμένο Πρότυπο (ΚΠ) της σωματιδιακής φυσικής είναι μια κβαντική θεωρία πεδίου (QFT) που περιγράφει τις ηλεκτρομαγνητικές, ασθενείς και ισχυρές αλληλεπιδράσεις χρησιμοποιώντας την ομάδα συμμετρίας $SU(3)_C \times SU(2)_L \times U(1)_Y$. Ταξινομεί τα στοιχειώδη σωματίδια βάσει των κβαντικών ιδιοτήτων τους, εξηγώντας τις θεμελιώδεις δυνάμεις, πλην της βαρύτητας. Η ανακάλυψη του μποζονίου Higgs το 2012 επιβεβαίωσε το τελευταίο προβλεπόμενο σωματίδιο του προτύπου. Τα στοιχειώδη σωματίδια στο καθιερωμένο πρότυπο είναι τα φερμιόνια και τα μποζόνια.

Φερμιόνια: Οι δομικοί λίθοι της ύλης: Τα φερμιόνια (σωματίδια με σπιν 1/2) ακολουθούν τη στατιστική Fermi-Dirac. Λόγω της απαγορευτικής αρχής του Pauli, φερμιόνια με ίδιους κβαντικούς αριθμούς δεν μπορούν να καταλάβουν την ίδια κβαντική κατάσταση. Χωρίζονται σε κουάρκ και λεπτόνια:

- **Κουάρκ:** Έξι γεύσεις (up, down, charm, strange, top, bottom), καθεμία με τρία χρωματικά φορτία, σχηματίζουν αδρόνια μέσω της ισχυρής αλληλεπίδρασης.
- **Λεπτόνια:** Περιλαμβάνουν τα φορτισμένα λεπτόνια (ηλεκτρόνιο, μίονιο, ταυ) και τα αντίστοιχα νετρίνα τους, που αλληλεπιδρούν μέσω της ασθενούς δύναμης.

Τα φερμιόνια οργανώνονται σε τρεις γενιές, καθεμία εκ των οποίων περιέχει ένα ζεύγος κουάρκ και ένα ζεύγος λεπτονίων, με τα βαρύτερα σωματίδια να ανήκουν στην τελευταία γενιά. Επιπλέον, σε κάθε φερμιόνιο αντιστοιχεί ένα αντισωματίδιο με αντίθετο ηλεκτρικό φορτίο, και πανομοιότυπο ως προς τις λοιπές ιδιότητες. Τα νετρίνα είναι ηλεκτρικά ουδέτερα, και επομένως διαφοροποιούνται από τα αντινετρίνα από μία ιδιότητα που σχετίζεται με το σπιν τους.

Μποζόνια: Οι φορείς των δυνάμεων: Το Καθιερωμένο Πρότυπο περιγράφει τις αλληλεπιδράσεις μέσω τεσσάρων βαθμωτών μποζονίων με μηδενικό σπιν:

- **Φωτόνιο (γ):** Άμαζος φορέας του ηλεκτρομαγνητισμού, επιδρώντας σε όλα τα φορτισμένα φερμιόνια (τα νετρίνα δεν “αντιλαμβάνονται” την ηλεκτρομαγνητική δύναμη).
- **Γλυσόνια (g):** Οκτώ άμαζα σωματίδια που αποτελούν τους φορείς της ισχυρής αλληλεπίδρασης, φέρουν χρωματικό φορτίο και εμφανίζουν αυτο-αλληλεπίδραση, οδηγώντας σε χρωματικό

περιορισμό (color confinement). Η ισχυρή αλληλεπίδραση είναι η ισχυρότερη από τις τέσσερις θεμελιώδεις αλληλεπιδράσεις και έχει εμβέλεια της τάξης του ατομικού πυρήνα.

- **Μποζόνια $W^\pm$ και Z :** Μποζόνια που έχουν μάζα και φέρουν την ασθενή αλληλεπίδραση, της οποίας η ασθενής φύση οφείλεται στο γεγονός ότι έχουν μεγάλη μάζα. Το μποζόνιο Z είναι ηλεκτρικά ουδέτερο και διατηρεί τη γεύση των σωματιδίων στις αλληλεπιδράσεις του (“neutral-current”), ενώ τα μποζόνια $W^\pm$ είναι ηλεκτρικά φορτισμένα και αλλάζουν τη γεύση των σωματιδίων (“charged-current”).
- **Μποζόνιο Higgs (H):** Σωματίδιο με spin 0 υπεύθυνο για τη θραύση της ηλεκτρασθενούς συμμετρίας και τη δημιουργία μάζας μέσω του μηχανισμού Higgs.

Εν κατακλείδι, τα στοιχειώδη σωματίδια που αποτελούν το ΚΠ, είναι: 12 φερμιόνια, 4 μποζόνια βαθμίδας (gauge bosons) και ένα βαθμωτό μποζόνιο.

Οι αλληλεπιδράσεις μεταξύ των σωματιδίων του Καθιερωμένου Προτύπου διατυπώνονται μαθηματικά στη Λαγκρανζιανή του ΚΠ, η οποία βασίζεται σε τοπικές συμμετρίες βαθμίδας.

Η Κβαντική Ηλεκτροδυναμική (QED) περιγράφει τις ηλεκτρομαγνητικές αλληλεπιδράσεις υπό την $U(1)$ βαθμωτή συμμετρία. Το πεδίο του φωτονίου διαμεσολαβεί στις αλληλεπιδράσεις μεταξύ φορτισμένων σωματιδίων, εξασφαλίζοντας τη διατήρηση του φορτίου.

Η Κβαντική Χρωμοδυναμική (QCD) εξηγεί τις ισχυρές αλληλεπιδράσεις μέσω της συμμετρίας $SU(3)$. Τα γλουόνια διαμεσολαβούν αυτές τις αλληλεπιδράσεις. Ένα φαινόμενο που εμφανίζεται στην ισχυρή αλληλεπίδραση είναι η ασυμπτωτική ελευθερία (asymptotic freedom). Στις χαμηλές ενέργειες η σταθερά σύζευξης a_s της ισχυρής αλληλεπίδρασης παίρνει τιμές της τάξης $O(1)$, ενώ στις υψηλές ενέργειες οι τιμές είναι της τάξης $O(0.1)$. Κατά συνέπεια, στις χαμηλές ενέργειες τα κουάρκ είναι ισχυρά δέσμια, ενώ στις υψηλές ενέργειες μπορούν να θεωρηθούν ελεύθερα σωματίδια, διότι η αλληλεπίδραση γίνεται ασθενέστερη.

Η Ηλεκτρασθενής Ενοποίηση, που διέπει τις μεταβολές γεύσης, χαρακτηρίζεται από παραβίαση της ισοτιμίας (parity violation). Τα αριστερόστροφα φερμιόνια αλληλεπιδρούν μέσω των ασθενών αλληλεπιδράσεων, ενώ οι μεταβάσεις των κουάρκ καθορίζονται από τον πίνακα Cabibbo-Kobayashi-Maskawa (CKM). Η ηλεκτρασθενής θεωρία ενοποιεί τις ηλεκτρομαγνητικές και ασθενείς αλληλεπιδράσεις υπό τη βαθμωτή συμμετρία $SU(2) \times U(1)$.

Θραύση ηλεκτρασθενούς συμμετρίας - μηχανισμός BEH:

Το Καθιερωμένο Πρότυπο προβλέπει τα στοιχειώδη φερμιόνια και τα ασθενή μποζόνια να είναι άμαζα, γεγονός που έρχεται σε αντίφαση με τα πειραματικά δεδομένα με αποτέλεσμα, το 1964 θεωρητικοί φυσικοί ανεξάρτητα, να προτείνουν έναν μηχανισμό, ο οποίος παραβιάζει αυθόρμητα την ηλεκτρασθενή συμμετρία. Ο μηχανισμός αυτός πήρε το όνομά του από τους φυσικούς R. Το δυναμικό BEH αποτελείται από μία διπλέτα μιγαδικών πεδίων με τέσσερις βαθμούς ελευθερίας, οι τρεις από τους οποίους δίνουν μάζα στα ασθενή μποζόνια ($W^\pm, Z$), ενώ η τέταρτη συνιστώσα είναι το βαθμωτό μποζόνιο Higgs. Το δυναμικό BEH αποτελείται από μία διπλέτα μιγαδικών πεδίων με τέσσερις βαθμούς ελευθερίας, οι τρεις από τους οποίους δίνουν μάζα στα ασθενή μποζόνια ($W^\pm, Z$) ενώ η τέταρτη συνιστώσα είναι το βαθμωτό μποζόνιο Higgs. Με αυτόν τον τρόπο, τα κβάντα της ασθενούς αλληλεπίδρασης αποκτούν μάζα και η θεωρία παραμένει επανακανονικοποιήσιμη.

Μηχανισμοί παραγωγής και διάσπασης του Higgs

Η ανακάλυψη του μποζονίου Higgs το 2012 στον LHC επιβεβαίωσε τον μηχανισμό διάσπασης της ηλεκτρασθενούς συμμετρίας του Καθιερωμένου Προτύπου και την αλληλεπίδρασή του με τα

φερμιόνια μέσω της σύζευξης Yukawa.

Η παραγωγή του Higgs πραγματοποιείται μέσω διαφόρων μηχανισμών, όπως σύντηξη γλουονίων (ggF), σύντηξη διανυσματικών μποζονίων (VBF), συσχετιζόμενη παραγωγή με διανυσματικά μποζόνια (VH) και παραγωγή συσχετιζόμενη με top/bottom κουάρκ (ttH, bbH).

Το Higgs μπορεί να διασπασθεί σε ζεύγος φερμιονίων, σε εικονικά μποζόνια ή μέσω βρόγχων σε φωτόνια ή γλουόνια. Η διάσπαση του Higgs ακολουθεί πειραματικά τα θεωρητικά προβλεπόμενα κλάσματα διάσπασης, ευνοώντας τα βαρύτερα σωματίδια όπως $b\bar{b}$, WW^* , και ZZ^* , σύμφωνα με το Καθιερωμένο Πρότυπο. Οι διαδικασίες αυτές επιτρέπουν ακριβείς ελέγχους του τομέα Higgs, ενισχύοντας την εγκυρότητα του Καθιερωμένου Προτύπου.

Αδυναμίες του καθιερωμένου προτύπου

Το Καθιερωμένο Πρότυπο (ΚΠ) έχει επιτύχει σε πολλές προβλέψεις του, με την πιο πρόσφατη να είναι η ανακάλυψη του σωματιδίου Higgs. Παρόλα αυτά, υπάρχουν ακόμα καίρια ζητήματα που περιμένουν να απαντηθούν. Το Καθιερωμένο Πρότυπο εντούτοις είναι μία θεωρία η οποία θα πρέπει να εμπλουτισθεί. Θεωρίες πέραν του Καθιερωμένου Προτύπου (Beyond Standard Model Theories) αναπτύσσονται και καλούνται να συμπληρώσουν τα κενά του ΚΠ. Κάποια από τα ανοιχτά ερωτήματα είναι

- **Βαρύτητα:** Δεν περιλαμβάνει τη βαρύτητα, η οποία περιγράφεται από τη θεωρία της γενικής σχετικότητας και δεν ενσωματώνεται εύκολα στο πλαίσιο της κβαντικής μηχανικής.
- **Μάζα των νετρίνων:** Το Καθιερωμένο Πρότυπο προβλέπει ότι η μάζα των νετρίνων είναι μηδενική, διότι δεν εμφανίζεται ο όρος δεξιόστροφου νετρίνου στην Lagrangian, κάτι που δεν συνάδει με τα πειραματικά δεδομένα.
- **Σκοτεινή ύλη και σκοτεινή ενέργεια:** Το Καθιερωμένο Πρότυπο δεν εξηγεί την σκοτεινή ύλη και την σκοτεινή ενέργεια, οι οποίες πιθανόν αποτελούν το μεγαλύτερο μέρος της μάζας και ενέργειας του γνωστού έως τώρα σύμπαντος.
- **Ασυμμετρία ύλης-αντιύλης:** Το Καθιερωμένο Πρότυπο προβλέπει τη δημιουργία ίσης ποσότητας ύλης και αντιύλης στο πρώιμο Σύμπαν. Κάτι τέτοιο όμως δεν ισχύει και παρατηρείται πλεονάζουσα ποσότητα ύλης, το οποίο υποδεικνύει ότι υπάρχει κάποιος μηχανισμός που ευνοεί την δημιουργία ύλης.

Μοντέλα πέραν του καθιερωμένου προτύπου

Τα μέχρι στιγμής πειραματικά άνω όρια επιτρέπουν ένα σημαντικό κλάσμα διάσπασης του μποζονίου Higgs σε σωματίδια πέρα από το ΚΠ. Η παρατήρηση τέτοιων διασπάσεων του μποζονίου H θα μπορούσε να οδηγήσει στην επέκταση του ΣΜ σε μια πιο πλήρη θεωρία. Πολλά μοντέλα φυσικής πέρα από το ΚΠ προβλέπουν έναν εκτεταμένο τομέα Higgs. Παραδείγματα αποτελούν το $t\bar{t}$ Higgs-doublet model (2HDM), 2HDM+s και NMSSM.

Το μοντέλο 2HDM+s:

Στο 2HDM+s ο τομέας Higgs περιέχει συνολικά επτά βαθμωτά και ψευδοβαθμωτά σωματίδια. Μια από τις βαθμωτές καταστάσεις μπορεί να ταυτιστεί με το μποζόνιο Higgs μάζας 125 GeV ενώ μία από τις ψευδοβαθμωτές καταστάσεις, συμβολίζεται με $^{\prime}a^{\prime}$, μπορεί να είναι αρκετά ελαφρύ ώστε να επιτρέπει την διάσπαση του $H \rightarrow aa$. Αυτές οι διασπάσεις μπορεί να συμβούν με κλάσμα διάσπασης $< 16\%$. Για σχεδόν όλους από τους τέσσερις τύπους και τις τιμές της ελεύθερης παραμέτρου β του 2HDM+s, κυρίαρχη διάσπαση του a είναι $a \rightarrow b\bar{b}$ για $M_b < M_a$.

Το Πείραμα στο CMS:

Ο Μεγάλος Επιταχυντής Αδρονίων (Large Hadron Collider, LHC) είναι ο μεγαλύτερος στον κόσμο, από άποψη μεγέθους αλλά και ενεργειακής κλίμακας, επιταχυντής σωματιδίων που έχει κατασκευαστεί ποτέ. Ο LHC βρίσκεται και λειτουργεί στις εγκαταστάσεις του ευρωπαϊκού κέντρου έρευνας στην πυρηνική φυσική (CERN), κοντά στη Γενεύη.

Ο LHC έχει σχεδιαστεί να επιταχύνει αδρόνια (κυρίως πρωτόνια). Δύο δέσμες σωματιδίων υψηλής ενέργειας επιταχύνονται προς αντίθετες κατευθύνσεις σε δύο ξεχωριστούς αγωγούς. Για την επιτάχυνση της δέσμης σωματιδίων, το CERN διαθέτει ένα σύμπλεγμα μικρότερων επιταχυντικών διατάξεων που έχουν ως σκοπό τη μερική επιτάχυνση της δέσμης πριν εισέλθει στον κεντρικό επιταχυντή, τον LHC, για το τελικό στάδιο της επιτάχυνσης και τη σύγκρουση των δύο δέσμεων.

Οι δέσμες πρωτονίων συγκρούονται σε τέσσερα προκαθορισμένα σημεία του δακτυλίου του επιταχυντή, όπου βρίσκονται τοποθετημένες οι ανιχνευτικές διατάξεις των τεσσάρων μεγαλύτερων πειραμάτων του LHC: ATLAS - A Toroidal LHC Apparatus, CMS - Compact Muon Solenoid, LHCb - LHC Beauty και ALICE - A Large Ion Collider Experiment. Ο κάθε ανιχνευτής σωματιδίων έχει κατασκευαστεί για τη μελέτη διαφορετικών τομέων στη φυσική υψηλών ενεργειών. Οι δύο πρώτοι ανιχνευτές, ATLAS και CMS, αφορούν πειράματα γενικού ενδιαφέροντος και επικεντρώνονται στην αναζήτηση νέας φυσικής και θεωριών εκτός του ΚΠ.

Ο ανιχνευτής CMS

Οι τέσσερις ανιχνευτικές διατάξεις από τις οποίες απαρτίζεται ο CMS είναι οι ακόλουθες:

- **Ανιχνευτές Τροχιάς - Tracker Detector:** Σύστημα ανίχνευσης τροχιών που δίνει πληροφορίες για το είδος και την ορμή των σωματιδίων που προέρχονται από τη σύγκρουση των δέσμεων.
- **Το Ηλεκτρομαγνητικό Θερμιδόμετρο (ECAL):** Θερμιδόμετρο για την ανίχνευση και μελέτη φωτονίων και ηλεκτρονίων.
- **Το Αδρονικό Καλορίμετρο** μετράει την εναπόθεση ενέργειας των φορτισμένων αλλά και ουδέτερων αδρονίων στον ανιχνευτή. Επιπλέον, μετρώντας την ασυμμετρία στο εγκάρσιο επίπεδο παρέχει μία έμμεση μέτρηση των διερχόμενων σωματιδίων που δεν αλληλεπιδρούν με το υλικό του ανιχνευτή, όπως τα νετρίνα.
- **Ανιχνευτές Μιονίων - Muon detectors:** Ένα ανιχνευτικό σύστημα υψηλής απόδοσης για τη μελέτη και ανίχνευση μιονίων.

Ο Σωληνοειδής Μαγνήτης (Solenoid Magnet) στον οποίο οφείλει εν μέρει το όνομα του ο CMS. Ο μαγνήτης παράγει μαγνητικό πεδίο έντασης περίπου 4 T παράλληλα στον άξονα της δέσμης, με αποτέλεσμα να καμπυλώνει τις τροχιές των σωματιδίων κάθετα στον άξονα αυτής.

Το Σύστημα Σκανδαλισμού και Ανάκτησης δεδομένων (TriDAC System) βελτιστοποιεί τον ρυθμό καταγραφής γεγονότων κατά τη συλλογή δεδομένων από τους ανιχνευτές, επιλέγοντας να κρατήσει τα γεγονότα που παρουσιάζουν φυσικό ενδιαφέρον. Το πρώτο επίπεδο Level-1 (L1) trigger αποτελείται από ειδικό εξοπλισμό (hardware- based system) που επεξεργάζεται ταχύως τα δεδομένα από τα καλορίμετρα και τους ανιχνευτές μιονίων. Τα επιλεγμένα γεγονότα στη συνέχεια περνούν στο δεύτερο επίπεδο που ονομάζεται High-Level Trigger (HLT), και αναλαμβάνει την ανακατασκευή γεγονότων, χρησιμοποιώντας ένα λογισμικό (software- based system) σχεδιασμένο για ταχεία επεξεργασία δεδομένων.

Φυσική συγκρούσεων πρωτονίων:

Η κατανομή της ορμής του πρωτονίου στα παρτόνια του, όπως τα κουάρκ και τα γκλουόνια, περιγράφεται από τις συναρτήσεις κατανομής ορμής (PDFs). Αυτές εκφράζουν την πιθανότητα ένα παρτόνιο να φέρει κλάσμα της ορμής του πρωτονίου και εξαρτώνται από την κλίμακα ενέργειας Q^2 . Για υψηλές τιμές του Q^2 , η ανάλυση επικεντρώνεται σε μικρότερες αποστάσεις μέσα στο πρωτόνιο, αποκαλύπτοντας με μεγαλύτερη λεπτομέρεια την εσωτερική του δομή, ενώ για μικρότερες τιμές οι PDFs είναι λιγότερο λεπτομερείς και επικεντρώνονται στα κουάρκ σθένους. Στον LHC η σκληρή σκέδαση γίνεται μεταξύ των παρτονίων των δύο πρωτονίων, τα οποία φέρουν μικρό ποσοστό της ορμής τους. Η ενέργεια κέντρου μάζας για αυτή τη διαδικασία είναι μικρότερη από αυτή των πρωτονίων και μπορεί να αναλυθεί χρησιμοποιώντας τις ενεργές διατομές για τα παρτόνια και τις αντίστοιχες συναρτήσεις πιθανότητας.

Η ανακατασκευή αδρονικών jets είναι κρίσιμη στην πειραματική φυσική, καθώς συνδέει τα ανιχνεύσιμα σωματίδια με τα αρχικά παρτόνια. ένα jet χωρίζεται σε τρία επίπεδα : 1. Το παρτονικό επίπεδο - Parton Level 2. Το σωματιδιακό επίπεδο - Particle Level 3. Το επίπεδο του ανιχνευτή - Detector Level

Υπάρχουν διάφοροι αλγόριθμοι για αυτή τη διαδικασία που χρησιμοποιούν δεδομένα από τα καλλορίμετρα και την ανίχνευση τροχιών για να προσδιορίσουν τα jets.

Επιπλέον, η διαδικασία **b-tagging** χρησιμοποιείται για την ταυτοποίηση των jets που προέρχονται από b κουάρκ, βασισμένη σε χαρακτηριστικά όπως η υψηλή μάζα και ο μεγάλος χρόνος ζωής των b κουάρκ. Ο αλγόριθμος b-tagging χρησιμοποιεί παραμέτρους όπως η απόσταση impact parameter και την ανάλυση της δευτερεύουσας κορυφής διάσπασης.

Ανάλυση:

Η διεργασία σήματος:

Το σήμα που μελετάται είναι η εξωτική διάσπαση του Higgs σε δύο ψευδοβαθμωτά σωματίδια a , το κάθε ένα από τα οποία αποδιεγείρεται άμεσα σε ένα ζεύγος b κουάρκ. Επιπλέον, η διάσπαση $a \rightarrow b\bar{b}$ είναι κυρίαρχη σε όλους τους τύπους του 2HDM+S, καθώς το κλάσμα της διάσπασης αυτής τείνει στη μονάδα, όταν $M_a > 10 \text{ GeV}$. Η μάζα του a να είναι σε συμφωνία με τους κινηματικούς περιορισμούς: $2M_b < M_a < M_h/2$.

Η παρούσα μελέτη λοιπόν εξερευνά την παραγωγή νέων ελαφριών μποζονίων, στο εύρος μάζας $12 \leq M_a \leq 60 \text{ GeV}$. Επομένως, γι' αυτό το εύρος μάζας του "a" μπορούμε να θεωρήσουμε ότι $BR(a \rightarrow b\bar{b}) \approx 1$. Στο σήμα, η παραγωγή του Higgs γίνεται σε συσχέτιση με το μποζόνιο Z ($\sigma_{Zh} = 0.88 \text{ pb}$), ενώ αυτό στη συνέχεια διασπάται είτε λεπτονικά, με κλάσμα διάσπασης περίπου ίσο με 10%, είτε σε νεutrino, με κλάσμα διάσπασης περίπου ίσο με 20%.

Οι λεπτονικές τελικές καταστάσεις χαρακτηρίζονται από πολύ καθαρές πειραματικές υπογραφές (2-lepton channel), ενώ το κανάλι χωρίς λεπτόνια στην τελική κατάσταση (0-lepton channel) μελετάται πρώτη φορά για την συγκεκριμένη εξωτική διάσπαση του Higgs.

Μελετώνται αρχικά κάποιες βασικές κινηματικές μεταβλητές και χαρακτηριστικά του σήματος σε επίπεδο generator. Βασικές παρατηρήσεις είναι ότι η εγκάρσια ορμή του ζεύγους ($b\bar{b}$) που προέρχεται από το ίδιο "a" μποζόνιο αυξάνεται καθώς η μάζα του μειώνεται, ενώ παράλληλα η γωνιακή απόσταση τους μειώνεται, κάτι που μπορεί να αποδειχθεί από την κινηματική. Μελετώντας το $\Delta R(b\bar{b})$ ως συνάρτηση του p_T παρατηρούμε ότι για μεγάλες τιμές εγκάρσιας ορμής του ζεύγους, αυτό θα μπορούσε να ανακατασκευαστεί εντός ενός κώνου ακτίνας $\Delta R = 0.8$ για $M_a \leq 40 \text{ GeV}$. Έτσι, ορίζουμε δύο περιοχές, την resolved για μάζες του "a" στο εύρος $20 \leq M_a \leq 60$ και την

merged για το εύρος μαζών $15 \leq M_a \leq 40$ GeV. Στην resolved κάθε b-quark θα ανακατασκευαστεί μέσω AK4 jet και θα ταυτοποιηθεί μέσω του DeepJet b-tagging αλγορίθμου, ενώ στην merged δύο b-quarks θα ανακατασκευαστούν εντός ενός AK8 jet και θα ταυτοποιηθούν μέσω του ParticleNet_MD double-btag discriminator.

Διεργασίες υποβάθρου

Οι διαδικασίες υποβάθρου που συνυπάρχουν με το σήμα που ερευνάται, εμπίπτουν στο φάσμα του ΚΠ. Η τελικές καταστάσεις τέτοιων διαδικασιών συνήθως “μιμούνται” το σήμα, με αποτέλεσμα η διαφοροποίησή τους από αυτό να είναι απαιτητική. Στην παρούσα ανάλυση, επικρατές υπόβαθρο είναι το Drell-Yan, και το υπόβαθρο tt+jets στο 2-lepton channel, ενώ η all hadronic QCD καθώς και τα υπόβαθρα V+jets αποτελούν μεγάλο υπόβαθρο στο 0-lepton channel.

Ορισμός των αντικειμένων της ανάλυσης (object configuration)

Στην συνέχεια γίνεται ο ορισμός των αντικειμένων (object configuration) της ανάλυσης με βάση τόσο τα γεωμετρικά κριτήρια αποδοχής του ανιχνευτή, όσο και των trigger. Περιγράφονται αναλυτικά τα κριτήρια αποδοχής για τα λεπτόνια, την MET, τα AK4 και AK8 jet. Για το b-tagging των AK4 jet χρησιμοποιούμε το medium working point του DeepJets αλγορίθμου, ενώ για τα AK8 jets η στρατηγική μας αναλύεται με λεπτομέρεια στην ανάλυση.

Κριτήρια επιλογής γεγονότων

Τα κριτήρια επιλογής γεγονότων βοηθούν στη μείωση του όγκου των δεδομένων και έπειτα στη μελέτη γεγονότων που έχουν μεγαλύτερη πιθανότητα να είναι το σήμα. Οπότε, θέτουμε τις απαιτήσεις για το λεπτονικό σύστημα: τον αριθμό λεπτονίων κατωφλίου της MET(0-lepton) εύρος αναλωίωτης μάζας λεπτονίων (2-lepton) και για το αδρονικό σύστημα. Για τον προσδιορισμό της τιμής κατωφλίου της MET έγινε τόσο μελέτη της απόδοσης του MET trigger για το σήμα μας, καθώς και περαιτέρω διερεύνηση για να επιτύχουμε απόρριψη “fake MET” γεγονότων. Για το αδρονικό σύστημα, απαιτούμε τουλάχιστον 3 b-tagged jets resolved περιοχή, ενώ τουλάχιστον ένα fat jet και ένα cross cleaned b-tagged jet στη merged περιοχή.

Πολυμεταβλητή Ανάλυση (MVA)

Στη συνέχεια, γίνεται χρήση του TMVA, που γενικά στοχεύει στην επίλυση προβλημάτων κατηγοριοποίησης (classification problems). Συγκεκριμένα, ο κύριος σκοπός είναι η διαφοροποίηση του σήματος από το συντριπτικό υπόβαθρο, με τη χρήση της τεχνικής των Boosted Decision Trees (BDTs). Ως μεταβλητές εισόδου ορίζονται έτσι ώστε να έχουν μεγάλη διαχωριστική ικανότητα μεταξύ σήματος και υποβάθρου. Επίσης θα πρέπει να είναι όσο το δυνατόν λιγότερο συσχετισμένες μεταξύ τους. Για το boosting του BDT επιλέχθηκε Gradient boosting καθώς προσαρμόστηκαν και οι υπερπαράμετροι του training. Ένας BDT εκπαιδεύτηκε για κάθε κανάλι και περιοχή, για τον οποίο ως σήμα ορίστηκε το άθροισμα όλων των “a” mass samples της αντίστοιχης περιοχής. Έτσι τελικά πήραμε την κατανομή του BDT (output BDT score) για την οποία έγινε μελέτη ως προς την παρατηρησιμότητα του σήματος, των καμπυλών ROC καθώς και έλεγχος για overtraining. Χρησιμοποιώντας τις output BDT κατανομές θα πραγματοποιηθεί στην συνέχεια η binned shape analysis. Πριν από αυτό για να μεγιστοποιήσουμε την παρατηρησιμότητα στην περιοχή σήματος, δηλαδή στην περιοχή της κατανομής BDT κοντά στη μονάδα, υλοποιούμε μία βελτιστοποίηση του binning του.

Μέθοδος Μέγιστης Πιθανοφάνειας

Έπειτα, με χρήση του RooFit ακολουθεί ο ορισμός των μοντέλων που αντιστοιχούν στην υπόθεση

ύπαρξης μόνο του υποβάθρου (model_0) και στην ύπαρξη και σήματος πέρα από το υπόβαθρο (model_1). Από το model_0 παράγουμε Monte Carlo binned ψευδοδεδομένα και πραγματοποιούμε προσαρμογή μέγιστης πιθανοφάνειας (extended binned maximum likelihood fit) του model_1 σε αυτά τα ψευδοδεδομένα. Έτσι, μετά από κάθε προσαρμογή, λαμβάνουμε τις τιμές των παραμέτρων που μεγιστοποιούν τη συνάρτηση πιθανοφάνειας. Οι παράμετροι αυτές είναι ο αναμενόμενος αριθμός γεγονότων σήματος (expected signal yield) και υποβάθρου κάθε διεργασίας (expected background yields). Στην περίπτωση μας, η παράμετρος ενδιαφέροντος (POI) είναι ο αναμενόμενος αριθμός γεγονότων σήματος, ενώ για τον αναμενόμενο αριθμό γεγονότων των διαφόρων υποβάθρων εισάγονται προσεγγιστικά οι συστηματικές αβεβαιότητες μέσω Gaussian κατανομών που περιορίζουν το επιτρεπτό εύρος τιμών τους κατά την προσαρμογή. Στη συνέχεια εκτελούμε κάποια τεστ για τον προσδιορισμό της ποιότητας της προσαρμογής.

Εξαγωγή άνω ορίων με χρήση Bayesian μεθόδου

Τέλος, χρησιμοποιώντας Bayesian στατιστική μέθοδο, συγκεκριμένα την υλοποίηση του Monte Carlo Markov Chain αλγορίθμου του RooStats και θεωρώντας ομοιόμορφη prior κατανομή σε ένα καθορισμένο αρκετά μεγάλο διάστημα για την παράμετρο ενδιαφέροντος, εξάγουμε διαστήματα εμπιστοσύνης (credible intervals) για την posterior κατανομή της παραμέτρου ενδιαφέροντος. Από την posterior προκύπτουν άμεσα τα άνω όρια με 95% επίπεδο εμπιστοσύνης για τον αναμενόμενο αριθμό γεγονότων σήματος και κατ' επέκταση για την ισχύ του αναμενόμενου σήματος. (expected signal strength). Έτσι, τρέχοντας τον MCMC αλγόριθμο για πολλά toy data , εξάγουμε τις τελικές μπάντες των άνω ορίων για την ισχύ του αναμενόμενου σήματος και κατ' επέκταση για το κλάσμα διάσπασης $\text{Br}(h \rightarrow aa \rightarrow 4b)$ για όλες τις υποθέσεις μάζας του "a" σε κάθε κανάλι και περιοχή. Το τελευταίο συμβαίνει εφόσον θεωρούμε $\sigma_{Zh} = \sigma_{Zh}^{\text{SM}}$.

Αποτελέσματα

Τα εξαγόμενα άνω όρια για την ισχύ του αναμενόμενου σήματος είναι σε συμφωνία με προηγούμενες αναλύσεις που έχουν διεξαχθεί για το λεπτονικό κανάλι. Συγκεκριμένα στην resolved περιοχή, το $\text{Br}(h \rightarrow aa \rightarrow 4b)$ φαίνεται να μειώνεται για μεγαλύτερες μάζες του "a", παίρνοντας τιμές αρκετά μικρότερες της μονάδας ($M_a < 30$). Ανάλογη τάση παρατηρείται στην resolved περιοχή στο κανάλι χωρίς λεπτόνια. Αντίθετα, για χαμηλές μάζες, λόγω του ότι η παρατηρησιμότητα του σήματος μειώνεται, τα άνω όρια λαμβάνουν τιμές μεγαλύτερες της μονάδας.

Στην merged περιοχή, η παρατηρησιμότητα του σήματος ανακτάται για χαμηλές μάζες του "a". Έτσι λαμβάνουμε χαμηλότερα άνω όρια για M_a στο διάστημα 15-25 GeV στο 0-lepton κανάλι, ενώ μόνο για την μάζα 15 GeV στο 2-lepton κανάλι.

Contents

1	Theoretical framework	1
1.1	The foundations of the Standard Model	1
1.1.1	The path to the Standard Model	1
1.1.2	Elementary Particles of the Standard Model	1
1.1.3	Path to the Standard Model Lagrangian formulation	4
1.1.4	Electroweak Symmetry Breaking and the Higgs Mechanism	8
1.1.5	The Higgs Boson at the LHC	11
1.2	Beyond the Standard Model	16
1.2.1	Shortcomings of the Standard Model	16
1.2.2	Theories Beyond the Standard Model	18
1.2.3	The Two-Higgs-Doublet Model (2HDM)	19
1.2.4	2HDM+s Model	19
2	The CMS Experiment at the LHC	21
2.1	The Large Hadron Collider	21
2.2	Compact Muon Solenoid	24
2.3	Electromagnetic calorimeter	26
2.4	The CMS Hadronic Calorimeter (HCAL)	28
2.5	The Superconducting Solenoid	30
2.6	The Muon System	31
2.7	Trigger and Data Acquisition System	32
3	Physics of pp collisions at the LHC	34
3.1	Hard scattering processes	34
3.2	Monte Carlo Event Generators	37
3.3	Particle reconstruction and identification	37
3.3.1	The Particle Flow Algorithm	37
3.3.2	Reconstruction of Jets	38
3.3.3	b-quark identification	40
3.4	Kinematic variables	41
4	Analysis	44
4.1	The $h \rightarrow aa \rightarrow 4b$ decay	44
4.1.1	2HDM+s Model	44
4.1.2	The $h \rightarrow 4b$ Decay Channel	46
4.2	Signal process	47

4.2.1	Generator level analysis	49
4.3	Background processes	53
4.4	Trigger selection	57
4.4.1	MET trigger efficiency	57
4.5	Physics object configuration	59
4.6	Event Selection Criteria	67
4.6.1	Event flow Tables	69
4.7	Multivariable Analysis	71
4.7.1	Toolkit for Multivariate Analysis	71
4.7.2	Boosted Decision Trees	72
4.7.3	BDT input variables	75
4.7.4	BDT response	88
4.7.5	Receiver Operating Characteristic Curve	88
4.7.6	BDT Cut Efficiencies	91
4.8	BDT binned shape analysis	95
4.8.1	BDT binning optimization	95
4.9	Maximum likelihood fit	98
4.9.1	Maximum Likelihood Estimation	99
4.9.2	Extended Maximum Likelihood	99
4.9.3	Binned Extended Maximum Likelihood (Binned EML)	100
4.9.4	Systematic uncertainties and fit ranges	101
4.9.5	Fit results	102
4.9.6	Goodness of Fit	110
4.10	Upper limits to the signal yield using Bayesian method	113
4.10.1	Introduction to Bayesian Statistics	113
4.10.2	Credible Intervals	114
4.10.3	Upper limits on the signal yield using Bayesian methods	116
5	Results: Model independent Upper Limits	118
5.1	Outlook	121
A	CMS MC Samples	123
B	Expected signal yield in all the "a" mass samples	125
C	AK8 Jet Characteristics	126
D	Wh contribution to the 0-lepton channel	128
E	BDT input variable distributions	133
E.0.1	The resolved regime of the 0-lepton channel	133
E.0.2	The merged regime of the 0-lepton channel	135
E.0.3	The resolved regime of the 2-lepton channel	138
E.0.4	The merged regime of the 2-lepton channel	140
F	Nuisance parameters and their uncertainties	144

G	Posterior distributions of the expected signal yield	145
H	Upper limits in logarithmic scale	151

Chapter 1

Theoretical framework

1.1 The foundations of the Standard Model

This chapter provides a description of the Standard Model, its limitations, and an introduction to the phenomenological foundations of extended Higgs sectors and focuses on the 2HDM model that is relevant to the main topic of this thesis.

1.1.1 The path to the Standard Model

The Standard Model (SM) is a comprehensive model that summarizes the current understanding of the basic building blocks of matter and the fundamental sub-atomic interactions, except gravity. More precisely, SM is a relativistic quantum field theory (QFT) based on the $SU(3) \times SU(2) \times U(1)$ symmetry group. Within the SM framework, elementary particles can be classified according to their quantum numbers, as explained in the following section.[1] The development of the Standard Model spans over a century of scientific discoveries and theoretical advancements.

The path to the SM began with J.J. Thomson's discovery of the electron in the 19th century, followed by the identification of the nucleus and neutron in the early 20th century. Quantum mechanics emerged in the 1920s, evolving into quantum field theory (QFT) and quantum electrodynamics (QED) by the 1940s. In the 1950s, the introduction of Yang-Mills theory incorporated non-Abelian gauge groups into QFT. This period also witnessed significant progress in accelerator technology, leading to the discovery of numerous subatomic particles. These discoveries prompted the development of the quark model and quantum chromodynamics (QCD). The SM theoretical foundations were formulated during the 1960s and 1970s when the electroweak theory as well as Higgs Mechanism were proposed. This fact motivated a multitude of ground-breaking experiments resulting in new observations.[2] For example, the charm quark was found in 1974, the bottom quark in 1977, the W and Z bosons in 1983, the top quark in 1995, and finally the Higgs boson in 2012. Most important particle discoveries that were the milestone to SM formulation and provided experimental proof are shown in the following schematic timeline(Figure: 1.1).

1.1.2 Elementary Particles of the Standard Model

Bosons: mediators of interactions

The Standard Model combines electromagnetic, strong, and weak interactions in a single relativistic quantum field theory. These interactions are mediated by vector bosons with spin 1 (in units

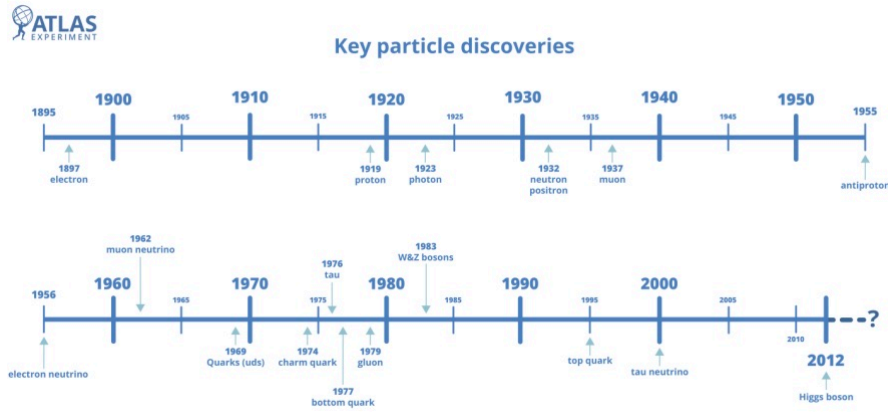


Figure 1.1: The timeline of SM elementary particles discovery [3]

of $\hbar$), known as gauge bosons. The three gauge groups of the SM correspond to three of the four fundamental forces: the electromagnetic force, the strong force, and the weak force.

- **The photon γ** , which mediates the electromagnetic force, remains massless, preserving the U(1) gauge symmetry associated with electromagnetism. In other words, photons couple to the electric charge of particles and since photon itself carries no charge, no direct interaction between photons, outside loops, is possible.
- **Gluons g** are the force carriers of the strong interaction, mediating the interactions between particles with color charges (red, green, and blue). Unlike photons in electromagnetism, gluons themselves possess color charge, enabling self-interaction. This unique property results in eight distinct gluon configurations, each carrying a superposition of color and anti-color charges. Despite being massless, gluons' self-interaction leads to two significant phenomena in strong force dynamics: confinement and asymptotic freedom. Confinement refers to the observation that the strong force between particles does not diminish with distance. Instead, the energy in the field increases linearly with separation, potentially creating new particle-antiparticle pairs if sufficient energy is available. This phenomenon restricts strongly interacting particles to exist only in color-neutral bound states called hadrons. Conversely, asymptotic freedom describes the weakening of the strong coupling at high energies, allowing quarks to behave as quasi-free particles in these conditions. This behavior enables the application of perturbation theory in high-energy regimes. The theory of Quantum Chromodynamics (QCD) provides a comprehensive framework for understanding these complex interactions within the context of relativistic quantum field theory. [4]
- **W and Z bosons** are the carriers of the weak interaction. They obey Bose-Einstein statistics, which means that bosons with the same quantum numbers can occupy the same quantum state. Their spin is an integer multiple of $\hbar$. $W^\pm$ and Z with masses of 80.4 GeV and 91.2 GeV respectively, are exceptionally massive elementary particles. Their substantial mass contributes to the short-range nature of the weak interaction. While the Z boson is involved in weak interactions of the "neutral-current" type, the $W^\pm$ bosons participate in weak interactions of the "charged-current". Thus, W boson-mediated interactions are characterized by flavor transitions, necessitated by charge conservation principles. These transitions

manifest in various ways: Charged leptons transform into neutrinos (and vice versa), involving a charge change from $\pm 1e$ to 0. Up-type quarks convert to down-type quarks, with a charge shift from $\pm 2/3e$ to $\mp 1/3e$. Down-type quarks transform into up-type quarks, changing charge from $\pm 1/3e$ to $\mp 2/3e$. Interestingly, the Standard Model allows quark transitions to occur across generations. For instance, an up-type u quark can transition to a down-type s quark, rather than being confined to $u \rightarrow d$ transitions. This phenomenon arises from the discrepancy between quark flavor eigenstates in weak interactions and their mass eigenstates. The mixing between these states is quantified by the CKM matrix.

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \quad (1.1)$$

- **Higgs boson** (detailed in sections 1.1.4-1.1.5) is a massive, 0-spin, scalar boson and describes the interactions within the Higgs field.

Fermions : Building blocks of matter

Fermions carry a spin of $1/2$ in units of $\hbar$ and follow the Fermi-Dirac statistics and subsequently the Pauli exclusion principle. The fermions are subdivided into leptons and quarks.

There are six flavours of quarks: up (u), down (d), charm (c), strange (s), top (t), and bottom (b). Each quark flavour comes in three colour charges: red, green, and blue. The u , c , and t quarks have a positive electric charge of $+2/3 e$ and are called up-type quarks, while the down-type quarks, d , s , and b , have $-1/3 e$ electric charges. Quarks participate in both electroweak and strong interactions, with the latter binding them into colour-neutral composite states known as hadrons (e.g., protons and neutrons in atomic nuclei).

Leptons, on the other hand, do not possess color charge and, therefore, do not engage in strong interactions. There are six leptons: the electron (e), muon (μ), and tau (τ), which carry electric charge, each accompanied by a corresponding neutral particle, the electron neutrino (ν_e), muon neutrino (ν_μ), and tau neutrino (ν_τ). All leptons participate in weak interactions. More precisely, all charged leptons feature an electric charge of $Q = -e$ and can interact via the electromagnetic and the weak force. The neutrinos, on the other hand, feature a charge of $Q = 0$ and can consequently only interact via the weak interaction, which only interacts with left-handed neutrinos. Hence, right-handed neutrinos are not part of the SM and have also not been observed so far.

There exist three fermionic generations, organized in isospin doublets with increasing mass. They consist of a charged lepton and its corresponding neutrino, and an up-type and a down-type quark for leptons and quarks, respectively.

- First generation : up and down quarks, electron, and electron neutrino
- Second generation: charm and strange quarks, muon, and muon neutrino
- Third generation: top and bottom quarks, tau, and tau neutrino

Every fermion has a corresponding anti-particle, which possesses opposite quantum numbers but with the same mass and spin [4], [5]

Subsequently, the SM is a comprehensive theory unifying all known elementary particles and forces, except gravity. These 17 SM particles: 12 fermions, 4 gauge bosons, and one scalar boson, as well as their properties, are summarized in Figure 3.4 . The Higgs boson will be discussed in detail in a later section.

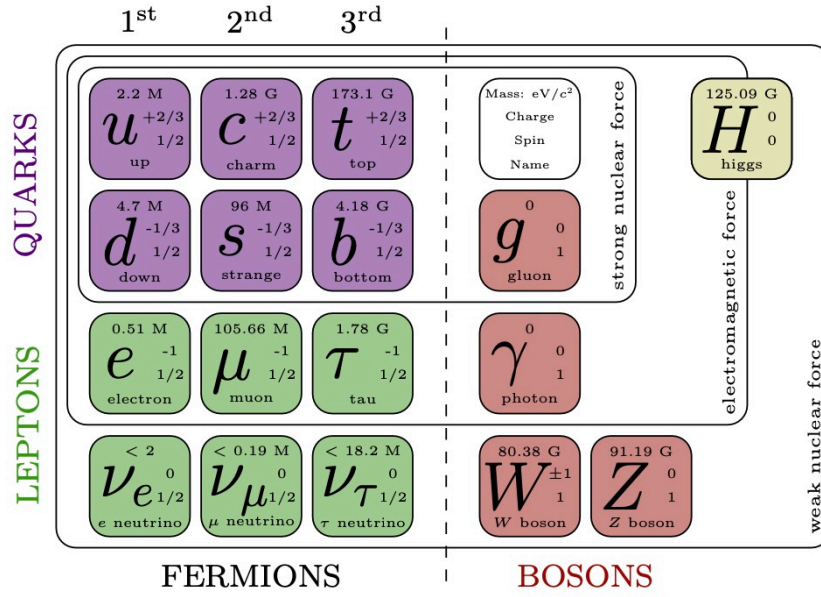


Figure 1.2: A visualization of all elementary particles in the SM and their properties

1.1.3 Path to the Standard Model Lagrangian formulation

As already mentioned, the SM defines a Lagrangian density that combines the electromagnetic, strong, and weak interactions in a single relativistic quantum field theory, and the interactions are mediated by the gauge bosons described in the previous section. In the framework of quantum field theory, particles and forces are described by fields. The Lagrangian density $\mathcal{L}$ characterizes the kinematics and interactions of these fields. It follows an analytical description of how the Lagrangian is derived for each of the three SM interactions.

Quantum Electrodynamics (QED)

Quantum Electrodynamics (QED) is the quantum field theory describing the interaction of the electromagnetic field with charged matter particles, formulated within a relativistic framework. The QED Lagrangian density, $\mathcal{L}_{QED}$, emerges from applying local U(1) gauge invariance to the Dirac Lagrangian:

$$\mathcal{L}_{Dirac} = i\bar{\psi}(x)\gamma^\mu\partial_\mu\psi(x) - m\bar{\psi}(x)\psi(x) \quad (1.2)$$

Here, $\psi(x)$ denotes the fermion field (Dirac spinor) and $\bar{\psi}(x)$ its adjoint. Under U(1) transformation:

$$\psi(x) \rightarrow \psi'(x) = \psi(x)e^{iq\alpha(x)} \quad \bar{\psi}(x) \rightarrow \bar{\psi}'(x) = \bar{\psi}(x)e^{-iq\alpha(x)} \quad (1.3)$$

where q represents charge and $\alpha(x)$ a phase. This yields:

$$\mathcal{L}' = \mathcal{L}_{Dirac} - q\bar{\psi}(x)\gamma^\mu\psi(x)\partial_\mu\alpha(x) \quad (1.4)$$

Gauge invariance is achieved by introducing a covariant derivative:

$$D_\mu = \partial_\mu + iqA_\mu(x) \quad (1.5)$$

$A_\mu(x)$ is a vector field transforming as $A_\mu(x) \rightarrow A_\mu(x) + \partial_\mu \theta(x)$. The QED Lagrangian becomes:

$$\mathcal{L}_{QED} = \bar{\psi}(x) i \gamma^\mu \partial_\mu \psi(x) - m \bar{\psi}(x) \psi(x) - q \bar{\psi}(x) \gamma^\mu \psi(x) A_\mu(x) \quad (1.6)$$

A kinetic term for the photon field is added:

$$\mathcal{L}_A = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \quad (1.7)$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the field strength tensor. The complete QED Lagrangian reads:

$$\mathcal{L}_{QED} = \bar{\psi} i \gamma^\mu \partial_\mu \psi - m \bar{\psi} \psi - q \bar{\psi} \gamma^\mu \psi A_\mu - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} \quad (1.8)$$

The current density $J^\mu = q \bar{\psi}(x) \gamma^\mu \psi(x)$ ensures charge conservation. The coupling strength $q = -e$ relates to the fine structure constant $\alpha = \frac{e^2}{4\pi}$. QED's coupling constant α varies with energy scale Q^2 , from $\sim \frac{1}{137}$ at low energies to $\sim \frac{1}{128}$ at the Z boson mass scale.

Quantum Chromodynamics (QCD)

Quantum Chromodynamics (QCD) describes the strong interaction between color-charged particles, mediated by gluons. In analogy to QED, quarks carry color charge (red, green, blue) and anti-quarks carry anti-color charge.

Similar to QED, the QCD Lagrangian density can be derived by enforcing gauge invariance under the SU(3) group. The resulting Lagrangian density is:

$$\mathcal{L}_{QCD} = i \bar{\psi}(x) \gamma^\mu D_\mu \psi(x) - m \bar{\psi}(x) \psi(x) - \frac{1}{4} G_{\mu\nu}^a G^{a\mu\nu}, \quad (1.9)$$

where the vector fields represent eight colored, massless, and electrically neutral gluons. The covariant derivative is $D_\mu = \partial_\mu - i g_s T^a G_\mu^a$, where g_s is the QCD coupling constant and T^a are the 3x3 matrices in color space related to the Gell-Mann matrices (t^a) by $T^a = \frac{t^a}{2}$. For $\mathcal{L}_{QCD}$ to be gauge-invariant, the mass term must be equal to zero.

The gluon field strength tensor is:

$$G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g_s f^{abc} G_\mu^b G_\nu^c, \quad (1.10)$$

where f^{abc} is the structure constant of the SU(3) group.

A crucial consequence of gluon self-interaction is the behavior of the strong coupling constant (α_s) with respect to Q^2 :

$$\alpha_s(Q^2) = \frac{12\pi}{(33 - 2n_f) \log \left(\frac{Q^2}{\Lambda_{QCD}^2} \right)}, \quad (1.11)$$

where n_f is the number of quark flavors with mass $m_q^2 \ll Q^2$. As Q^2 increases ("asymptotic freedom"), α_s decreases, allowing perturbative QCD calculations. Conversely, at low Q^2 , the coupling becomes large, entering a non-perturbative regime at Λ_{QCD} , leading to "color confinement" where quarks and gluons form color-neutral hadrons.

Weak Interaction

The weak interaction, initially proposed to explain nuclear β decays, violates parity (P) conservation, as demonstrated by the Wu experiment. This led to the development of the "vector-axial" (V-A) theory.

Dirac spinors are decomposed into left-handed and right-handed components:

$$\psi(x) = \psi_L(x) + \psi_R(x) \quad (1.12)$$

These components are obtained using projection operators P_L and P_R :

$$\psi_L(x) = P_L \psi(x) = \frac{1}{2}(1 - \gamma^5) \psi(x) \quad (1.13)$$

$$\psi_R(x) = P_R \psi(x) = \frac{1}{2}(1 + \gamma^5) \psi(x) \quad (1.14)$$

The V-A theory relies on the distinct parity properties of vector currents ($\bar{\psi}(x)\gamma^\mu\psi(x)$) and axial-vector currents ($\bar{\psi}(x)\gamma^\mu\gamma^5\psi(x)$). Vector currents conserve parity, while axial-vector currents transform with a sign change under parity transformation.

To introduce parity violation into the weak interaction, the term $\bar{\psi}(x)\gamma^\mu\psi(x)$ from the QED Lagrangian is modified as:

$$\bar{\psi}(x) \left[\gamma^\mu \frac{1}{2}(1 - \gamma^5) \right] \psi(x) = \frac{1}{2} [\bar{\psi}(x)\gamma^\mu\psi(x) - \bar{\psi}(x)\gamma^\mu\gamma^5\psi(x)] \quad (1.15)$$

The right-hand side shows the V-A structure, explicitly violating parity conservation. This formulation also implies that weak interactions involve only left-handed (LH) fermions and right-handed (RH) antifermions.

Beyond parity violation, the weak interaction violates charge conjugation (C) and parity (P) conservation. This is incorporated into the Standard Model through the Cabibbo-Kobayashi-Maskawa (CKM) matrix, which describes quark mixing.

Electroweak Unification

Electroweak theory unites the electromagnetic and weak interactions, formulated with $SU(2) \times U(1)$ gauge invariance. This framework involves two conserved quantities: weak isospin (I_3) associated with $SU(2)$ and hypercharge (Y) associated with $U(1)$. [1]

Left-handed and right-handed fermions are treated differently. LH fermions (ψ_L) are weak isospin doublets with non-zero I_3 and Y values. RH fermions (ψ_R) transform as singlets with zero isospin and non-zero hypercharge. The covariant derivatives also differ for doublets and singlets [6]:

$$D_{\mu,L}\psi_L = \left(\partial_\mu + igW_\mu^i \frac{\sigma_i}{2} + ig'B_\mu \frac{Y}{2} \right) \psi_L \quad (1.16)$$

$$D_{\mu,R}\psi_R = \left(\partial_\mu + ig'B_\mu \frac{Y}{2} \right) \psi_R \quad (1.17)$$

where σ_i are Pauli matrices ($SU(2)$ generators), and g and g' are coupling constants.

Electroweak unification links I_3 , Y , and electric charge q through the Gell-Mann–Nishijima formula:

$$q = I_3 + \frac{1}{2}Y \quad (1.18)$$

This integrates electromagnetic and weak interactions under a single electroweak force. Mixing of the vector fields W_μ^i and B_μ yields four electroweak gauge bosons:

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2) \quad (1.19)$$

$$Z_\mu = W_\mu^3 \cos \theta_W - B_\mu \sin \theta_W \quad (1.20)$$

$$A_\mu = W_\mu^3 \sin \theta_W + B_\mu \cos \theta_W \quad (1.21)$$

These correspond to $W^\pm$, Z , and the photon (γ), respectively. The mixing angle is the Weinberg angle θ_W . Weak interactions are classified as charged current (mediated by $W^\pm$) and neutral current (mediated by Z).

The electroweak Lagrangian density is:

$$\mathcal{L}_{EW} = -\frac{1}{4}W_{\mu\nu}^i W^{i\mu\nu} - \frac{1}{4}B_{\mu\nu} B^{\mu\nu} + \bar{\psi}_L i\gamma^\mu D_{\mu,L} \psi_L + \bar{\psi}_R i\gamma^\mu D_{\mu,R} \psi_R \quad (1.22)$$

After applying the mixing, interactions with vertices like WWZ , $WW\gamma$, $WWWW$, $WWZ\gamma$, $WWZZ$, and $WW\gamma\gamma$ emerge due to the non-abelian $SU(2)$ transformation.

While electroweak unification was a major achievement, it raises an important problem: any mass term in the Lagrangian would break gauge invariance.

sectionThe Standard Model Lagrangian before Higgs mechanism introduction

The Standard Model (SM) Lagrangian is characterized by the combined symmetry $SU(2)_L \times U(1)_Y \times SU(3)_C$. Taking into account the three SM interactions discussed above, it is expressed as follows:

$$\begin{aligned} \mathcal{L}_{SM} = & -\frac{1}{4}G_{\mu\nu}^a G^{a\mu\nu} - \frac{1}{4}W_{\mu\nu}^i W^{i\mu\nu} - \frac{1}{4}B_{\mu\nu} B^{\mu\nu} \\ & + i\bar{L}_i D_\mu \gamma^\mu L_i + i\bar{e}_{Ri} D_\mu \gamma^\mu e_{Ri} + i\bar{Q}_i D_\mu \gamma^\mu Q_i \\ & + i\bar{u}_{Ri} D_\mu \gamma^\mu u_{Ri} + i\bar{d}_{Ri} D_\mu \gamma^\mu d_{Ri} \end{aligned} \quad (1.23)$$

The first three terms describe the dynamics of the gauge fields and form the Yang-Mills part of the Lagrangian. Here, μ and ν are spacetime indices ranging from 0 to 3. The field strength tensors for the gauge fields are defined as follows:

The tensor for the eight gluon fields G_μ^a , corresponding to the generators T^a of the $SU(3)_C$ color group, is given by:

$$G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g_s f^{abc} G_\mu^b G_\nu^c, \quad (1.24)$$

where $a, b, c = 1, \dots, 8$, and g_s is the coupling constant for $SU(3)_C$.

The tensor for the gauge field B_μ , associated with the generator Y of the $U(1)_Y$ hypercharge group, is:

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu. \quad (1.25)$$

The tensor for the three gauge fields W_μ^i , corresponding to the generators $I^i = \tau^i/2$ of $SU(2)_L$, where τ^i are Pauli matrices, is:

$$W_{\mu\nu}^i = \partial_\mu W_\nu^i - \partial_\nu W_\mu^i + g' \epsilon^{ijk} W_\mu^j W_\nu^k, \quad (1.26)$$

where $i, j, k = 1, 2, 3$, and g' is the coupling constant for $SU(2)_L$.

The covariant derivative, which describes interactions between fields, is defined as:

$$D_\mu = \partial_\mu - ig I_i W_\mu^i - ig' \frac{Y}{2} B_\mu - ig_s T_a G_\mu^a. \quad (1.27)$$

The remaining terms in Equation (1) represent the fermionic part of the Lagrangian. This includes left-handed $SU(2)$ doublets for charged leptons and neutrinos:

$$L = (\nu_{eL}, e_L)^T, \quad (1.28)$$

where $e = e, \mu, \tau$ and $\nu_e = \nu_e, \nu_\mu, \nu_\tau$. Similarly, there are left-handed doublets for up-type and down-type quarks:

$$Q = (u_L, d_L)^T, \quad (1.29)$$

and right-handed singlets for up-type quarks (u_R), down-type quarks (d_R), and charged leptons (l_R).

In this formulation of the Standard Model Lagrangian, mass terms for gauge bosons and fermions are not included. For QCD interactions, mass terms of the form $-m_q q \bar{q}$ can be introduced without breaking $SU(3)$ gauge symmetry since gluons remain massless. However, in electroweak interactions, introducing mass terms directly would violate local gauge invariance under $SU(2)_L \times U(1)_Y$. For instance, a mass term like $-m_f f = m(f_L f_R + f_R f_L)$ cannot be added because f_L (part of a doublet) and f_R (a singlet) transform differently under $SU(2)_L$, breaking its symmetry.

To complete the Standard Model and give mass to gauge bosons ($W^\pm, Z$) and fermions while preserving gauge symmetry, spontaneous symmetry breaking (SSB) via the Higgs mechanism is required. This mechanism introduces a scalar Higgs boson and breaks the electroweak symmetry to provide masses to particles.

1.1.4 Electroweak Symmetry Breaking and the Higgs Mechanism

The unified theory of Glashow, Weinberg, and Salam successfully describes the electroweak force and its carriers. However, a significant problem emerges: the corresponding particles remain massless. While this is true for the photon, gauge bosons and fermions actually have mass. In 1964, three independent models were developed to provide a solution to this obstacle: one by Robert Brout and François Englert, another by Peter Higgs, and a third by Gerard Guralnik, C.R. Hagen, and Tom Kibble. The Brout-Englert-Higgs-Guralnik-Hagen-Kibble mechanism, or BEH mechanism for short, is based on the ideas of the Goldstone model. A milestone of this model is the idea of spontaneous symmetry breaking. It occurs when the equations of motion and the Lagrangian of a theory are symmetric under certain transformations, while the vacuum state does not obey this symmetry. Instead, it "spontaneously" selects a state that breaks the symmetry of the equations. This phenomenon is characterized as spontaneous because it appears without any external intervention. Using spontaneous symmetry breaking, the Higgs mechanism generates mass terms for the $W^\pm$ and Z bosons as well as fermions. According to this theory, every particle that interacts

with the Higgs field acquires mass. Conversely, photons - which do not interact with the field - remain massless. [2] [7]

In more detail, the process of electroweak symmetry breaking introduces a scalar field that breaks the Lagrangian's symmetry. To generate mass terms for the three gauge bosons $W^\pm$ and Z while ensuring the photon remains massless, it is crucial to break the $SU(2)_L$ symmetry while preserving the $U(1)_Y$ symmetry intact.

This symmetry breaking necessitates a configuration of scalar fields with a minimum of three degrees of freedom. The most straightforward approach to this problem involves introducing an $SU(2)_L$ doublet of scalar fields Φ , possessing 4 degrees of freedom:

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \quad (1.30)$$

Furthermore, the Higgs field is associated with a potential, commonly referred to as the "Mexican hat potential," given by:

$$V(\phi) = -\mu\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2 \quad (1.31)$$

$V(\phi)$ is gauge invariant under rotations of ϕ^+ and ϕ^0 . For $\mu > 0$ and $\lambda > 0$, this potential exhibits a global minimum at the vacuum expectation value $|\phi| = v = \sqrt{\frac{\mu}{2\lambda}}$ and a local maximum at $\phi = 0$. The Higgs field's behavior is energy-dependent. At high energy scales, the full potential, including its inherent symmetries, is relevant and conserved. However, at lower energies, the Higgs field becomes restricted to a minimum-energy region, specifically a valley near the global minimum. In this regime, the effective potential can be approximated around a single point $|\phi|$ within the valley, and the overall symmetry of the initial potential is no longer evident. This phenomenon is known as spontaneous symmetry breaking [7].

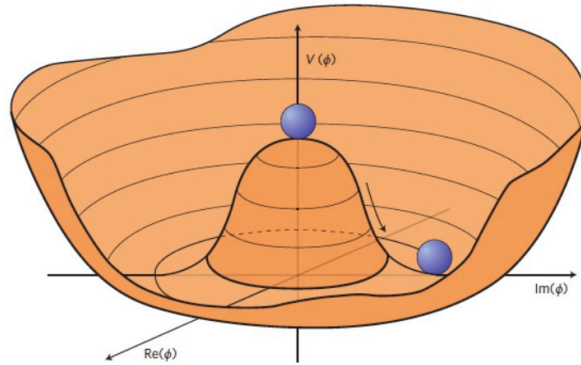


Figure 1.3: An illustration of the Higgs potential with $\mu^2 < 0$

The spin of the doublet 1.30 is assumed to be zero, thus, Klein-Gordon's equation dictates the propagation and so, Lagrangian density can be written as:

$$\mathcal{L}_\phi = (D^\mu\phi(x))^\dagger (D_\mu\phi(x)) - V(\phi(x)) = |D_\mu\phi(x)|^2 - V(\phi(x)), \quad (1.32)$$

The expression $|\phi|^2 = \phi^\dagger \phi = (\phi^+, \phi^0) \cdot \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$ defines the squared magnitude of the scalar field doublet. The term D_μ represents the $SU(2) \times U(1)$ covariant derivative which is shown in equation 1.30.

Considering the Higgs boson as an excitation of the Higgs field, the field can be expressed as:

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \quad (1.33)$$

Substituting this into the Higgs potential (Eq.1.34) yields:

$$V(\phi) = -\mu^2 H^2 + \lambda v H^3 + \frac{1}{4} \lambda H^4 \quad (1.34)$$

The first term in this equation is related to the Higgs boson mass, which is given by $m_H = \sqrt{-2\mu^2}$ and is experimentally determined to be approximately 125.30 GeV. The remaining terms describe the Higgs boson's self-interactions.

Returning to the Lagrangian density of the Higgs field (Eq. 1.35), and substituting the covariant derivative, we get:

$$\mathcal{L}_\phi = \left| \left(\partial_\mu + igW_\mu^i \frac{\sigma_i}{2} + ig'B_\mu \frac{Y}{2} \right) \phi(x) \right|^2 - V(\phi(x)) \quad (1.35)$$

Considering the potential's minimum and expanding the above expression, one obtains:

$$\mathcal{L}_\phi = \frac{1}{8} v^2 g^2 (W_\mu^1)^2 + \frac{1}{8} v^2 g^2 (W_\mu^2)^2 + \frac{1}{8} v^2 (gW_\mu^3 - g'B_\mu)^2 + 0(gW_\mu^3 + g'B_\mu)^2 \quad (1.36)$$

Now, considering the physical states of the EW bosons (Eqs. 1.22-1.24) and using the fact that $\tan \theta_W = \frac{g'}{g}$, the masses of the EW bosons are obtained:

$$m_W = \frac{1}{2} gv \quad (1.37)$$

$$m_Z = \frac{1}{2} v \sqrt{g^2 + g'^2} \quad (1.38)$$

$$m_A = 0 \quad (1.39)$$

These fulfill the relation $m_Z = \frac{m_W}{\cos \theta_W}$, with the photon (A_μ) remaining massless, as expected. The coupling of the Higgs to the gauge bosons (g_V) is proportional to the square of their mass:

$$g_V = \frac{2m_V^2}{v} \quad (1.40)$$

Finally, considering the Yukawa interaction with the Higgs field the masses of the fermions can be derived. An additional gauge-invariant interaction term is added to the Lagrangian density in the form of:

$$\mathcal{L}_{Yukawa} = -g_Y \bar{\psi}(x) \Phi \psi(x), \quad (1.41)$$

and the fermion masses can be derived from $m_f = \frac{1}{\sqrt{2}}g_Y^f v$, where g_Y^f is the Yukawa coupling constant of a Fermion to the Higgs boson, which is proportional to the fermion mass. For quarks, this interaction leads to weak flavor mixing, described by CKM matrix (eq:1.3). [6] [8]

So, there have to be introduced in SM Lagrangian three more terms: the one characterizes the Yukawa coupling and describes the interaction of the fermion fields with the Higgs field, another term for the Higgs interaction with the boson field and a last term for Higgs field potential. The compact form of the SM Lagrangian is written as:

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^a F_a^{\mu\nu} + \bar{\psi}(i\gamma^\mu D_\mu - m)\psi + (Y_{ij}\bar{\psi}_i\phi\psi_j + h.c.) + |D_\mu\phi|^2 - V(\phi). \quad (1.42)$$

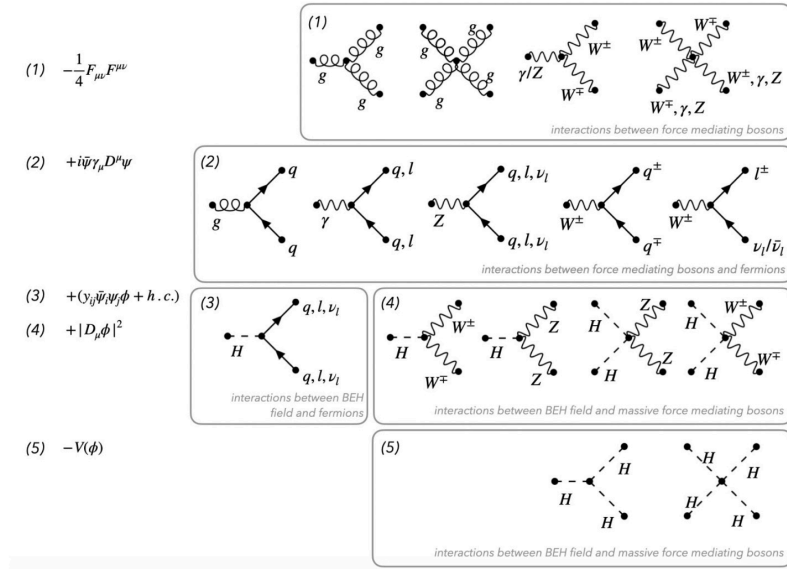


Figure 1.4: The interaction of SM particles representative by the five terms in the SM Lagrangian in the format of Feynman diagrams

1.1.5 The Higgs Boson at the LHC

The discovery of a scalar particle in agreement with the Higgs boson in 2012 was the final accomplishment of the SM and one of the greatest triumphs of both theory and experiment. It confirmed the mechanism in which bosons gain their masses, the electroweak symmetry-breaking (EWSB), and fermions via Yukawa interaction. Its properties show remarkable consistency with the SM predictions for the observed Higgs mass.[9] [10] In particular, the Higgs boson signal yields as well as all the measured coupling strengths to fermions and vector bosons are compatible with the SM, thus upholding the SM's validity up to the TeV scale with unparalleled precision.

Production Mechanisms of Higgs boson

The challenging task of observing the Higgs boson requires its production under controlled, high-energy conditions achievable only in particle accelerators like the LHC, Tevatron, and LEP. The

primary Higgs production mechanisms depend on its couplings, which are proportional to the mass of the participating particle (see Equations 1.37-1.39). Consequently, the main interactions of the Higgs boson are with heavier particles such as W and Z bosons and third-generation fermions. By exploiting these interactions, Higgs production is achieved at the LHC through four main mechanisms, which will be discussed below. Figure 1.5 shows the Feynman diagrams of these mechanisms. [11]

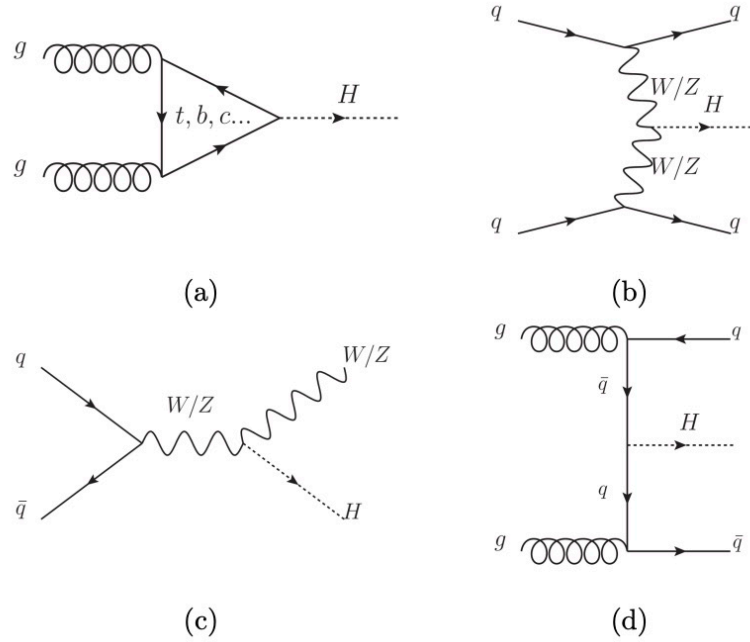


Figure 1.5: Main production Higgs boson production mechanisms at the LHC at tree level (a) ggF, (b) VBF, (c) VH and (d) $t\bar{t}H$ or $b\bar{b}H$

Since the Higgs couplings to Standard Model particles are proportional to their masses, it interacts very weakly with the proton's constituents (u, d quarks and gluons). In contrast, its couplings to heavier particles (W, Z bosons and top quark) are stronger. Therefore, Higgs production in proton accelerators like the LHC first requires the production of a heavy particle. We focus on the top quark, the heaviest known particle. The only way to produce a top quark after a proton collision is through the $g \rightarrow t\bar{t}$ channel. To preserve color conservation, one of the top quarks must interact with a second gluon, changing its color.[12] Thus, the involvement of two gluons from the colliding proton beams leads to the primary Higgs boson production mechanism, called "gluon-gluon fusion" or "gluon fusion". As shown in Figure 1.4(a), this production mechanism is dominant, corresponding to the largest cross-section across the entire center-of-mass energy range. For a Higgs boson of mass 125 GeV and energy $\sqrt{s} = 13$ TeV, the gluon fusion process has a cross-section of 49 pb. The contribution of lighter quarks to production via gluon fusion is rarer, as the Higgs interacts with them to a much lesser degree. [13]

• Vector Boson Fusion (VBF):

As mentioned, the absence of color charge in W and Z bosons prevents their participation in the gluon fusion mechanism. In this case, the Higgs boson can be produced in two ways: either in

association with a vector boson (associated production) or through Vector Boson Fusion (VBF). In the VBF mechanism, quarks from two different incoming protons emit either two Z bosons or W^+ and W^- bosons. These form ZZ or W^+W^- pairs, which then fuse to produce the Higgs. The VBF process involving W bosons is three times more likely than with Z bosons due to differences in their mass and couplings to fermions. For a center-of-mass energy of 13 TeV, the VBF cross-section is 3.8 pb. [8] The VBF mechanism in the LHC produces a distinct signature: after emitting W and Z bosons, the quarks continue moving in the direction of the initial colliding proton beams, creating two high-energy back-to-back jets detected farther from the collision point. Additionally, there is no "color charge exchange" between quarks in VBF, resulting in significantly less gluon radiation in the central detector region. This makes the VBF signal cleaner and easier to identify compared to the more common gluon fusion mechanism.

- **Associated production with vector bosons:**

The production of the Higgs boson in association with W or Z bosons is the next most significant mechanism at the LHC, denoted as Wh and Zh respectively. In this process, a quark-antiquark pair annihilates to create an off-shell W or Z boson, which then emits a Higgs boson and continues as a real boson (Figure 1.4(c)). Additionally, Zh production is often described by the Feynman diagram in Figure 1.6, where gluon interaction precedes the process. Despite the contribution of gluon fusion in this case, Zh production is less likely. [12] Since associated production requires the presence of an antiquark - which in the LHC would be a "sea quark" - the corresponding cross-section is smaller than that of the VBF channel, despite involving the same Higgs coupling. This is evident in Figure 1.7, which also shows that Wh production is more probable than Zh production. For $\sqrt{s} = 13$ TeV, the cross-sections for these two manifestations of this mechanism are 1.37 pb and 0.88 pb respectively.

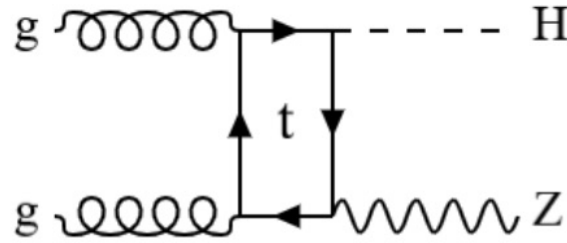


Figure 1.6: NLO Feynman diagram of ZH associated production

- **Production in association with heavy quarks:**

Finally, Higgs production in association with heavy quarks has a much smaller cross-section compared to the previous mechanisms. Due to the strong coupling of the Higgs with third-generation quarks, this process occurs largely through top quarks (Figure 1.4(d)), usually denoted as $t\bar{t}h$. In this mechanism, each of the two gluons produces a $t\bar{t}$ pair. The t from one pair annihilates with the $\bar{t}$ from the other, forming the Higgs boson. Unlike the ggF channel, $t\bar{t}h$ production doesn't involve a virtual loop, allowing direct study of the Yukawa coupling between the Higgs and the

top quark. [14] [8] To a lesser extent, the Higgs boson is produced in association with the comparatively lighter bottom quark. At $\sqrt{s} = 13$ TeV, $t\bar{t}h$ production has a cross-section of 0.51 pb, while $b\bar{b}h$ production has 0.49 pb. Another manifestation of this mechanism is the production, associated with a single top quark. This process is divided into three production channels, depending on the nature of the participating W boson:

- s-channel: Involves a virtual time-like W boson (0.003 pb)
- t-channel: Exchanges a virtual space-like W boson (0.07 pb)
- tW-channel: The top quark is produced in association with the W boson, followed by Higgs production: $pp \rightarrow tWh$ (0.02 pb)

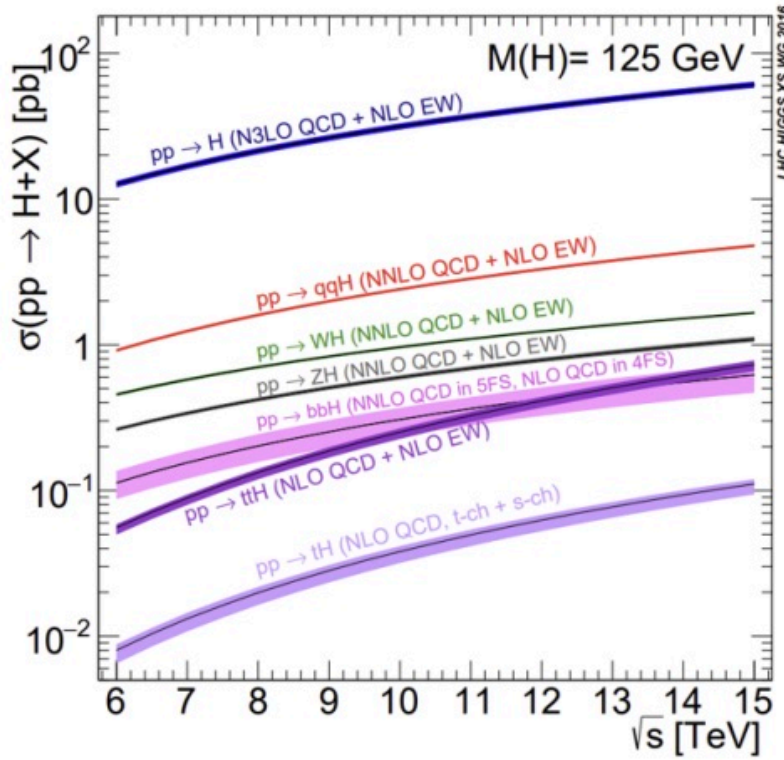


Figure 1.7: Production cross-section of the SM Higgs boson, with a mass of 125 GeV, for different production processes, depending on the centre-of-mass energy of p-p collisions

SM decay modes of Higgs

Regarding the decay of the Higgs boson, it can decay through multiple channels due to its interactions with leptons, quarks, and vector bosons, each with distinct branching fractions shown in Figure 1.9. In general, the decay of the Higgs boson prefers heavier final-state particles, as the

matrix element for each decay mechanism is proportional to the mass of the particle coupling with the Higgs. The decay modes are briefly outlined below, and corresponding Feynman diagrams are shown in Figure 1.8

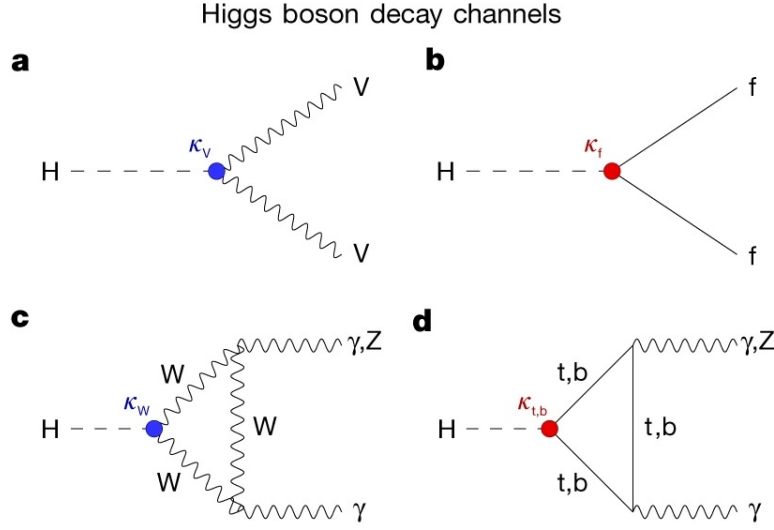


Figure 1.8: Decay modes of Higgs boson

- **Decays to fermions:**

The Higgs boson decays to fermion pairs when kinematically allowed, i.e., when $M_h > 2M_f$. The decay width for the fermionic channel is described by:

$$\Gamma(h \rightarrow f\bar{f}) = N_C \frac{M_f^2 M_h}{8\pi v^2} \beta^3, \quad (1.43)$$

where $\beta = (1 - 4M_f^2/M_h^2)^{1/2}$ is the velocity of the final state fermions, and N_c is the color factor: $N_c = 3$ for quarks and $N_c = 1$ for leptons. This width is proportional to the Higgs mass but is primarily influenced by the fermion mass. Thus, the Higgs boson's decays to first-generation fermions, i.e. $b\bar{b}$ and $\tau\bar{\tau}$ pairs, are significant. These decays are predominant but challenging to isolate in hadron collider environments due to the difficulty in constructing their final state signatures. Searching for their decay into second-generation fermions can be more challenging as the branching ratio is small.

- **Decays to vector bosons:**

The mass constraints of the W and Z bosons forbid Higgs boson's on shell decay into vector bosons. However, the decay branching ratios to the on shell– off shell WW^* , ZZ^* pairs are still considerable. In addition, the decay to a pair of ZZ^* bosons plays an important role at the LHC in the Higgs boson due to a very clean experimental signature in the four-lepton final state.

- **Loop-induced decay processes :**

The Higgs boson decay into massless particles, including gluons and photons, is possible through a loop-induced process. The Higgs boson to gg decay takes up a considerable proportion, but its direct measurement is difficult. Meanwhile, the Higgs boson decaying to diphoton through a vector-boson loop or a fermion loop provides clear modeling of a peak structure on the diphoton invariant mass and is the golden channel to probe the Higgs boson at the LHC.

Taking into account all possible decay modes, a Higgs boson with mass 125.09 GeV, the Standard Model predicts a total decay width of [15]:

$$\Gamma_h^{\text{tot}} = 4.100^{+0.73\%}_{-0.73\%}(\text{THU})^{+0.99\%}_{-0.97\%}(\text{PU}, m_q)^{+0.60\%}_{-0.61\%}(\text{PU}, a_s) \quad (1.44)$$

In this context, THU denotes the theoretical uncertainty, and PU represents the parametric uncertainty. The latter relates to the error arising from parameters such as the quark masses (m_q) and the QCD coupling constant (α_s). From this value, we can calculate the particle's lifetime:

$$\tau_h = \frac{\hbar}{\Gamma_{\text{tot}}} \Rightarrow \tau_h = (1.61 \pm 0.02) \times 10^{-22} \text{ s} \quad (1.45)$$

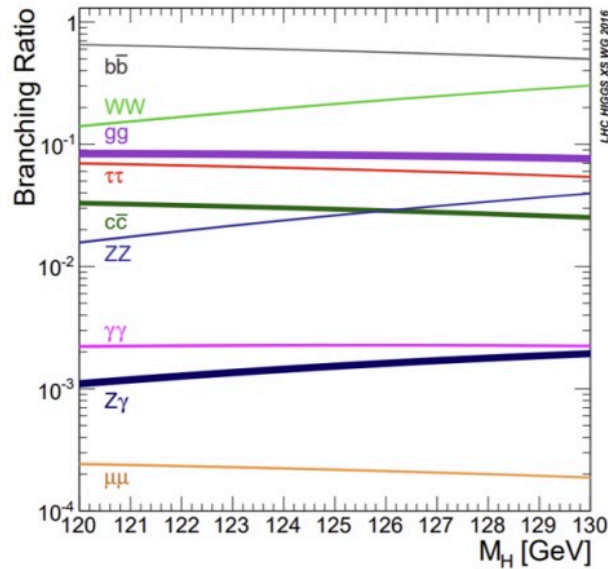


Figure 1.9: Branching ratios of the SM Higgs boson decay in a mass range near 125 GeV at a centre-of-mass energy of 13 TeV

1.2 Beyond the Standard Model

1.2.1 Shortcomings of the Standard Model

The SM has been tested experimentally to a very high precision and over a wide energy range with remarkable success. However, its incompleteness becomes apparent when confronted with several observed phenomena and theoretical challenges.

- **Absence of Gravity in the Unification of Forces**

The SM successfully describes the strong, weak, and electromagnetic forces. However, it excludes gravity, the fourth fundamental force. Developing a consistent quantum theory of gravity remains one of the greatest challenges in theoretical physics. A successful theory of everything would need to unify all four forces, including gravity, within a single framework.

- **Dark Matter and Dark Energy**

Galaxy rotation curves deviate significantly from Newtonian predictions, implying the existence of a DM halo. Similarly, the accelerating expansion of the universe necessitates the concept of DE, with proposed explanations ranging from a cosmological constant to dynamic scalar fields. Compelling evidence indicates that dark matter (DM) and dark energy (DE) constitute over 85% of the universe's mass-energy density, yet the SM offers no viable particle candidates for these entities.

Furthermore, the observed imbalance between matter and antimatter presents a significant challenge. The SM predicts their equal production in the early universe, contradicting the observed matter-dominated cosmos. While CP violation within the SM offers a potential mechanism for baryogenesis, its magnitude is insufficient to explain the observed asymmetry.

- **The Neutrino Mass Problem**

Experimental evidence for neutrino oscillations demonstrates that neutrinos possess non-zero mass, a fact that contradicts the original formulation of the SM. The SM's Lagrangian lacks a right-handed neutrino term, leading to massless neutrinos. The discovery of neutrino oscillations, where neutrinos change flavor as they propagate, necessitates an extension to the SM to incorporate neutrino mass.

- **The Hierarchy Problem and the Planck Scale**

The SM faces a significant theoretical challenge known as the hierarchy problem. This problem arises from the sensitivity of the Higgs boson mass to quantum corrections. The Higgs mass receives contributions from virtual particles at all energy scales, including the Planck scale ($\Lambda_{\text{Planck}} \sim 10^{19}$ GeV), where quantum gravity effects are expected to become dominant. This leads to quadratically divergent corrections, requiring an unnatural fine-tuning of the bare Higgs mass to maintain its observed value. The fine-tuning problem can be expressed as:

$$m_{h,\text{SM}}^2 = m_{h0}^2 + \Delta m_h^2, \quad (1.46)$$

where

$$\Delta m_h^2 \propto \Lambda_{\text{Planck}}^2. \quad (1.47)$$

To stabilize the Higgs mass against these large corrections, new physics beyond the SM is required at or below the TeV scale.

The issues described above suggest that the SM, while remarkably successful within its domain, represents an incomplete picture of fundamental physics. The search for solutions to these problems has led to the development of numerous Beyond the Standard Model (BSM) theories. These theories are the focus of intense research efforts.

1.2.2 Theories Beyond the Standard Model

In the context of BSM physics, the Higgs sector is of particular interest, as a relatively simple enlargement would lead to a much richer phenomenology that may solve some of the aforementioned issues. Recent experimental analyses have set an upper limit of approximately 16% on the branching fraction into exotic final states at 95% CL of about 16% [16]. Furthermore, the striking disparity between the Higgs boson's total decay width (on the order of MeV) and its mass (on the order of GeV) fuels intense interest in searching for these exotic decays. The substantial difference between these scales arises from the relatively weak Yukawa couplings between the Higgs and the fermions it decays into within the SM. This disparity hints that a BSM final state could potentially possess a significantly larger partial decay width, leading to a non-negligible branching fraction compared to decays into SM particles. These considerations provide a strong impetus for investigating BSM decays of the Higgs boson. Indeed, numerous theoretical frameworks predict the existence of these exotic Higgs decays, with models often involving extended Higgs sectors, dark matter candidates, and scenarios invoking neutral naturalness.

Dark Matter Models: There are many theories within the BSM field that propose the connection of the Higgs boson with dark matter. According to some of them, the Higgs could decay into dark matter particles, which interact strongly with CP particles and are difficult to detect. Such models study objects such as the Higgs Portal, and "Dark Photons" [3].

Neutral Naturalness: Neutral Naturalness theories talk about the "hierarchy problem" in SM. According to this, the mass of the Higgs boson should be as large as the mass scale it interacts with, the Planck scale. These models attempt to address this problem by proposing new particles that are neutral under the SM gauge groups. By this term, we mean particles that do not interact through the three fundamental forces described by SM. Examples of such theories are "Twin Higgs" and "Composite Higgs" models.

Extended Higgs Sectors: Some BSM theories propose the existence of a richer spectrum of Higgs particles, leading to what are known as extended Higgs sectors. Some prominent examples include:

- **Next-to-Minimal Supersymmetric Standard Model (NMSSM):** The Minimal Supersymmetric Standard Model is a theoretical extension of SM that incorporates supersymmetry. This is a symmetry that relates bosons (integer spin) to fermions (half-integer spin). According to this, each particle of the SM corresponds to a super-particle with different spin. The term "minimal" indicates that it refers to the simplest supersymmetric expansion of the CP, which introduces the minimum number of new particles and parameters. So, Next-to-Minimal Supersymmetric Standard Model (NMSSM) extends the Minimal Supersymmetric Standard Model (MSSM) by adding a singlet field that can mix with the Higgs doublets. This leads to a more complex Higgs sector with additional neutral and charged Higgs particles.
- **Two-Higgs-Doublet Models (2HDMs):** In the Two Higgs Doublet Model the scalar sector is extended, as the name indicates, with one additional doublet. This enlargement results in five physical Higgs states, two of them charged, and three of them neutral. Out of the neutral particles, one is a pseudo-scalar, and the others are scalars. Specifically, 2HDMs predict two scalar, neutral, CP-even bosons (denoted as h and H , with H being the heavier), one pseudoscalar CP-odd boson (a), and two charged scalar bosons ($H^\pm$). Such extensions open up new decay channels for the observed Higgs boson, including decays into exotic particles. A more in-depth discussion of these models will be presented later.

1.2.3 The Two-Higgs-Doublet Model (2HDM)

The Higgs sector consists of an $SU(2)_L$ doublet Φ_1 with hypercharge $Y_{\Phi_1} = +1/2$. The 2HDM extension is one of the simplest extensions of the Standard Model (SM) and is based on the simple addition of another doublet Φ_2 . The vacuum expectation values v_1, v_2 for the respective Higgs fields define one of the most important free parameters of the 2HDM, β , as follows [17]:

$$\tan \beta = \frac{v_2}{v_1}. \quad (1.48)$$

Another important parameter of the 2HDM is the rotation angle α ($-\pi/2 \leq \alpha \leq \pi/2$), which defines the scalar mass eigenstates through the real components of $\Phi_{1,2}$.

After electroweak symmetry breaking in this model, five scalar physical states result:

- 2 CP-even, neutral Higgs bosons: h and H , where H is conventionally heavier. Since the SM Higgs boson is CP-even and neutral, it can be identified with either h or H .
- 2 charged Higgs bosons: $H^\pm$.
- 1 CP-odd pseudoscalar particle a .

At the lowest order, there are four types of 2HDMs, characterized by the coupling of each group of fermions to the doublets $\Phi_{1,2}$ without Flavor Changing Neutral Currents (FCNCs). FCNCs are processes where the "flavor" of a fermion changes without changing its charge. FCNCs at tree-level (first-order interactions, without loops) occur more often in models with multiple Higgs doublets. This is problematic as it can lead to large and experimentally unobserved rates of such flavor-changing processes. By invoking discrete Z_2 symmetries, the couplings of the fermions with the two doublets $\Phi_{1,2}$ can be controlled. Such a symmetry can be defined so that different types of fermions couple only with one of the two Higgs doublets [17]. This restriction prevents the simultaneous coupling of fermions with the two doublets, which could cause FCNCs. With the use of the appropriate Z_2 symmetry to avoid FCNCs, the four types of 2HDM result, which are characterized by the coupling of each group of fermions with the doublets $\Phi_{1,2}$. The ratio of the

	Type-1	Type-2	Type-3 (lepton-specific)	Type-4 (flipped)
Up-type quarks	Φ_2	Φ_2	Φ_2	Φ_2
Down-type quarks	Φ_2	Φ_1	Φ_2	Φ_1
Charged leptons	Φ_2	Φ_1	Φ_1	Φ_2

Table 1.1: The Higgs doublets to which various types of fermions couple in the four types of 2HDM without FCNC at the lowest order.

Yukawa couplings of the pseudoscalar boson of the 2HDM in relation to those of the Higgs in the SM is a function of $\tan \beta$ and the 2HDM type, and are found in Table 1.2.

1.2.4 2HDM+s Model

Assuming temporarily that the Higgs particle with a mass of 125 GeV of the SM is identified with one of the scalar bosons h or H^0 of the 2HDM, we are led to the following exotic decays, as these are allowed by the 2HDMs:

	Type-1	Type-2	Type-3 (lepton-specific)	Type-4 (flipped)
Up-type quarks	$\cot \beta$	$\cot \beta$	$\cot \beta$	$\cot \beta$
Down-type quarks	$-\cot \beta$	$\tan \beta$	$-\cot \beta$	$\tan \beta$
Charged leptons	$-\cot \beta$	$\tan \beta$	$\tan \beta$	$-\cot \beta$

Table 1.2: The ratio of the Yukawa couplings of the pseudoscalar boson a in the 2HDM, compared to those of the Higgs boson in the SM, for the four types of 2HDM without FCNC at the lowest order.[17]

$$h \rightarrow aa, \quad H^0 \rightarrow hh, aa \quad \text{and} \quad h \rightarrow Za \quad (1.49)$$

where h , a decay into fermions or gauge bosons of the SM. However, the existing data for 2HDM+s do not allow a large variety regarding the exotic decays of the Higgs.

To avoid such restrictions, we initially assume that the 2HDM is close to the decoupling limit

$$a \rightarrow \beta - \frac{\pi}{2}, \quad (1.50)$$

where the lightest state of the model, h , can be identified with the observed Higgs boson with a mass of 125 GeV of the SM. All properties of h depend on the parameters $\tan \beta$ and α , as well as on the type of fermionic couplings.

The next step is the addition of a singlet scalar field of the form:

$$S = \frac{1}{\sqrt{2}}(S_R + iS_I), \quad (1.51)$$

to which a vacuum expectation value corresponds. This field does not make direct Yukawa couplings, as it couples with $\Phi_{1,2}$ only in the potential. Therefore, all of its interactions with fermions of the SM come from its mixing with the Higgs doublets $\Phi_{1,2}$. However, this should be done to a small extent so as not to alter the nature of the Higgs in the SM [18].

With these two assumptions, many exotic decays of the Higgs emerge, of the form:

$$h \rightarrow ss/aa \rightarrow X\bar{X}Y\bar{Y} \quad \text{and} \quad h \rightarrow aZ \rightarrow X\bar{X}Y\bar{Y},$$

where $s(a)$ are (pseudo)scalar mass eigenstates, which are composed of $S_R(S_I)$. X, Y are fermions or gauge bosons of the SM [17].

In the real and imaginary component of Equation (1.41), separate masses can be assigned, while their mixings with the 2HDM sector can be characterized as disturbances. Therefore, in the 2HDM+S there are two pseudoscalar states a, S_I as well as five scalar states $h, H^0, H^\pm$ and S_R [18].

Chapter 2

The CMS Experiment at the LHC

2.1 The Large Hadron Collider

The LHC is the world's largest and highest-energy particle accelerator. Located on the border of Switzerland and France, it is composed of a 27 kilometer ring of superconducting magnets with a number of accelerating structures that boost the energy of charged particles, such as protons.[19] These particles are accelerated to high speeds in order to collide with each other. Besides protons, ions of lead (Pb) are also injected in the LHC, hence p-Pb and Pb-Pb collisions can take place too. The LHC is designed to reach collision energies of 14 TeV, equivalent to accelerating protons to almost the speed of light. A schematic view of the LHC is depicted in figure 2.1.

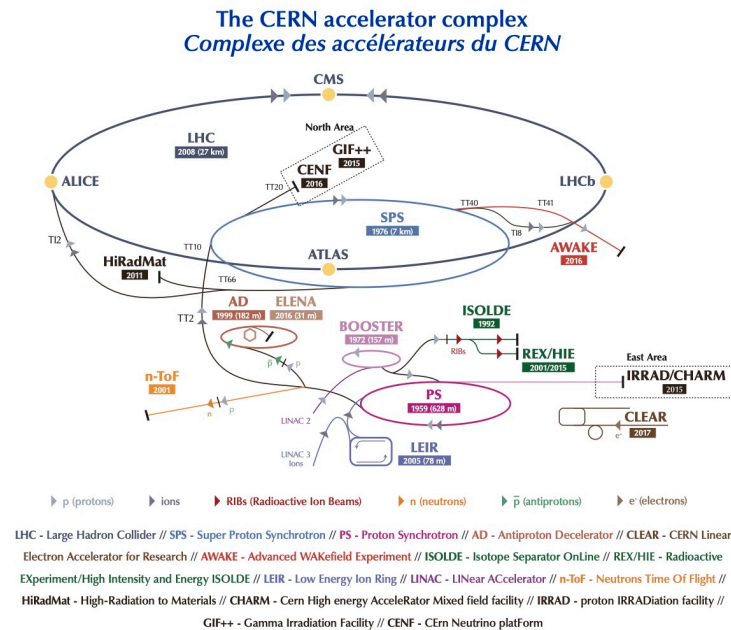


Figure 2.1: A schematic view of the LHC

Proton collisions occur at four particle detectors on the LHC:

- CMS (Compact Muon Solenoid)
- ATLAS (A Toroidal LHC Apparatus)
- ALICE (A Large Ion Collider Experiment)
- LHCb (Large Hadron Collider beauty).

ATLAS and CMS are two general-purpose detectors with the same research scope and physical objectives. A significant achievement was reached in 2012 when both ATLAS and CMS collaborations independently discovered the Higgs boson. ALICE is a heavy-ion detector primarily focused on studying quark-gluon plasma at extreme energy densities. LHCb explores the differences between matter and antimatter through b-meson physics. [20]

The CERN acceleration complex is a series of interconnected machines, each progressively increasing the energy of particles before transferring them to the next stage. This process begins with the Linac4, a linear accelerator operational since 2020, which accelerates negative hydrogen ions to 160 MeV. These ions are then prepared for injection into the Proton Synchrotron Booster, an integral part of the LHC injection chain. The acceleration process in Linac4 utilizes radiofrequency cavities, which create charged cylindrical conductors. These conductors accelerate the ions by alternately attracting and repelling them as they pass through. Quadrupole magnets help maintain the beam's focus. The acceleration in Linac4 occurs in four distinct phases: Firstly, a radio-frequency quadrupole (RFQ) accelerates ions to 3 MeV. Then, Drift tube linacs (DTLs) further increase energy to 50 MeV and Coupled-cavity drift tube linacs (CCDTLs) boost energy to 100 MeV. Finally, Pi-mode structures (PIMS) provide the final acceleration to 160 MeV. [21]

As the ions are injected from Linac4 into the Proton Synchrotron Booster, they are stripped of their electrons. The Booster then accelerates the resulting protons to 2 GeV before sending them to the Proton Synchrotron (PS), which further increases their energy to 26 GeV. This method enhances both particle accumulation and beam quality. The protons then enter the Super Proton Synchrotron (SPS), CERN's second-largest accelerator, where they reach energies of 450 GeV. Finally, they are injected into the LHC's two beam pipes, where they can achieve energies up to 6.8 TeV.

To increase the probability of proton collisions, the protons in the accelerator are squeezed into a very small space before colliding. Therefore, the collisions at the collision points involve proton bunches. The collision of proton bunches (bunch crossing) is every 25 nanoseconds or a collision frequency of 40 MHz. A typical proton bunch at the beginning of an LHC run contains approximately 1.2×10^{11} protons, with each beam comprising 2808 such bunches. This high proton density often results in a phenomenon known as in-time pileup, where multiple proton-proton interactions occur during a single bunch crossing. The severity of in-time pileup is quantified by the average number of interactions per bunch crossing. Figure 2.2 illustrates this concept by showing stacked pileup distributions in the CMS detector for each year of operation, covering both Run 2 and the initial part of Run 3. Another form of pileup, called out-of-time pileup, can also occur. This happens when the detector's signal readout time exceeds the interval between consecutive bunch crossings, causing signals from different crossings to overlap and combine. These pileup effects present significant challenges for data analysis and reconstruction in high-energy physics experiments at the LHC, requiring sophisticated techniques to disentangle the multiple interactions and accurately measure particle properties.

In addition to the number of concurrent collisions within a single bunch crossing, several machine-dependent parameters are crucial for physics analyses. These include the center-of-mass energy of the collisions, as well as both instantaneous and integrated luminosity.[22]

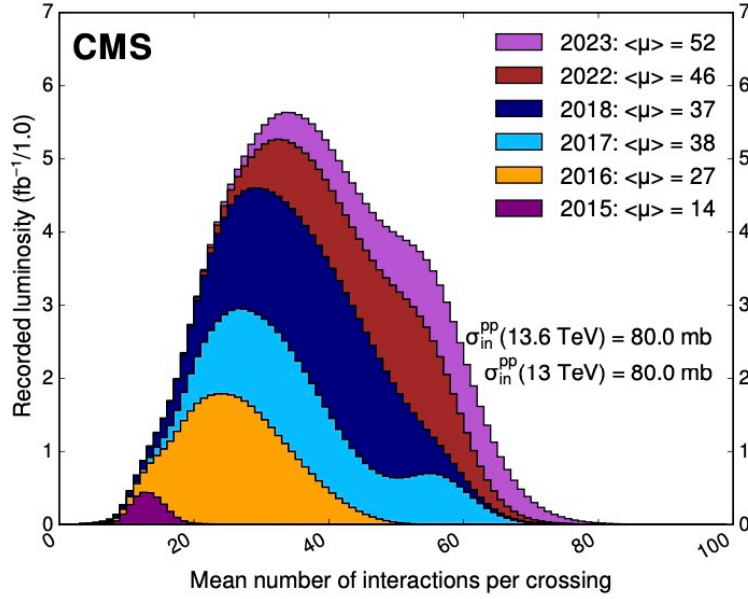


Figure 2.2: average number of interactions per bunch crossing

The production rate of a physical process is determined by the cross-section σ for a specific process and the instantaneous luminosity L :

$$dN = \sigma L dt \quad (2.1)$$

The total number of events produced is obtained by integrating this equation over time. Given that the cross-section depends solely on the center-of-mass energy, which remains constant during a fill, the number of events is expressed as:

$$N = \sigma \int dt L = \sigma \mathcal{L} \quad (2.2)$$

where $\mathcal{L}$ represents the integrated luminosity.

The instantaneous luminosity is influenced by various beam parameters:

$$L = \frac{\gamma N_p^2 k f}{4\pi\epsilon\beta^*} F \quad (2.3)$$

Here, γ is the relativistic gamma factor, N_p denotes the number of particles per bunch, k is the number of bunches per beam, and f is the revolution frequency of the bunches. The normalized beam emittance ϵ quantifies the beam's spread in phase space, while the beta function β^* describes the beam focusing at the interaction point. The form factor F accounts for the reduction in effective area due to the non-zero collision angle of the beams.

Luminosity, measured in $\text{cm}^{-2}\text{s}^{-1}$, serves as a key indicator of accelerator performance. Higher luminosity enables the collection of more data, facilitating the observation of rare processes. The Run 2 data-taking period (2015–2018) was operated at the centre-of-mass energy of 13 TeV and

collected data corresponding to an integrated luminosity of around 140 fb^{-1} . In the year 2024, the Run 3 period (2022–2025) is ongoing, with the LHC's increased luminosity and energy. The peak luminosity can reach around $2 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, and the energy at a centre-of-mass energy of 13.6 TeV. The High-Luminosity LHC (HL-LHC) project is a substantial upgrade planned to start in 2029 following the long shutdown 3 (LS3). The HL-LHC aims to significantly increase the luminosity, potentially by a factor of 10, extending the LHC's operational life and discovery potential out to 2041. Figure 2.3 shows the timeline of the LHC and HL-LHC programs.[23]

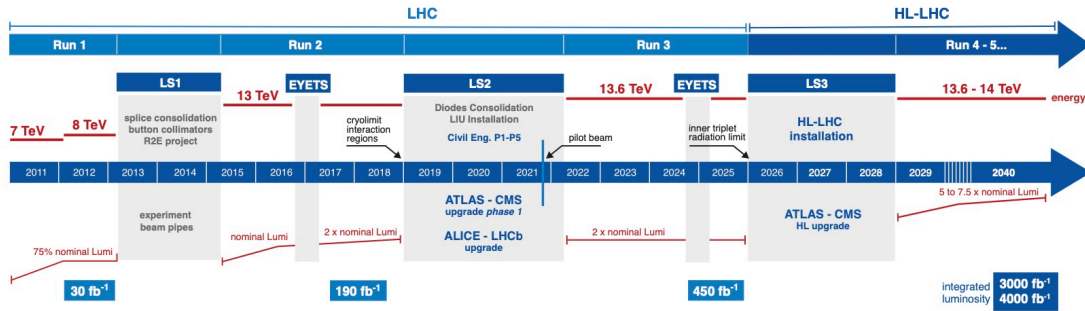


Figure 2.3: Integrated luminosity collected by the CMS experiment

2.2 Compact Muon Solenoid

Located in Cessy, France, the CMS detector is one of two general-purpose colliders whose physics targets include precision measurement of the SM and the search for new physics. The CMS detector is 21 m long, 15 m in diameter and weighs about 14000 tons [20]. It constitutes the following sub-detectors:

- the inner tracking system including the silicon pixel detector and silicon stripe detector
- the electromagnetic calorimeter
- the hadronic calorimeter,
- the muon system.

Each sub-detector consists of a small and individual hardware unit covering most of the area around the collision point. All sub-detectors are inside a superconducting solenoid at 3.8 T, except for the muon system and a minor part of the hadron calorimeter. A 3-D render of the CMS detector as well as a slice of it are depicted in figures 2.4 and 2.5 to visualize the sub-detectors as well as the particles detected in each of them. [20]

CMS Tracking system

The tracking system of the CMS detector is designed to measure the trajectories of charged particles. When charged particles pass through the inner tracker, the large magnetic field induces a curved trajectory, which enables a precise measurement of the particle momenta. Given the

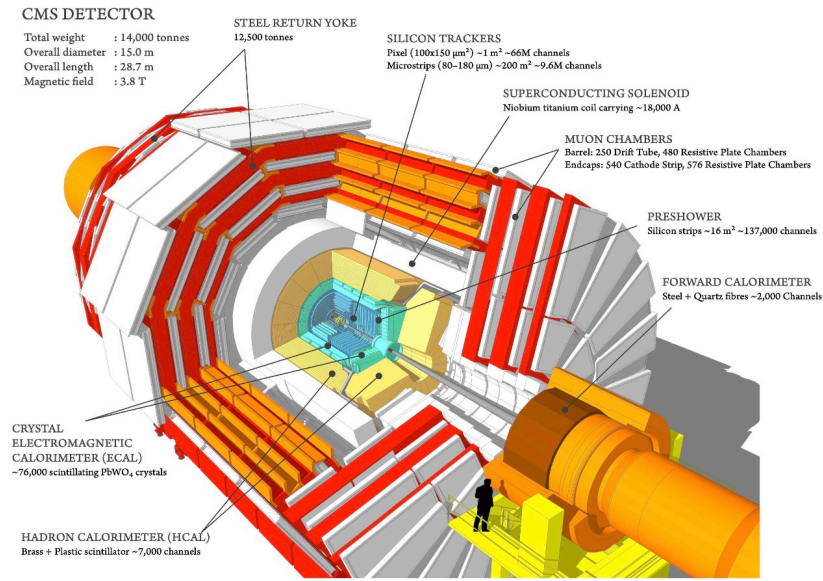


Figure 2.4: A 3-D render of the CMS detector

tracker's impressive accuracy of $10 \mu\text{m}$, only a limited number of position measurements are required to achieve satisfactory trajectory reconstruction [24]. However, the tracker's proximity to the interaction point exposes it to an intense particle flux, presenting several significant challenges: The detector must possess both high granularity and rapid response capabilities for distinguishing between the multitude of particles produced in each collision event. Also, components must be engineered to withstand the harsh, high-radiation environment near the interaction point. To address these challenges effectively, the tracker detector is constructed exclusively from silicon. Silicon detectors offer an optimal combination of high spatial resolution, fast response times, and excellent radiation tolerance, making them ideal for the demanding conditions near the LHC collision point. Figure 2.6 depicts a layout of the CMS tracker. The tracker is composed of two main distinct sub-regions:

- **Silicon Pixel Detectors:** At the heart of the tracking system, closest to the collision point, lie the silicon pixel detectors. These high-precision devices feature individual pixels measuring $100 \times 150 \mu\text{m}^2$. Covering a total surface area of approximately 1 m^2 , the pixel detector array comprises an impressive 6.6×10^7 pixels, enabling extremely fine-grained particle tracking. A silicon pixel is depicted in figure 2.7 [25]
- **silicon strip tracker:** Enveloping the pixel detector is the silicon strip tracker. This component utilizes strips ranging from 80 to $180 \mu\text{m}$ in width, spread across a vast surface area of 200 m^2 . A silicon strip is depicted in figure 2.8 With approximately 9.6×10^6 strips, this tracker provides extensive coverage for particle detection. The strip tracker is further subdivided into the Inner Barrel (TIB), Inner Disks (TID), Outer Barrel (TOB), and Outer Endcaps (TEC), as depicted in Figure 2.6

The inner tracker provides coverage for a pseudorapidity range of $|\eta| < 2.4$. This configuration allows for particle tracking through electrical signals generated by their interactions with the

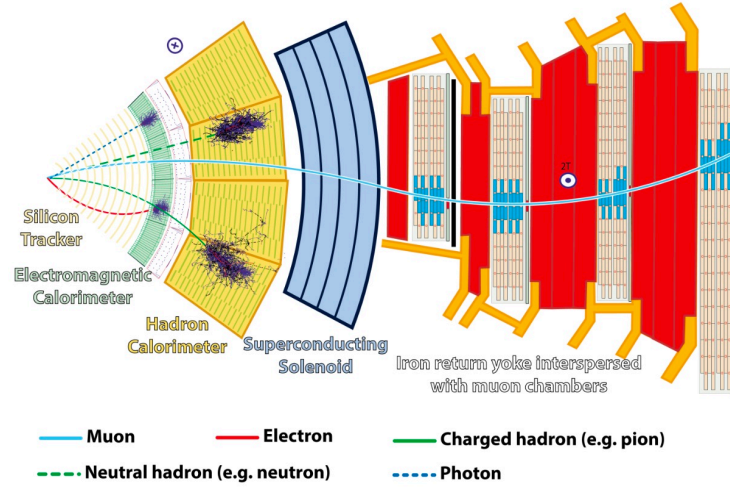


Figure 2.5: A transverse slice of the CMS detector and the particles detected by each sub-detector.

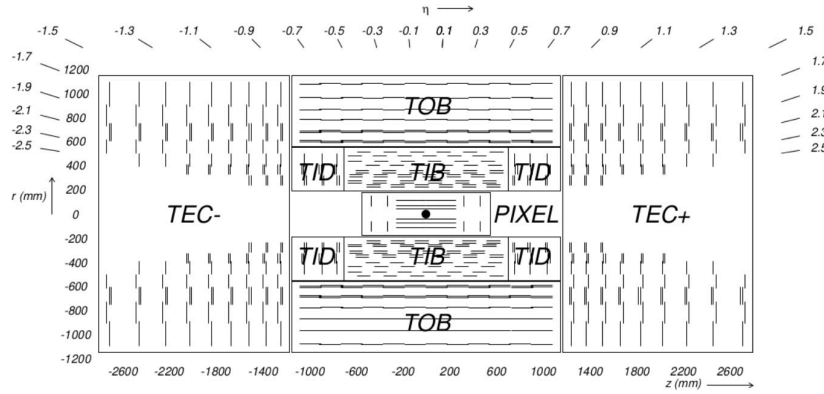


Figure 2.6: An r-z cross section of a half of the CMS detector, showing the layout of the inner tracker system

detector material. The material thickness varies as a function of pseudorapidity: approximately $0.4X_0$ at $\eta = 0$, around $1 X_0$ at $|\eta| = 1.6$ and decreasing to $0.6 X_0$ at $|\eta| = 2.5$, where X_0 represents the radiation length. During the LHC's year-end technical stop in 2016-2017, during the Run 2 data-taking period, the pixel detector underwent an upgrade. This enhanced detector, known as the Phase I pixel detector, incorporates: an additional pixel layer in both the barrel and endcap sections and so, extended pseudorapidity coverage from $|\eta| = 2.5$ to $|\eta| = 3$ [24]

2.3 Electromagnetic calorimeter

Electromagnetic Calorimeter (ECAL) is a crucial component in its particle detection arsenal of CMS. This device is engineered to measure the energies of electrons and photons with exceptional precision. At the heart of the ECAL lies an array of 75,848 lead tungstate (PbWO_4) crystals, chosen for their remarkable properties. These crystals boast a high density of 8.28 g/cm^3 , a short radiation

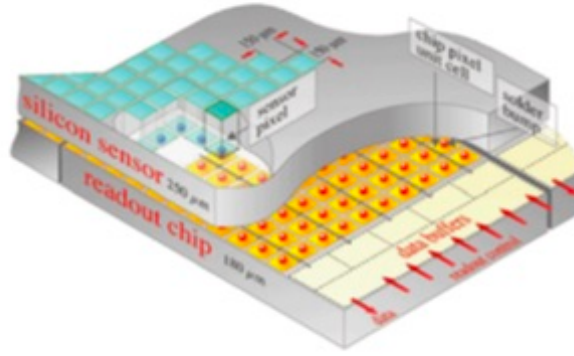


Figure 2.7: CMS silicon pixel detector

length of 0.89 cm, and a compact Molière radius of 2.19 cm. Their rapid response time, collecting 99% of scintillation light within a mere 100 nanoseconds, coupled with high radiation tolerance, makes them ideal for the challenging environment of the Large Hadron Collider (LHC).

The ECAL's design is a testament to precision engineering, comprising a barrel section and two endcaps. The barrel, covering a pseudorapidity range of $|\eta| < 1.48$, houses crystals measuring 23 cm in length with a front face of $2.2 \times 2.2 \text{ cm}^2$. The endcaps extend the coverage to $|\eta| < 3.0$, utilizing slightly shorter 22 cm crystals with a larger front face of $2.86 \times 2.86 \text{ cm}^2$. This configuration results in a total crystal volume of 11 m^3 , weighing an impressive 92 tons.

When high-energy particles traverse the ECAL, they initiate a cascade of interactions. Photons generate electron-positron pairs, while charged particles emit Bremsstrahlung radiation. These processes create a shower of secondary particles, continuing until their energy is fully absorbed by the crystals. The resulting scintillation light, proportional to the initial particle's energy, is captured by avalanche photodiodes, enabling precise energy measurements. [26] [27]

To enhance its capabilities, particularly in distinguishing between single high-energy photons and closely spaced photon pairs from neutral pion decays, a preshower detector is positioned in front of the endcaps. This additional layer, covering $1.65 < |\eta| < 2.6$, consists of lead absorbers and silicon strip sensors, adding approximately three radiation lengths to the detector's thickness. A geometric view as well as a cutaway diagram of ECAL is depicted in figure 2.9 .

The ECAL's performance is characterized by its energy resolution, typically expressed as:

$$\frac{\sigma}{E} = \frac{2.8\%}{\sqrt{E/\text{GeV}}} \oplus \frac{12\%}{E/\text{GeV}} \oplus 0.3\%$$

This equation encapsulates the detector's response across different energy ranges. The first term represents statistical fluctuations in the shower development and photon detection. The second term accounts for electronic noise in the readout system, while the constant term reflects various systematic effects such as calibration errors and potential crystal damage over time.

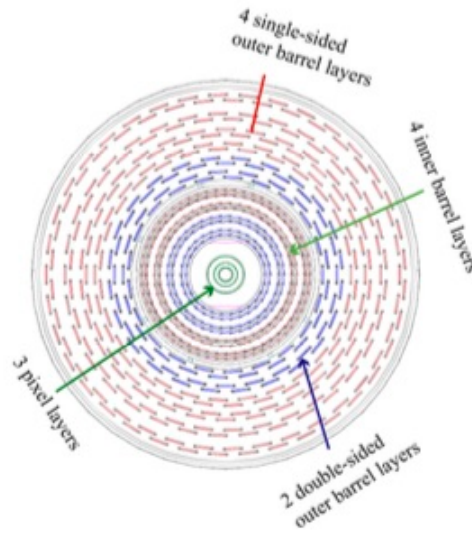


Figure 2.8: CMS silicon strip tracker, perpendicular render.

2.4 The CMS Hadronic Calorimeter (HCAL)

The Hadronic Calorimeter (HCAL) is designed to measure the energy of strongly interacting particles. Working in tandem with the Electromagnetic Calorimeter (ECAL), the HCAL forms an integral part of the CMS calorimetry system, enabling precise measurements of particle jets and missing transverse momentum [28].

The HCAL's primary function is to capture and measure the energy of hadrons - particles composed of quarks and gluons. These include protons, neutrons, pions, and other composite particles that interact via the strong nuclear force. Additionally, the HCAL plays a vital role in the indirect detection of weakly interacting particles like neutrinos, which escape the detector without leaving a direct signal.

Constructed as a sampling calorimeter, the HCAL consists of alternating layers of absorber material (brass) and active material (plastic scintillators). When incoming hadrons interact with the dense brass absorbers, they initiate hadronic showers - cascades of secondary particles. These secondary particles then traverse the scintillator layers, producing rapid light pulses. Specialized optical fibers collect these light pulses and transmit them to photodetectors for amplification and measurement. The total amount of light collected in a given region correlates directly with the energy of the initial particle, allowing for precise energy reconstruction.

The HCAL's architecture comprises four distinct subsystems, each designed to maximize coverage and performance:

1. Barrel (HB): Covering the pseudorapidity range $|\eta| < 1.4$, the HB extends radially from 1.77 m to 2.95 m. It contains 5.82 nuclear interaction lengths at $|\eta| = 0$, increasing to 10.6 at $|\eta| = 1.4$.
2. Endcap (HE): Extending the coverage to $1.3 < |\eta| < 3.0$, the HE ensures comprehensive detection in the forward regions.
3. Outer (HO): Positioned outside the solenoid in the barrel region, the HO captures shower

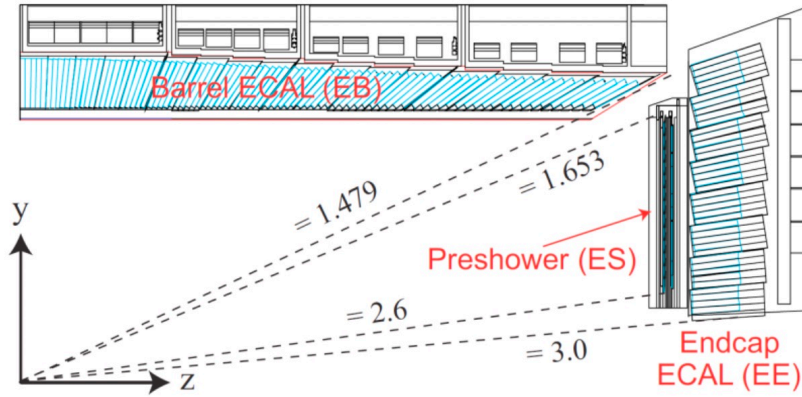


Figure 2.9: A geometric view of ECAL

tails that might escape the inner layers. It extends to $|\eta| < 1.26$ and significantly enhances the detector's thickness to at least 11.8 interaction lengths at $\eta = 0$.

4. Forward (HF): Located 11.15 meters from the interaction point, the HF uses a Cherenkov-based design to cover the extreme forward region of $3 < |\eta| < 5.2$. Its unique construction, utilizing steel absorbers and quartz fibers, ensures radiation hardness in the high-flux environment close to the beam.

The HCAL's segmentation varies with pseudorapidity. For $|\eta| < 1.6$, the calorimeter is divided into cells of 0.087×0.087 in η and ϕ . Beyond $|\eta| = 1.6$, the cell size doubles in both dimensions. A layout of HCAL is depicted in figure 2.10

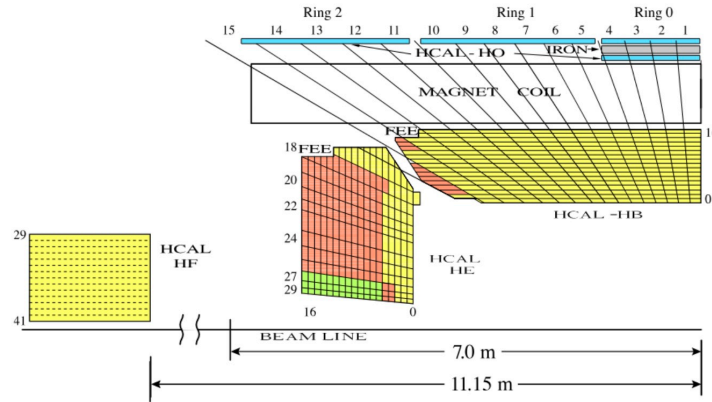


Figure 2.10: A layout of HCAL

The energy resolution of the HCAL, which depends on both the ECAL and HCAL responses due to shower initiation in the ECAL, is typically parametrized as:

$$\frac{\sigma}{E} = \frac{84.7\%}{\sqrt{E}} \oplus 7.4\%$$

where E is in GeV. This resolution, while less precise than that of the ECAL due to the sam-

pling nature of the HCAL, is sufficient for accurate jet energy measurements and missing energy calculations.

The HCAL's comprehensive design and coverage make it an indispensable tool in CMS physics analyses. Its ability to accurately measure hadronic energy, coupled with its near-hermetic coverage, enables precise reconstruction of complex events, identification of various particle types, and exploration of new physics phenomena at the energy frontier of particle physics [28].

2.5 The Superconducting Solenoid

The massive superconducting solenoid magnet of the CMS detector generates a powerful magnetic field that bends the trajectories of charged particles, allowing for precise momentum measurements and charge determination.

The CMS solenoid is a feat of engineering, boasting impressive dimensions and capabilities. It has an inner bore of 6.3 meters, an inner diameter of 5.9 meters, and stretches 12.9 meters in length. This enormous cylinder is constructed using niobium-titanium (NbTi) coils, chosen for their superb superconducting properties. When operating at peak performance, the solenoid generates a formidable magnetic field of 3.8 Tesla, storing a colossal energy of 2.6 gigajoules.

To achieve and maintain its superconducting state, the solenoid is cooled to an extremely low temperature of 4.7 Kelvin using approximately 6000 liters of liquid helium. This cryogenic environment is crucial for eliminating electrical resistance in the coils, allowing for the generation of such an intense magnetic field without prohibitive energy losses.

The strength of the magnetic field is a critical factor in the CMS detector's performance. The radius r of a charged particle's helical trajectory in a magnetic field is given by the equation:

$$r = \frac{p}{q \cdot B}$$

where p is the particle's momentum, q is its charge, and B is the magnetic field strength. The strong 3.8 T field enables CMS to measure particles with very high transverse momenta, a key capability for studying high-energy physics phenomena.

Surrounding the solenoid is a massive iron yoke, weighing an impressive 12,500 tonnes. This yoke serves multiple crucial functions. Firstly, it acts as a return path for the magnetic flux, effectively containing and shaping the magnetic field. This design reduces stray magnetic fields and enhances the overall field uniformity within the detector volume. The yoke consists of five wheels and two endcaps, arranged to allow for modular assembly and maintenance of the detector.

Beyond its role in magnetic field management, the iron yoke provides essential mechanical support to the entire experimental structure. Its design allows for relative ease of movement, simplifying the assembly and potential disassembly of various subdetectors. This feature is particularly valuable for a complex apparatus like CMS, where accessibility for maintenance and upgrades is a significant consideration.

Interestingly, the magnetic field extends beyond the solenoid itself. In the region outside the solenoid, encompassing the muon system, the field strength is approximately 2 Tesla, oriented in the opposite direction to the internal field. This extended field is crucial for the muon detection system, which is integrated into the iron yoke structure. The yoke acts as an absorber for other particles, leaving primarily muons to reach the outermost detector layers. Four interleaved layers of muon chambers within the yoke allow for independent measurements of muon momentum, com-

plementing the data from the inner tracking system and enhancing the overall particle identification capabilities of CMS.

2.6 The Muon System

The CMS experiment boasts one of the most sophisticated muon identification systems at the LHC. True to its name, CMS places significant emphasis on precise muon measurement, which is crucial for various physics analyses. The muon system serves a threefold purpose: muon identification, momentum measurement, and triggering.

Muons, being minimum ionizing particles, can traverse the entire detector volume with minimal energy loss. This unique property allows the muon system to be positioned outside the solenoid, alleviating space constraints and enabling a robust design. The system provides excellent coverage, extending to a pseudorapidity of $|\eta| < 2.4$, and is capable of precisely measuring muon trajectories with energies up to and beyond 1 TeV.

The muon system's architecture is dictated by the solenoid's cylindrical shape, resulting in a barrel section and two planar endcap regions. It is interleaved with the steel return yoke of the solenoid, which serves a dual purpose: it channels the magnetic flux and acts as a hadron absorber, enhancing muon identification capabilities.

CMS employs three types of gaseous detectors in its muon system, each optimized for specific regions and functions: Drift Tubes (DTs), Cathode Strip Chambers (CSCs), and Resistive Plate Chambers (RPCs).

DTs are utilized in the barrel region ($|\eta| < 1.2$), where particle flux is relatively low. Four stations of DTs are positioned at increasing radii from the interaction point. These tubes are filled with an Ar-CO₂ gas mixture and contain sensitive wires acting as anodes, with cathode strips placed along the tube walls.

CSCs are employed in the endcap regions ($0.9 < |\eta| < 2.4$), where magnetic fields are inhomogeneous and particle rates are higher. CSCs are filled with an Ar-CO₂-CF₄ gas mixture. Their segmented cathode design, perpendicular to the anode wires, enables three-dimensional measurements with shorter drift times compared to DTs.

RPCs are installed in both barrel and endcap regions, covering $|\eta| < 1.6$. While offering lower spatial resolution, RPCs provide rapid response and excellent timing resolution, crucial for identifying the primary bunch crossing. They consist of gaseous parallel plate detectors with strips between two anode-cathode Bakelite plates.

The operating principle of these gaseous detectors involves the ionization of gas by passing particles. The resulting ions are accelerated in a strong electromagnetic field, initiating an avalanche that is detected as an electrical signal. This mechanism allows for efficient muon detection and precise trajectory reconstruction.

The muon system's performance is characterized by its spatial resolution: approximately 100 μm for DTs and CSCs, and 1 cm for RPCs. This high-precision measurement, combined with the information from the inner tracking system, enables excellent muon momentum resolution and efficient triggering.[29] The expansive muon system, covering 25,000 m² of detection planes, demanded cost-effective and reliable components. Despite its size, the system achieves remarkable performance, with muons typically registering hits in four stations. These hits are combined to create high-purity muon tracks, distinguishing muons from all other particles that are stopped by the calorimeters.

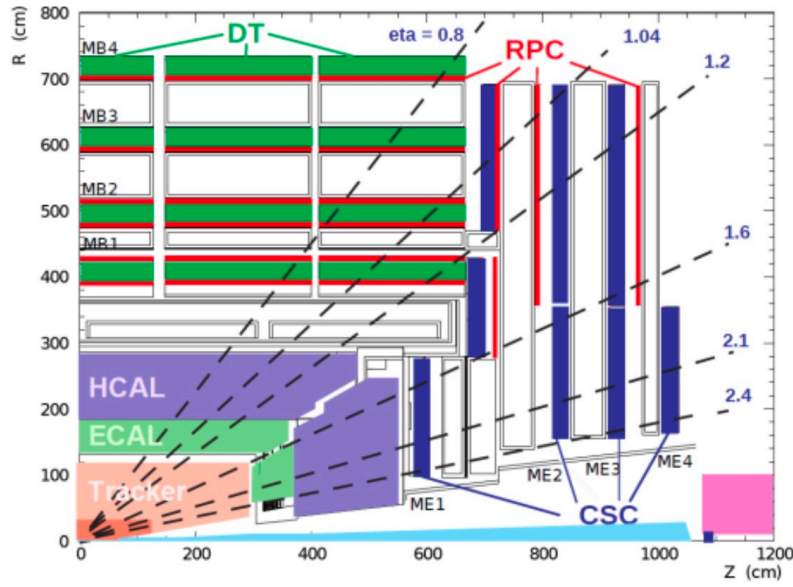


Figure 2.11: Cross-sectional view of the CMS detector in the (r,z) plane. The muon detector layout is depicted in detail. The drift tube stations (DTs) are labeled as MB (“Muon Barrel”) and the cathode strip chambers (CSCs) are labeled ME (“Muon Endcap”). The resistive plate chambers (RPCs) for the barrel and endcap regions, are labeled too.

2.7 Trigger and Data Acquisition System

CMS experiment faces an extraordinary challenge in data management. At the LHC’s design luminosity of $10^{34} \text{ cm}^{-2}\text{s}^{-1}$ and a center-of-mass energy of 13 TeV, proton-proton collisions occur at a staggering rate of approximately 1 GHz, with bunch crossings every 25 ns (40 MHz). Each event generates roughly 1 MB of data, resulting in a potential data flow of several terabytes per second. This volume vastly exceeds current storage and processing capabilities, necessitating a sophisticated trigger and data acquisition system.

The CMS Trigger and Data Acquisition System (TriDAS) is designed to address this challenge. Its primary function is to rapidly identify and select physically significant events while discarding the overwhelming majority of less interesting ones, predominantly soft multi-jet scattering processes. The TriDAS accomplishes this task through a two-tiered approach: the Level-1 (L1) Trigger and the High-Level Trigger (HLT).[30]

The L1 Trigger, implemented on custom electronics, forms the first line of defense against the data deluge. It processes partial information from the muon detectors and calorimeters to make rapid decisions about event significance. The L1 Trigger identifies signatures consistent with muons, electrons, photons, jets, τ -leptons, and missing transverse energy. Operating with a maximum latency of 4 μs per event, it reduces the event rate from 40 MHz to a maximum of 100 kHz. To minimize latency, the L1 Trigger hardware is situated near the experimental cavern.

Events passing the L1 Trigger are then subjected to the more sophisticated HLT. Unlike its hardware-based predecessor, the HLT is a software trigger running on a dedicated computing farm. It has access to the complete detector readout and employs algorithms similar to those used in of-

fine reconstruction, albeit optimized for speed. The HLT further reduces the event rate to approximately 1000 Hz, suitable for permanent storage.

The HLT operates through a series of paths, each consisting of sequential algorithms built from modules representing various object reconstruction and selection steps. These paths are designed with increasing complexity, allowing for efficient use of computing resources. Events are processed through these paths until they are either rejected or pass all steps. This approach ensures that the most computationally intensive operations are performed only on the most promising events.

Both L1 and HLT paths incorporate prescale factors, allowing for the collection of events that would otherwise be lost due to rate limitations. These prescaled events are invaluable for testing, efficiency measurements, and calibration purposes.

The TriDAS not only selects events for storage but also plays a crucial role in monitoring detector performance. It reserves a small sample of rejected events for this purpose and operates the Detector Control System (DCS), which oversees all detector components and the experiment's infrastructure.

To manage the immense data volume produced by CMS, the Worldwide LHC Computing Grid (WLCG) was established. This distributed computing infrastructure is organized in tiers, with the Tier 0 center at CERN serving as the central hub for initial data processing and distribution. Tier 1 and Tier 2 centers, located globally, provide additional storage and processing capabilities, enabling researchers worldwide to access and analyze CMS data. In conclusion, the CMS Trigger

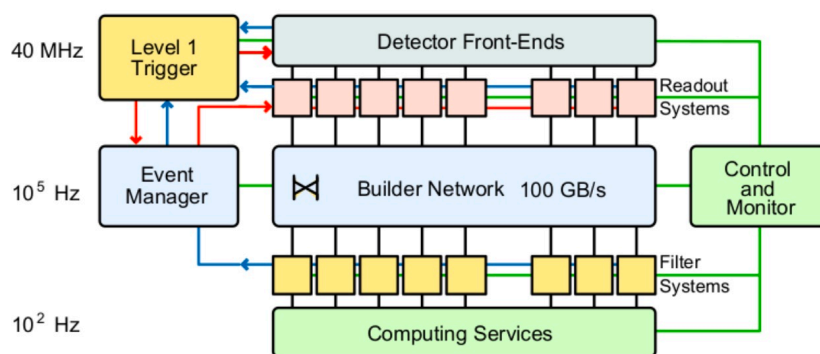


Figure 2.12: Architecture of the CMS DAQ system

and Data Acquisition System balances the need for high event selectivity with the preservation of rare and potentially groundbreaking physics processes.

Chapter 3

Physics of pp collisions at the LHC

3.1 Hard scattering processes

At its core, the proton consists of three valence quarks (uud), but its true complexity arises from the dynamic state of gluons and quark-antiquark pairs known as sea quarks. These components, collectively referred to as partons, are in constant interaction, shaping the proton's structure through processes like gluon splitting and quark-antiquark pair annihilation.

The key feature relevant to proton collisions is the distribution of the proton's four-momentum among its constituent partons [31]. For a parton of gluon type or light quarks ($i = g, u, \bar{u}, d, \bar{d}$), the probability density of carrying a fraction x of the proton's momentum is given by $f_i(x, \mu_F^2)$, where μ_F is the factorization scale.

The factorization scale μ_F separates the hard scattering component of the parton interaction, which can be computed using perturbative methods, from the low-energy region where perturbation theory breaks down, and the process is instead described by probability densities. In this sense, μ_F plays a role similar to the renormalization parameter μ_R , with the difference that it is associated with low-energy infrared (IR) divergences rather than ultraviolet (UV) divergences related to renormalization.

In proton-proton collisions, the distribution of the proton's momentum among its constituent partons is described by Parton Distribution Functions (PDFs). These functions, denoted as $f_{i/p}(x_i, Q^2)$, represent the probability density for a parton i to carry a fraction x_i of the proton's momentum p at an energy scale Q^2 [32]. The energy scale Q^2 is crucial, as it determines the resolution at which the proton's internal structure is probed. [33].

The DGLAP (Dokshitzer-Gribov-Lipatov-Altarelli-Parisi) evolution equations play a crucial role in describing how the parton distribution functions (PDFs) evolve with the energy scale Q^2 . These equations enable the calculation of quark and gluon structure functions, which ultimately produce proton, neutron, and deuteron structure functions. The factorization theorem in QCD allows for the separation of the hard scattering process from the soft, non-perturbative structure of the proton, confining the latter to universal, measurable PDFs. This factorization has been experimentally confirmed and is essential for making predictions about processes at high-energy colliders like the LHC. This principle holds true for certain processes, resulting in well-understood PDFs across a wide range of momentum fractions and scales, as shown in Figure 3.1. [34]

At high Q^2 values, smaller distances within the proton are analyzed, revealing a more detailed picture that includes sea quarks and gluons. Conversely, at lower energy scales, the PDFs primarily

reveal the valence quarks.

It is common to set $\mu_F = \mu_R = Q$, where Q represents the characteristic energy scale of the studied process.

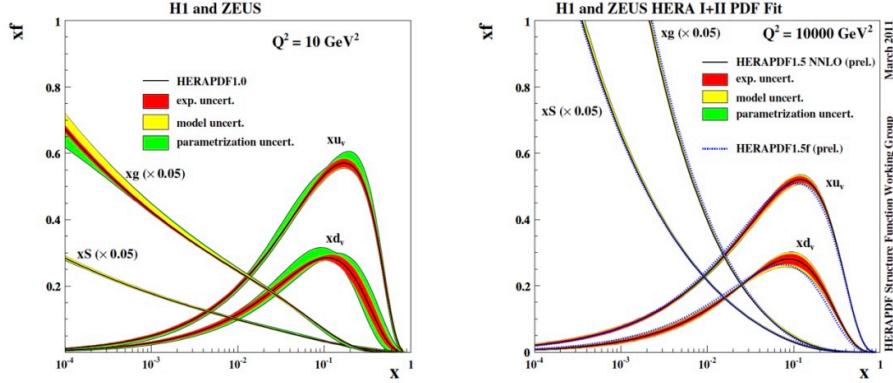


Figure 3.1

Figure 3.2: The PDFs from HERAPDF at $\mu_F^2 = 10$ (Left) and at $\mu_F^2 = 10000$ (right).[35]

The center-of-mass energy $\sqrt{s}$ plays a significant role in proton collisions at accelerators like the LHC. For protons accelerated to the same energy E_p and moving in opposite directions, the center-of-mass energy is given by $\sqrt{s} = 2E_p$. However, the actual hard scattering occurs between partons from the two protons, with four-momenta $p_i = x_i P_1$ and $p_j = x_j P_2$. The center-of-mass energy for this parton-level interaction is $\hat{s} = (p_i + p_j)^2 = x_i x_j s$

Cross-section calculations for hard scattering events involve two key components: the partonic scattering cross-section $\hat{\sigma}$ and the PDFs $f_{i,p}$. A generalized expression for the process $pp \rightarrow X$ can be formulated as: Hard scattering refers to high-energy collisions between partons (quarks or gluons) inside protons. These interactions occur at very short distances, much smaller than the proton's radius. Hard scattering processes are characterized by high momentum transfer (Q) between colliding partons, where Q is much larger than the QCD scale ($\Lambda_{QCD} \approx 1$ GeV) and interaction timescales of order ($\delta t \approx 1/\sqrt{s}Q^2$), allowing the use of perturbative QCD calculations. Thus, hard scattering provides the ability to probe the internal structure of protons at small distance scales, specifically its constituents: quarks and gluons.[36]

The cross-section calculations for hard scattering events incorporate

$$\sigma(pp \rightarrow X) = \sum_{i,j} \int dx_1 dx_2 f_{i,p}(x_1, \mu_F^2) f_{j,p}(x_2, \mu_F^2) \hat{\sigma}_{ij}(x_1 x_2 s, \mu_R^2, \mu_F^2) \quad (3.1)$$

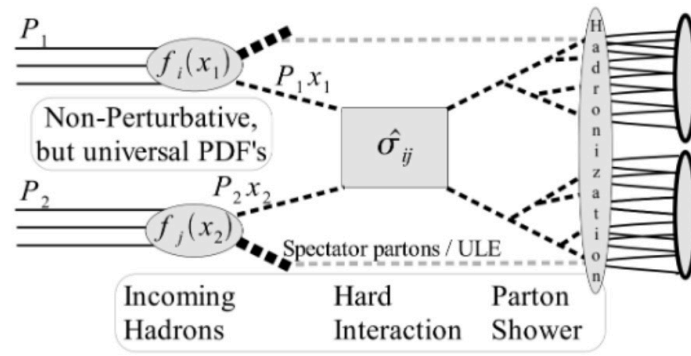


Figure 3.3

Figure 3.4: Schematic representation of hard scattering of the partons in pp collisions.

The summation covers all possible initial-state partons, where x_1 and x_2 are the longitudinal momentum fractions of the partons. $\sqrt{x_1 x_2 s}$ represents the center-of-mass energy of the parton-parton system, while μ_F and μ_R denote the factorization and renormalization scales, respectively.

As discussed in the previous chapter, particle production in hadron-hadron collisions involves several complex processes: parton-parton scatterings, initial-state radiation (ISR), final-state radiation (FSR), and beam-beam remnants (BBR). Furthermore, the high parton densities in pp collisions often lead to multiple parton interactions (MPI). The combination of particle production from BBR, ISR, FSR, and MPI is collectively termed the underlying event (UE).[37]

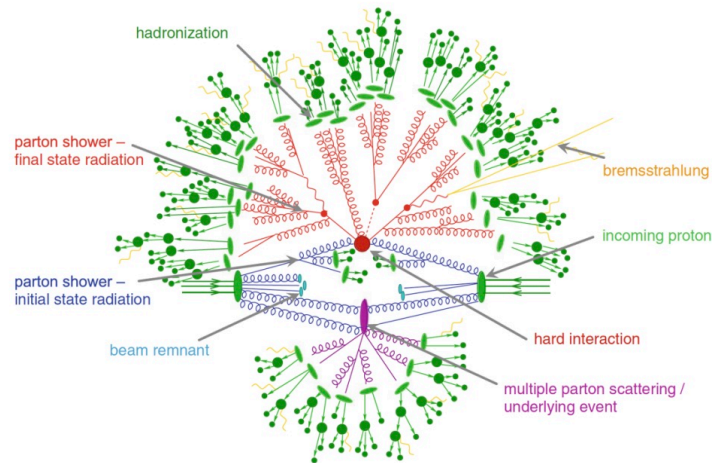


Figure 3.5: Illustration of a typical proton-proton collision event. The incoming partons are in blue, red shows hard scattering, underlying events are in purple, and hadronization is shown in green.[38]

Accurate simulations of these processes are crucial for precision measurements of the Standard Model and beyond at the LHC. As it will be discussed in the next section, Monte Carlo simula-

tions model the evolution of parton showers, accounting for the probabilistic nature of quantum processes involved in secondary particle creation from gluon emissions.

3.2 Monte Carlo Event Generators

Monte Carlo (MC) event generators like HERWIG, PYTHIA, and SHERPA simulate high-energy collisions, playing a crucial role in data analysis and experimental planning. These generators describe physics from short distances, based on perturbative QCD, up to hadron formation and decay, which relies on QCD-inspired models. Short-distance physics involves simulating QCD and QED radiation via shower algorithms, handling collinear and soft emissions, and incorporating concepts such as angular ordering and initial-state radiation. [39]

Hadronization models describe the non-perturbative transition from partons to hadrons, with popular models including the string model (Lund model) and the cluster model. Models for soft hadron physics address aspects like the underlying event, Bose-Einstein correlations, and color-reconnection effects.

Modern generators employ matrix element corrections (MEC), ME+PS (Matrix Elements and Parton Shower matching), and NLO+PS (Next-to-Leading Order + Parton Shower) techniques to improve accuracy. Additionally, the treatment of massive quarks, color information, and electromagnetic corrections is considered. For Beyond-the-Standard-Model physics, relevant hard processes and decay chains are implemented, with specific handling of short-lived particle decays.

In conclusion, generator-level datasets are produced in two ways: using a general-purpose generator to simulate the event and the hadronisation (Eg: Pythia, Herwig, Tauola) or using a Matrix Element (ME) generator to deliver the event at the parton level and then a general-purpose generator to hadronise the event (eg: Powheg, MadGraph5_aMCatNLO, Alpgen).

After the four-vectors of the final-state particles have been generated by the Monte Carlo generators, a detailed simulation of the detector is performed with GEANT4, which accurately models the interaction of particles with the detector's material and geometry. This process covers a wide range of electromagnetic, hadronic, and optical interactions over an extensive energy range, from a few hundred eV to the TeV scale.

3.3 Particle reconstruction and identification

3.3.1 The Particle Flow Algorithm

As discussed in the previous chapter, the produced particles from the collisions can be identified based on their interaction signatures with the different subsystems: charged particles “hit” the tracking system, electrons and photons deposit energy in the ECAL, charged and neutral hadrons are absorbed by the HCAL, and muons uniquely leave an imprint on the muon chambers. Event reconstruction in the CMS experiment is essential for transforming raw detector data into meaningful physics objects, enabling effective analysis of particle collisions. At the heart of CMS event reconstruction lies the particle-flow (PF) algorithm [40], which combines information from all subdetectors to identify and reconstruct individual particles within each event. This leads to the creation of PF candidates, providing a global description of the event. The reconstruction processes for electrons, muons, secondary vertices (SVs), jets, and missing transverse energy. The Particle Flow (PF) algorithm initiates by reconstructing tracks from pixel and strip hits in the in-

ner tracker. This process is facilitated by the Combinatorial Track Finder (CTF), which employs the Kalman filter, as detailed in References and. The Kalman filter's primary function is to connect hits across successive tracker layers and fit them to a helical curve. It imposes adjustable constraints on the maximum number of missing hits, the minimum number of sequential hits, and other parameters to ensure that reconstructed tracks meet a minimum quality standard.

Following track reconstruction, the algorithm proceeds to group all identified tracks into vertices. A vertex represents a point in space where a collision or decay occurs, resulting in emitted particles. Vertices are reconstructed by extrapolating tracks backward in time to determine their intersection points. Each bunch crossing can contain multiple proton-proton collisions and intermediate particle decays, leading to multiple vertices. The vertex with the highest sum of transverse momentum squared is designated as the primary vertex, while others are classified as secondary. Typically, secondary vertices are disregarded in event analysis, although displaced secondary vertices are utilized in specific reconstructions, such as b-jets.

Before particle reconstruction, another crucial step involves identifying clusters of energy deposits in the electromagnetic calorimeter (ECAL) and hadronic calorimeter (HCAL). These clusters help differentiate between electrons and photons, as well as between charged and neutral hadrons.

With tracks, vertices, and clusters reconstructed, the algorithm proceeds to identify muons. Muons are characterized by a clean and distinctive signature. They are reconstructed by aligning a track in the inner tracker with a track in the muon system, resulting in what is known as a global muon. It is also possible to reconstruct muons using only the inner tracker (tracker-only muons) or the muon system alone (standalone muons).

Subsequently, electrons and isolated photons are identified, as their reconstruction relies on ECAL energy deposits. An electron candidate is a charged track in the tracker that matches a comparable ECAL energy deposit, while a photon candidate is an ECAL deposit without a corresponding track. Additional criteria are applied to confirm these identifications.

Finally, any remaining particles are classified as hadrons. However, the algorithm also aims to reconstruct non-isolated photons and muons originating from hadron decays. Any non-isolated ECAL deposit not linked to a track is assigned as a photon, and any HCAL deposit not linked to a track is considered a neutral hadron. Any remaining HCAL deposit linked to a track is classified as a charged hadron. More details about hadronic jet reconstruction are provided in the next subsection. Furthermore, details about requirements for object selection in the current analysis are provided in the section of Object Configuration .

3.3.2 Reconstruction of Jets

A jet is a collection of hadrons and other particles produced by the hadronization of a quark or gluon, confined within a cone of defined radius ΔR . Quarks and gluons from inelastic scattering, being colored partonic states, cannot exist as free particles due to QCD constraints. Instead, they hadronize into colorless bound states, converting their energy and causing a narrow particle shower[1].

Jets are categorized into two types:

- Prompt jets originating from the primary collision
- Pile-up jets from secondary collisions of the same crossing beams

Jet reconstruction is crucial for analyzing final hadronic states as jets carry information about the initial partons. A jet is divided into three levels as shown in Figure 3.6:

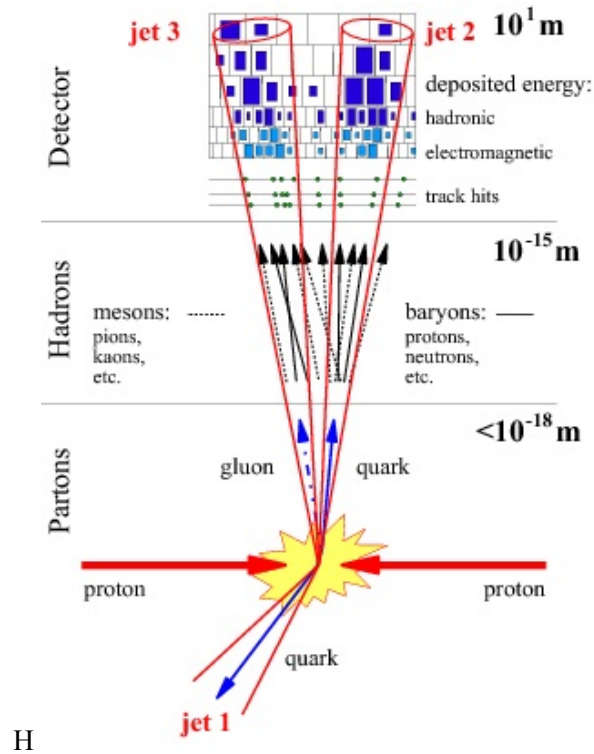


Figure 3.6: Representation of an event, where the point of proton collision is shown, the production of the initial partons, the cascade caused by their hadronization, and finally their detection through the interaction of stable particles with the detection devices.

1. Parton Level: The initial parton (quark or gluon) from which the jet originated
2. Particle Level: Particles produced from the initial parton after hadronization
3. Detector Level: The jet as detected by the interaction of shower particles with detector systems

The goal of jet reconstruction is to extract information about the initial parton from detector-level data [41]. Various algorithms have been developed to effectively reconstruct and study jets, combining information from calorimeters and the tracking system.

Key characteristics of these algorithms include:

- Determining jet size (e.g AK4 or AK8 jets)
- Being IRC safe (Infrared and Collinear safe):

An important part of jet reconstruction is energy calibration, which adjusts the correlation between particle energy and detector signal. This involves initial calibration with test beams and additional calibration during experiment operation using well-studied physical processes.

The CMS has developed a multi-level jet energy correction system. Correction factors are obtained from simulations based on test beam data, collision data, and simulations tuned to collision data [?]. The corrections are applied in the following order:

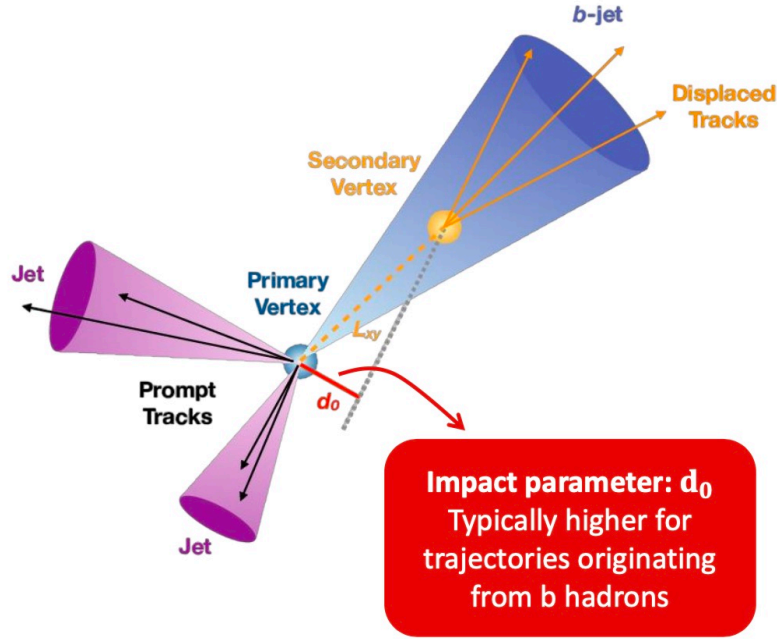


Figure 3.7: Impact parameters and SVs in b-tagging

1. Offset correction for pile-up and electronic noise
2. Relative η correction to make jet response flat as a function of η_{jet} for a specific p_T
3. Absolute p_T correction to make jet response flat as a function of $p_{T_{jet}}$

CHS, PUPPI, and Soft Drop algorithms, used in the current analysis, are further referred to in the section.

3.3.3 b-quark identification

b-tagged jets originating from the hadronization of b-quarks, serve as a crucial signature in numerous physics analyses, including the one discussed in this thesis. Due to the relatively long lifetime of b-hadrons ($\tau = 1.5$ ps), these particles travel a significant distance before decaying, leading to the formation of a secondary vertex that can be reconstructed using data from the CMS tracker. The secondary vertex (SV) information plays a fundamental role in b-tagging techniques, necessitating precise reconstruction.[42] SVs are reconstructed from tracks with large impact parameters relative to a primary vertex using the inclusive vertex finding (IVF) algorithm. This method also considers the angular distributions of tracks, as illustrated in Figure 3.7 and removes tracks that could create ambiguity in primary and secondary vertex identification. [43]The selected tracks are then clustered to reconstruct a secondary vertex. The reconstructed secondary vertices, along with other discriminating features such as track multiplicity, are utilized to determine jet flavors. Description of DeepJet algorithm used for b-tagging of jets, as well as of ParticleNet_MD used for double b-tagging of large-radius jets are further referred to in the section.

3.4 Kinematic variables

In the LHC, proton collisions at a center-of-mass energy of 13 TeV are characterized as follows: two protons are accelerated to an energy of $E_{\text{proton}} = 6.5$ TeV each and collide with opposite momenta along the z -axis. The four-momenta of the protons can be written as:

$$P_1^\mu = (E_{\text{proton}}, 0, 0, E_{\text{proton}}), \quad (3.2)$$

$$P_2^\mu = (E_{\text{proton}}, 0, 0, -E_{\text{proton}}), \quad (3.3)$$

where at these energies, the proton mass is negligible compared to its kinetic energy.

In the CMS experiment, the coordinate system is established with its origin at the interaction point. The x -axis is oriented radially inward toward the LHC center, the y -axis extends vertically upward, and the z -axis aligns with the beam direction, pointing towards the Jura mountains [44].

The radial angle is expressed as:

$$r = \sqrt{x^2 + y^2} \quad (3.4)$$

The azimuthal angle ϕ is measured relative to the z -axis, as it is depicted in Figure 3.8.

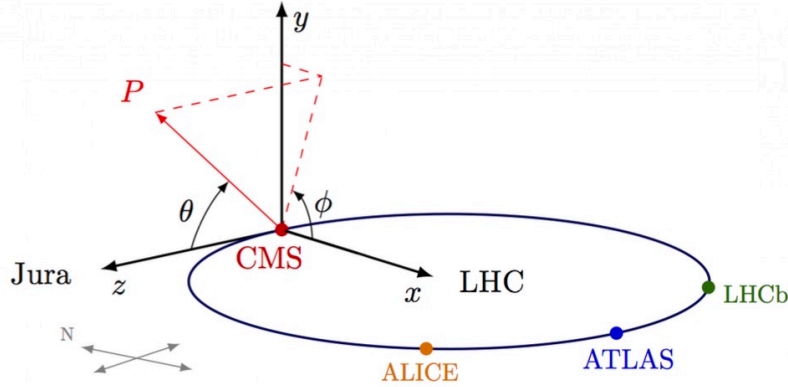


Figure 3.8: Illustration of CMS coordinate system.

The three-momentum of each particle $\vec{p} = (p_x, p_y, p_z)$ is decomposed into its longitudinal component, p_z , along the z -axis and its transverse component, p_T . The transverse momentum vector and its magnitude are given by:

$$\vec{p}_T \equiv (p_x, p_y, 0) = (p_T \cos \phi, p_T \sin \phi, 0), \quad (3.5)$$

$$p_T \equiv |\vec{p}_T| = \sqrt{p_x^2 + p_y^2}. \quad (3.6)$$

The azimuthal angle ϕ is expressed in terms of the momentum components as:

$$\phi = \tan^{-1} \left(\frac{p_y}{p_x} \right). \quad (3.7)$$

Both the transverse momentum p_T and angle ϕ are Lorentz invariant under longitudinal boosts, unlike the component p_z . The polar angle θ is also not invariant:

$$\theta = \tan^{-1} \left(\frac{p_T}{p_z} \right). \quad (3.8)$$

Since particle collisions occur near the CMS center, it is useful to describe particle trajectories in terms of angular coordinates. However, using the polar angle θ from spherical coordinates is not ideal for high-energy relativistic physics, as it complicates transformations between different reference frames. Instead, collider physics employs rapidity, defined as:

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right) \quad (3.9)$$

This quantity y , measures the relativistic velocity of a particle. It has the property that under a longitudinal boost it changes only by a constant, so that changes Δy remain invariant.

For massless particles ($m = 0$), where $p_z = |\vec{p}| \cos \theta = E \cos \theta$, rapidity simplifies to:

$$y = \frac{1}{2} \ln \left(\frac{1 + \cos \theta}{1 - \cos \theta} \right) = \ln \left(\cot \frac{\theta}{2} \right). \quad (3.10)$$

The difference in rapidity (Δy) between two particles remains invariant under Lorentz transformations. However, determining rapidity requires measuring both the energy (E) and the longitudinal momentum component (p_z). To simplify measurements, from this relation with θ , we define a geometric quantity called pseudorapidity η , given by:

$$\eta = -\ln \tan \frac{\theta}{2}. \quad (3.11)$$

For massless particles, $\eta = y$, while for particles with mass, variations in $\Delta \eta$ are not Lorentz invariant under boosts. Figure ?? shows values of η for different angles θ .

Also, at relativistic speeds, pseudorapidity approximates rapidity. Furthermore, the azimuthal angle difference $\Delta \phi$ is also Lorentz-invariant because it is measured transversely to the z-axis. Consequently, the distance between two physics objects in (η, ϕ) plane is defined in a Lorentz-invariant manner using:

$$(\Delta R)^2 = (\Delta \eta)^2 + (\Delta \phi)^2 \quad (3.12)$$

This metric is useful for defining cones around particles or evaluating the spatial separation between two objects in a relativistically invariant way. For fixed ΔR , larger values of η correspond to smaller angular openings $\Delta \phi$.

Unobservable momenta are inferred indirectly through missing transverse momentum (p_T^{miss}), defined as:

$$\vec{p}_T^{\text{miss}} = - \sum_i \vec{p}_{T,i}, \quad i = (\text{leptons}, \text{jets}) \quad (3.13)$$

where missing transverse energy (E_T^{miss}) quantifies energy lost due to undetected particles: Similarly, missing transverse energy (MET), denoted as E_T^{miss} , is defined as the energy lost from an event due to undetectable particles. It is equal to the magnitude of the negative vector sum of the transverse momenta of all detected particles:

$$E_T^{\text{miss}} \equiv |\vec{p}_T^{\text{miss}}| = \left| - \sum_i \vec{p}_{T,i} \right|, \quad i = \text{leptons, jets}. \quad (3.14)$$

”Fake” MET refers to an apparent imbalance in transverse energy, which, however, is not caused by undetectable particles. Instead, some of the causes of fake MET include:

energy mis-measurements i.e., errors or inaccuracies in energy measurements by the detector can lead to incorrect estimates of particle energies or gaps in detector coverage, caused when certain regions of the detector may fail to properly detect particles, creating "blind spots" in the measurement. Distinguishing between real and fake MET is crucial for accurate data analysis in high-energy physics experiments. Additional quantities used in collision physics include the scalar sum of the transverse momenta of hadronic jets, known as H_T :

$$H_T = \sum_j |\vec{p}_{T,j}|, \quad j = \text{jets} \quad (3.15)$$

and the scalar sum of the transverse momenta of all detectable particles, both hadronic and leptonic in nature, known as S_T :

$$S_T = \sum_i |\vec{p}_{T,i}|, \quad i = \text{leptons, jets}. \quad (3.16)$$

Another important quantity is the invariant mass, m , which for a particle with four-momentum $p^\mu = (E, \vec{p})$ is equal to the rest mass of the particle and is defined as:

$$m = \sqrt{p^\mu p_\mu} = \sqrt{E^2 - \vec{p}^2}. \quad (3.17)$$

The distribution of the invariant mass of a set of particles in the final state of a decay, when calculated for a large number of events, should exhibit a peak at the rest mass of the decayed particle. For example, in the decay $Z \rightarrow l^+ l^-$, the invariant mass of the two leptons will be:

$$m_{ll} = \sqrt{\left(\sum_{i=l^\pm} p_i^\mu \right)^2} = \sqrt{(E_1 + E_2)^2 - (\vec{p}_1 + \vec{p}_2)^2} \quad (3.18)$$

and the distribution of m_{ll} will show a maximum at the mass m_Z of the Z boson.

However, there are cases where not all decay products in the final state are detectable particles. This occurs in decays involving neutrinos, such as $W \rightarrow \nu_l l$. In such cases, the invariant mass cannot be determined, and instead, we use a corresponding quantity called the transverse mass, m_T , which is calculated only from the transverse components of momentum according to the relation:

$$m_T = \sqrt{(E_{T,1} + E_{T,2})^2 - (\vec{p}_{T,1} + \vec{p}_{T,2})^2}. \quad (3.19)$$

The quantity $E_T = \sqrt{m^2 + p_T^2}$ is analogously the transverse energy. If $m_T = m$, then the momenta p_1 and p_2 lie purely in the transverse plane ($\eta = 0$), while if $m_T = 0$, the momenta p_1 and p_2 are longitudinal. In other cases, $0 \leq m_T \leq m$.

Chapter 4

Analysis

4.1 The $h \rightarrow aa \rightarrow 4b$ decay

This analysis examines the $h \rightarrow aa \rightarrow 4b$ decay in a model-independent manner; no assumptions were made about model-dependent parameters, and so our final results are also model-independent. However, this exotic decay is favored within the framework of the Two-Higgs Doublet + Scalar (2HDM+s) model. A more detailed reference to the 2HDM+s is provided in this section. The motivation for studying exotic decays in the $h \rightarrow 4b$ channel will also become clear.

4.1.1 2HDM+s Model

As it was discussed in section 1.2, the 2HDM+s model introduces two CP-odd pseudoscalar bosons (A, S_I) and five CP-even scalar bosons ($h, H^0, H^\pm, S_R$). The S_I and S_R states correspond to components of the additional scalar singlet field S . One can choose a singlet-like pseudoscalar state of the form:

$$a = \cos \theta_a S_I + \sin \theta_a A, \quad \theta_a \ll 1 \quad (4.1)$$

where θ_a is the mixing angle between the doublet-sector and singlet-sector pseudoscalars. This state, lighter than the SM Higgs boson, allows for two possible exotic decays:

- $h \rightarrow Za$, where $M_a < M_h - M_Z \approx 35 \text{ GeV}$
- $h \rightarrow aa$, where $M_a < M_h/2 \approx 63 \text{ GeV}$

The decay of "a" into SM fermions occurs when the A couplings are multiplied by the factor $\sin \theta_a$.

Figures 4.1 - 4.4 show the branching ratio $BR(a \rightarrow \text{SM})$ of the pseudoscalar "a" to SM particles as a function of its mass M_a for various possible decays. These ratios depend on the crucial parameter $\tan \beta$ and the 2HDM type. Second and third-order radiative corrections for quark decays are considered. The four types of 2HDM and their corresponding $BR(a \rightarrow \text{SM}) - M_a$ diagrams are [18] :

- Type I: All fermions couple to Φ_2 . Branching ratios are independent of $\tan \beta$.
- Type II: Up-type quarks couple to Φ_2 , while leptons and down-type quarks couple to Φ_1 . Decays to down-type quarks are enhanced for $\tan \beta > 1$ and suppressed for $\tan \beta < 1$.

- Type III: Leptons couple to Φ_1 , while all quarks couple to Φ_2 . Leptonic decays are enhanced for $\tan \beta > 1$, while quark decays are enhanced for $\tan \beta < 1$.
- Type IV: Down-type quarks couple to Φ_1 , while up-type quarks and leptons couple to Φ_2 . For $\tan \beta < 1$, decays to up-type quarks and leptons are enhanced relative to down-type quarks.

Among these types, only Type I is independent of $\tan \beta$. When $\tan \beta = 1$, all types reduce to Type I.

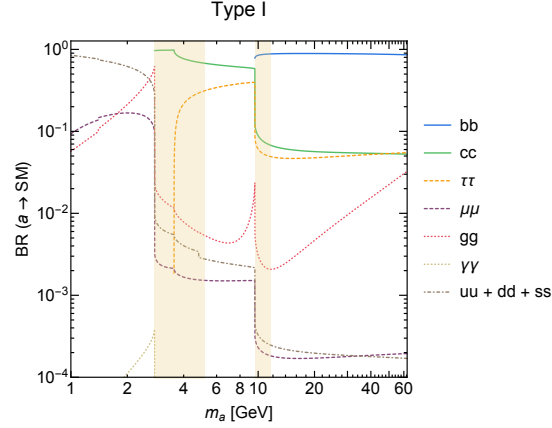


Figure 4.1: Branching ratios of a singlet-like pseudoscalar in the 2HDM+s for Type I Yukawa couplings. Decays to quarkonia likely invalidate our simple calculations in the shaded regions.

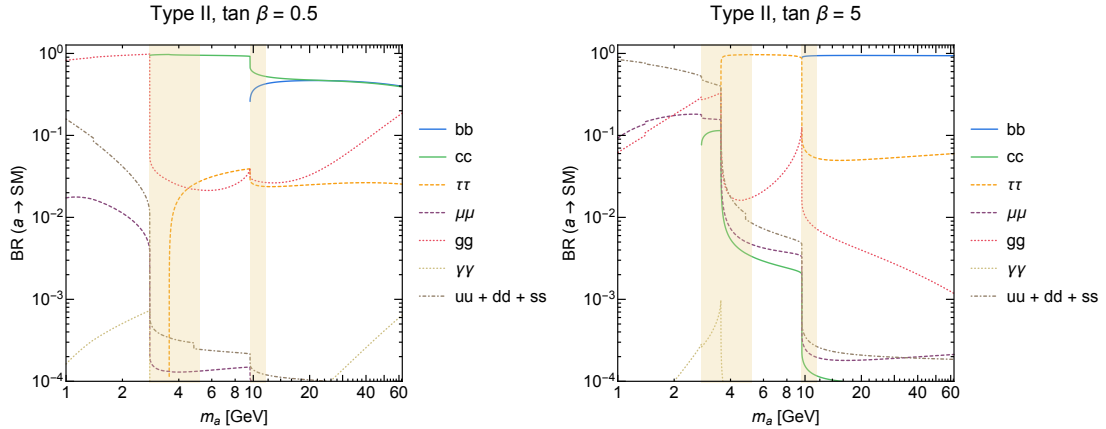


Figure 4.2: Branching ratios of a singlet-like pseudoscalar in the 2HDM+s for Type II Yukawa couplings. $\tan \beta = 0.5$: decays to up-type quarks and leptons are enhanced (left). $\tan \beta = 5$: decays to down-type quarks are enhanced (right).

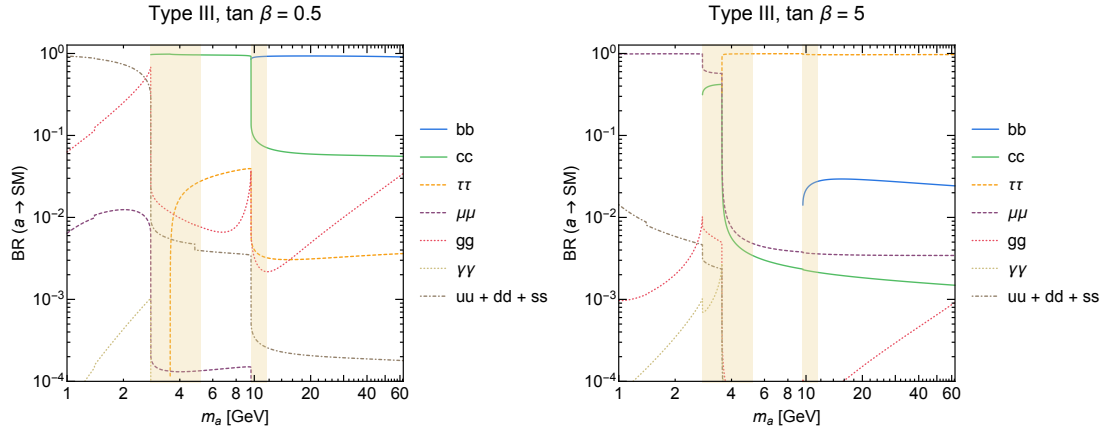


Figure 4.3: Branching ratios of a singlet-like pseudoscalar in the 2HDM+s for Type III Yukawa couplings. $\tan\beta = 0.5$: decays to quarks are enhanced (left). $\tan\beta = 5$: decays to leptons are enhanced (right).

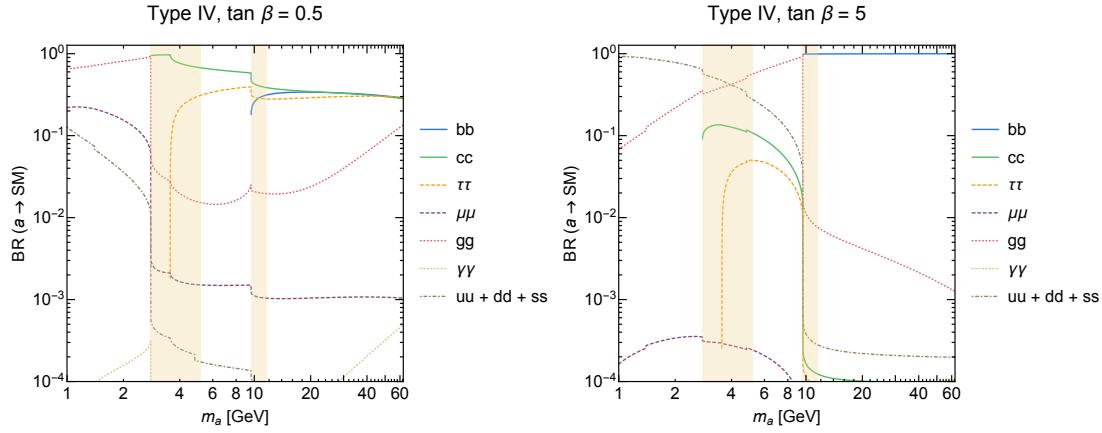


Figure 4.4: Branching ratios of a singlet-like pseudoscalar in the 2HDM+s for Type IV Yukawa couplings. $\tan\beta = 0.5$: decays to up-type quarks and leptons are enhanced (left). $\tan\beta = 5$: decays to down-type quarks are enhanced (right).

4.1.2 The $h \rightarrow 4b$ Decay Channel

In the Standard Model (SM), the Higgs boson can decay into four b quarks through the channel $h \rightarrow ZZ^* \rightarrow b\bar{b}b\bar{b}$. The branching ratio for this process is calculated as:

$$\text{Br}(h \rightarrow 4b) = \text{Br}(h \rightarrow ZZ^*) \times \text{Br}(Z \rightarrow b\bar{b})^2 \quad (4.2)$$

This value is on the order of 10^{-4} , making it a very rare final state when originating from this channel. Additionally, the b quarks from the on-shell Z boson are not collimated due to Z's large mass. The resulting experimental signature is not distinctive and is accompanied by a strong QCD background that is difficult to constrain. The combination of the low branching ratio and the persistent QCD background does not favor the study of the $h \rightarrow 4b$ channel in the SM.

However, this channel can be studied more effectively in Beyond Standard Model (BSM) scenarios, where the Higgs initially decays into two new particles, each of which then decays into two b quarks. Various BSM theories support this process. In all these models, the $X \rightarrow b\bar{b}$ decay (where X is a new particle) is the dominant mechanism for specific parameter values, providing strong motivation for researching this decay in BSM models [18]. The model of interest for this analysis is the 2HDM+s (Two-Higgs Doublet Model + scalar). In this model, the decay $h \rightarrow aa$ (where "a" is the "mostly-singlet" pseudoscalar) is possible. As shown in Figures 4.1 - 4.4 the $a \rightarrow b\bar{b}$ process (depicted by a blue line) is the dominant decay mechanism for a , especially when decays to down-type quarks are favored. This channel is only allowed when $M_a > 2M_b$, ($M_b = 4.183$ GeV), so the mass of "a" should be approximately greater than 10 GeV [18].

4.2 Signal process

The signal being studied is the exotic decay of the Higgs into two pseudo-scalar particles "a", each of them decaying directly into a pair of ($b\bar{b}$) quarks. Therefore, the mass of "a" must be in accordance with the kinematic constraints: $2M_b < M_a < M_h/2$. This study thus explores the production of new light bosons, within the kinematically allowed mass range of $12 < M_a < 60$ GeV. As discussed earlier, the motivation for studying this channel is quite strong, given that the branching fraction of the Higgs decay into exotic particles corresponds to an upper limit of 16%, a promising percentage [16]. Furthermore, the decay $a \rightarrow b\bar{b}$ is dominant in all types of 2HDM+s, as the branching fraction for this decay tends towards one when $M_a > 10$ GeV. Therefore, for the mass range we are studying, we can consider that $\text{BR}(a \rightarrow b\bar{b}) \approx 1$. In the signal, the production of the Higgs occurs in association with the Z boson ($\sigma_{Zh} = 0.89 \text{ pb}$) [13], while it decays leptonically via the channel $Z \rightarrow l^-l^+$, which corresponds to a branching fraction of about 10% or invisibly $Z \rightarrow \nu\nu$, which corresponds to a branching fraction of about 20% [15]. The leptonic final states are characterized by very clean experimental signatures. Using the events where the Z decays in this way, we can reduce the strong backgrounds that overshadow our signal, such as the production of b -jets through QCD processes. Therefore, the motivation for this research is enhanced due to the leptonic decay of the Z , which allows us to significantly distinguish the signal from the accompanying background. On the other hand, 0-lepton channel has a much higher QCD background, but has never been investigated before. A combined analysis for Zh using both the 0-lepton and 2-lepton channels is also possible in the future. The above description of the signal is illustrated in the Feynman diagram of the image 4.5.

Moreover, $hW \rightarrow a(b\bar{b})a(b\bar{b})$, where Higgs is produced in association with W boson ($\sigma_{Wh} = 1.37$ pb) acts as signal contamination in ZH analysis [45]. Its contribution to the signal is mainly significant in the 0-lepton channel, caused mainly by the hadronic decay of W boson ($\text{Br}(W \rightarrow qq) = 67.3\%$) and by W boson leptonic decay where the lepton is misidentified. Percentages of this contamination to the final expected signal yield will be provided later on. In the 2-lepton channel no such signal contamination is observed.

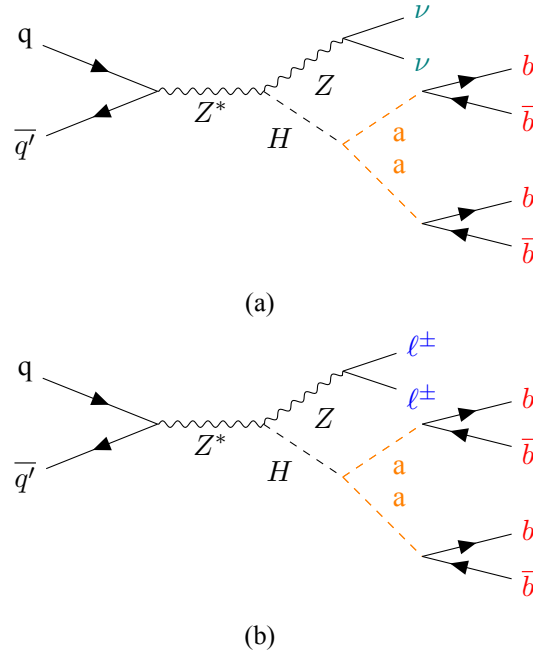


Figure 4.5: Feynman Diagrams of signal process $Zh \rightarrow a(bb)a(bb)$. Higgs is produced in association with Z boson and Z decays: (a) invisibly (b) leptonically.

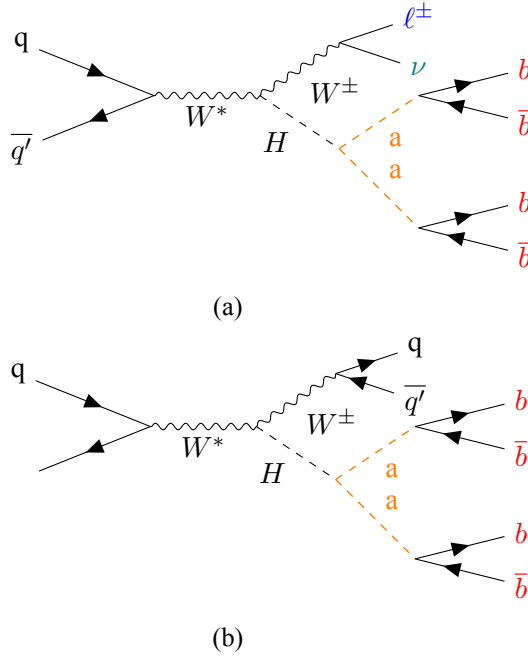


Figure 4.6: Feynman Diagram of process $Wh \rightarrow a(bb)a(bb)$. Higgs is produced in association with W boson and W decays: (a) leptonically (b) hadronically. This processes acts as signal contamination in the Zh (0-lepton) channel.

Signal events with associated VH production followed by the decay $H \rightarrow aa \rightarrow 4b$ are simulated at LO using the MADGRAPH5 aMC@NLO (version 2.6.5) generator. The cross sections are normalized to the values that are calculated in Refs and include NNLO QCD and NLO EW corrections.[46]

4.2.1 Generator level analysis

For this analysis, events generated using the PYTHIA software were utilized. PYTHIA is a Monte Carlo event generator that creates datasets with a very large number of generated events. Since the generation method is random, these data are accompanied by statistical fluctuations, which naturally decrease as the number of Monte Carlo events increases. For particle identification, PYTHIA software uses the Particle Data Group (PDG) codes, a standardized set of identification numbers for various particles [47]. Additionally, various codes are used that contain information such as the status of a particle and its mother particle identity. Below we examine some useful kinematic characteristics of the signal sample at the generator level, for three possible "a" masses. At first, the Higgs p_T , Higgs η and $\Delta R(aa)$ are shown in Figure 4.7. As it was expected, Higgs kinematics are independent of the a-boson mass. It can be shown that $\Delta R(aa)$ decreases as "a" mass increases:

$$\begin{aligned}
 p_H^\mu \cdot p_{H\mu} &= (p_{a_1} + p_{a_2})^\mu (p_{a_1} + p_{a_2})_\mu \\
 &\rightarrow m_H^2 = 2m_a^2 + 2(E_{a_1}E_{a_2} - p_{a_1}p_{a_2} \cos \Delta\varphi(a_1, a_2)), \quad m_b \ll m_a \\
 &\rightarrow m_H^2 \approx 2m_a^2 + 2p_{a_1}p_{a_2} (1 - \cos \Delta\varphi(a_1, a_2)) \\
 &\rightarrow \cos \Delta\varphi(a_1, a_2) = 1 - \frac{m_H^2 - 2m_a^2}{2p_{b_1}p_{b_2}}
 \end{aligned} \tag{4.3}$$

As it is shown in figure 4.9 the $p_T(b)$ slightly decreases when "a" mass increases so the term m_a^2 in the numerator dominates, $\frac{m_H^2 - 2m_a^2}{2p_{b_1}p_{b_2}}$ decreases and $\cos \Delta\varphi(a_1, a_2)$ increases, which means $\Delta\varphi(a_1, a_2)$ becomes smaller.

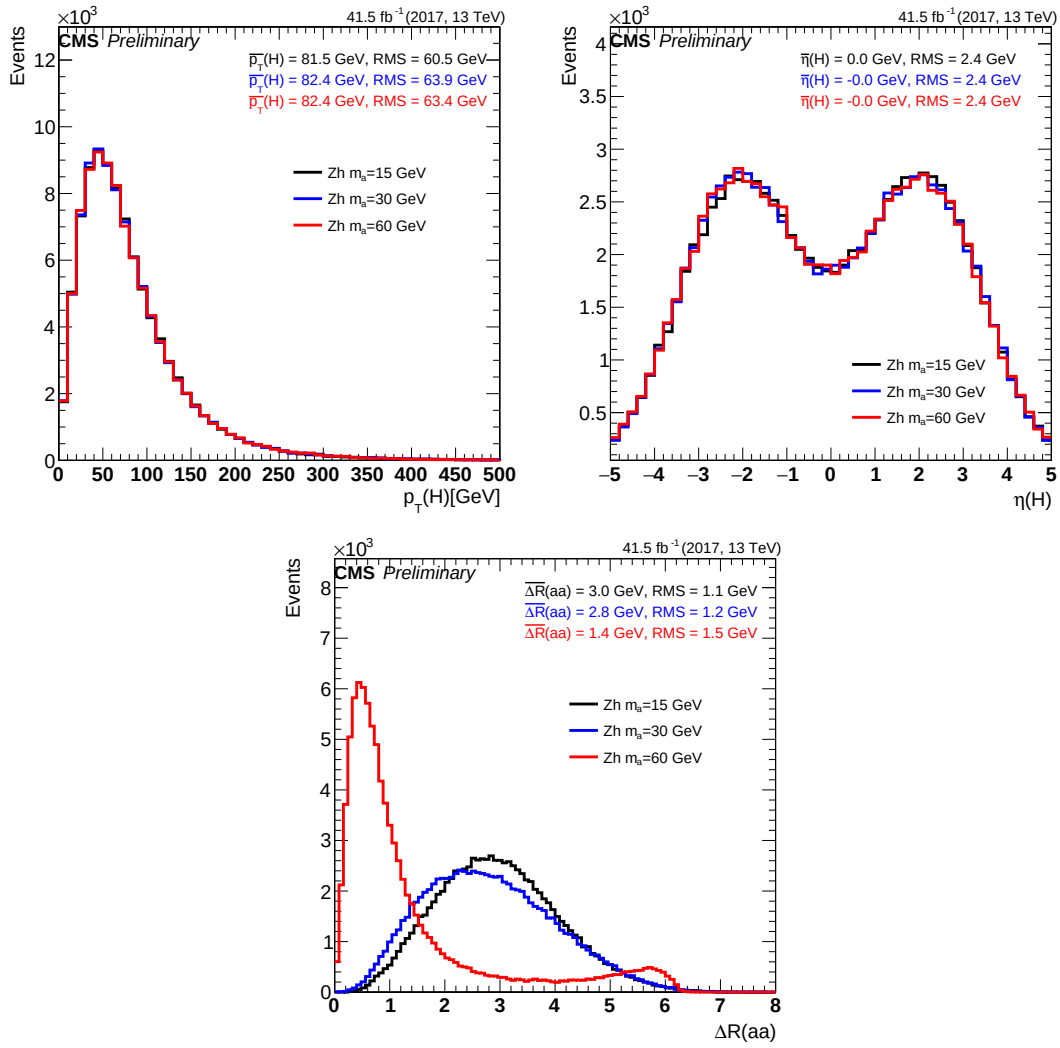


Figure 4.7: Higgs p_T (top left), Higgs η (top right) and $\Delta R(aa)$ (bottom) for signal samples of $m_a = 15, 30, 60$ GeV

On the other hand, as it is depicted in Figure 4.10, as "a" mass increases, the "a" boson p_T decreases and so, p_T the (bb) pair coming from the same "a" boson decay decreases. We further observe that, as well as "a" mass decreases, the (bb) pair is centered to the transverse plane as $\eta(bb)$ is centered to zero, while for high "a" masses the distribution has peaks around $|\eta| = 2$. Similarly, as before,

$$\begin{aligned}
 p_a^\mu \cdot p_{a\mu} &= (p_{b_1} + p_{b_2})^\mu (p_{b_1} + p_{b_2})_\mu \\
 \rightarrow m_a^2 &= 2m_b^2 + 2(E_{b_1}E_{b_2} - p_{b_1}p_{b_2} \cos \Delta\varphi(b_1, b_2)) \\
 \rightarrow m_a^2 &\approx 2p_{b_1}p_{b_2} (1 - \cos \Delta\varphi(b_1, b_2)) \\
 \rightarrow \cos \Delta\varphi(b_1, b_2) &= 1 - \frac{m_a^2}{2p_{b_1}p_{b_2}}
 \end{aligned} \tag{4.4}$$

So for smaller "a" masses $\cos \Delta\varphi(b_1, b_2)$ increases and thus $\Delta\varphi(b_1, b_2)$ becomes smaller. All the

above explains that for smaller "a" masses, $\Delta R(b_1, b_2)$ decreases, as the (bb) pair becomes more collimated, which is shown in Figure 4.9 .

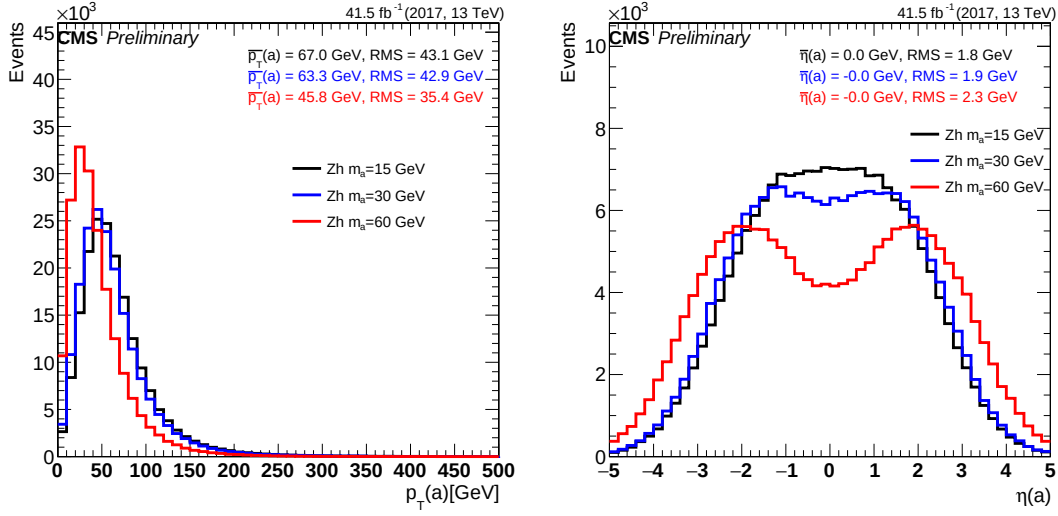


Figure 4.8: "a" boson p_T (left), "a" boson η (right) for signal samples of $m_a = 15, 30, 60$ GeV

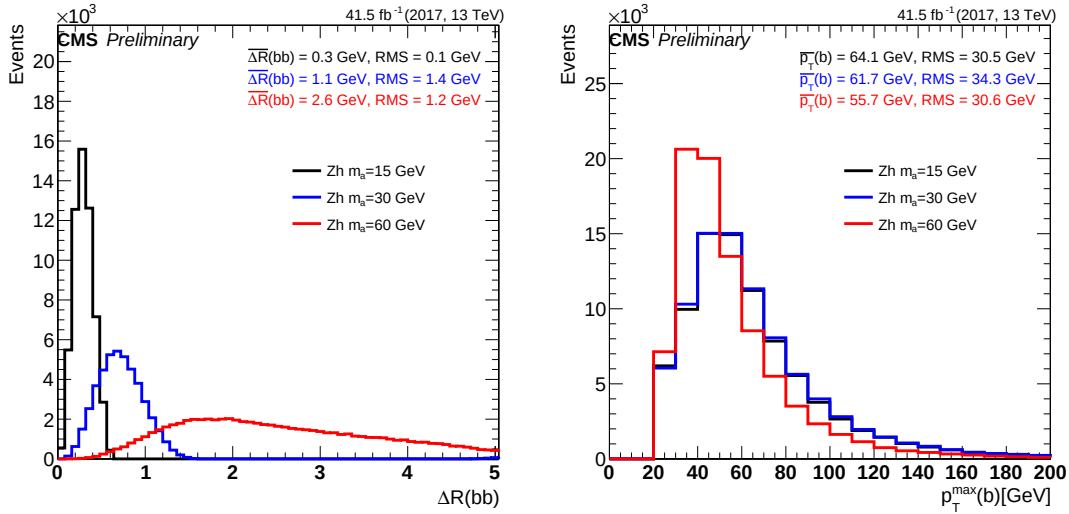


Figure 4.9: $\Delta R(bb)$ of (bb) pair coming from the same "a" boson decay (left) and p_T of b-quark with the highest transverse momentum (right) for signal samples of $m_a = 15, 30, 60$ GeV

Definition of regimes:

The analysis focuses on two regimes, the resolved and the merged one.

At first, we define the resolved regime for "a" mass samples in the range $[20, 60]$ GeV, which has already been investigated through the Vh channel in past CMS analysis Ref [48], [49]. It has been proved that for "high" "a" masses, especially for $m_a > 30$ GeV, adequate sensitivity can be achieved through classic jet reconstruction techniques (AK4 jets) in this regime.

A novel possibility with CMS will be to exploit jet substructure techniques in large radius jets, so as to enhance further the sensitivity for low masses of the "a" boson. For the time being, we define the merged regime for "a" masses in the range [15,40] GeV, where we use both a large radius jet (AK8 jet) in order to identify a collimated (bb) pair, as well as AK4 jets for the other pair. Important characteristics of AK8 jets, such as soft drop mass, as well as already existing CMS ParticleNet-MD discriminants, are applied in this regime to increase sensitivity. With full RUN2 statistics, requirement of two AK8 jets may become possible.

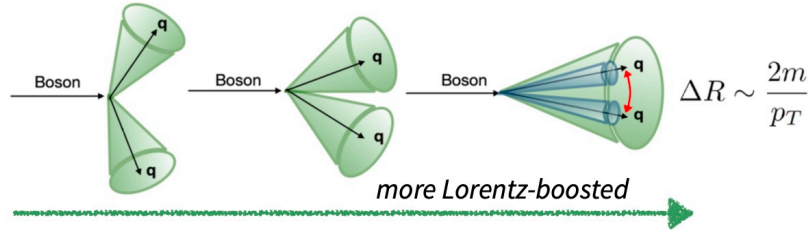


Figure 4.10: As "a" boson becomes more boosted, (bb) pair tends to be more collimated, so the two b quarks can be reconstructed as sub-jets of a single large radius jet

The above definition of regimes can be justified by the following generator level plots. In Figure 4.11 we observe $\langle \Delta R(bb) \rangle$ of (bb) pair coming from the same mother "a" boson, as a function of $p_T(bb)$. From this plot, it is derived that for $p_T > 150$ GeV, which is the selection cut for AK8 jets (as it will be discussed in the section of object configuration), for all samples with "a" masses $m_a \leq 40$ GeV, $\langle \Delta R(bb) \rangle$ is small enough to be reconstructed within the cone of a single AK8 jet.

Furthermore, it is distributed the efficiency defined as $\text{eff} = \frac{N_{bb}^{\Delta R < 0.8}}{N_{bb}^{\text{total}}}$, i.e. the ratio of the number of (bb) pairs with $\Delta R(bb) < 0.8$ divided by the total number (bb) pairs from the same A boson, as a function of $p_T(bb)$. In this plot, it is shown that for all samples with "a" masses $m_a \leq 40$ GeV the efficiency has already reached its plateau at $p_T = 150$ GeV.

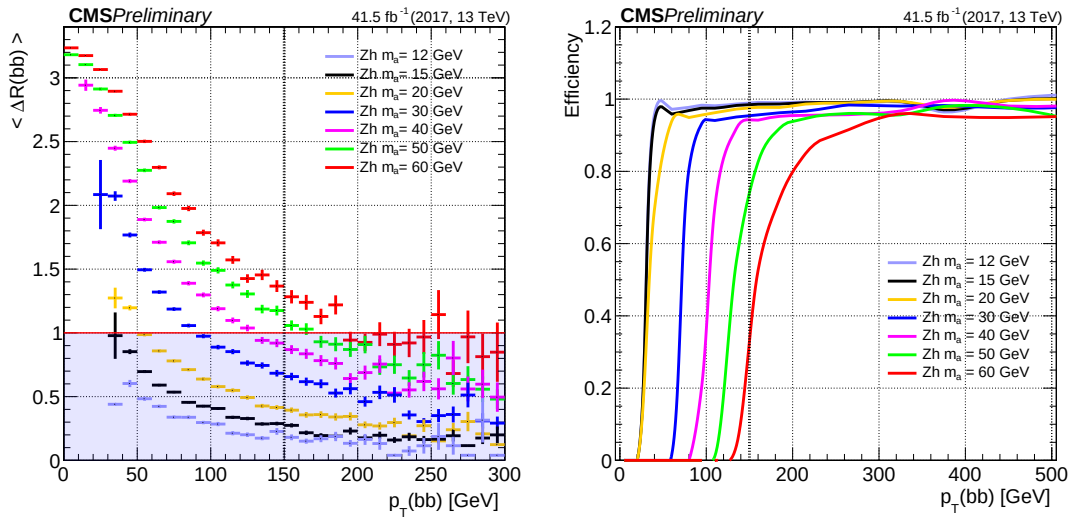


Figure 4.11: $\langle \Delta R(bb) \rangle$ as a function of $p_T(bb)$ (left), selection efficiency of (bb) pair: $\frac{N_{bb}^{\Delta R < 0.8}}{N_{bb}^{total}}$ as a function of $p_T(bb)$ (right) for signal samples of m_a in the range [12,60] GeV.

4.3 Background processes

The background processes that coexist with the signal being investigated fall within the spectrum of SM. The final states of such processes typically "mimic" the signal, making their differentiation from it challenging. A detailed presentation of all signal and background processes relevant to this analysis with their cross-sections, weight, and statistics is provided in Table 4.1. Moreover, the Monte Carlo (MC) datasets of these backgrounds are listed in the Table of Appendix A along with the respective cross-sections and k-factors. MC samples for various background processes are generated using the MadGraph5 at LO, PYTHIA [50], and POWHEG [51] generators.

Below is a brief description of each process, as well as its Feynman diagram presented.

- **$t\bar{t}$ + Jets:**

At the LHC, one of the dominant processes is the production of top quark pairs. These pairs are produced through strong interactions of QCD, either via the gluon-gluon fusion mechanism ($gg \rightarrow t\bar{t}$) or from quark-antiquark annihilation ($qq \rightarrow t\bar{t}$).

The top quarks decay almost exclusively into a W boson and a b quark ($t \rightarrow Wb$). Depending on the decay of the W bosons, the $t\bar{t}$ production processes are distinguished into three channels [15]:

- (a) $t\bar{t}$ hadronic: Both W bosons decay into quark-antiquark pairs. This channel corresponds to a branching fraction of 45.5%.
- (b) $t\bar{t}$ semileptonic: One W boson decays into a quark-antiquark pair ($W \rightarrow q\bar{q}$) and the other decays into a lepton and the corresponding neutrino ($W \rightarrow l\bar{\nu}$). The corresponding branching fraction is 43.5%.
- (c) $t\bar{t}$ dileptonic: Both W bosons decay into a lepton and the corresponding neutrino. The branching fraction for this channel is the smallest, at 10.5%.

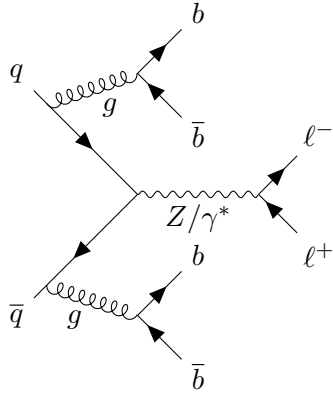
However, this dominant, irreducible background from $t\bar{t} + \text{jets}$ is simulated at next-to-leading order (NLO) accuracy with the POWHEG2.0 generator, and the splitting of it is not based in W boson decay mode. The events are categorized depending on the parton-level flavor content of additional particle-level jets not originating from the decay of the $t\bar{t} + \text{jets}$ system. Events that contain at least one additional particle-level jet matched to a b hadron are labeled as $t\bar{t} + b\bar{b}$. Events that contain at least one additional particle jet matched to a c hadron and no b hadron are labeled as $t\bar{t} + c\bar{c}$. The $t\bar{t} + b\bar{b}$ and $t\bar{t} + c\bar{c}$ categories, with heavy-flavor (HF) production, are referred to as the " $t\bar{t} + HF$ " background. This background mostly arises from events with gluon splitting, $g \rightarrow cc$ or bb . The remaining $t\bar{t} + \text{jets}$ events are referred to as the " $t\bar{t} + LF$ " background, where "LF" refers to light flavor. The " $t\bar{t} + LF$ " category also includes events with no additional particle-level jets. One source of this background is " $t\bar{t} + qq$ ($q = u, d, s$) events in which light-flavored jets are erroneously identified as b jets.

- **DY + Jets:**

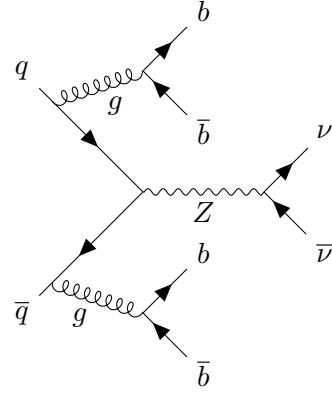
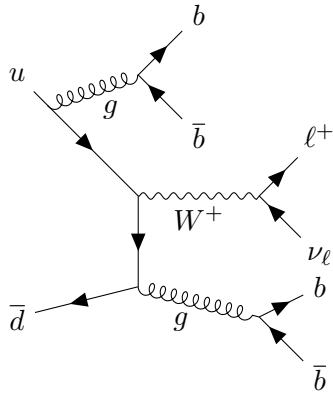
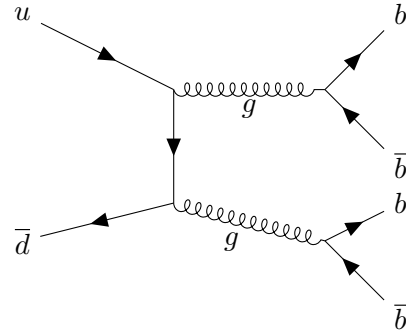
The Drell-Yan (DY) background is one of the primary processes that occur during hadron collisions in accelerator facilities. It was proposed in 1970 by Tung-Mow Yan and Sidney Drell to describe the production of lepton-antilepton pairs in high-energy hadron collisions. In this process, a quark from one hadron and an antiquark from another hadron annihilate to produce a pair of oppositely charged leptons. The production of such pairs occurs through the exchange of a virtual photon or Z boson. [41] Therefore, the leptons in the final state of the Drell-Yan process "mimic" the leptonic final state of the signal being investigated in 2-lepton channel, when Z boson is produced in association with b quark jets.

- **Z + Jets and W + Jets :**

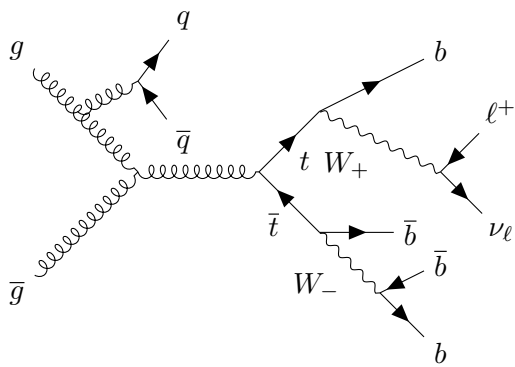
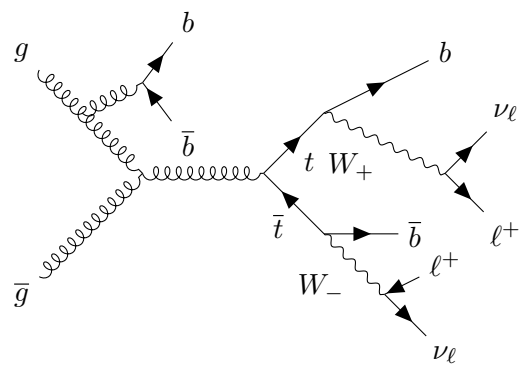
The W + jets and Z + jets backgrounds mostly arise when a W or Z boson is produced in association with b quark jets. When Z decays to neutrino pair, Z + Jets becomes a considerable background source in 0-lepton channel. Similarly, $W \rightarrow \ell\nu + \text{Jets}$ contributes to background in 0-lepton channel when the lepton is missed. In 2-lepton channel, contribution of these background processes is negligible and therefore ignored. The V+jets background is simulated at leading order (LO) accuracy with MADGRAPH5 aMC@NLO (version 2.4.2) [29], with cross sections normalized to NLO .

(a) $DY + Jets$

:

(b) $Z \rightarrow \nu\bar{\nu} + jets$ (c) $W \rightarrow \ell\nu + jets$ 

(d) QCD all hadronic

(e) $t\bar{t} \rightarrow (\ell\nu)(q\bar{q}) + b\bar{b}$ (f) $t\bar{t} \rightarrow (\ell\nu)(\ell\nu) + b\bar{b}$

Process	$\sigma(\text{pb})$	N_{stat}	Weight
$ZH : m_a = 12 \text{ GeV up to } m_a = 60 \text{ GeV}$	0.8839	75000	0.49
$WH : m_a = 12 \text{ GeV up to } m_a = 60 \text{ GeV}$	1.37	75000	0.76
$t\bar{t} : \text{Hadronic}$	378.93	235719999	0.07
$t\bar{t} : \text{Semileptonic}$	366.29	355332000	0.04
$t\bar{t} : \text{Dileptonic}$	88.51	106724000	0.03
$Z \rightarrow \ell\ell (m_{\ell\ell} > 50 \text{ GeV})$	4895	205238822	1.17
$Z \rightarrow \nu\nu (H_T : 100 \text{ to } 200)$	344.83	1898397	0.58
$Z \rightarrow \nu\nu (H_T : 200 \text{ to } 400)$	95.53	17349597	0.22
$Z \rightarrow \nu\nu (H_T : 400 \text{ to } 600)$	13.2	13963690	0.019
$Z \rightarrow \nu\nu (H_T : 600 \text{ to } 800)$	3.15	4418971	0.023
$Z \rightarrow \nu\nu (H_T : 800 \text{ to } 1200)$	1.45	1513585	0.029
$Z \rightarrow \nu\nu (H_T : 1200 \text{ to } 2500)$	0.35	267125	0.038
$Z \rightarrow \nu\nu (H_T : 2500 \text{ to } \infty)$	0.0085	176201	0.001
$W \rightarrow \ell\nu (H_T : 70 \text{ to } 100)$	1637.13	3.053	177.1
$W \rightarrow \ell\nu (H_T : 100 \text{ to } 200)$	1628.66	1.885	701.1
$W \rightarrow \ell\nu (H_T : 200 \text{ to } 400)$	435.237	42602407	0.42
$W \rightarrow \ell\nu (H_T : 400 \text{ to } 600)$	59.1811	6904016	0.36
$W \rightarrow \ell\nu (H_T : 600 \text{ to } 800)$	14.5805	1002075	0.42
$W \rightarrow \ell\nu (H_T : 800 \text{ to } 1200)$	6.65621	8436624	0.16
$W \rightarrow \ell\nu (H_T : 1200 \text{ to } 2500)$	1.60809	4955636	0.01
$W \rightarrow \ell\nu (H_T : 2500 \text{ to } \infty)$	0.0389136	1185699	0.001
$W \rightarrow \ell\nu (\text{inclusive})$	61526.7	78981243	32.33
$QCD(H_T : 300 \text{ to } 500)$	351302	55090472	261.9
$QCD(H_T : 500 \text{ to } 700)$	31630	62026995	21.48
$QCD(H_T : 700 \text{ to } 1000)$	6802	48666395	5.83
$QCD(H_T : 1000 \text{ to } 1500)$	1206	14485988	3.46
$QCD(H_T : 1500 \text{ to } 2500)$	120.4	12941701	0.32
$QCD(H_T : 2500 \text{ to } \infty)$	20.2	5940307	0.14
SingleT_s	3.36	1388200	0.01
SingleT_t_antitop	26.38	69921000	0.02
SingleT_t_top	44.33	12990300	0.01
SingleT_tW_antitop	35.85	5674000	0.06
SingleT_tW_top	35.85	5649000	0.26
TTGJets	3.697	3534208	0.01
TGJets	2.967	1965000	0.06
TTWJetsLNu	0.2043	7460671	0.03
TTZJets2L2Nu	0.2529	5696006	0.02
WplusH_HToBB_WToLNu	0.03286	4862115	0.00
WminusH_HToBB_WToLNu	0.0205509	4824500	0.00
ZH_HToBB_ZToLL	0.04865	4757700	0.00
WW	118.7	15634000	0.05
WZ	47.13	7889000	0.13
ZZ	16.523	2706000	0.25
ZZZ	0.01398	9720000	0.02
WZZ	0.05565	10196000	0.05
WWZ	0.1651	1011000	0.60
WWW	0.2086	10020500	0.40

Table 4.1: All signal and background processes used in the current analysis. Their statistics, weight and cross-section is provided as well.

- **QCD Multi-jet all Hadronic :**

The large total cross section for QCD multi-jet interactions makes it a dominant source of background in 0-lepton channel. The final state consists of many jets originating from initial or final state radiation. Misidentification of jets can result in the production of "fake" b-tagged jets. Due to Insufficient MC statistics, that leads to large weights this process generally needs data driven estimations.

- **Other Background processes :**

Small backgrounds from diboson, triboson, single-top quark, rare electroweak processes, as well as the SM $VhH \rightarrow bb$ process are merged all together in this category and have a small but not negligible contribution to total background in both 0-lepton and 2-lepton channels.

4.4 Trigger selection

The list of triggers used to search for our signal is listed in Table 4.2. MET/MHT paths were included in the 0-lepton channel, while standard double e and μ paths were included in the 2-lepton channel.[46],[52]

Channel	Primary dataset	Trigger path (2017)
0-lepton	JetMET	HLT_PFMET120_PFMHT120_IDTight HLT_PFMET120_PFMHT120_IDTight_PFHT60
2-lepton	DoubleMuon	Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8 Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8
	DoubleEG	Ele23_Ele12_CaloIdL_TrackIdL_IsoVL DoubleEle33_CaloIdL_MW

Table 4.2: Primary datasets and trigger paths for year 2017

4.4.1 MET trigger efficiency

MET trigger efficiency estimated for the Zh signal sample in the 0-lepton channel.

Efficiency in our case is defined as the ratio of signal events containing a neutrino pair at generator level and passing the first or second trigger divided by the total signal events containing a neutrino pair at generator level.

A common issue when calculating efficiency is the correct assignment of statistical uncertainty to this efficiency. Let's consider that k events out of a total of N events pass the cuts. To determine the uncertainties in each bin of a histogram, one merely applies the same rule independently to each bin i . Two common methods to propagate uncertainties in each bin are Poisson errors and binomial errors [53] . In the first method, by using the standard error propagation formula we derive to :

$$\delta\epsilon' = \epsilon' \sqrt{\left(\frac{\delta k}{k}\right)^2 + \left(\frac{\delta N}{N}\right)^2} = \frac{k}{N} \sqrt{\frac{1}{k} + \frac{1}{N}} = \sqrt{\frac{k^2(N+k)}{N^3}}. \quad (4.5)$$

But in some cases, this formula is problematic; for example, if $k=0$, then $\delta\epsilon = 0$, which is obviously nonphysical. In addition, if $k = N$, then $\delta\epsilon' = 1 \pm \frac{\sqrt{2}}{\sqrt{N}}$, the error range extends past 1, even though

by definition it is impossible that $\epsilon > 1$. In the second method, “binomial” error calculation is based on the knowledge that the application of a cut can be considered a binomial process, with probability of “success” (i.e. true efficiency) ϵ . Since the true efficiency is not known a priori, an estimate: $\epsilon' = \frac{k}{N}$ into this equation of standard deviation of the binomial distribution, $\sigma_k = \sqrt{\epsilon(1-\epsilon)N}$, and then divided through by N , yields the result: $\delta\epsilon' = \left(\frac{1}{N}\right) \sqrt{k\left(1 - \frac{k}{N}\right)}$. Again, for $k=0$ or $k=N$, then the error results to be zero [53].

It is crucial to derive a formula that leads to reasonable results even in limiting cases. Let's consider the probability that k events out of N will pass the cut, given the condition that the true efficiency is ϵ , which follows a binomial distribution.

$$P(k | \epsilon, N, I) = \frac{N!}{k!(N-k)!} \epsilon^k (1-\epsilon)^{N-k}. \quad (4.6)$$

What needs to be done is to determine efficiency and its statistical error or, in other words, the probability that the true efficiency is between ϵ and $\epsilon + \delta\epsilon$: $P(\epsilon | k, N, I)$. This probability derives from applying Bayes' theorem to take the inverse probability of (4.6):

$$P(\epsilon | k, N, I) = \frac{P(k | \epsilon, N, I)P(\epsilon | N, I)}{Z} \quad (4.7)$$

where Z is a constant to be determined by normalization, and $P(\epsilon | N, I)$ is the probability for the true efficiency to be between ϵ , $\epsilon + \delta\epsilon$ before considering the data. Here is the most important point: since a binomial process is being dealt with, such that ϵ must be in the inclusive range $[0,1]$, no reason would exist to favor one value of the efficiency over another.

Thus,

$$P(\epsilon | N, I) = \begin{cases} 1 & \text{if } 0 \leq \epsilon \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (4.8)$$

So, the use of probability theory allows one to include the knowledge that the efficiency must be between zero and one, which assigns zero probability to those values of ϵ which are obviously impossible. Then the normalization factor Z has to be derived from $\int_{-\infty}^{\infty} P(\epsilon | k, N, I) d\epsilon = 1$ and in the case of a uniform prior distribution, it results to be the beta function of (1,1) mode, $B(x,1,1)$. So, finally, the derived form is:

$$P(\epsilon | k, N, I) = \frac{\Gamma(N+2)}{\Gamma(k+1)\Gamma(N-k+1)} \epsilon^k (1-\epsilon)^{N-k}. \quad (4.9)$$

It can be easily calculated that the above equation has its maximum at $\epsilon = \frac{k}{N}$, so that's the most probable value of the efficiency given k and N . The measure of the uncertainty in the efficiency measurement can be extracted from the distribution $P(\epsilon | k, N, I)$: a common way is 68.3% confidence interval. [53]

The Efficiency plot

In figure 4.13 efficiency is plotted as a function of E_T^{miss} . Then, graph points are fitted with a function which is the cumulative distribution function of Exponentially Modified Gaussian (EMG) distribution:

$$f(x; \mu, \sigma, \lambda) = \frac{\lambda}{2} \cdot e^{\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2x)} \cdot \operatorname{erfc}\left(\frac{\mu + \lambda\sigma^2 - x}{\sqrt{2}\sigma}\right) \quad (4.10)$$

where erfc is the complementary error function defined as:

$$\text{erfc}(x) = 1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \quad (4.11)$$

In this functional form, the parameter determines the point of 50% of the plateau efficiency and the σ gives the resolution. We use this parametrization in order to benchmark the efficiency at the plateau, the turn-on points (bias) and resolution for trigger[54]

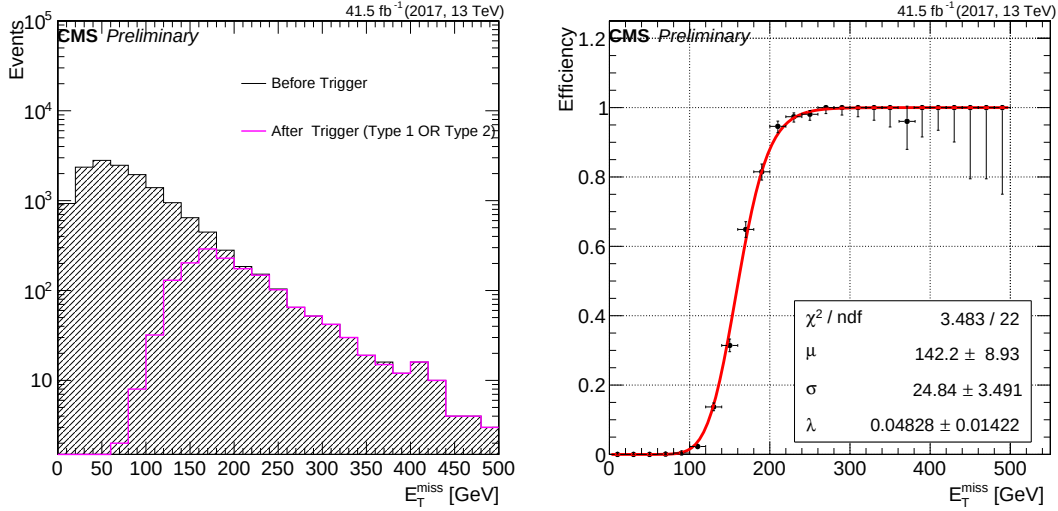


Figure 4.13: Signal events (for $M_a = 20$ GeV) as a function of MET before and after Type 1 OR Type 2 MET trigger requirement (left) MET trigger efficiency(right).

As we can see, in figure 4.13 (right) the MET cut at 170 GeV corresponds to trigger efficiency about 70%, so it is below the efficiency plateau, which starts at 220 GeV. An offline threshold at 170 GeV is set for the time being, until we include the full Run 2 statistics.

4.5 Physics object configuration

In the following, we describe the strategy adopted to pre-select the events and the physics objects of our analysis.

Sample cleanup filters are applied to minimize the contamination from beam-gas interactions, beam-halo and calorimeter noise. The usage of these filters is advisable to minimize mis-reconstruction effects [46].

Primary Vertices (PV) are vertices which are required to be reconstructed inside a cylindrical region of length 24 cm and radius 2 cm surrounding the nominal interaction point. They are in addition required to have a minimum of 4 degrees of freedom regarding the vertex fit. When multiple PV's are present in an event, the vertex with the highest $\sum_{\text{tracks}} p_T^2$ is chosen.

Leptons (electrons and muons) must be reconstructed in the tracker acceptance (i.e. $|\eta| < 2.5$ for electrons and $|\eta| < 2.4$ for muons). Offline p_T cuts are summarized below in Table 4.6. Quality and isolation requirements are furthermore imposed on the leptons. Both ECAL and tracker driven

electrons, not reconstructed in the barrel to endcap transition region (i.e. $1.4442 < |\eta| < 1.5660$) are considered in our analysis.

Electrons are required to fulfill the so-called tight working point (tightest-id) [46]. It consists of a series of requirements on the supercluster shape and geometric distance between the electron track and the supercluster. The ratio of energy deposits in HCAL and ECAL is also used in the definition of the tight id. The tight id imposes tight cuts (also with respect to the trigger) and adds a veto on potential electrons from conversions as well as a requirement on $|1/E - 1/p|$, where ρ is the average energy density due to pileup and underlying event. [37]

Table 4.3 summarizes the identification criteria used for electrons, requiring tight-id. The iso-

Requirement	Barrel	Endcap
$\sigma_{in\eta}$	0.0104	0.0353
$\Delta\eta_{in}$	0.00255	0.00501
$\Delta\phi_{in}$	0.022	0.0236
H/E	$0.026 + 1.15/E + 0.0324 \cdot \rho/E$	$0.0188 + 2.06/E + 0.183 \cdot \rho/E$
$ 1/E - 1/p $	0.159	0.0197
Missing hits	≤ 1	≤ 1
Pass conversion veto	yes	yes
PF rel iso	$0.0287 + 0.506/p_T$	$0.0445 + 0.963/p_T$

Table 4.3: Requirements for Tight-id Electron Selection for year 2017

lation of the electron candidates is computed from the flux of particle flow candidates found within a cone of $R = 0.3$ built around the lepton direction [46]. The flux of particles is computed independently for the charged hadrons, neutral hadrons, and photon candidates. When dealing with electron candidates, the neutral flux is corrected by using the average energy density due to pileup and underlying event in the central region of the detector (ρ) and an effective area (A_{eff}) correction which normalizes this estimator in such a way that the isolation is independent of the number of pileup interactions. The electron isolation is therefore defined as:

$$I_{rel} = \frac{1}{p_T} [I_{ch} + \max(I_{nh} + I_g - A_{eff} \cdot \rho, 0)] \quad (4.12)$$

Muons are required to be either tracker or standalone muons. Requirements are furthermore imposed on the quality of the track fit, number of hits in the pixel, tracker, and number of matched muon segments for the muons to be considered for the lepton candidates. The requirements are summarized in Table 4.4. For the muons, the “delta-beta” correction is applied to correct for pileup contamination. This correction is achieved by subtracting half the sum of the p_T of the charged particles in the cone of interest but with particles not originating from the primary vertex. The muon isolation is therefore defined as:

$$I_{rel} = \frac{1}{p_T} [I_{ch} + \max(I_{nh} + I_g - 0.5 \cdot I_{chPU}, 0)] \quad (4.13)$$

In current analysis, the muon candidates must have PF iso tight to be considered as isolated as defined in Table 4.4.

Jets are reconstructed from PF candidates using the anti- k_T algorithm [55] with a specified distance parameter R . This algorithm follows an iterative clustering approach, where each step

Requirement	Tight
Global	true
PF	true
isTracker	
track	
muonBestTrack (ptError/pt)	
Muon arbitration	
χ^2/dof (global)	<10
χ^2/dof (inner)	
Valid pixel hits	>0
Pixel layers with measurement	
Tracker layers with measurement	>5
Valid muon hits	>0
Matched muon stations	>1
$d_0(vtx)$ (cm)	0.2
$d_z(vtx)$ (cm)	0.5
PF rel iso	<0.15
track rel iso	<0.05

Table 4.4: Requirements for muon Tight-ID Selection

either combines two input constituents into one or merges a constituent with the beam, effectively identifying it as a jet. At each iteration, the anti- k_T distance is evaluated for every pair of particles (i, j) as:

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

and for a particle and the beam as:

$$d_{iB} = p_{T,i}^{-2}$$

After computing these distances, the merging step is executed. The clustering continues until no further particles remain. The application of the anti- k_T algorithm with $R = 0.4$ produces a collection of small- R jets, referred to as AK4 jets or simply jets. Larger jets are reconstructed using a clustering parameter of $R = 0.8$, resulting in the AK8 jet collection aka fat jets.

The particles originating from pileup are mitigated through the charged hadron subtraction (CHS) algorithm [56] for AK4 jets and the pileup per-particle identification (PUPPI) [57] algorithm for AK8 jets. The CHS algorithm first identifies charged PF candidates originating from pileup and removes them before the jet clustering. After AK4 jets are clustered, an event-by-event correction based on the jet area is applied to the jet four-momentum to remove the remaining energy from pileup, mostly from the neutral particles. In the PUPPI algorithm, charged particles identified to be originating from pileup are discarded, and a weight between zero and one is assigned to each neutral particle as the probability of the particle originating from the primary vertex. The four-momenta is rescaled per particle to form the new PF candidate collection, which is used as input to cluster AK8 jets.

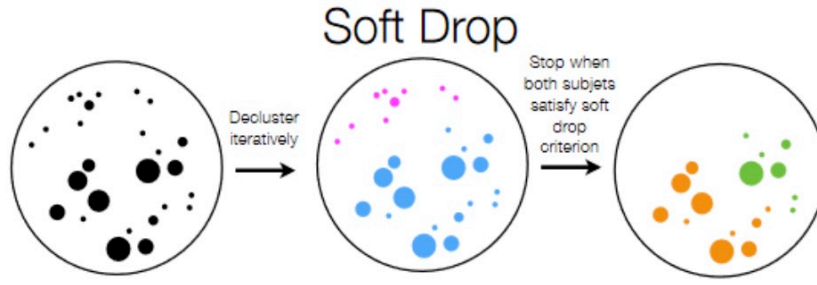


Figure 4.14: Soft drop algorithm procedure.

The Soft Drop algorithm [58] is employed to refine large- R jets by eliminating soft radiation contributions in a systematic manner. The procedure begins by declustering the jet into its constituents, which are then reclustered using the Cambridge/Aachen clustering algorithm. This method resembles the anti- k_T algorithm, with the key difference being that the transverse momentum exponent in the clustering equations is modified from -2 to 0 .

This process generates a clustering tree that encodes the merging history of jet constituents. The Soft Drop procedure iterates through the reversed merging history, systematically splitting each branch into two sub-branches. At each step, the following criterion is evaluated:

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} \Delta R^\beta > z_{\text{cut}}(R_0), \quad (4.14)$$

where the parameters are set to $\beta = 0$ and $z_{\text{cut}} = 0.1$. The jet radius R takes values of 0.8 for AK8 jets.

If the condition is not met, it implies that one branch contains significantly softer radiation and is discarded. The algorithm then proceeds with the remaining branch. This iterative process continues until the criterion is satisfied. The final two surviving branches are identified as the sub-jets of the jet, and the Soft Drop mass is defined as their invariant mass. So, in order to have fat-jets with well reconstructed soft drop mass we demand each fat jet to contain at least two sub-jets [59].

We consider for the analysis all AK4 jets with $p_T > 20$ GeV and $|\eta| < 2.5$ passing the PF-tight requirements. In addition, we perform jet-lepton cross-cleaning by rejecting a jet if it is found close to a tight ID and isolated electron or muon, within a cone of $\Delta R = 0.4$.

We also consider fat jets with $p_T > 150$ GeV and $|\eta| < 2.5$, containing exactly two sub-jets. In addition, we perform fat jet-lepton cross-cleaning by rejecting a fat jet if it is found close to a tight ID and isolated electron or muon, within a cone of $\Delta R = 0.8$. We also cross-clean fat jets from jets by rejecting a jet if found close to a fat jet, within a cone of $\Delta R = 0.8$. The following distribution shows the matching efficiency of fat jets to a generator level (bb) pair coming from

the same "a" boson, as a function of fat jet p_T . It is defined as $\text{eff} = \frac{N_{\text{fjet}}^{\Delta R(\text{fjet}, \text{bb}_{\text{gen}}) < 0.8}}{N_{\text{fjet}}^{\text{total}}}$. For low p_T values, higher efficiency is observed for low "a" masses, while for $p_T > 200$ GeV the matching efficiency seems to become equal for all "a" masses. It must be noted that for high p_T values a (bb) pair becomes collimated enough to be reconstructed by a single AK4 jet, but since double b-tag discriminants do not exist yet for AK4 jets, this method is not applicable for the time being.

b-tagged Jets

The DeepJet, a deep neural network algorithm [60], is for b-tagged AK4 jets identification. The DeepJet algorithm leverages a broad set of both high-level and low-level input variables, including

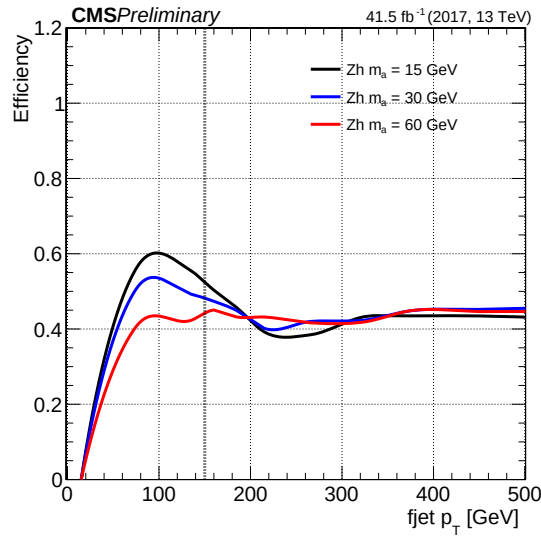


Figure 4.15: Matching efficiency of fat jets to a generator level (bb) pair coming from the same "a" boson, as a function of p_T .

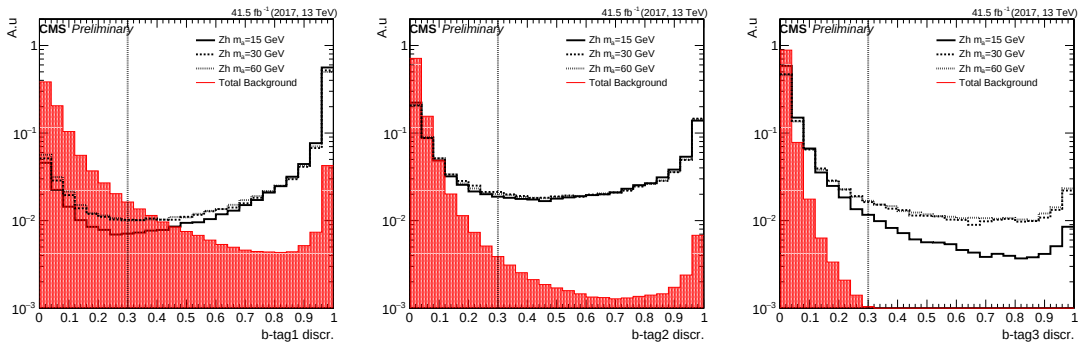


Figure 4.16: DeepJet b-tag discriminants of first three jets (ranged by their b-tag score) .

jet and track kinematics, jet track multiplicity, and primary vertex information, leading to a better performance than previously developed b-tagging algorithms (e.g., DeepCSV).

The Medium WP of DeepJet discriminant is chosen as baseline ($b\text{-tag} > 0.304$), which corresponds to 79% b-tagging efficiency .

Boosted-jet flavour tagging: The boosted-jet flavour tagging algorithms aim to identify a type of large-radius jets initiated from a highly Lorentz-boosted heavy resonance decaying to heavy-flavour quarks such as (bb) or (cc). They are crucial to several physics measurements, including searching for the SM Higgs bosons in their HF decay modes or a BSM resonance decaying to such final states, as in our case.[61]

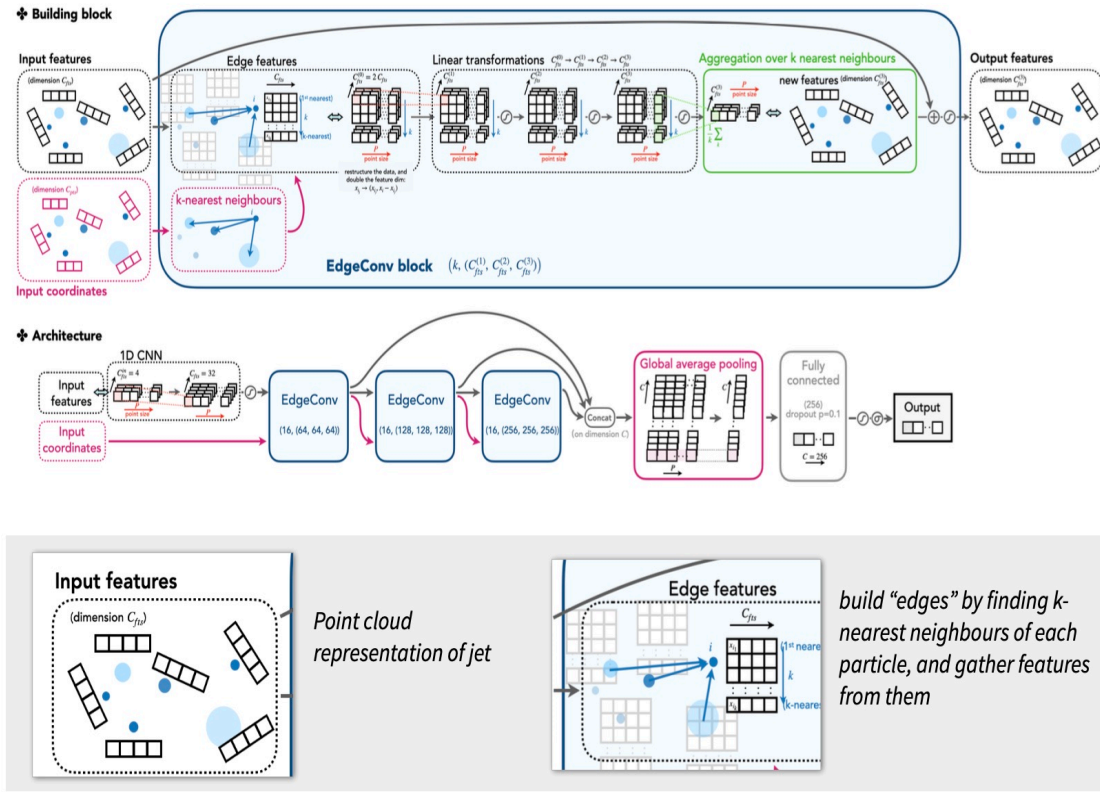


Figure 4.17: ParticleNet GNN architecture.

In general, ParticleNet is graph neural network-based particle identification algorithm where a jet is represented as a particle cloud, analogous to the point cloud representation of 3D shapes used in computer vision.[61] Its purpose is to identify hadronic decays of highly Lorentz-boosted top quarks and W, Z, and Higgs bosons and classify different decay modes. The “ParticleNet” NN architecture [62] is used to process the input particle-flow candidates and secondary vertices in a permutation-invariant way.

ParticleNet-MD tagger is one of the latest and most effective algorithms used in CMS. It is of more general purpose, since it is a mass-decorrelated particle identification algorithm. It is designed for identifying two-prong hadronic decays of highly Lorentz-boosted particles. The tagger is trained on a set of signal jets including $X \rightarrow b\bar{b}$, $X \rightarrow c\bar{c}$, $X \rightarrow q\bar{q}$, and background QCD jets, where X is a variable-mass spin-0 particle. Jets from both signal and background samples are reweighted to yield flat distributions in both p_T and soft-drop mass (m_{SD}), so as to decorrelate the trained tagger variable with the jet soft-drop mass.

The ParticleNet-MD algorithm outputs four probability-like scores: $p(X \rightarrow b\bar{b})$, $p(X \rightarrow c\bar{c})$, $p(X \rightarrow q\bar{q})$, and $p(\text{QCD})$. [61]. These scores are used to construct the ParticleNet-MD discriminant variables, depending on the mode of the resonant decay that needs to be distinguished from the QCD.

If a single resonant decay mode needs to be distinguished, the discriminant variable is defined as the ratio of the probability score of this decay mode, divided by its sum with the probability score of the QCD.

It is also possible to define more inclusive discriminants, that distinguish resonant decays of more than one possible decay modes at a time, e.g to distinguish a resonant decay that can be either $X \rightarrow b\bar{b}$, or $X \rightarrow c\bar{c}$ from the QCD background. To do so, one has to derive the sum of probability scores of the possible decay modes and divide it by its sum with the QCD probability score.

In order to form a notation of these discriminants, we annotate as X_i the probability score of decay mode i where i is bb , cc or qq . The probability score of the QCD is simply notated as QCD. Thus, a discriminant for a single resonant decay mode will be formed as $X_i \text{VsQCD}$, where i is bb , cc or qq .

From the above, it comes out that a double b-tag discriminant is defined as:

$$X_{bb} \text{VsQCD} = \frac{p(X \rightarrow b\bar{b})}{p(X \rightarrow b\bar{b}) + p(\text{QCD})} \quad (4.15)$$

A more general double resonant decay discriminant can be defined as a discriminant between any type of double resonant decay modes (leading to hard or heavy flavour decay products) and QCD processes (gluon splitting or gluon radiation).

As mentioned before, to define a discriminant for more than one possible decay mode, the sum of probability scores of all these modes needs to be included in both the numerator and the denominator. For this reason, to derive a notation for them, we use the "+" symbol between the possible decay modes. So, this discriminant is defined as :

$$X_{bb+X_{cc}+X_{qq}} \text{VsQCD} = \frac{p(X \rightarrow b\bar{b}) + p(X \rightarrow c\bar{c}) + p(X \rightarrow q\bar{q})}{[p(X \rightarrow b\bar{b}) + p(X \rightarrow c\bar{c}) + p(X \rightarrow q\bar{q})] + p(\text{QCD})} \quad (4.16)$$

The Figures 4.18 and 4.19 show the distributions of these two discriminants for the leading fat jet and the sub-leading fat jet. Since better signal efficiency was observed for the mixed discriminator, in the configuration of fat jets we applied a cut at the value 0.5 of their mixed discriminant. This was done in order to reject background events, without rejecting much signal at the same time. It was investigated that this cut value corresponds to the best signal efficiency and background rejection at the same time, yielding a signal efficiency about 67% (for $m_a = 20$ GeV) and background rejection 80%, for the leading fat jet. So, the main purpose of this discriminant is to apply a first cut to reduce background without affecting signal yield that much.

Missing Transverse Energy (E_T^{miss}) is determined based on the imbalance of transverse momentum from all reconstructed particle-flow candidates and is given by:

$$\vec{E}_T^{\text{miss}} = - \sum_{i=1}^{N_{\text{PF}}} \vec{p}_{T,i}$$

Jet energy scale corrections can influence the estimation of E_T^{miss} by introducing an additional term that accounts for the net momentum balance of clustered energy after applying corrections to individual jets. This refined estimator is commonly referred to as the Type-I E_T^{miss} correction and is expressed as:

$$\vec{E}_T^{\text{miss, Type-I}} = \vec{E}_T^{\text{miss}} - \sum_{j=1}^{N_{\text{jets}}} \delta \vec{p}_{T,j}$$

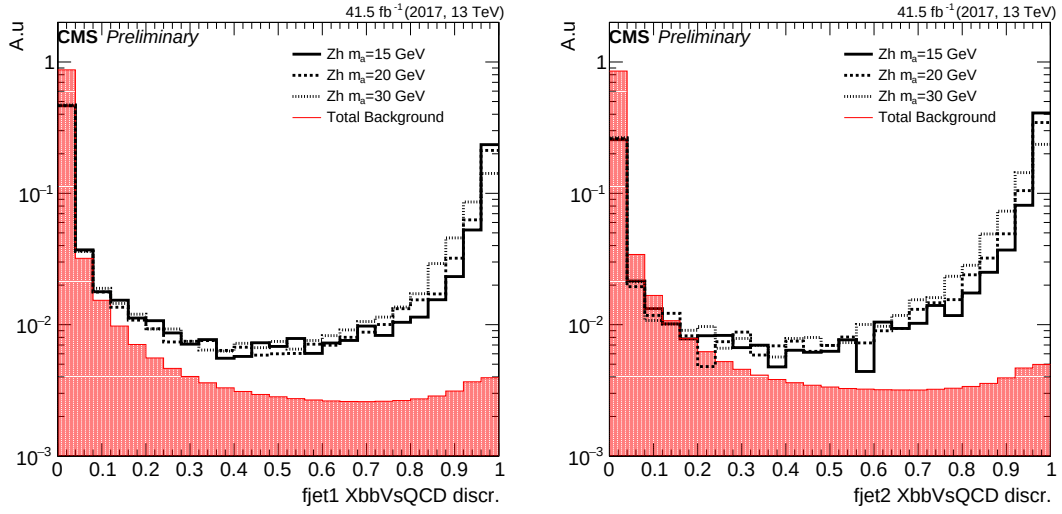


Figure 4.18: ParticleNet-MD double-b tag discriminant of leading fat jet(left) and sub-leading fat jet (right) for signal and total background superimposed.

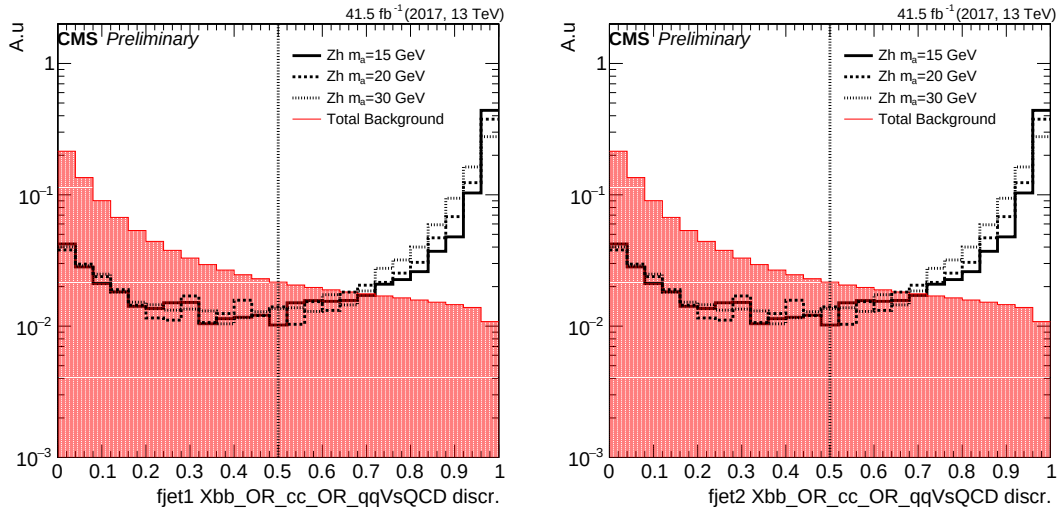


Figure 4.19: ParticleNet-MD double resonant decay discriminant of leading fat jet(left) and sub-leading fat jet(right) for the signal and the total background superimposed.

The Type-I correction is computed using the CHS algorithm [56], following the guidelines established by the JetMET group.

As far as MET is concerned, noteworthy is the contribution of fake MET, attributed to incorrect measurements during detection. In as already mentioned, in 2-lepton channel, we are working with reconstructed electrons and muons, and not τ leptons. The reason is the instability of the τ particle and therefore its immediate decay, which always includes neutrinos in the final state.

The following table summarizes all pre-selection criteria applied to physics objects

Leptons	Electrons	$p_T > 15 \text{ GeV}$, $ \eta < 2.4$, tight-Id and tight-iso
	Muons	$p_T > 10 \text{ GeV}$, $ \eta < 2.5$, tight-Id and tight-iso
Jets	AK4	$p_T > 20 \text{ GeV}$, $ \eta < 2.5$, PF-tight
	AK8	$p_T > 150 \text{ GeV}$, $ \eta < 2.5$, $N_{\text{subjets}} \geq 2$ Xbb+Xcc+XqqVsQCD disc.>0.5

Table 4.5: Object configuration

Electrons	Muons
leading $p_T > 25 \text{ GeV}$	leading $p_T > 20 \text{ GeV}$
sub-leading $p_T > 15 \text{ GeV}$	sub-leading $p_T > 10 \text{ GeV}$

Table 4.6: lepton p_T cuts applied to the leptons in the 2-lepton channel.

4.6 Event Selection Criteria

The selection of the events to be used for our search consists of the following requirements:

1. at least two opposite-charge,same flavor leptons in the 2-lepton channel (their p_T cuts are listed in Table 4.6) or exactly zero leptons in the 0-lepton channel.
2. A cut on the invariant mass of the two leptons to be within 20 GeV window around the Z-boson mass, in the 2 lepton channel or a cut on $E_T^{\text{miss}} > 170 \text{ GeV}$ in the 0-lepton channel.
3. At least one fat jet and at least one b-tagged jet in the merged regime and at least 3 b-tagged jets in the resolved regime. These requirements for the hadronic system are the same in the 0-lepton channel and in the 2-lepton channel.

		Selection Cuts	
		0-lepton	2-lepton
Leptonic system		Exactly zero leptons	$N_l \geq 2$, OSSF (e^+e^- OR $\mu^+\mu^-$)
Dileptonic invariant mass window		-	$80 < m_{\ell\ell} < 100 \text{ GeV}$
MET cut		$E_T^{\text{miss}} > 170 \text{ GeV}$	-
Hadronic system	Merged	$N_{\text{fj}} \geq 1$ and $N_b \geq 1$	
	Resolved	$N_b \geq 3$	

Table 4.7: Event selection cuts

The selection of MET cut is very crucial in the 0-lepton channel, in order to get rid of fake MET in the final selected events. Besides MET trigger efficiency plots, shown in the previous paragraph, we performed a further investigation. Figure 4.20 shows MET of signal events ($m_a = 20$) with Z decaying at generator level to neutrinos or (bb) or (qq) of lighter flavor, superimposed. It is obvious that for MET values greater than 170 GeV, fake MET coming from misidentified quarks has almost diminished.

Furthermore, it's shown in Figure 4.21, we plotted MET as a function of fat-jet p_T in the 0-lepton channel. For events with fat-jet $p_T = 150$ GeV, the corresponding MET is about 170 GeV, and so an offline cut was applied to this value.

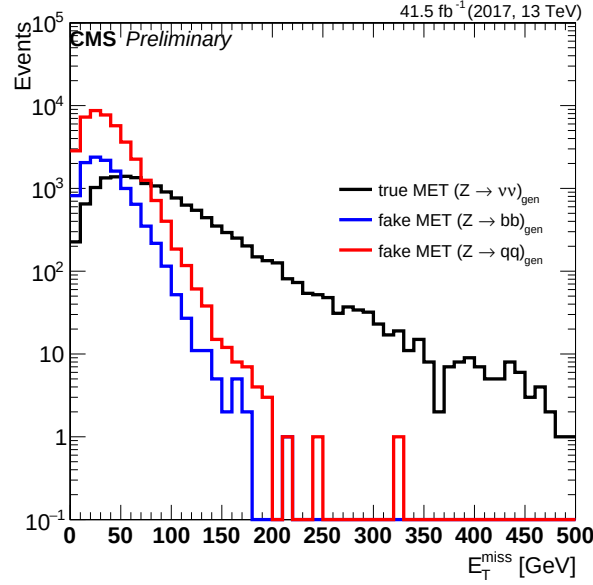


Figure 4.20: MET coming from signal events with $Z \rightarrow \nu\nu$ or $Z \rightarrow b\bar{b}$ or $Z \rightarrow q\bar{q}$ of lighter flavor at generator level.

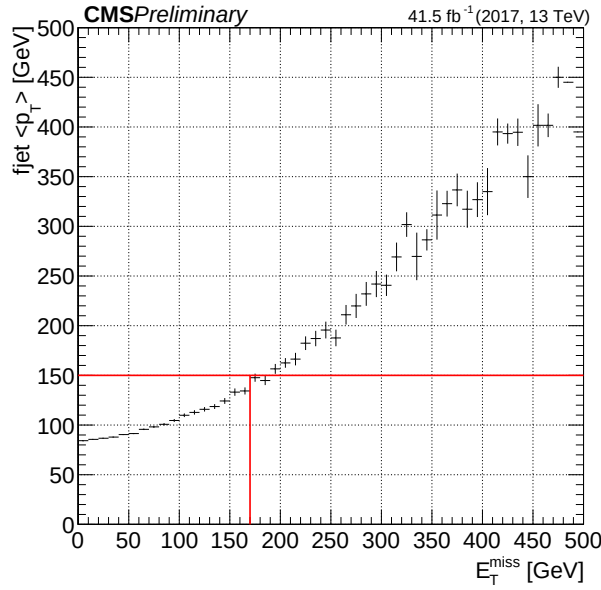


Figure 4.21: fat jet p_T as a function of MET in 0-lepton channel before pre-selection criteria

4.6.1 Event flow Tables

In tables 4.8 and 4.9, the event flow table is presented for both the signal and the background in the 2-lepton and in the 0-lepton channel, respectively. For the signal, we have the indicative cases of the mass of "a": $m_a = 15, 30$, and 60 GeV, while all the backgrounds are considered. Each column of the table corresponds to a selection criterion, and the number of events that survive after its application is displayed for all the studied processes. For each step, the corresponding efficiency has been calculated, which is equal to the percentage of the initial events that survive after the application of each criterion. This quantity is given by the relation:

$$eff = \frac{N_{step-i}}{N_{initial}}, \quad (4.17)$$

where N_{step-i} is the number of events that successfully pass step i of the selection process, and $N_{initial}$ is the initial number of events. In analyses like the present one, it is important to make a theoretical estimation of the number of events expected to survive the selection criteria. This quantity is denoted as $N_{expected, final}$ and is calculated using the following equation:

$$N_{expected, final} = N_{expected} \cdot eff = N_{expected} \cdot \frac{N_{final\ step}}{N_{initial}} \quad (4.18)$$

In this equation, $N_{expected}$ represents the initial number of expected events before applying the selection criteria. This number is given for each process by the equation:

$$N_{expected} = \sigma_{process} \cdot L_{int} \quad (4.19)$$

where $\sigma_{process}$ is the cross-section of the process, and $L_{int} = 41.5 \times 10^3 \text{ pb}^{-1}$ is the integrated luminosity.

Thus, Equation can be rewritten as:

$$N_{\text{expected, final}} = \sigma_{\text{process}} \cdot L_{\text{int}} \cdot \frac{N_{\text{final step}}}{N_{\text{initial}}} \quad (4.20)$$

At this point, it is important to mention that the factor

$$\frac{\sigma_{\text{process}} \cdot L_{\text{int}}}{N_{\text{initial}}} \quad (4.21)$$

corresponds to the weight of the process, which quantifies its contribution to the analysis.

So, having just the initial statistical events of each process, as well as its cross-section and efficiency, the expected event yield of that process can be calculated straightforwardly.

As it was previously referred, $\text{Wh} \rightarrow \text{a}(\text{bb})\text{a}(\text{bb})$ process acts as signal contamination to the Zh process, in the 0-lepton channel. In order to quantify this contamination, we calculate the expected event yield of Zh and Wh processes also separately. In the Table 4.10 the final event yield of Zh signal as well as its contribution to total signal yield is shown for various "a" masses. More details are provided in the appendix.

Process	Raw	$N_l \geq 2$	Z Mass Window	Merged: $N_{fjet} \geq 1$ and $N_b \geq 1$	Resolved: $N_b \geq 3$
$t\bar{t} + LF$	2.31×10^7	387786 $eff = 1.6\%$	52361 $eff = 0.22\%$	1917 $eff = 0.008\%$	824 $eff = 0.004\%$
$t\bar{t} + c\bar{c}$	9.822×10^6	15317 $eff = 0.15\%$	2033 $eff = 0.02\%$	165 $eff = 0.0016\%$	177 $eff = 0.0018\%$
$t\bar{t} + b\bar{b}$	1.557×10^6	15600 $eff = 1.5\%$	2111 $eff = 1\%$	123 $eff = 0.009\%$	271 $eff = 0.017\%$
$Z \rightarrow \ell\ell$	1.717×10^8	4.333×10^7 $eff = 25.3\%$	$.821 \times 10^7$ $eff = 22.2\%$	5179 $eff = 0.003\%$	1908 $eff = 0.001\%$
Other Bkgs	1.335×10^7	168564 $eff = 1.26\%$	80594 $eff = 0.6\%$	429 $eff = 0.003\%$	164 $eff = 0.0012\%$
Signal $m_a = 15$ GeV	36677	1036 $eff = 2.8\%$	997 $eff = 2.7\%$	35.2 $eff = 0.1\%$	19.5 $eff = 0.05\%$
Signal $m_a = 30$ GeV	36687	957 $eff = 2.6\%$	923 $eff = 2.5\%$	30.8 $eff = 0.08\%$	104.6 $eff = 0.29\%$
signal $m_a = 60$ GeV	36685	1008 $eff = 2.7\%$	947 $eff = 2.6\%$	24.5 $eff = 0.07\%$	136 $eff = 0.37\%$

Table 4.8: Event yields expected for background and signal after the selection cuts in the 2-lepton channel

Process	Raw	Exactly 0 leptons	$MET > 170$ GeV	Merged: $N_{fjet} \geq 1$ and $N_b \geq 1$	Resolved: $N_b \geq 3$
$t\bar{t} + LF$	2.3107×10^7	1.643×10^7 $eff = 71.1\%$	407213 $eff = 1.8\%$	114298 $eff = 0.49\%$	25792 $eff = 0.11\%$
$t\bar{t} + c\bar{c}$	9.822×10^6	8.389×10^6 $eff = 85.4\%$	119469 $eff = 1.2\%$	41136 $eff = 0.42\%$	14652 $eff = 0.15\%$
$t\bar{t} + b\bar{b}$	1.558×10^6	1.205×10^6 $eff = 71.1\%$	27580 $eff = 1.8\%$	9845 $eff = 0.63\%$	6321 $eff = 0.4\%$
QCD	1.1086×10^{12}	1.1084×10^{12} $eff = 99.9\%$	2.0524×10^7 $eff = 0.0018\%$	256233 $eff = 2.3 \times 10^{-5}\%$	46835 $eff = 4.22 \times 10^{-6}\%$
$Z \rightarrow \nu\nu$	1.896×10^7	1.896×10^7 $eff = 99.9\%$	1.601×10^6 $eff = 8.44\%$	18466 $eff = 0.09\%$	1308 $eff = 0.007\%$
$W \rightarrow \ell\nu$	8.6673×10^7	5.614×10^7 $eff = 64.7\%$	1.808×10^6 $eff = 2\%$	34640 $eff = 0.39\%$	1657 $eff = 0.02\%$
Other Bkgs	1.335×10^7	1.075×10^7 $eff = 80\%$	114692 $eff = 0.9\%$	17985 $eff = 0.13\%$	3387 $eff = 0.025\%$
Signal $m_a = 15$ GeV	93525	84137 $eff = 89.9\%$	1052 $eff = 1.12\%$	209.6 $eff = 0.22\%$	38.4 $eff = 0.04\%$
Signal $m_a = 30$ GeV	93535	83968 $eff = 89.8\%$	1089 $eff = 1.16\%$	229.1 $eff = 0.24\%$	192.2 $eff = 0.2\%$
Signal $m_a = 60$ GeV	93529	84102 $eff = 89.9\%$	1002 $eff = 1.07\%$	160 $eff = 0.17\%$	221 $eff = 0.23\%$

Table 4.9: Event yields expected for background and signal after the analysis selection cuts in the 0-lepton channel

ZH contribution to total signal			
Regime	$m_a = 15$	$m_a = 30$	$m_a = 60$
Merged	116 (56%)	124 (54.2%)	99 (62%)
Resolved	27 (56%)	124 (64%)	147 (65.4%)

Table 4.10: Zh signal yield and its contribution (%) to total signal yield for various "a" masses, in the two regimes of the 0-lepton channel.

4.7 Multivariable Analysis

Multivariate Analysis (MVA) refers to a set of statistical methods used to analyze data involving multiple variables simultaneously. In the field of experimental particle physics, MVA combines information from multiple variables to achieve the best possible separation between signal and background events.

4.7.1 Toolkit for Multivariate Analysis

In this study, the Toolkit for Multivariate Analysis (TMVA) was used for the multivariate analysis. TMVA is a software package integrated within the ROOT framework and is a powerful tool for high-energy physics data analysis. TMVA provides various machine learning techniques that can be applied to multivariate analyses. TMVA provides a robust environment for processing and evaluating various multivariate classification and regression techniques. It supports a wide range of algorithms, including neural networks, support vector machines, and boosted decision trees, among others. Users can input data in the form of ROOT trees or text files, allowing for flexible handling of datasets where each event can have individual weights. TMVA facilitates the training, testing, and performance evaluation of models while also offering visualization tools to help users interpret results effectively. The supported methods respond to supervised learning only. This means that

the multivariate discriminator uses data samples, and their input information is mapped in feature space to the desired outputs. After training, multivariate discriminator is tested and evaluated on new samples, and it acts in accordance with its training. In our analysis, TMVA will be used for classification: to separate ZH signal from overwhelming SM background procedures. Boosted Decision Trees (BDTs) method will be applied for this binary classification.

4.7.2 Boosted Decision Trees

Boosted Decision Trees (BDTs) is a machine learning method used for classifying data into two or more categories based on their features. Their fundamental principle relies on the creation of a decision flow diagram, of each decision tree (DT), consisting of nodes. Each node includes a splitting criterion, i.e., a rule based on one of the input variables. As it is depicted in Figure 4.23, a DT flow diagram has a tree-like structure. Each event starts at the root of the tree [63]. At each node, the event is evaluated based on a splitting criterion, which is by default the Gini index, while there exist other options like cross entropy. The Gini index is a metric used to measure the purity of a dataset in the context of classification. It quantifies how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. A lower Gini index indicates a purer node, meaning that most of the samples belong to a single class. In DTs, the Gini index of the variable that has the highest discriminating power in each node (i.e., the ability to effectively separate signal from background) is employed as a criterion for splitting nodes during the construction of decision trees. The same variable may or may not be used on several nodes, while others may be used less or never. If the event satisfies the criterion, it follows one node of the tree. If it does not satisfy the criterion, it follows the other node. After a series of decisions, the event is classified into one of the two final categories. The final categories correspond to the leaves of the tree. In this analysis, the leaves represent the signal (S) state or the background (B) state, so each leaf node is labeled with the letter "S" for signal or "B" for background, that depends on the majority of events that end up in each node. The term "boosted" refers to the process of boosting, a machine learning technique designed to improve the performance of Decision Trees (DTs). Essentially, DTs are trained sequentially, and the main idea is to assign greater weight to the errors made by the previous trees in the sequence. Each DT in the sequence is a "weak learner," meaning it makes decisions that are almost random. These trees are usually very simple, as they have a limited number of branches. With boosting, one DT is created at a time, where each subsequent tree is trained to correct the errors of the previous ones. For this reason, the model assigns greater weight to the events that were misclassified through a function, so the next tree focuses on them. Once all the DTs are trained, their predictions are combined, taking into account their respective weighting coefficients. Consequently, BDT is an ensemble learning method that combines multiple weak learners to create a strong predictive model.

There exist different boosting techniques, with Adaptive boosting and Gradient boosting the most common for BDTs. Adaptive Boosting (AdaBoost) and Gradient Boosting differ significantly in their approach to weight assignment during the training process of weak learners. In AdaBoost, the algorithm dynamically adjusts the weights of training instances after each iteration, increasing the weights for misclassified samples while decreasing those for correctly classified ones. This mechanism allows subsequent weak learners to focus on the more challenging observations, effectively correcting the mistakes of their predecessors. The final prediction in AdaBoost is made by aggregating the predictions of all weak learners through a weighted majority vote, where stronger learners contribute more to the final outcome based on their accuracy. In contrast, Gradient Boost-

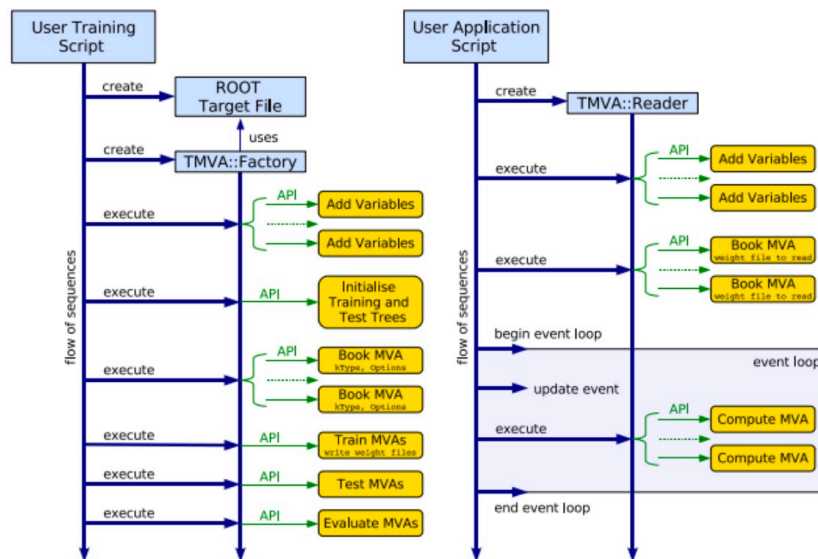


Figure 4.22: **Left: The typical workflow of a TMVA training application progresses from top to bottom.** The user can utilize a ROOT macro or similar options as their user script. The first step involves creating a ROOT TFile, which the TMVA Factory uses to store output histograms and trees. The Factory manages the interaction with the various TMVA modules and is the only TMVA object that is directly created and owned by the user. Initially, the user registers the discriminating variables, which must comply with branches in the training trees. Next, the user selects MVA methods by specifying a type identifier and a unique name. The TMVA analysis then advances by sequentially invoking the training, testing, and performance evaluation methods of the Factory. The results from training all booked methods are saved in custom weight files, while evaluation histograms are stored in the output file. These can be analyzed using specific macros included with TMVA [63] **Right: The flow of a typical TMVA analysis application also moves from top to bottom.** The MVA methods validated during the previous training and evaluation steps are now employed to classify data with unknown signal and background compositions. The process begins with creating a Reader class object, which functions as an interface for accessing the method's responses, similar to how the Factory operated during training and performance evaluation. The discriminating variables and references to locally declared memory placeholders are registered with this Reader. The selected MVA methods are then booked along with their corresponding weight files as arguments, which fully configures them for use. Finally, the user executes an event loop where, for each event, input variable values are copied to designated memory locations, and the MVA response values (and sometimes associated errors) are calculated.

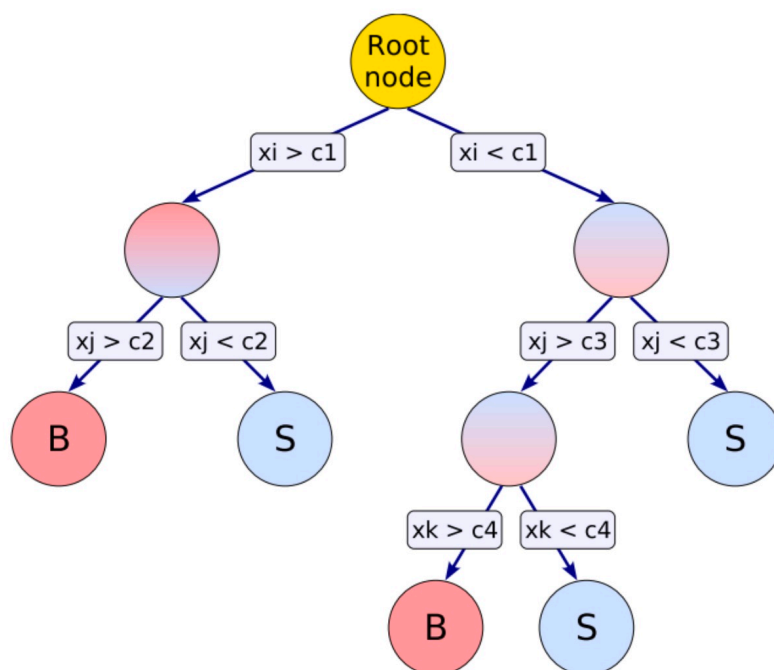


Figure 4.23: Decision flow diagram of a DT. The leafs of the tree correspond to signal(S) or background(B) state.

ing does not alter the weights of individual training instances directly. Instead, it builds new models based on the residual errors from previous models, focusing on minimizing a specified loss function—such as mean squared error or log loss—through gradient descent. Each new learner is trained on the errors made by the combined ensemble of previous learners, thereby correcting these errors iteratively without changing instance weights. This fundamental difference in weight assignment leads to distinct operational characteristics and performance outcomes for each boosting method. While both methods are powerful ensemble techniques, Gradient Boosting’s flexibility with loss functions and advanced optimization features often result in superior performance compared to AdaBoost in various scenarios. [64].

Hyper-parameters also play a crucial role in balancing overfitting and underfitting in Gradient Boosted Decision Trees (BDTs). Properly tuned hyperparameters can significantly impact the model’s ability to generalize to unseen data. Key hyperparameters include the learning rate, number of trees, and maximum depth of trees. The learning rate controls the contribution of each tree to the final prediction; a smaller learning rate generally leads to a more conservative model that requires more trees to achieve good performance, reducing the risk of overfitting. Conversely, a higher learning rate can accelerate training but may lead to overfitting if not managed carefully. The number of trees determines how many models are combined; while more trees can improve performance, they also increase the risk of overfitting if the model becomes too complex. The maximum depth of each tree influences its complexity—deeper trees can capture more intricate patterns in the data but are also more prone to overfitting. Limiting tree depth helps create simpler models that generalize better, thus preventing underfitting. Therefore, careful tuning of these

hyperparameters is essential to strike a balance between underfitting and overfitting, ultimately enhancing the model's predictive performance on new data. The following table summarizes the hyperparameters used in BDT training in each regime of each channel.

Parameter	0-lepton		2-lepton	
	Resolved	Merged	Resolved	Merged
Boost Type	Grad	Grad	Grad	Grad
Separation Type	Gini Index	Gini Index	Gini Index	Gini Index
Shrinkage	0.1	0.1	0.09	0.09
NTrees	200	200	150	100
MaxDepth	3	3	3	3
MinNodeSize	5.0%	5.0%	5.0%	5.0%
nCuts	20	20	20	20
Node Purity Limit	0.5	0.5	0.5	0.5

Table 4.11: Hyper-parameter Values used in the BDT training procedure in the 0-lepton and in the 2-lepton channel

4.7.3 BDT input variables

For the training of the analysis, variables with high discriminative power between the signal and the background were selected. At the same time, variables independent of the mass of the exotic "a" boson were preferred to ensure that the training could be applied to all mass hypotheses. Exceptions are the soft drop mass of the fat jet, as well as $\Delta R(\text{fjet}, \text{jet})$, which corresponds to $\Delta R(a, a)$ at the generator level, but both seemed to have a high discriminative power.

In this way, a single BDT discriminator is constructed for all signal cases in the range [12,40] GeV in the merged regime and [20,60] GeV in the resolved one. Table 4.12 lists the input variables used in the MVA in each channel and regime, along with a brief description of each one.

There follows a more detailed description of these variables:

- M_H : The invariant mass of the higgs candidate formed the three or four b-tagged jets in the resolved regime. In case there exist three b-tagged jets a non b-tagged 4th jet its also taken into account in the calculation. In the merged regime higgs invariant mass is reconstructed by the fat jet and the one or two (cross-cleaned) b-tagged jets. In case there is one b-tagged and a non b-tagged jet the second is taken also into account.
- p_T^H : The transverse momentum of the higgs candidate formed by summing the transverse momentum of the three or four b-tagged in the resolved regime. In case there exist three b-tagged jets a non b-tagged 4th jet, its also taken into account in the calculation. In the merged regime transverse momentum of the higgs candidate is formed by summing the transverse momentum of the fat jet and the one or two (cross-cleaned) b-tagged jets. Again, in case there is one b-tagged and a non b-tagged jet the second is taken also into account.
- H_T : The scalar sum of the transverse momentum activity of all jets in the resolved regime. In the merged regime, is the scalar sum of the transverse momentum of fat jet and all (cross-cleaned) jets.
- E_T^{miss} : The missing transverse energy.

- $|\Delta\phi(E_T^{\text{miss}}, j)|^{\text{min}}$: The minimum azimuthal opening angle between the E_T^{miss} : direction and any jet (or the fat jet in case of the merged regime). This is a useful variable, since in case of fake met from jet, the azimuthal opening angle between them will be small.
- $|\Delta\phi(E_T^{\text{miss}}, \ell)|^{\text{min}}$: the minimum azimuthal opening angle between the E_T^{miss} : direction and any lepton (in 2-lepton channel). This is a useful variable, since in case of fake met from lepton, the azimuthal opening angle between them will be small.
- b-tag Discriminants: In the resolved regime, the b-tag discriminant of the jet with the 3rd highest b-tag score (b-tag3) is used as input variable. The b-tag Discriminant of the jet with the 1st highest b-tag score is used as well. In the merged regime, the b-tag discriminant of the cross-cleaned jet with the 1st highest b-tag score is used.
- p_T^{b1} : Transverse momentum of the leading b-tagged jet
- $|\Delta\phi(\text{ZH})|$: Azimuthal opening angle between the reconstructed Z boson direction and the Higgs candidate direction. It is expected to take high values (near π) for signal, as H and Z decay back-to-back (2-lepton channel).
- N_j : The total number of jets in the resolved regime and the total number of cross-cleaned jets in the merged regime.
- $\Delta m(\text{b}, \text{b})^{\text{min}}$: This variable represents the minimum distance in the invariant mass between two pairs of b-jets in all possible pairings. It is used to find the pair of b-jets that best matches the "a" boson mass, helping to identify the correct b-jet combinations and distinguish signal events from background. (defined in the resolved regime)
- $\Delta R(\text{b}, \text{b})^{\text{ave}}$: Distance in the eta-phi plane between any b-tag pair, averaged over all possible combinations per event. (defined in the resolved regime)
- p_T^Z : Transverse momentum of the Z boson formed by summing the transverse momentum of the two leptons (in 2-lepton channel)
- $\Delta R(\ell, \ell)$: distance in the eta-phi plane between the two leptons (in 2-lepton channel)
- There are also 5 more variables defined only in the merged regime and have to do with fat jet characteristics. These are the p_T of fat jet, its soft drop mass, the opening angle between the fat jet and the leading b-tagged jet and finally the ParticleNet-MD discriminants, discriminating double resonant decay from QCD background.

There follows a presentation of the distributions of some of these variables after the application of all selection cuts in each channel and regime. In the following histograms, distributions of the signal of two possible 'a' masses ($m_a = 15$ GeV and $m_a = 30$ GeV in the merged regime and $m_a = 30$ GeV and $m_a = 60$ GeV in the resolved regime) are superimposed with the stacked distributions of all backgrounds that are present in each channel. The signal distributions are scaled by a factor for visibility purposes.

Figures 4.26, 4.28, 4.30 and 4.32 show the corresponding distributions of the training sample for these variables. The correlation matrices of these variables are also presented in figures 4.27, 4.29, 4.31 and 4.33. It is evident that there is no strong correlation between them, a crucial characteristic for a multivariate analysis. In fact, independent variables contribute in different ways to

the classification of an event as either signal or background. Each variable contains unique information, and their combination enables the model to make more accurate predictions regarding the classification of each event into one of the two categories. Finally, ranking of the input variables in terms of importance to the classification procedure is applied. In tables 4.13 - 4.16, the ranking results are shown.

Starting with the merged regime of the 0-lepton channel, the sum of "a" mass point samples of m_a [12,40] GeV is used as the signal sample. The QCD samples of $m_T < 300$ GeV were not included in the background sample due to large weights and low statistics. In total, the statistical events are 2.3k signal events and 5.7M background events, 50% of which are used for training and 50% for testing.

Variable	Description	0-lepton		2-lepton	
		Merged	Resolved	Merged	Resolved
M_H	Invariant mass of the Higgs candidate	✓	✓	✓	✓
p_T^H	Transverse momentum of the Higgs candidate	✓	✓	✓	✓
H_T	Scalar sum of the transverse momentum activity of all decay products of the Higgs candidate	✓	✓	✓	✓
E_T^{miss}	Missing transverse energy	✓	✓	✓	✓
p_T^{fjet}	Fat-jet transverse momentum	✓		✓	
p_T^{b1}	Leading b-tagged jet transverse momentum	✓	✓	✓	✓
$\Delta R(b, b)^{\text{ave}}$	Distance in the eta-phi plane between any b-tag pair, averaged over all possible combinations per event		✓		✓
$\Delta R(\text{fjet}, j)$	Distance in the eta-phi plane between fat-jet and leading cross-cleaned b-tagged jet	✓		✓	
$\Delta m(b, b)^{\text{min}}$	Minimum distance in $m(b1b2) - m(b3b4)$ among the possible pairing scenarios		✓		✓
m_{SD}	Fat-jet soft drop mass	✓		✓	
p_T^Z	Transverse momentum of the Z boson formed by summing the transverse momentum of the two leptons			✓	✓
XbbVsQCD	ParticleNet-MD double b-tag discriminant	✓		✓	
Xbb+Xcc+XqqVsQCD	ParticleNet-MD double resonant decay discriminant	✓		✓	
b-tag1	b-tag discriminator of jet with the highest DeepJet discriminator value	✓		✓	
b-tag3	b-tag discriminator of jet with the third highest DeepJet discriminator value	✓	✓	✓	✓
$ \Delta\phi(ZH) $	Azimuthal opening angle between the reconstructed Z boson direction and the Higgs candidate direction			✓	✓
	$\Delta R_{\ell\ell}$ Distance in the eta-phi plane between leptons			✓	✓
$ \Delta\phi(E_T^{\text{miss}}, j) ^{\text{min}}$	Minimum azimuthal opening angle between the E_T^{miss} and any jet or fat-jet			✓	✓
$ \Delta\phi(E_T^{\text{miss}}, \ell) ^{\text{min}}$	minimum azimuthal opening angle between the E_T^{miss} and any lepton			✓	✓
N_j	Number of jets	✓	✓	✓	✓

Table 4.12: MVA input variables

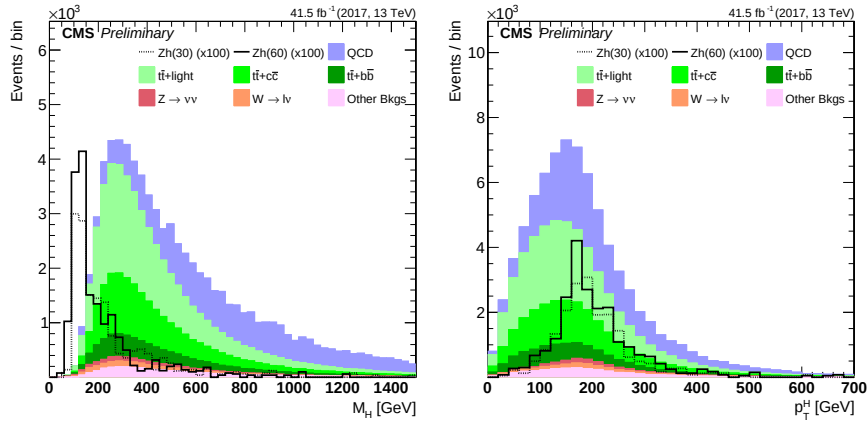
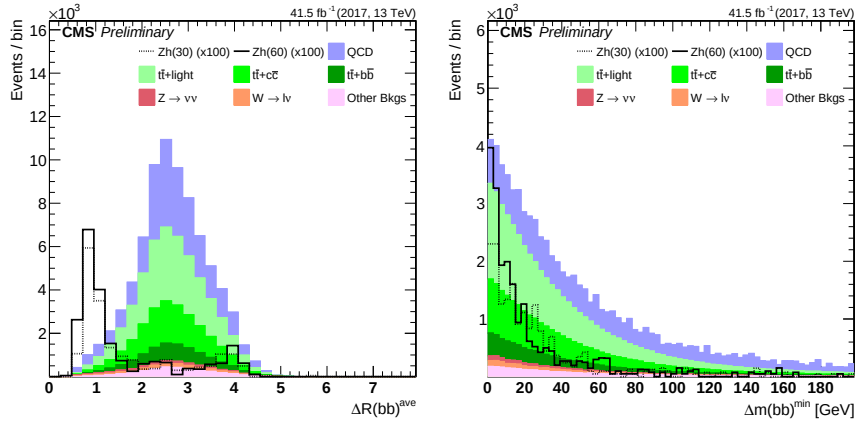
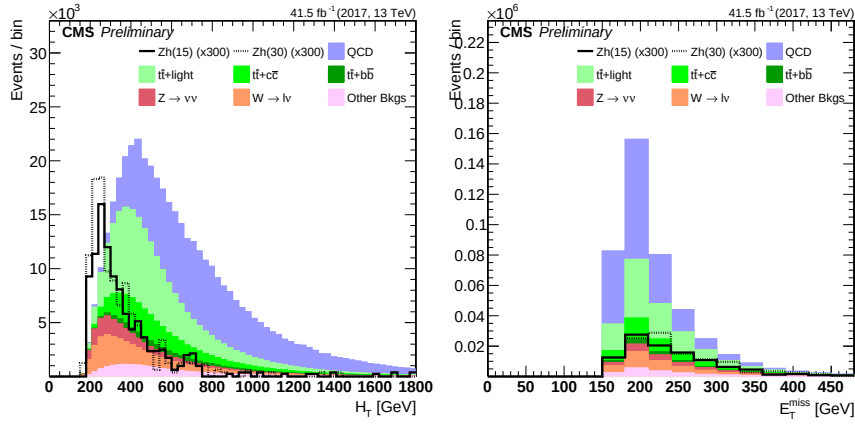
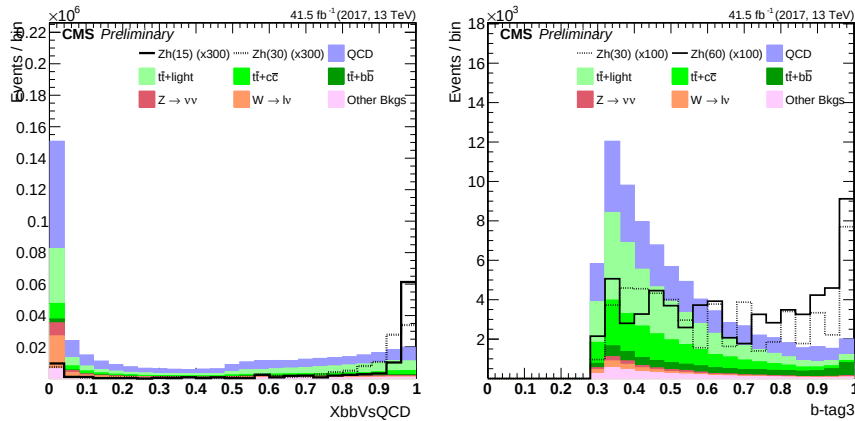
(a) M_H (left) and p_T^H (right) in the resolved regime.(b) $\Delta R(bb)^{ave}$ (left) and $\Delta m(bb)^{min}$ (right) in the resolved regime.(c) H_T (left) and E_T^{miss} (right) in the merged regime.(d) $b\text{-tag3}$ in the resolved regime (left) and $X_{bbVsQCD}$ in the merged regime (right).

Figure 4.24: Some of the input variable distributions in the resolved regime and in the merged regime of the 0-lepton channel.

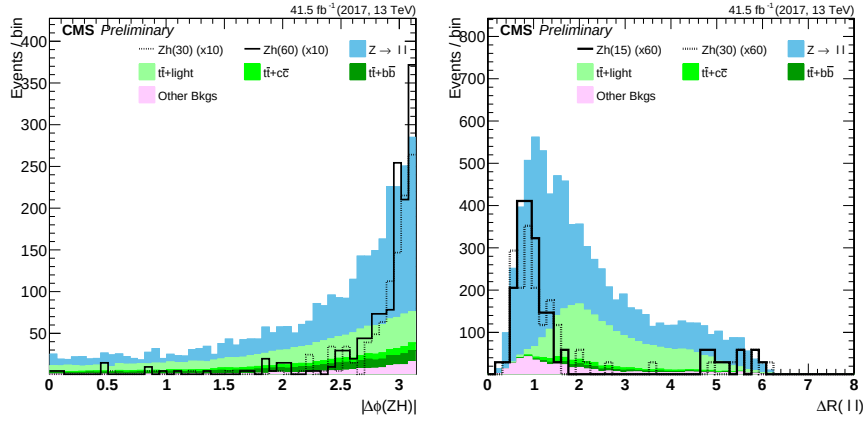
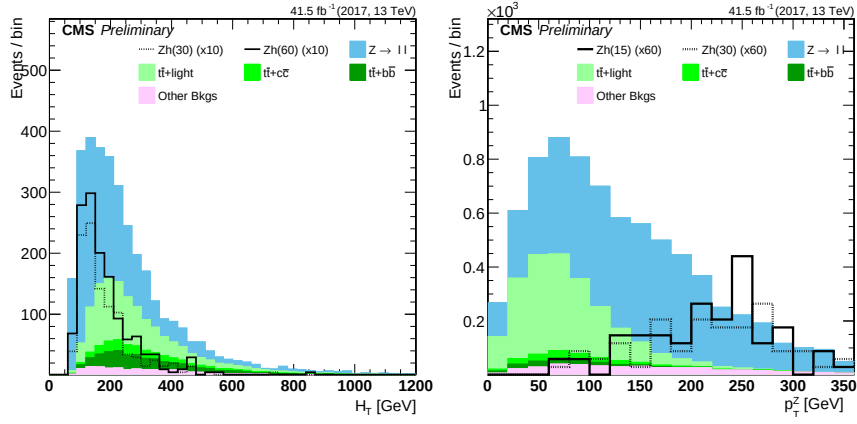
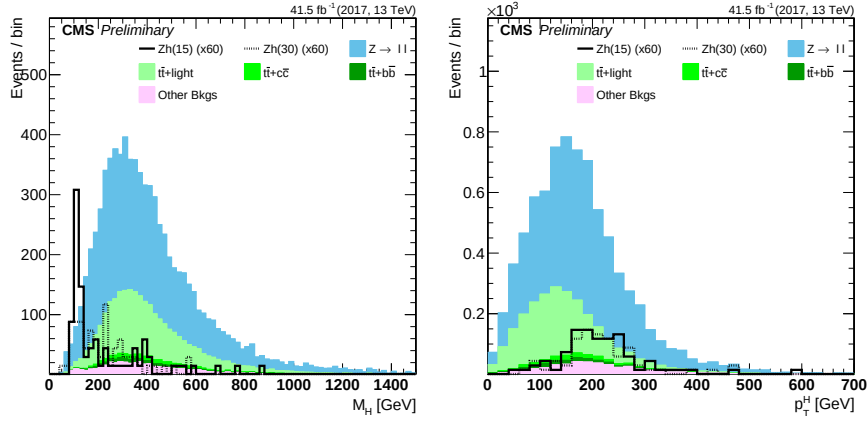
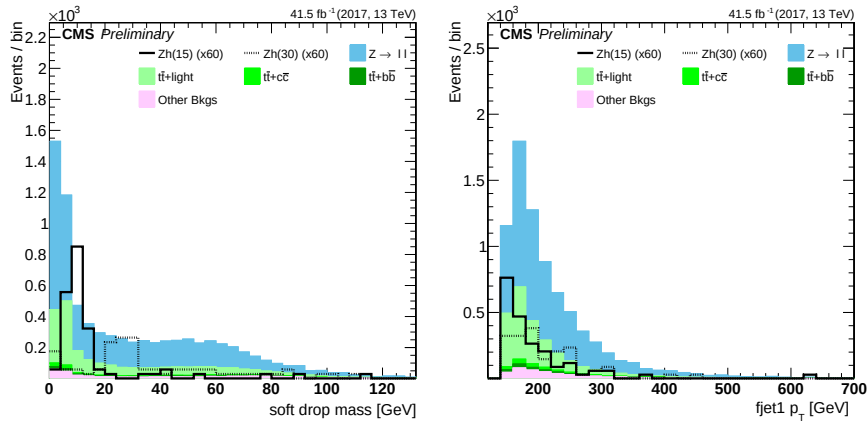
(a) $|\Delta\phi(\text{HZ})|$ (left) and $\Delta R(\text{ll})$ (right) in resolved regime.(b) H_T (left) and p_T^Z (right) in resolved regime.(c) M_H (left) and p_T^H (right) in the merged regime.(d) fat jet p_T (left) and m_{SD} (right) in the merged regime.

Figure 4.25: Some of the input variable distributions in the resolved regime and in the merged regime of the 2-lepton channel.

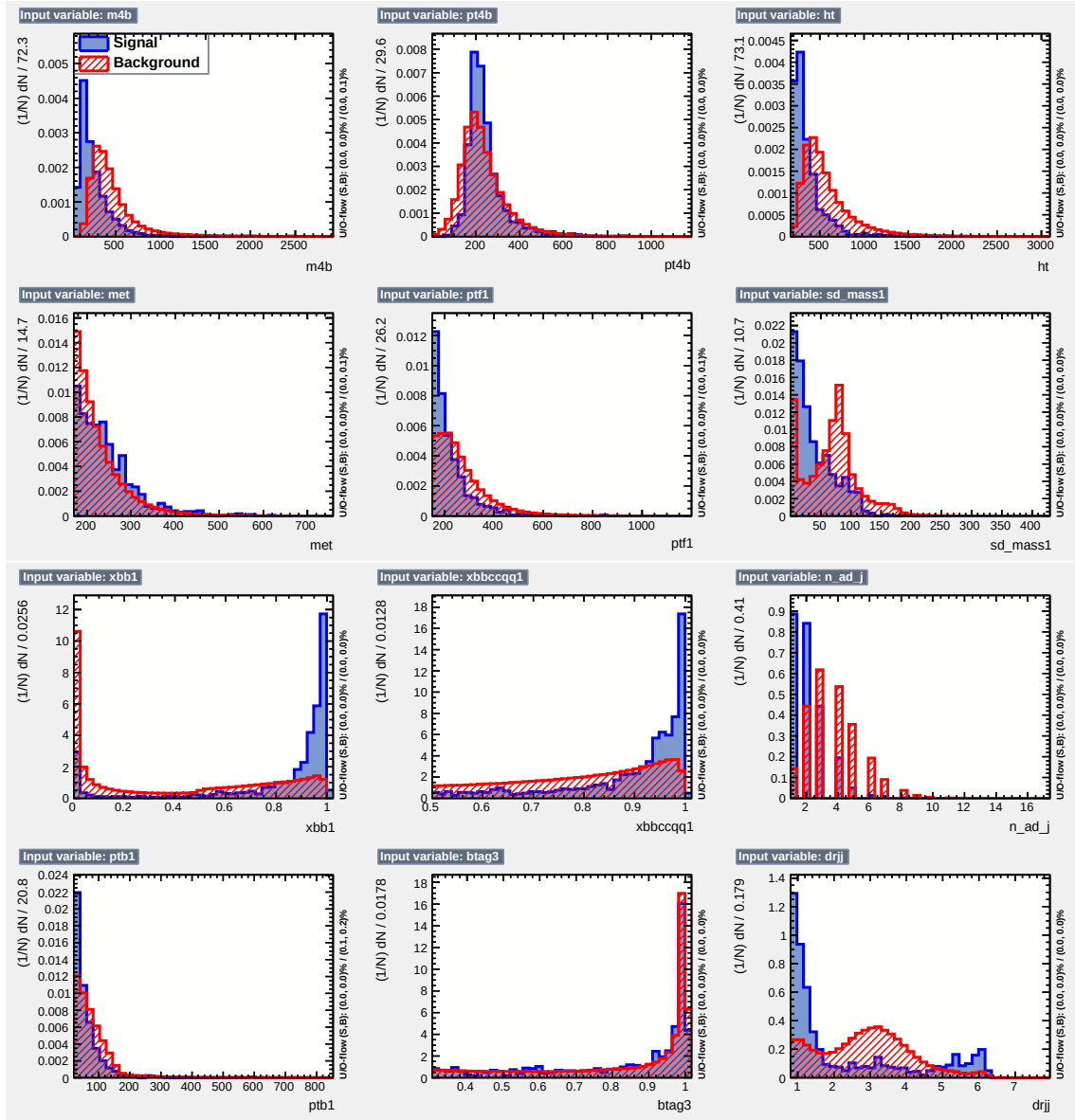


Figure 4.26: Distributions of the variables used for training the BDT in the merged regime of the 0-lepton channel. Red color is the sum of all backgrounds and blue is the sum of signal with $m_\alpha = [12, 15, 20, 25, 30, 35, 40]$

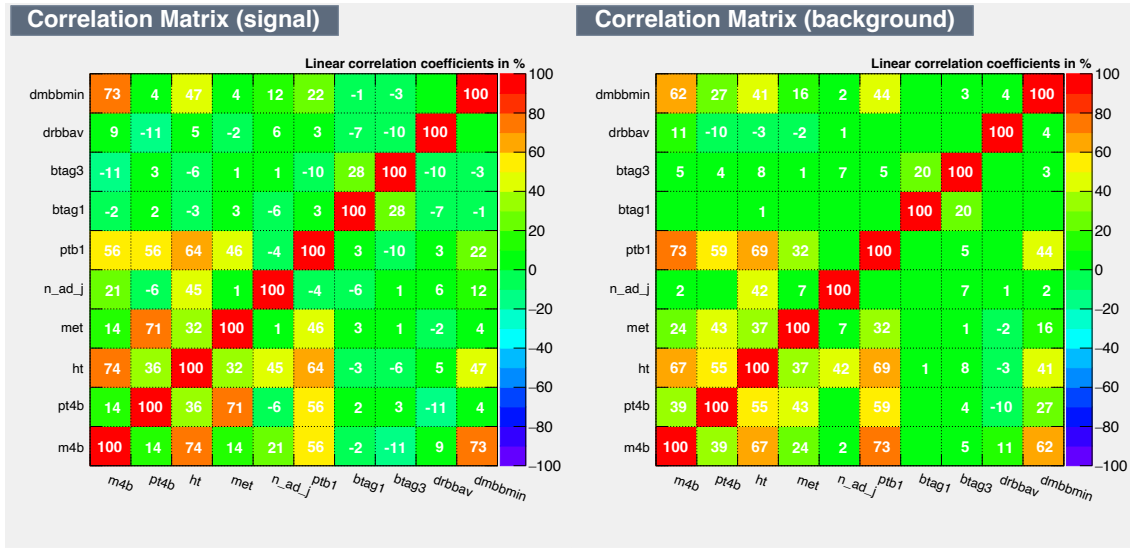


Figure 4.27: Correlation matrices of the training variables in the merged regime of the 0-lepton channel. On the left is the signal sample and on the right is the background sample.

Rank	Variable	Importance (%)
1	XbbVsQCD	14.8
2	m _{SD}	11.5
3	$\Delta R(fjet,j)$	11.3
4	N _j	10
5	b-tag3	9.4
6	Xbb+Xcc+XqqVsQCD	8.8
7	M _H	8
8	H _T	7.8
9	p _T ^H	6.3
10	p _T ^{fjet}	5.9
11	E _T ^{miss}	4.6
12	p _T ^{b1}	1.5

Table 4.13: Ranking of the MVA input variables in terms of importance in the merged regime of the 0-lepton channel

In the resolved regime of the 0-lepton channel, the sum of $M_a = [20, 60]$ GeV "a" boson mass point samples used as the signal sample. QCD samples of $H_T < 300$ GeV were again not included in the background. In total, the statistical events are 3.2K signal and 1.4M background events, 50% of which are used for training and 50% for testing.

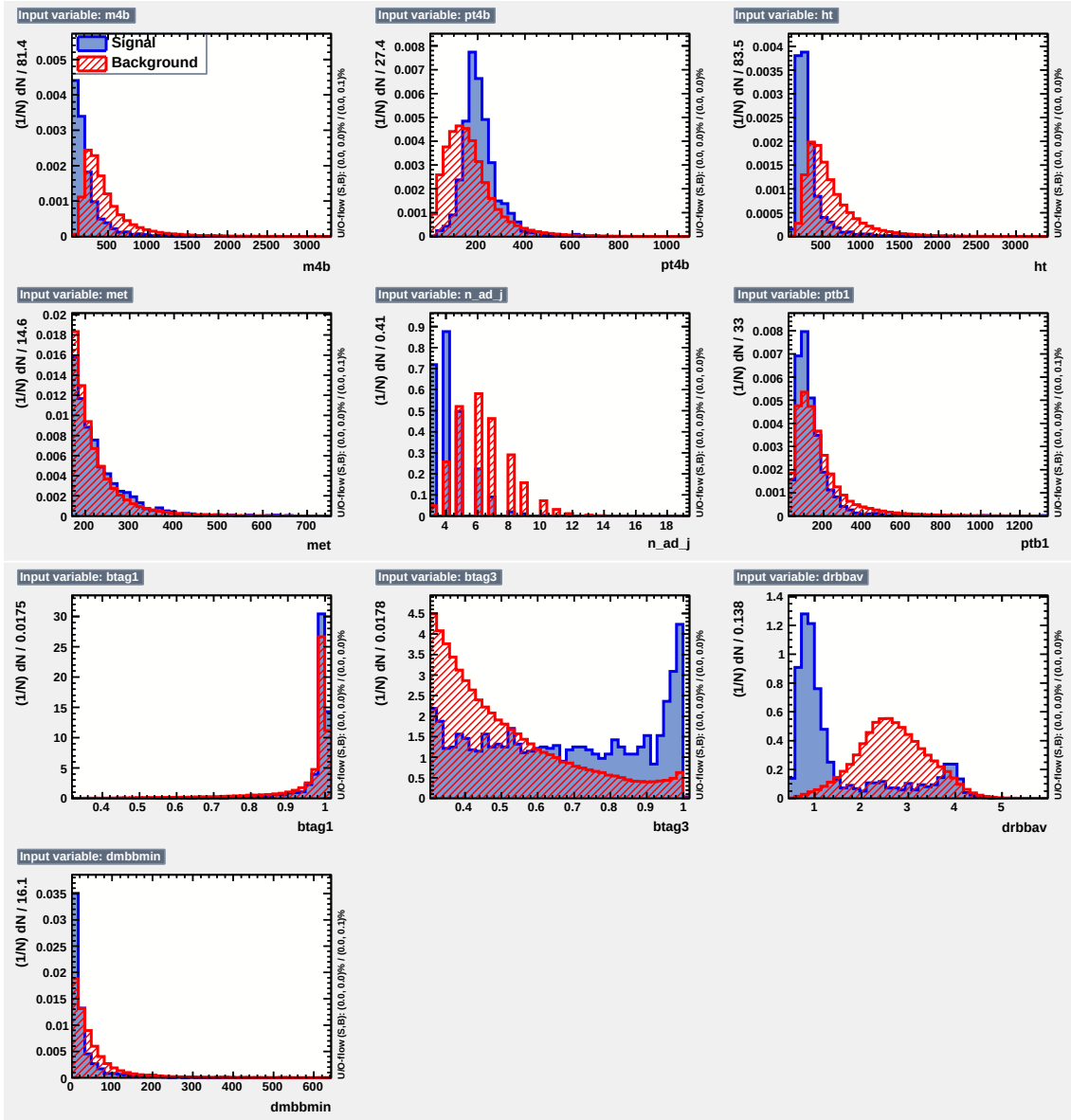


Figure 4.28: Distributions of the variables used for training the BDT in the resolved regime of the 0-lepton channel. Red color is the sum of all backgrounds and blue is the sum of signal with $m_a = [20, 25, 30, 35, 40, 45, 50, 55, 60]$ GeV

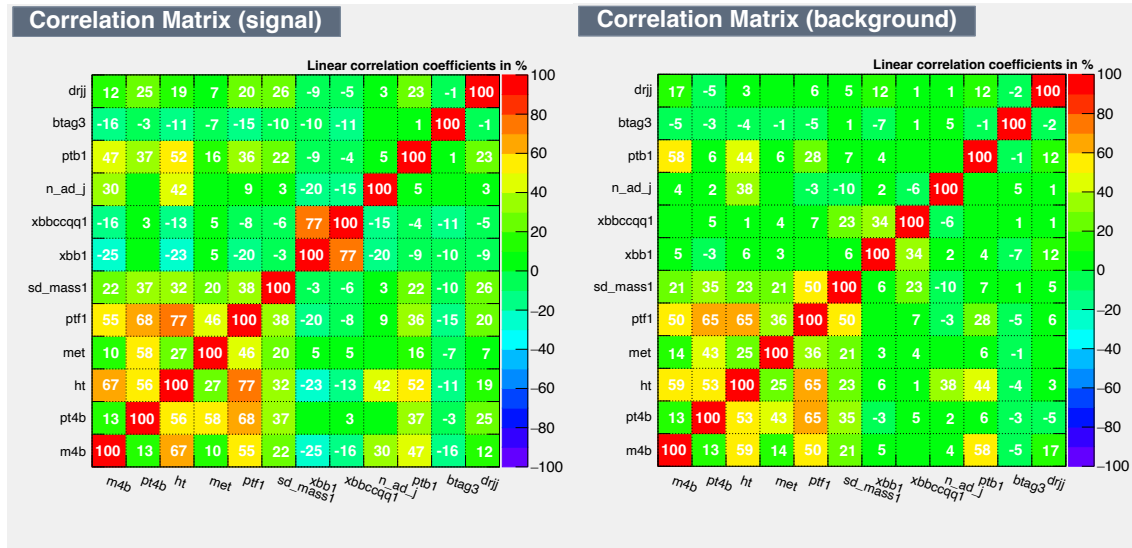


Figure 4.29: Correlation matrices of the training variables in the resolved regime of the 0-lepton channel. On the left is the signal and on the right is the background.

Rank	Variable	Importance (%)
1	$\Delta R(b, b)^{ave}$	16.9
2	M_H	11.8
3	b-tag3	11.6
4	N_j	11
5	p_T^b	9.6
6	p_T^H	9.2
7	$\Delta m(b, b)^{min}$	8.5
8	b-tag1	8.2
9	H_T	7.3
10	E_T^{miss}	5.9

Table 4.14: Ranking of the MVA input variables in terms of importance in the resolved regime of the 0-lepton channel

In the merged regime of the 2-lepton channel, the sum of "a" boson mass point samples of $m_a [12, 40]$ is used as the signal sample. In total, the statistical events are 396 signal and 252 K background events, 50% of which are used for training and 50% for testing.

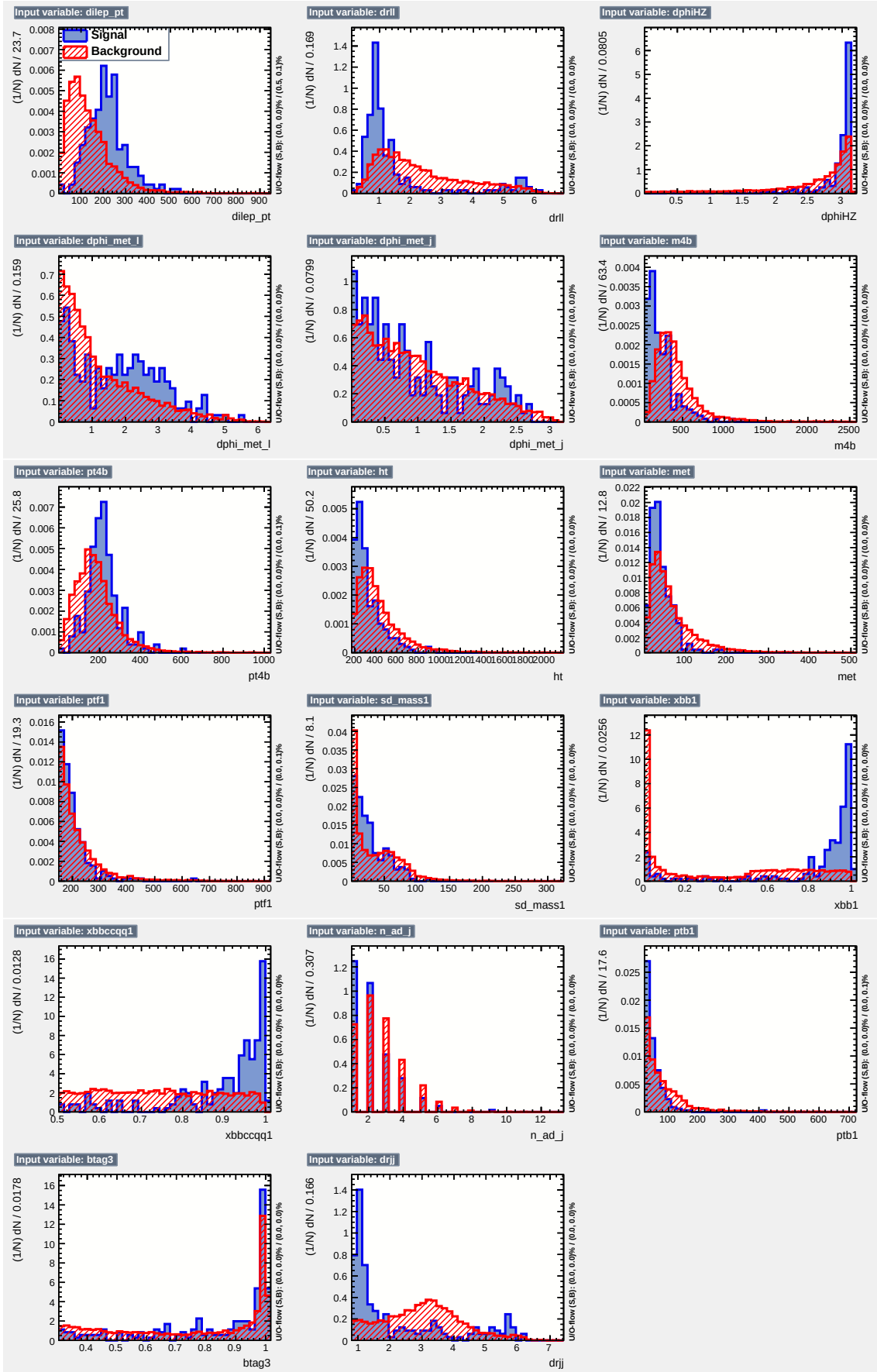


Figure 4.30: Distributions of the variables used for training the BDT in the merged regime of the 2-lepton channel. Red color is the sum of all backgrounds and blue is the sum of signal with $M_a = [12, 15, 20, 25, 30, 35, 40]$

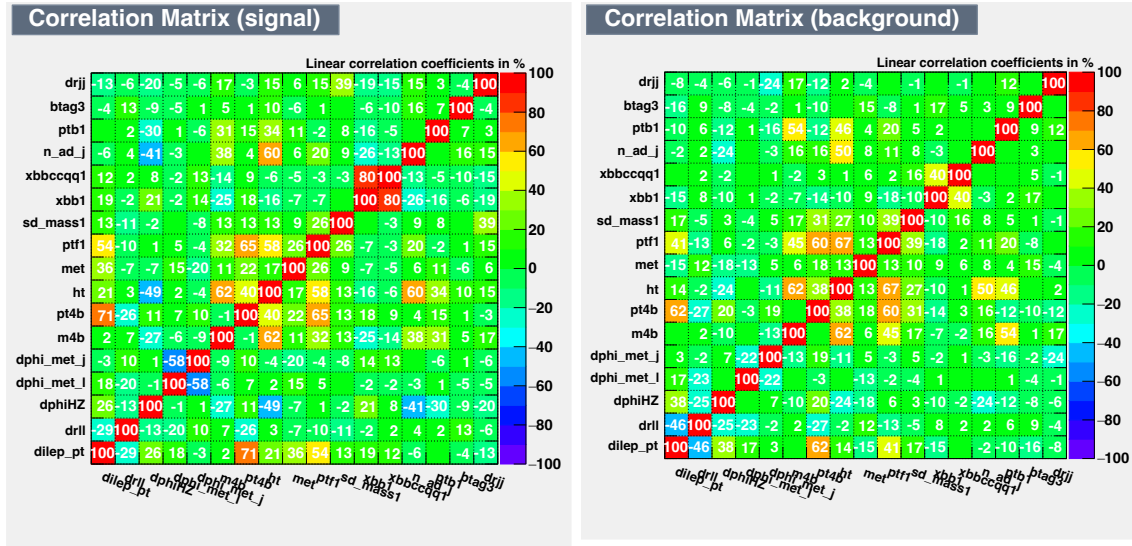


Figure 4.31: Correlation matrices of the training variables in the merged regime of the 2-lepton channel. On the left is the signal and on the right is the background.

Rank	Variable	Importance (%)
1	XbbVsQCD	14.2
2	$\Delta R(\text{fjet}_j)$	8.7
3	M_H	8
4	Xbb+Xcc+XqqVsQCD	7.3
5	p_T^Z	7.1
6	$ \Delta\phi(E_T^{\text{miss}}, j) ^{\text{min}}$	6.8
7	N_{jets}	6
8	b-tag3	5.7
9	m_{SD}	5.6
10	E_T^{miss}	5.3
11	$\Delta R_{\ell\ell}$	5
12	p_T^{b1}	4.5
13	$ \Delta\phi(ZH) $	4.1
14	H_T	3.4
15	p_T^{fjet}	2.3
16	$ \Delta\phi(E_T^{\text{miss}}, \ell) ^{\text{min}}$	5.1
17	p_T^H	2

Table 4.15: Ranking of the MVA input variables in terms of importance in the merged regime of the 2-lepton channel

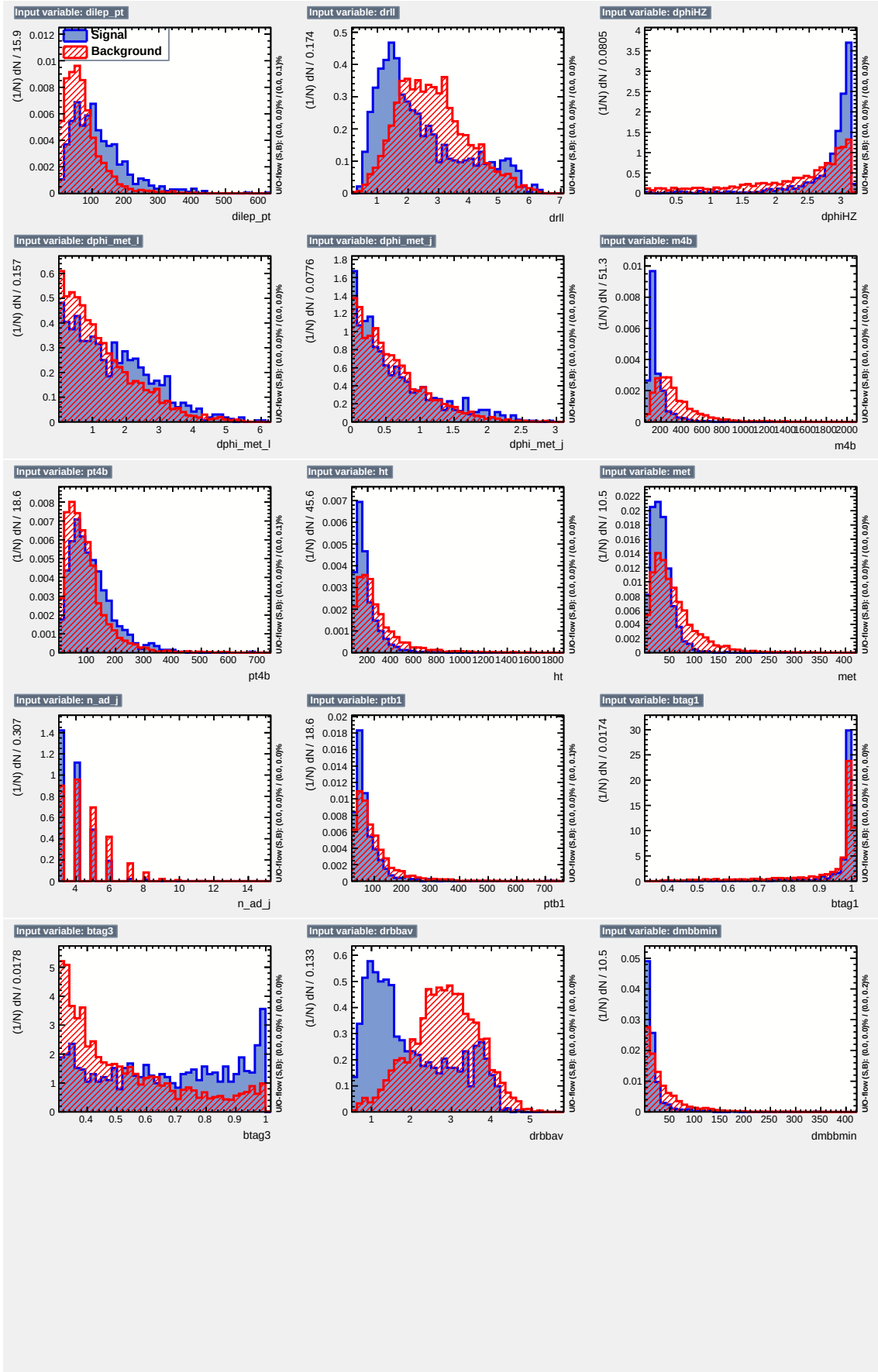


Figure 4.32: Distributions of the variables used for training the BDT in the resolved regime of the 2-lepton channel. Red color is the sum of all backgrounds and blue is the sum of signal with $M_a = [20,25,30,35,40,45,50,55,60]$ GeV

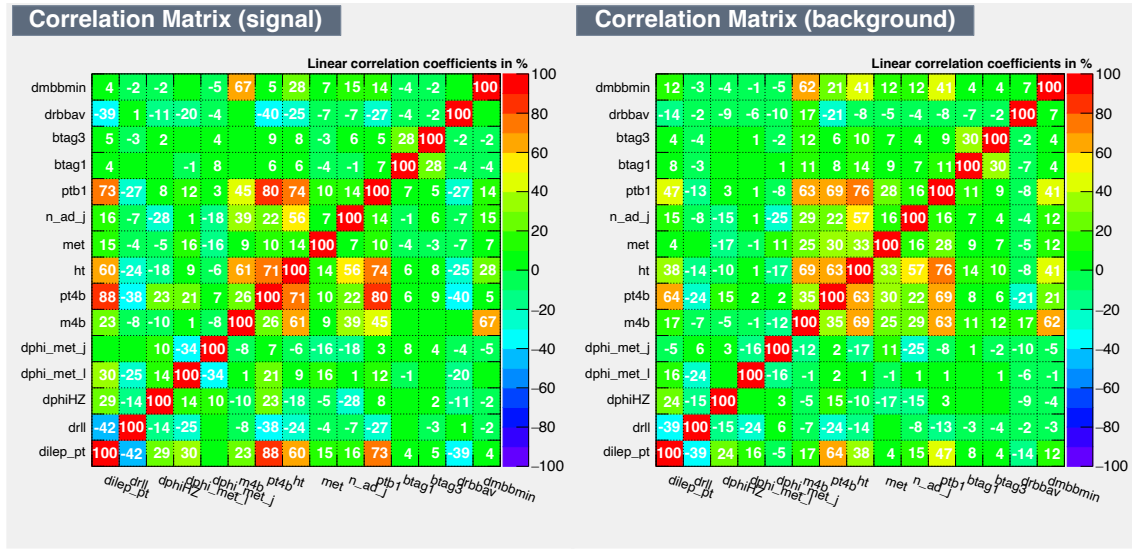


Figure 4.33: Correlation matrices of the training variables in the resolved regime of the 2-lepton channel. On the left is the signal and on the right is the background.

Rank	Variable	Importance (%)
1	$\Delta R(b,b)^{\text{ave}}$	15.3
2	M_H	9.7
3	E_T^{miss}	8.2
4	b-tag3	7.7
5	$ \Delta\phi(E_T^{\text{miss}}, j) ^{\text{min}}$	7.6
6	N_j	6.6
7	p_T^Z	6.5
8	$ \Delta\phi(E_T^{\text{miss}}, \ell) ^{\text{min}}$	6.3
9	$ \Delta\phi(ZH) $	6.2
10	p_T^H	6
11	H_T	5.7
12	p_T^{b1}	4.6
13	$\Delta m(b, b)^{\text{min}}$	4.4
14	$\Delta R_{\ell\ell}$	4.3
15	b-tag1	4

Table 4.16: Ranking of the MVA input variables in terms of importance in the resolved regime of the 2-lepton channel

Finally, in the resolved regime of the 2-lepton channel, the sum of m_{a} boson mass point samples of M_{a} [20,60] GeV is used as the signal sample. In total, the statistical events are 2.1 K signal and 57 K background events, 50% of which are used for training and 50% for testing.

4.7.4 BDT response

The BDT Response is the result obtained from applying multivariate analysis. It is a numerical value generated by the BDT (Boosted Decision Tree) model that quantifies the likelihood of an event belonging to the signal or background category. This value typically ranges between two extremes and is often normalized within the range $[-1, 1]$.

- A BDT response close to 1 indicates that the event is classified as signal.
- A BDT response close to -1 suggests that the event belongs to the background.

Figures show the BDT response for the training sample and the testing sample superimposed. By overlaying the two samples, the analysis checks for overtraining.

Overtraining occurs when the model focuses excessively on the training dataset instead of learning the features that distinguish between signal and background events. As a result, the model memorizes the training data, including its fluctuations, and fails to generalize to new datasets. Overtraining is often caused by:

A large number of parameters in the model relative to the available training data. The model's high flexibility, which allows it to adapt too closely to the noise and variations in the training sample.

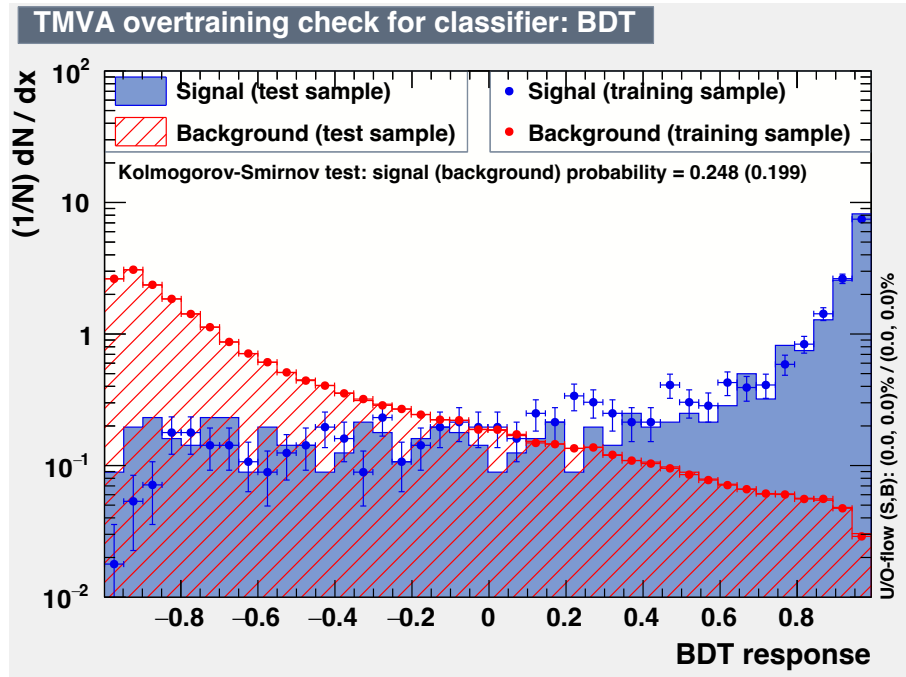
The data points of the training sample do not significantly deviate from those of the testing sample. This confirms that overtraining has not occurred in this case. Also, the Kolmogorov-Smirnov test is applied, which compares the CDF of training and testing BDT distributions for both signal and background samples. In any case, $K-S \text{ test} > 0.05$, which is a threshold indicating overfitting.

4.7.5 Receiver Operating Characteristic Curve

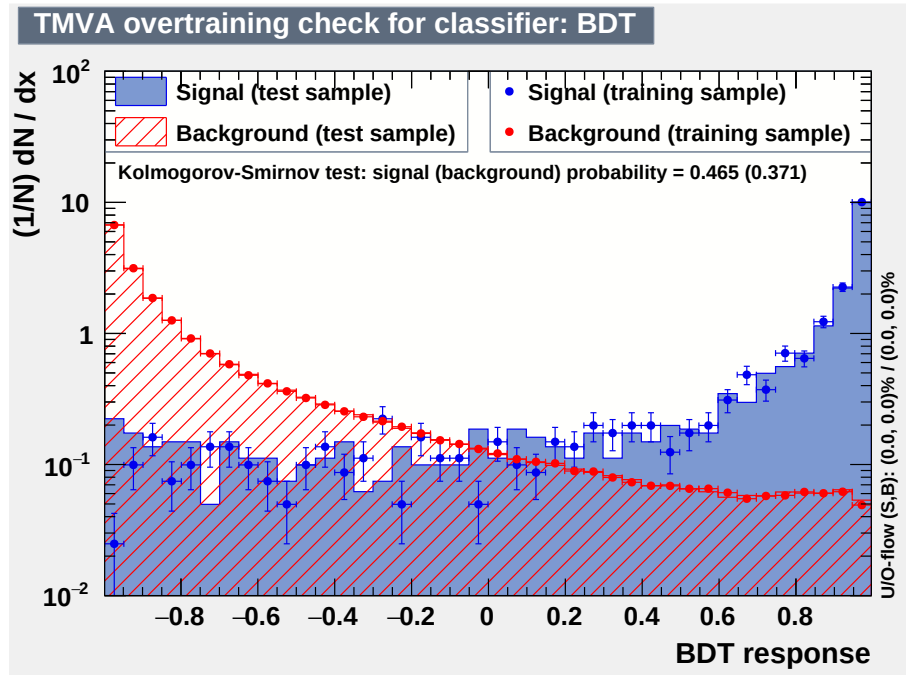
Receiver Operating Characteristic (ROC) curves are graphical representations used to evaluate the performance of a binary classifier, such as the BDT. These curves are constructed by plotting the relationship between signal efficiency and background rejection.

- **Signal efficiency** (x-axis): Represents the percentage of signal events correctly identified by the model.
- **Background efficiency**: Indicates the percentage of background events that the classifier incorrectly categorizes as signal events.
- **Background rejection** (y-axis): Reflects the percentage of background events that the model correctly classifies as background. It is complementary to background efficiency and is calculated using the formula:

$$\text{Background Rejection} = 1 - \text{Background Efficiency} \quad (4.22)$$

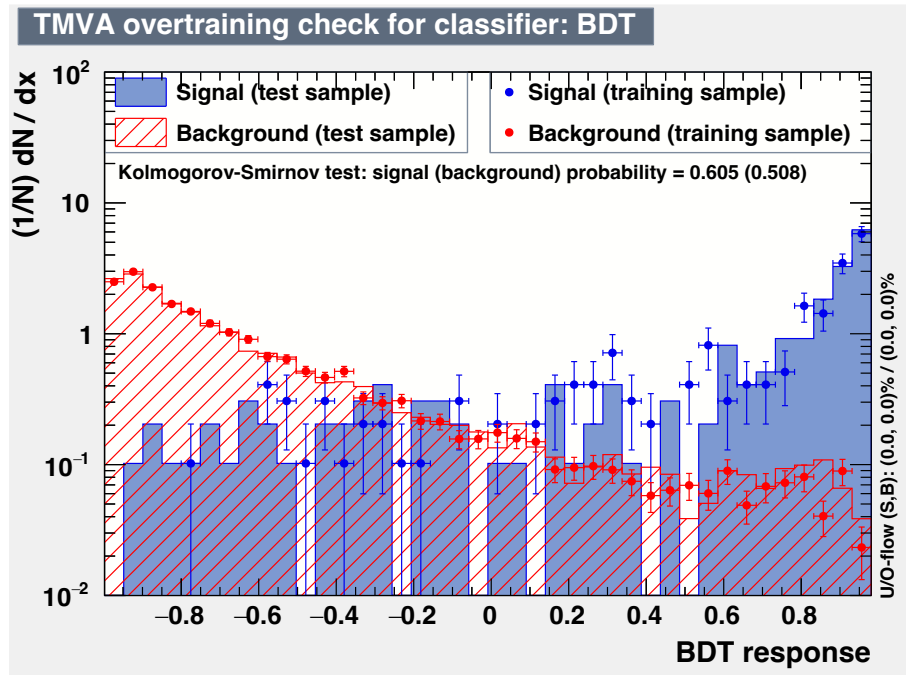


(a)

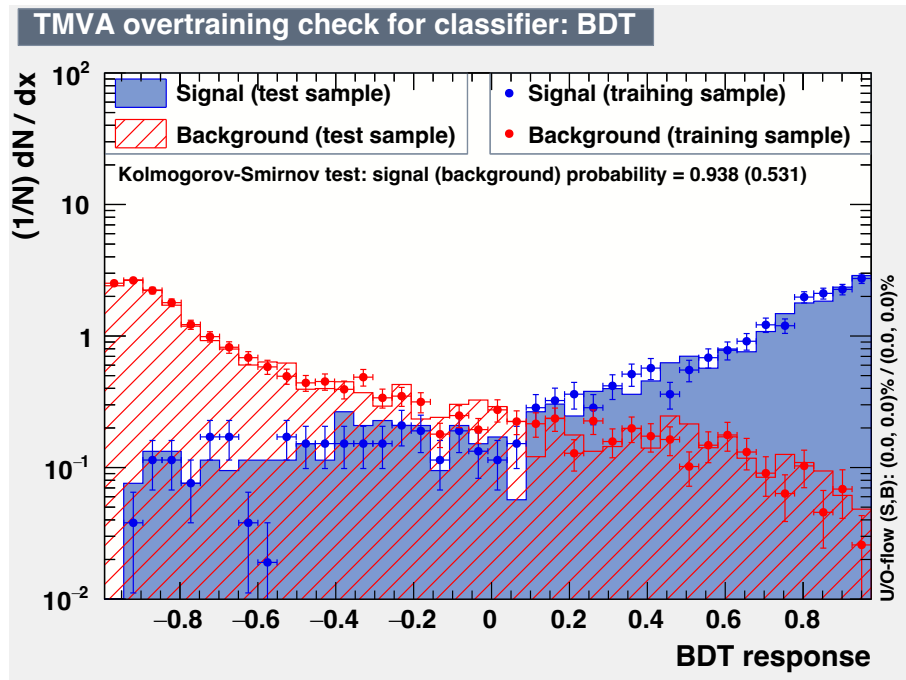


(b)

Figure 4.34: BDT response for train sample and test sample superimposed: in (a) merged regime (b) resolved regime, 0-lepton channel.



(a)



(b)

Figure 4.35: BDT response for train sample and test sample superimposed: (a) in the merged regime (b) in the resolved regime of the 2-lepton channel.

The ROC curve shows how effectively the model distinguishes between signal and background events. Each point on the ROC curve corresponds to a different BDT cut applied to the BDT out-

put, aiming to retain as much signal as possible while rejecting as much background as possible. Essentially, the ROC curve illustrates the trade-off between signal efficiency and background rejection.

For a good model, the ROC curve should approach the top-right corner of the graph, indicating high background rejection and high signal efficiency simultaneously.

In addition, the Area Under Curve (AUC) is often used as a performance metric. Higher AUC values (close to 1) are desirable, as they indicate better classifier performance. For the ROC curves shown in Figures , the AUC is between 0.94 and 0.96 for the test samples, signifying good classification performance.

Besides the ROC curve of the testing sample, the ROC curve of the training sample is superimposed in each figure, as an over-training check. No over-training is observed, as ROC curves corresponding to training and testing samples tend to be identical.

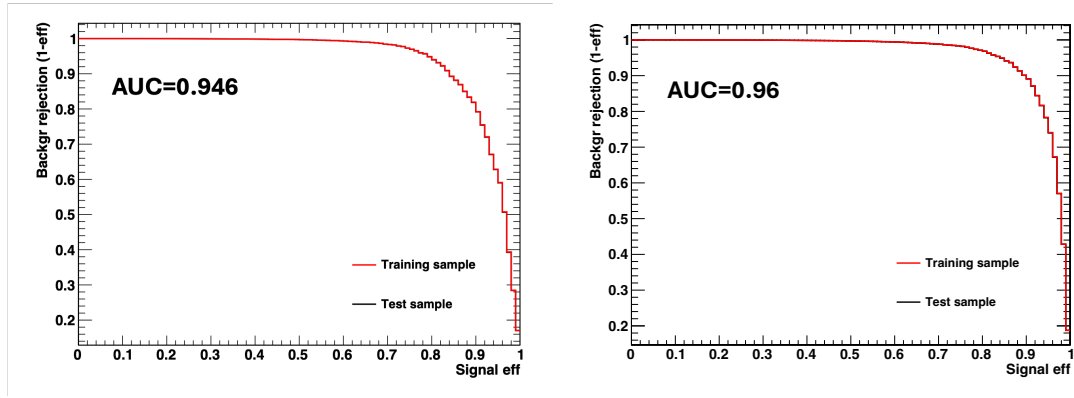


Figure 4.36: Roc curves for train sample and test sample superimposed: in the merged regime (left) and in the resolved regime (right) of the 2-lepton channel.

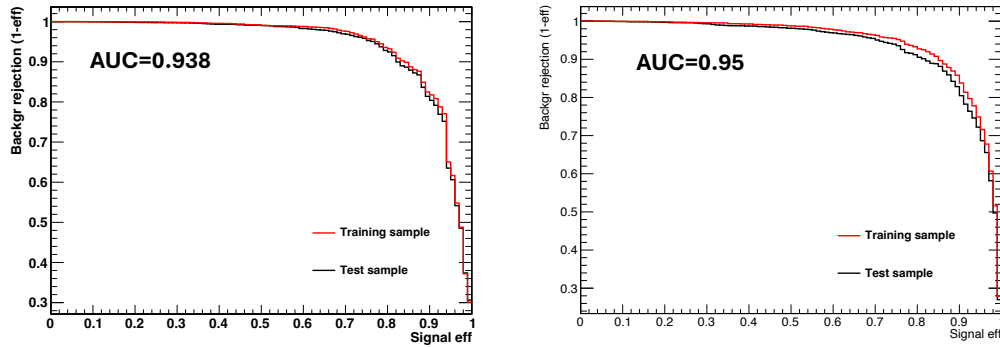


Figure 4.37: Roc curves for train sample and test sample superimposed: in the merged regime (left) and in the resolved regime (right) of the 2-lepton channel.

4.7.6 BDT Cut Efficiencies

The "cut efficiencies" graph illustrates how various quantities (signal and background efficiency, signal purity, significance) change with the cutoff value (BDT cut) applied to the output of the

BDT. Specifically, this graph presents the following curves:

- **Signal efficiency:** Reflects the proportion of signal events that are correctly categorized as signal after applying the BDT cut. As the cutoff value decreases, more events are classified as signal events, and the signal efficiency increases.
- **Background efficiency:** Indicates the percentage of background events that are incorrectly categorized as signal events, depending on the BDT cut value.
- **Signal purity:** Represents the proportion of events identified as signal events that truly belong to this category. Higher BDT cut values correspond to higher purity, as the model becomes more "selective" about which events it categorizes as signal. At the same time, high signal purity values suggest that signal events dominate the selected sample.
- **Significance or sensitivity:** Provides an estimate of how well the signal can be distinguished from the background, depending on the BDT cut value. In this study, we focus on the significance value as determined when the optimal BDT cut value is applied.

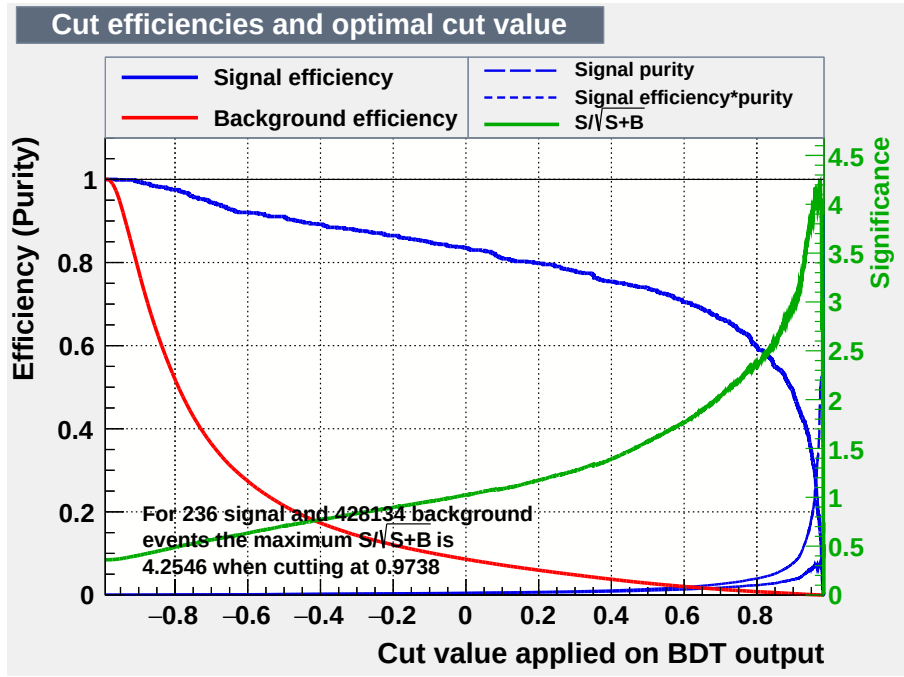
Significance is defined by the relationship:

$$\text{Significance} = \frac{S}{\sqrt{S+B}}. \quad (4.23)$$

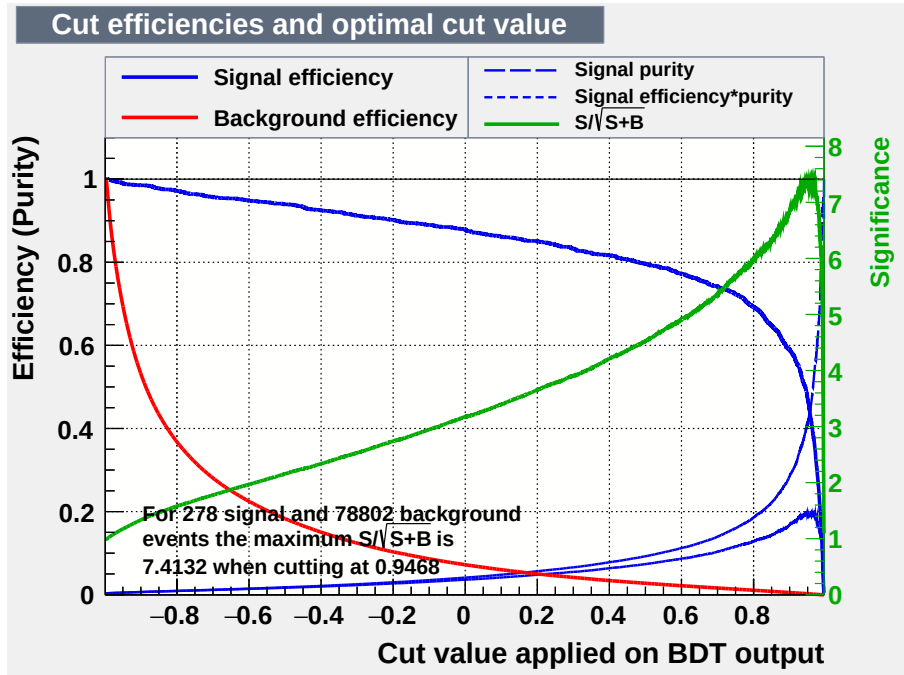
S represents the number of expected signal events detected after applying the BDT cut, and B represents the number of expected background events detected under the same conditions. The denominator, $\sqrt{S+B}$, corresponds to the statistical uncertainty in the total number of signal and background events. The greater the number of background events, the stronger the uncertainty, making it more challenging to distinguish the signal from the background. Essentially, the significance quantifies how well a signal event can be distinguished from the background.

Regarding the optimal cut value, this refers to the point on the graph where the significance is maximized. At this specific BDT cut value, the signal is most effectively separated from the background. This maximization is critical for potential discoveries, as it increases the likelihood that the observed signal event is real.

Figures 4.38 and 4.39 correspond to the merged and the resolved regime of the 0-lepton and of the 2-lepton channel respectively. The number of expected signal events applied is corresponding to the signal "a" mass hypothesis with the maximum expected yield. The final number of expected background events is the sum of the expected yield of all SM background processes. However, these results are just a figure of merit, as the method that will be followed in the analysis is BDT binned shape analysis.

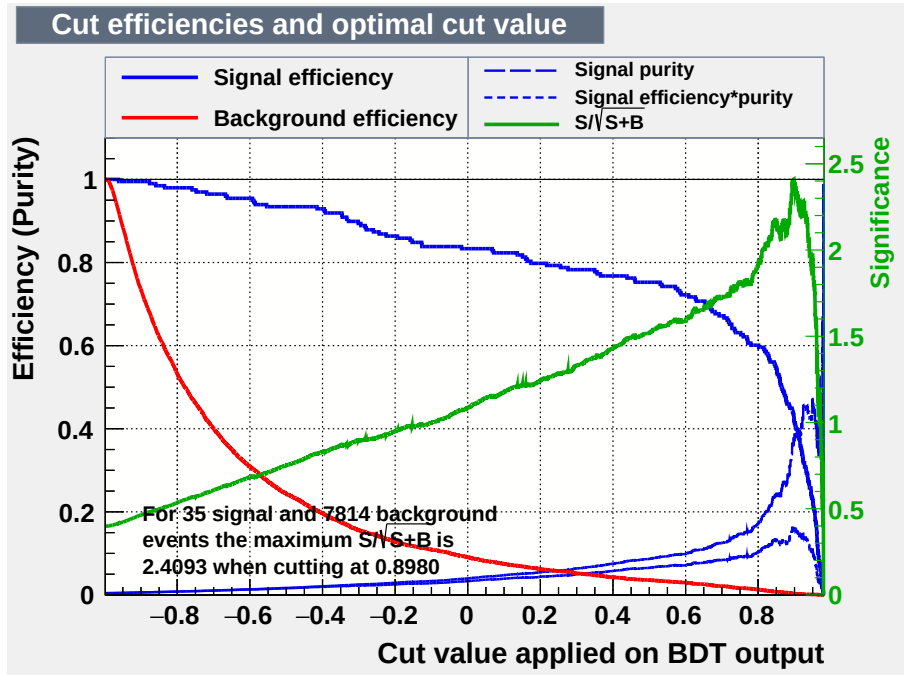


(a)

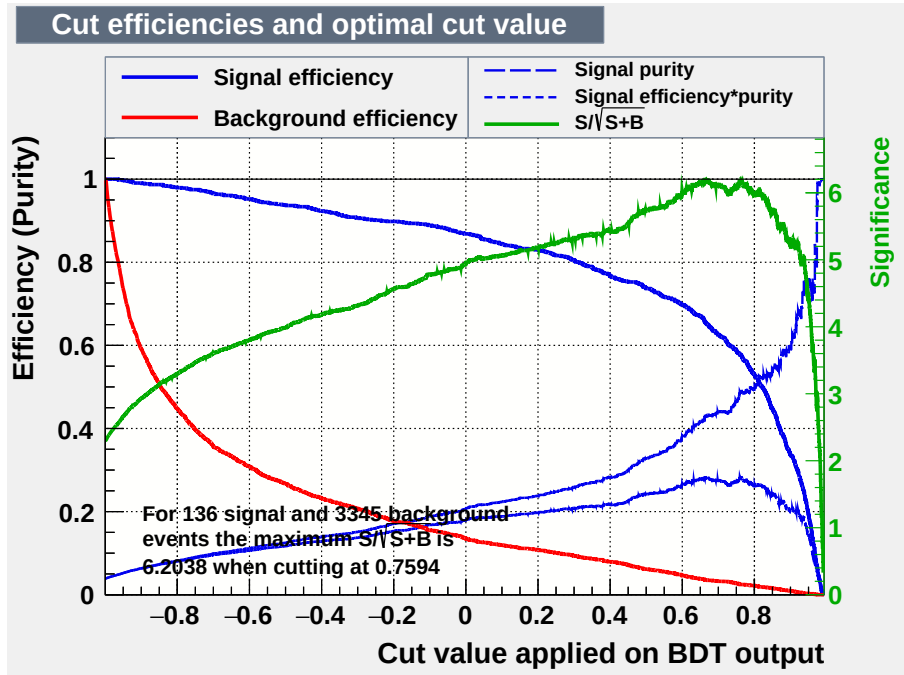


(b)

Figure 4.38: Cut efficiencies and optimal cut value for the BDT response channel: (a) In the merged regime of the 0-lepton channel: There are used 236 expected signal events corresponding to expected yield of $m_a = 12$ GeV and 428134 expected background events. The maximum significance is approximately: 4.3 which occurs when the BDT cut value is around 0.95 (b) In the resolved regime of the 0-lepton channel: There are used 278 expected signal events corresponding to expected yield of $m_a = 50$ GeV and all 78802 expected background events. The maximum significance is approximately: 7.4 which occurs when the BDT cut value is around 0.95



(a)



(b)

Figure 4.39: Cut efficiencies and optimal cut value for the BDT response: (a) In the merged regime of the 2-lepton channel: There are used 35 expected signal events corresponding to expected yield of $m_a = 12$ GeV and all 7814 expected background events. the maximum significance is approximately: 2.4 which occurs when the BDT cut value is around 0.89 (b) In the resolved regime of the 2-lepton channel: There are used 136 expected signal events corresponding to expected yield of $m_a = 60$ GeV and all 3345 expected background events. The maximum significance is approximately: 6.2 which occurs when the BDT cut value is around 0.76.

4.8 BDT binned shape analysis

4.8.1 BDT binning optimization

Before proceeding to the binned maximum likelihood fit on the BDT variable shape, and extracting upper limits on signal yield, we have to optimize the binning of the BDT distribution in order to maximize significance and maintain numerical stability of the fitting procedure as well. Significance, defined as $S/\sigma(B)$, is used as a metric for finding the optimal binning of BDT. We want to maximize significance in the signal-dominated area of the BDT, i.e., in the high BDT bins. We rebin the BDT distribution, such that the two highest-value BDT bins yield the maximum possible significance. What we do is scan the BDT distribution, starting from 1 and moving downwards, and compute the significance of all possible rebinning combinations of the two highest BDT bins. We want to find the lower bin edges of the two highest BDT bins that yield the maximum quadratic significance of all possible rebinning combinations. In other words, we want to maximize the quantity:

$$\sqrt{\left(\frac{S}{\sigma_B}(\text{bin_N-1})\right)^2 + \left(\frac{S}{\sigma_B}(\text{bin_N})\right)^2} \quad (4.24)$$

For numerical stability, we also require at least one event of each process to exist in each bin. Optimization is performed using the signal sample with the maximum events in each channel and regime. After that, since we have chosen the two highest-value BDT bins, we proceed to the selection of the rest such that one or two lower BDT bins also contribute to total significance while the lower one or two bins are background dominated.

The procedure of computing the quadratic significance over all possible two highest-BDT bins combinations is depicted in Figures 4.40-4.43. The significance in each bin of the rebinned BDT distribution is shown in Tables 4.17-4.20.

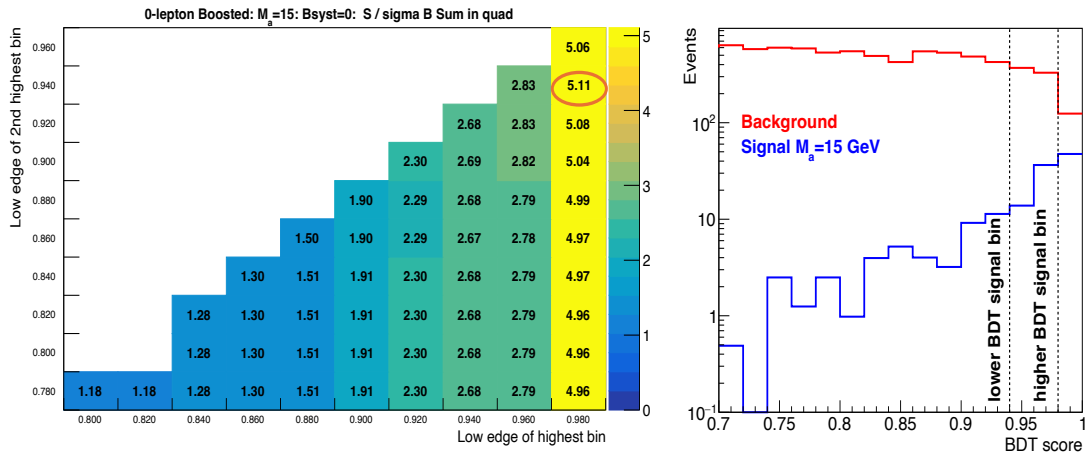


Figure 4.40: Binning optimization in the merged regime of the 0-lepton channel. On the left: The quadratic sum of significance of two highest BDT bins shown for many possible combinations of lower edges of the two highest BDT bins. The maximum possible value is encircled. On the right: BDT plot of the signal and the total background superimposed, shown for high BDT values. The lower edges of the two highest BDT bins are shown with dashed lines.

$S/(B)$	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
$M_a(15)$	0.02	0.04	0.09	0.33	1.42	4.90
$M_a(30)$	0.02	0.06	0.13	0.51	1.23	3.29

Table 4.17: Significance in each BDT bin after binning optimization, in the merged regime of the 0-lepton channel

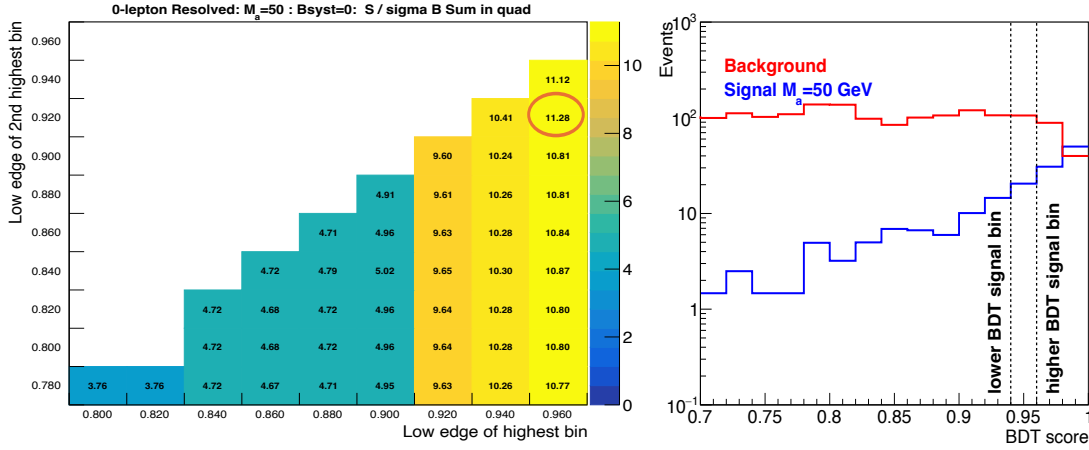


Figure 4.41: Binning optimization in the resolved regime of the 0-lepton channel. On the left: The quadratic sum of significance of two highest BDT bins shown for many possible combinations of lower edges of the two highest BDT bins. The maximum possible value is encircled. On the right: BDT score of the signal and the total background shown for high BDT values. The lower edges of the two highest BDT bins are shown with dashed lines.

Process	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
$S/\sigma(B) M_a(30)$	0.05	0.06	0.42	0.66	3.38	7.78
$S/\sigma(B) M_a(60)$	0.05	0.04	0.29	1.09	3.19	10.65

Table 4.18: Significance in each BDT bin after binning optimization, in the resolved regime of the 0-lepton channel

Process	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
$S/\sigma(B)M_a(15)$	0.03	0.02	0.03	0.27	2.06
$S/\sigma(B)M_a(30)$	0.06	0.13	0.13	0.23	1.35

Table 4.19: Significance in each BDT bin after binning optimization in the merged regime of the 2-lepton channel.

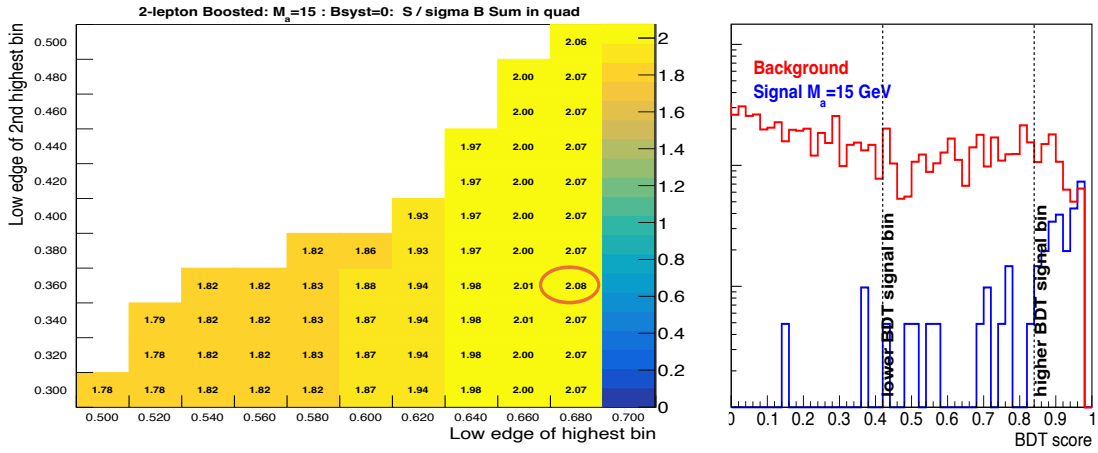


Figure 4.42: Binning optimization in the merged regime of the 2-lepton channel. On the left: The quadratic sum of significance of two highest BDT bins shown for many possible combinations of lower edges of the two highest BDT bins. The maximum possible value is encircled. On the right: BDT score of the signal and the total background shown for high BDT values. The lower edges of the two highest BDT bins are shown with dashed lines.

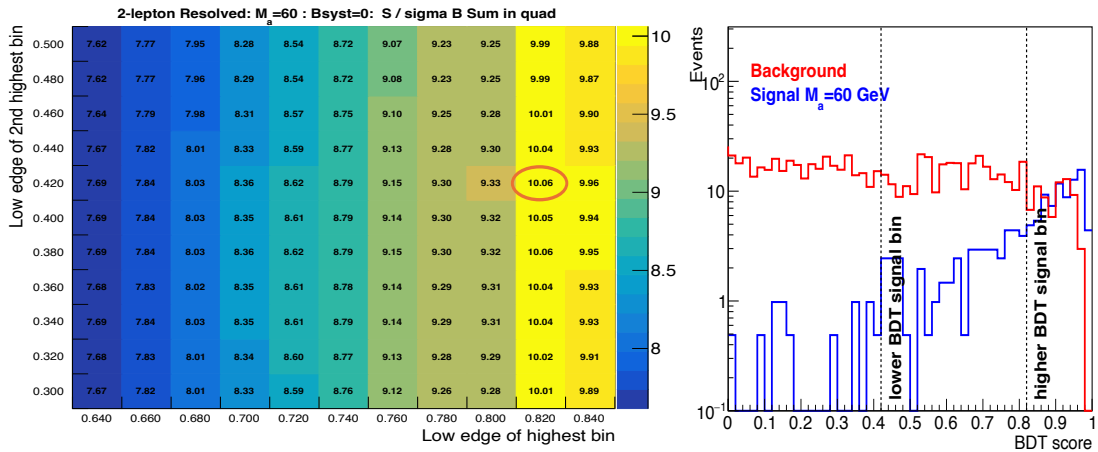


Figure 4.43: Binning optimization in the resolved regime of the 2-lepton channel. On the left: The quadratic sum of significance of two highest BDT bins shown for many possible combinations of lower edges of the two highest BDT bins. The maximum possible value is encircled. On the right: BDT plot of the signal and the total background shown for high BDT values. The lower edges of the two highest BDT bins are shown with dashed lines.

Process	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
$S/\sigma(B)M_a(30)$	0.06	0.22	0.48	1.89	6.87
$S/\sigma(B)M_a(60)$	0.06	0.17	0.31	2.61	9.72

Table 4.20: Significance in each BDT bin after binning optimization, in the resolved regime of the 2-lepton channel.

4.9 Maximum likelihood fit

In order to proceed to the description of the maximum likelihood fit procedure, let's define at first the two following hypotheses:

1. **Null Hypothesis** H_0 : This hypothesis includes only background processes referring to the Standard Model (SM), meaning there is only background (B).
2. **Alternative Hypothesis** H_1 : This hypothesis includes additional physical phenomena beyond the Standard Model (SM), meaning in addition to the background, there is also a signal (S+B).

Each hypothesis is represented by a model, H_0 hypothesis by model_0 and H_1 hypothesis by model_1. The theoretical distributions of these models are the following :

$$f_{\text{model}_0}(x) = \sum_{\text{bkg}} N_{\text{bkg}}^{\text{exp}} \cdot \text{pdf}_{\text{bkg}}(x) \Rightarrow$$

$$f_{\text{model}_0}(x) = \begin{cases} N_{t\bar{t}+\text{light}} \cdot \text{pdf}_{t\bar{t}+\text{light}} + N_{t\bar{t}+c\bar{c}} \cdot \text{pdf}_{t\bar{t}+c\bar{c}} + N_{t\bar{t}+b\bar{b}} \cdot \text{pdf}_{t\bar{t}+b\bar{b}} + N_{\text{QCD}} \cdot \text{pdf}_{\text{QCD}} \\ + N_{\text{Wjets}} \cdot \text{pdf}_{\text{Wjets}} + N_{\text{Zjets}} \cdot \text{pdf}_{\text{Zjets}} + N_{\text{EW}} \cdot \text{pdf}_{\text{EW}} & \text{(0-lepton)} \\ N_{t\bar{t}+\text{light}} \cdot \text{pdf}_{t\bar{t}+\text{light}} + N_{t\bar{t}+c\bar{c}} \cdot \text{pdf}_{t\bar{t}+c\bar{c}} + N_{t\bar{t}+b\bar{b}} \cdot \text{pdf}_{t\bar{t}+b\bar{b}} \\ + N_{\text{DY}} \cdot \text{pdf}_{\text{DY}} + N_{\text{EW}} \cdot \text{pdf}_{\text{EW}} & \text{(2-lepton)} \end{cases} \quad (2.22)$$

For the hypothesis H_1 , the model is:

$$f_{\text{model}_1}(x) = N_{\text{signal}}^{\text{exp}} \cdot \text{pdf}_{\text{signal}}(x) + \sum_{\text{bkg}} N_{\text{bkg}}^{\text{exp}} \cdot \text{pdf}_{\text{bkg}}(x) \Rightarrow$$

$$f_{\text{model}_1}(x) = \begin{cases} N_{\text{sig}} \cdot \text{pdf}_{\text{sig}} + N_{t\bar{t}+\text{light}} \cdot \text{pdf}_{t\bar{t}+\text{light}} + N_{t\bar{t}+c\bar{c}} \cdot \text{pdf}_{t\bar{t}+c\bar{c}} + N_{t\bar{t}+b\bar{b}} \cdot \text{pdf}_{t\bar{t}+b\bar{b}} \\ + N_{\text{QCD}} \cdot \text{pdf}_{\text{QCD}} + N_{\text{Wjets}} \cdot \text{pdf}_{\text{Wjets}} + N_{\text{Zjets}} \cdot \text{pdf}_{\text{Zjets}} + N_{\text{EW}} \cdot \text{pdf}_{\text{EW}} & \text{(0-lepton)} \\ N_{\text{sig}} \cdot \text{pdf}_{\text{sig}} + N_{t\bar{t}+\text{light}} \cdot \text{pdf}_{t\bar{t}+\text{light}} + N_{t\bar{t}+c\bar{c}} \cdot \text{pdf}_{t\bar{t}+c\bar{c}} + N_{t\bar{t}+b\bar{b}} \cdot \text{pdf}_{t\bar{t}+b\bar{b}} \\ + N_{\text{DY}} \cdot \text{pdf}_{\text{DY}} + N_{\text{EW}} \cdot \text{pdf}_{\text{EW}} & \text{(2-lepton)} \end{cases} \quad (2.22)$$

Where x is the observable variable, in our case the BDT response. Each probability density $\text{pdf}_{\text{process}}$, is by definition normalized such that they have unitary integral, while the normalization factor of each process N_{process} equal to $\sigma_{\text{process}} \cdot L_{\text{int}}$. The model definition as well as all the next steps of the analysis were conducted using the RooFit toolkit [64]

In our analysis, since we do not use real data, we generate binned MC data i.e. pseudo-data from the model_0. For the MC generation, we produce random pseudo-experimental values of the BDT score variable. The total number of MC events is equal to the sum of the expected number of events from all background sources, while in each bin of the distribution, there will be statistical fluctuations around the number of expected background events in that bin. We then apply the same analysis to the MC data as we do to the actual data. Our purpose is to determine an upper limit on the signal yield that may exist. So, the first step of this procedure is to fit the model_1, that corresponds to (S+B) hypothesis to our generated datasets, and obtain the signal yield value derived from the fit.

4.9.1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a statistical method used to estimate the parameters of a probability distribution by maximizing the likelihood function. Given a dataset, and assuming the data points are independent and identically distributed (i.i.d.), the likelihood function is given by:

$$\mathcal{L}(\theta) = \prod_{i=1}^N f(x_i|\theta), \quad (4.25)$$

where:

- $f(x_i|\theta)$ is the probability density function (PDF) parameterized by θ ,
- θ represents the set of parameters to be estimated.

Instead of maximizing $\mathcal{L}(\theta)$ directly, it is common to maximize the log-likelihood function:

$$\ln \mathcal{L}(\theta) = \sum_{i=1}^N \ln f(x_i|\theta). \quad (4.26)$$

This approach is widely used in parameter estimation.

4.9.2 Extended Maximum Likelihood

Extended Maximum Likelihood (EML) is an extension of standard Maximum Likelihood Estimation (MLE) that accounts for the total number of observed events in a dataset. Unlike standard MLE, which only fits the shape of a probability density function (PDF), EML also estimates the normalization (yield) of different event components. In our case, EML is required for two reasons. Firstly, in standard ML fits, the PDF is normalized to unit, meaning it only fits the shape of the distribution. However, in EML, the fit also determines the expected event yields, where $N_{\text{signal}}^{\text{exp}}$ and each $N_{\text{bkg}_i}^{\text{exp}}$ are free parameters in the fit. Secondly, the extended likelihood function accounts for fluctuations in the total observed event count using a Poisson constraint, which is critical for a realistic physics analysis where the total event count fluctuates due to statistical uncertainties.

The extended likelihood function is defined as:

$$L = \frac{e^{-N_{\text{exp}}} N_{\text{exp}}^{N_{\text{obs}}}}{N_{\text{obs}}!} \prod_{i=1}^{N_{\text{obs}}} f(x_i | \theta) \quad (4.27)$$

where:

- $f(x)$ is the probability density function (PDF) of the model (*model_1* in our case),
- N_{obs} is the observed number of events,
- N_{exp} is the expected number of events, given by:

$$N_{\text{exp}} = \int f(x) dx. \quad (4.28)$$

And so, the first term accounts for the Poisson probability of observing N_{obs} events given the expectation value N_{exp} .

4.9.3 Binned Extended Maximum Likelihood (Binned EML)

For binned data, Binned EML adapts the above approach by integrating the PDF over histogram bins and so, accounts for discrete event counts in histogram bins. In this case, the likelihood function is given by:

$$\mathcal{L} = \prod_i \frac{e^{-\mu_i} \mu_i^{n_i}}{n_i!}, \quad (4.29)$$

where:

- n_i is the observed number of events in bin i ,
- μ_i is the expected number of events in bin i , given by contributions from different processes:

$$\mu_i = \sum_j N_j P_j(x_i), \quad (4.30)$$

where N_j is the total expected number of events for process j , and $P_j(x_i)$ is the probability of process j contributing to bin i , computed as:

$$P_j(x_i) = \int_{\text{bin } i} \text{PDF}_j(x), dx. \quad (4.31)$$

To estimate the parameters, we minimize the negative log-likelihood (NLL) function:

$$-\ln \mathcal{L} = \sum_i [\mu_i - n_i \ln \mu_i + \ln n_i!]. \quad (4.32)$$

Substituting μ_i :

$$-\ln \mathcal{L} = \sum_i \left[\left(\sum_j N_j P_j(x_i) \right) - n_i \ln \left(\sum_j N_j P_j(x_i) \right) + \ln n_i! \right]. \quad (4.33)$$

Since the term $\ln n_i!$ is constant for a given dataset, it does not affect the minimization and is often omitted:

$$-\ln \mathcal{L} = \sum_i \left[\left(\sum_j N_j P_j(x_i) \right) - n_i \ln \left(\sum_j N_j P_j(x_i) \right) \right]. \quad (4.34)$$

The main difference from unbinned EML is that instead of evaluating $f(x)$ at each event, we integrate $f(x)$ over each bin.

In practice, minimization is performed on the negative log-likelihood (NLL) function:

$$-\ln \mathcal{L} = \sum_i [N_{exp_i} - n_i \ln \mu_i + \ln n_i!]. \quad (4.35)$$

The term $\ln n_i!$ is constant and does not affect the minimization process. The expected counts μ_i are computed by integrating the PDF over each bin.

Minimization is typically performed using a gradient-based minimization algorithm by MIDGRAD and error estimation based on the curvature of the likelihood function or asymmetric error estimation by HESSE or Minos, respectively. [64]

4.9.4 Systematic uncertainties and fit ranges

Systematic uncertainties in background estimation arise from various sources such as normalization, shape, and statistical uncertainties. These are incorporated into the fitting procedure using nuisance parameters that modify the background prediction within its uncertainty range. Proper treatment of these uncertainties ensures that they are propagated correctly to the final results. The likelihood function is extended to include nuisance parameters θ that model systematic variations:

$$L(\mu, \theta) = \prod_i P(N_i | \mu S_i + B_i(\theta)) \times \pi(\theta), \quad (4.36)$$

where:

- μ is the signal strength parameter.
- S_i and $B_i(\theta)$ are the expected signal and background contributions in bin i .
- $B_i(\theta)$ depends on nuisance parameters θ .
- $\pi(\theta)$ is a prior constraint on θ , typically modeled as a Gaussian or log-normal distribution.

Including systematics typically increases the uncertainty in the extracted signal yield and can degrade search sensitivity.

In our case, we have done a raw assumption of total systematic uncertainty, arising from the various sources, for each background process. This assumption derives from estimations of other researches [48],[49]. These uncertainties are listed in the appendix. We included these systematics as Gaussian constraints to the yield of each process. These Gaussian pdf distributions have mean value $\mu = N_{exp}^{process}$ and $\sigma = \text{uncertainty}(\%) \times N_{exp}^{process}$.

All fit parameters take as initial values their expected yields: $N_{in}^{process} = N_{exp}^{process}$. We define the range of each background process to vary between $\pm 3\sigma$ the expected yield of the corresponding background process.

On the other hand, the parameter N_{signal} takes values in a wide range, so that its best fit value can be unbiasedly derived by the fit. The expected signal yield range should be large enough to include the whole derived posterior distribution, which is different for each "a" mass sample, as it will be discussed later on. The smallest signal yield range was defined for $m_a = 60$ GeV in the resolved regime, 2 lepton channel: $[-0.8N_{\text{exp}}^{\text{signal}}, 2N_{\text{exp}}^{\text{signal}}]$, while the largest signal yield range was defined for $m_a = 15$ GeV in the resolved regime, 0 lepton channel: $[-20N_{\text{exp}}^{\text{signal}}, 30N_{\text{exp}}^{\text{signal}}]$.

Although a negative signal yield is nonphysical, it ensures the likelihood function behaves correctly, by allowing it to fluctuate symmetrically around zero and providing unbiased statistical inferences. Negative signal yield values account for downward fluctuations in data, which improves statistical robustness. When computing upper limits, negative best-fit values help maintain proper coverage.

4.9.5 Fit results

The following fit results are extracted by fitting model_1, (including its systematic uncertainties) to the generated Asimov dataset of model_0. Asimov datasets are built as binned datasets, in which the event count in each bin is set to the expected event yield for the chosen model parameters. [65]

In the upper subplots of Figures 4.44 and 4.45, the black dots represent the Asimov data points, and the blue line corresponds to the post-fit distribution. The individual distributions of background and signal yields (for mass $m_a = 60$ GeV in the resolved regime and $m_a = 15$ GeV in the merged regime) are shown with colorful lines. In the bottom subplots, the pulls in each bin are also shown, defined as:

$$\text{pull} = \frac{N_{\text{fit}} - N_{\text{in}}}{\delta N_{\text{fit}}}. \quad (4.37)$$

When pulls have a value close to zero and error close to one, it indicates a good fit quality.

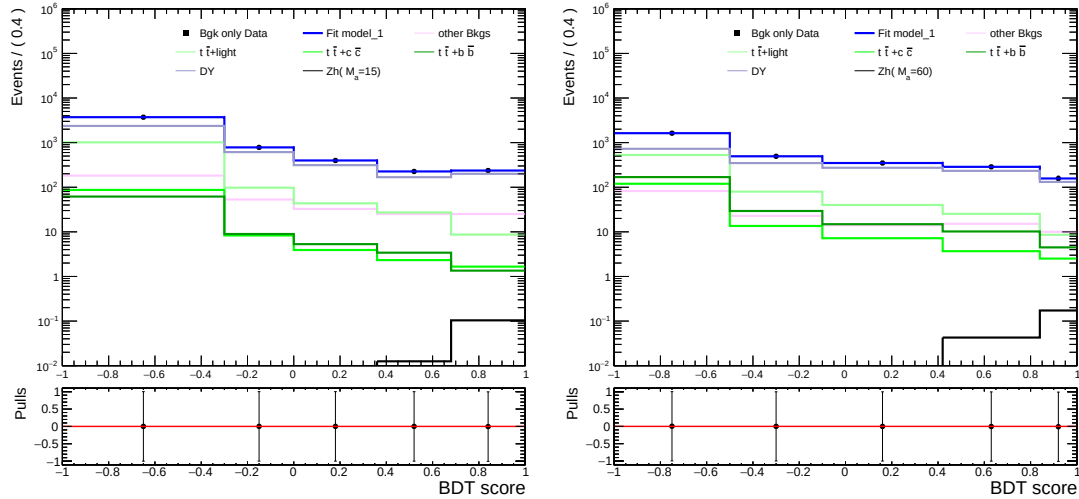


Figure 4.44: Extended binned ML fit to the BDT score distribution using background Asimov dataset, in the merged regime (left) and in the resolved regime (right) of the 2-lepton channel.

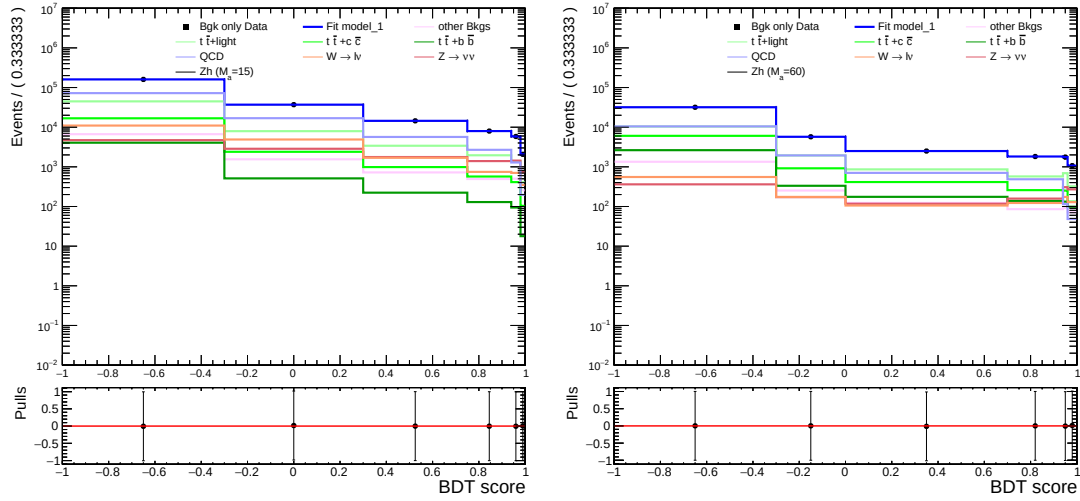


Figure 4.45: Extended binned ML fit to the BDT score distribution using background Asimov dataset, in the merged regime (left) and in the resolved regime (right) of the 0-lepton channel. In this case, since the fit converges to negative signal yield value and the plot is logarithmic, signal distribution is not shown.

The Figures 4.60 and 4.60 depict the projection of the negative log-likelihood onto the POI (the expected signal yield). The projection of NLL onto the POI helps to identify the best-fit value of the POI, since it is the value where NLL reaches its minimum. We see that in any case, the best-fit value of the POI is very close to zero, taking very small positive or negative values. Useful insights can be derived from the profiled NLL. If it is flat over a large POI range, the parameter is poorly constrained, or if the NLL has multiple minima, the fit might be unstable, which does not happen in our case. Also, if the profiled NLL is symmetric around the minimum, uncertainties are approximately Gaussian.

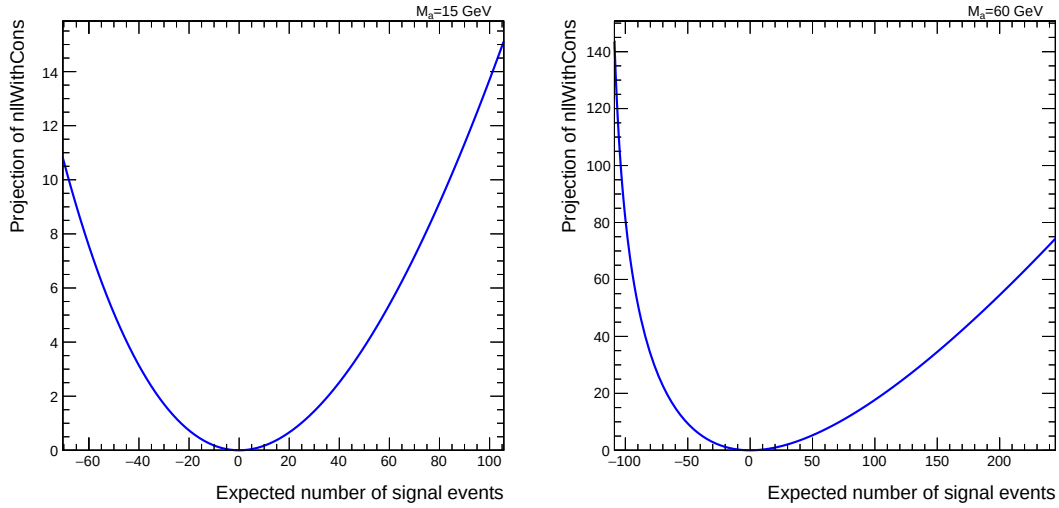


Figure 4.46: The negative log likelihood (NLL) vs the parameter of interest (POI). The NLL is minimized for signal value 0.46 in the merged regime (left) and 0.22 in the resolved regime (right) of the 2-lepton channel.

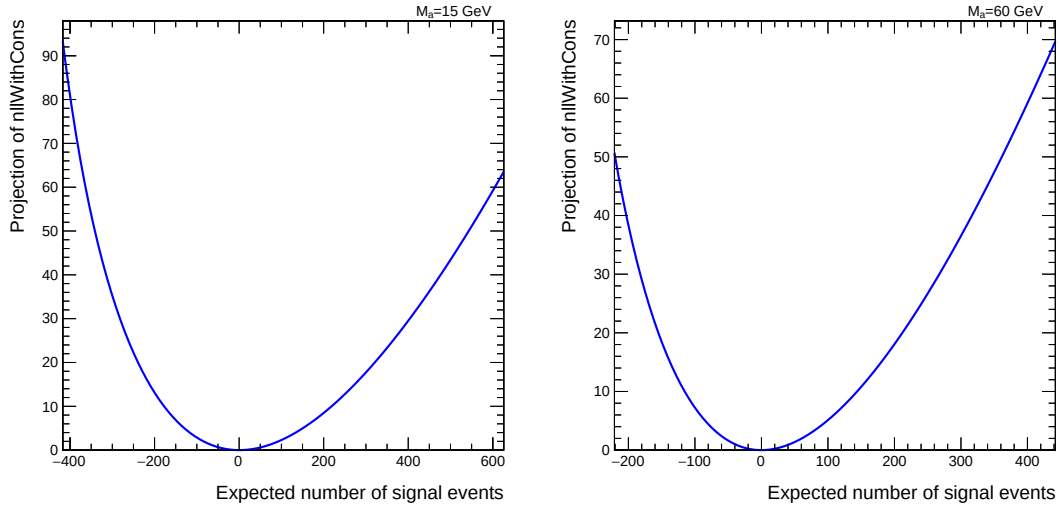


Figure 4.47: The negative log likelihood (NLL) vs the parameter of interest (POI). The NLL is minimized for signal value -0.32 in the merged regime (left) and -0.26 in the resolved regime (right) of the 0-lepton channel.

The ML fit distributions to a random binned toy dataset are also shown in Figures , as well as the profile log-likelihoods to the expected signal yield in Figures . This time, as the generated MC toy dataset is free to have Poissonian variations around the expected background yield in each bin, it allows the fit to converge to higher signal yields. The pulls in each bin vary around zero but are still small enough, and fit quality is still acceptable.

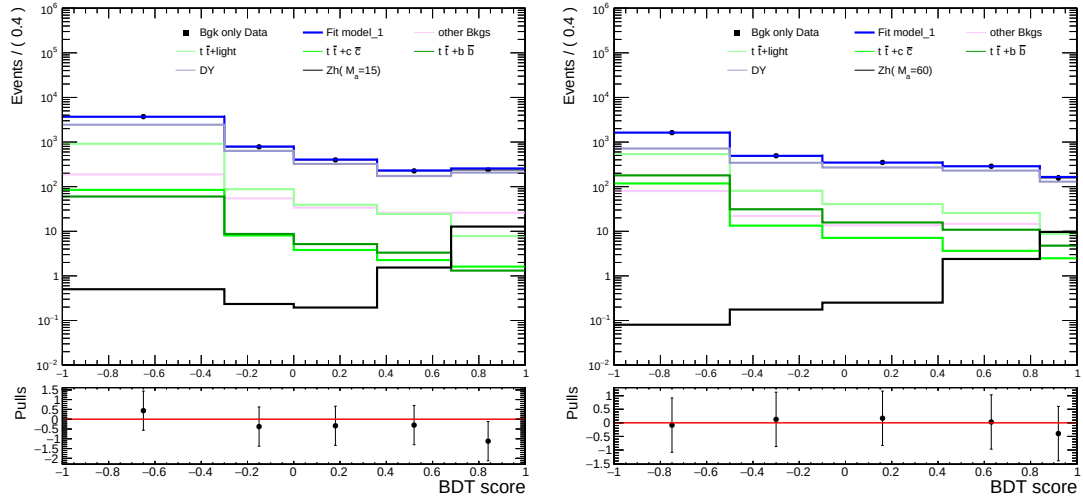


Figure 4.48: Extended binned ML fit to the BDT score distribution using a generated binned toy dataset, in the merged regime (left) and in the resolved regime (right) of the 2-lepton channel.

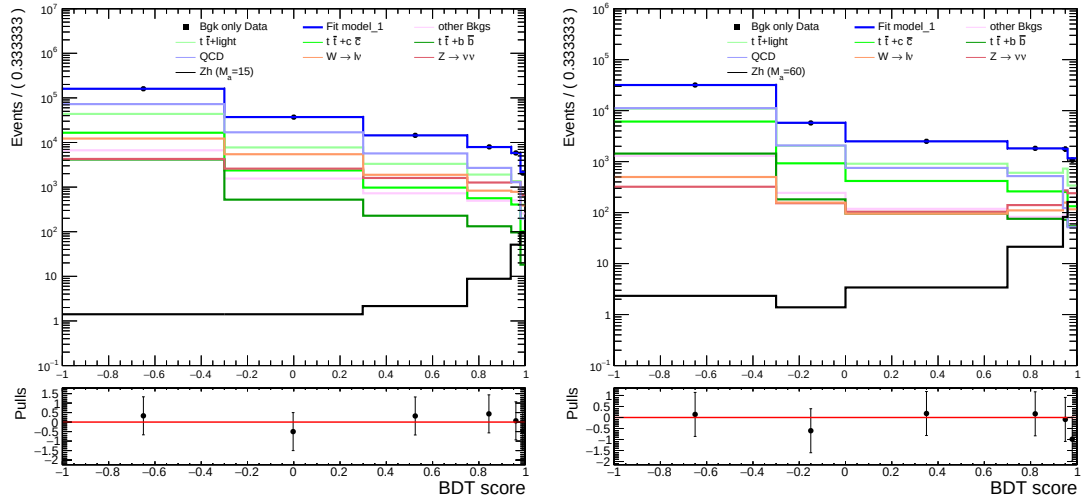


Figure 4.49: Extended binned ML fit to the BDT score distribution using background Asimov dataset, in the merged regime (left) and in the resolved regime (right) of the 0-lepton channel. Since the fit converges to negative signal yield value and the plot is logarithmic, signal distribution is not shown in this case.

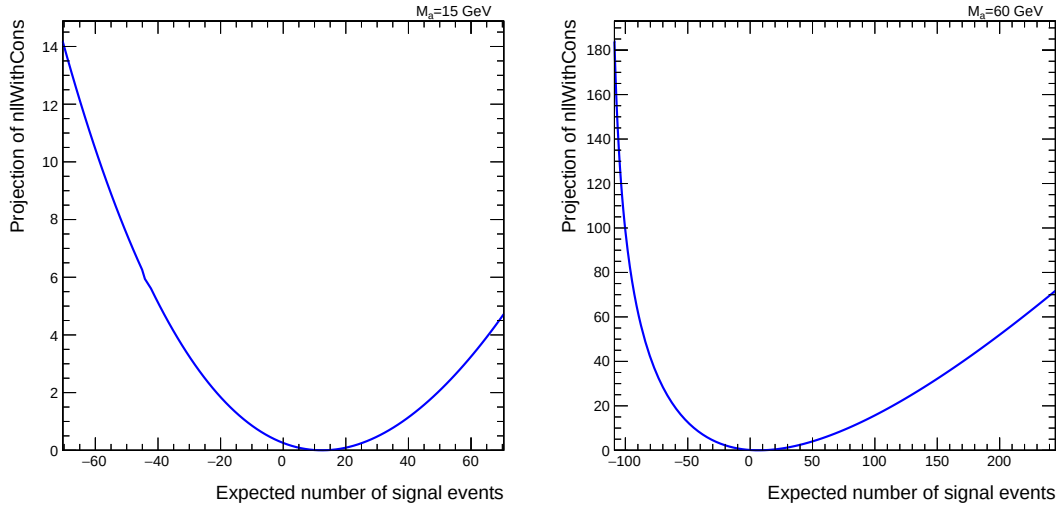


Figure 4.50: The negative log likelihood (NLL) vs the parameter of interest (POI). The NLL is minimized for signal value 6.7 in the merged regime (left) and 11.1 in the resolved regime (right) of the 2-lepton channel.

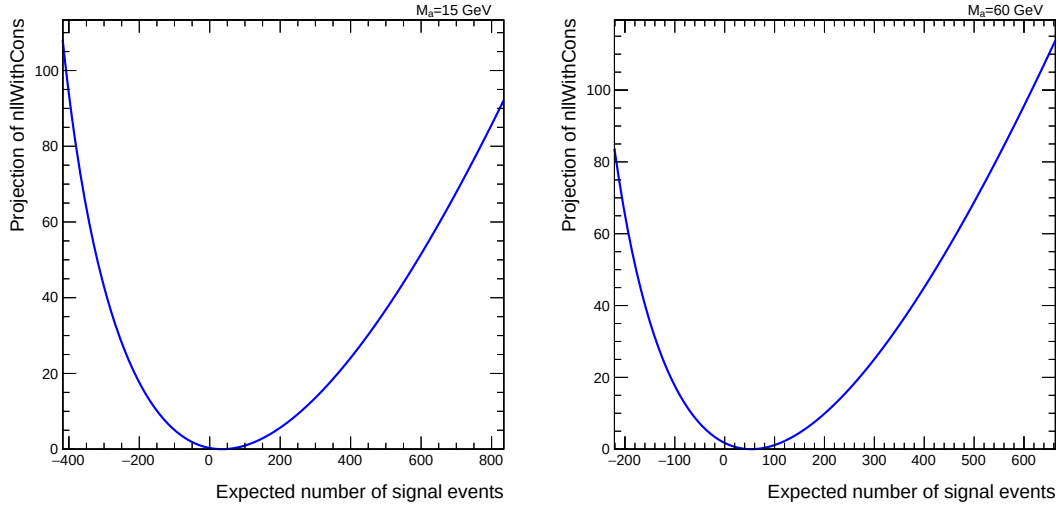


Figure 4.51: The negative log likelihood (NLL) vs the parameter of interest (POI). The NLL is minimized for signal value 37.2 in the merged regime (left) and 52.1 in the resolved regime (right) of the 0-lepton channel.

The following correlation matrices depict correlations between all the nuisance parameters and the POI as well. This information is useful, as correlations between nuisance parameters can significantly impact the fit procedure in RooFit, influencing the stability, uncertainty estimation, and behavior of likelihood-based tests. When nuisance parameters are correlated, their uncertainties propagate to the parameter of interest (POI). If the POI and nuisance parameters are highly correlated, the likelihood surface may become flat or ill-conditioned in certain directions, which is not the case as we see in Figures 4.47 and 4.46. Very strong correlations are considered when

$|\rho| > 0.8$, which in our case is only observed between $N_{\text{exp}}^{\text{WJets}}$ and $N_{\text{exp}}^{\text{QCD}}$ in the merged regime of the 0-lepton channel. Also, mid to high correlations are observed between $N_{\text{exp}}^{\text{QCD}}$ and $N_{\text{exp}}^{\text{ZJets}}$ and between $N_{\text{exp}}^{\text{tt+bb}}$ and $N_{\text{exp}}^{\text{tt+cc}}$ or $N_{\text{exp}}^{\text{tt+bb}}$ in some cases. Strong correlations between $N_{\text{exp}}^{\text{signal}}$ and nuisance parameters are not observed at all. However, mid correlations ($|\rho| > 0.3$) between $N_{\text{exp}}^{\text{signal}}$ and $N_{\text{exp}}^{\text{DY}}$ or $N_{\text{exp}}^{\text{QCD}}$ exist in some cases.

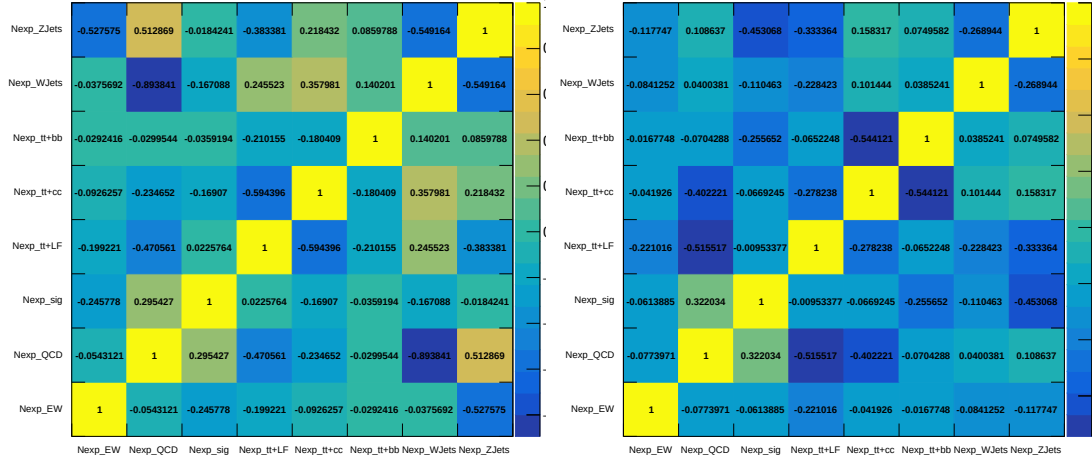


Figure 4.52: Correlation matrix of the fit variables in the merged regime (top) and in the resolved regime (bottom) of the 0-lepton channel.

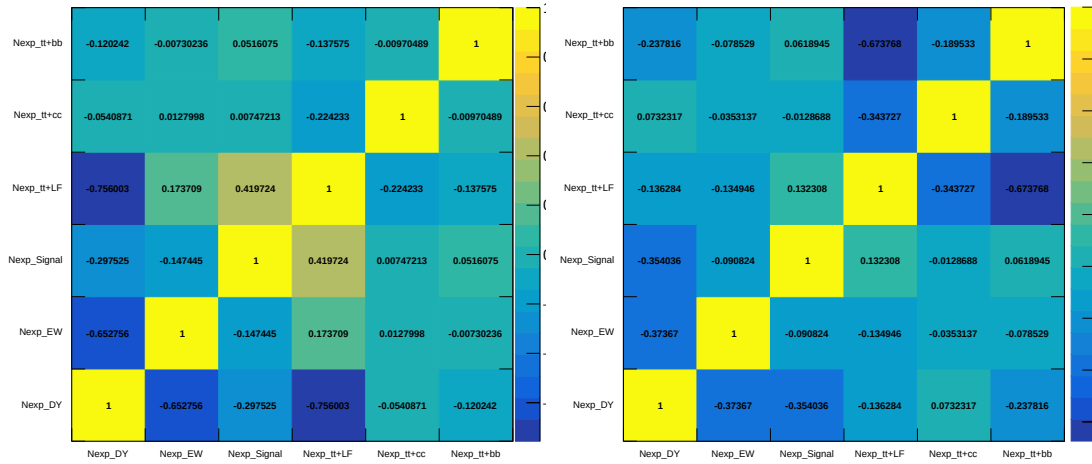


Figure 4.53: Correlation matrix of the fit variables in the merged regime (top) and in the resolved regime (bottom) of the 2-lepton channel.

In the following tables, post-fit results are shown in each regime of each channel. Results are presented in each bin of the BDT.

We observe that, in the 0-lepton channel, $Z \rightarrow vv$ and $t\bar{t} + \text{LF}$ are the most dominant backgrounds in the highest significance bins, while QCD contributes the least to the background of the highest significance bins. In the 2-lepton channel, DY is the most dominant background in the highest significance bins. Error propagation is done through the covariance matrix based on the following formula:

Let $C_i(b)$ be the bin content of process i in bin b and cov_{ij} the covariance between $N_{\text{exp},i}$ and $N_{\text{exp},j}$.

The normalized bin content for process i is:

$$\tilde{C}_i(b) = \frac{C_i(b)}{N_{\text{exp},i}}$$

The bin-wise uncertainty for process i is:

$$\sigma_i^2(b) = \sum_j \text{cov}_{ij} \cdot \tilde{C}_i(b) \cdot \tilde{C}_j(b)$$

Consequently, the total bin-wise uncertainty is:

$$\sigma_{\text{total}}^2(b) = \sum_i \sigma_i^2(b)$$

And so, the total background error for each bin is set as:

$$\Delta B(b) = \sqrt{\sigma_{\text{total}}^2(b)}$$

Process	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
$t\bar{t} + \text{light}$	1792.32 ± 85.60	68.42 ± 11.13	36.24 ± 6.81	20.40 ± 2.68	6.54 ± 1.29
$t\bar{t} + c\bar{c}$	153.17 ± 12.26	5.77 ± 1.70	3.50 ± 0.91	1.84 ± 0.41	1.18 ± 0.27
$t\bar{t} + b\bar{b}$	109.06 ± 9.09	6.45 ± 2.04	4.61 ± 0.61	2.74 ± 0.80	0.88 ± 0.04
$Z \rightarrow \ell\ell$	4165.00 ± 30.90	439.83 ± 19.77	281.16 ± 12.33	135.32 ± 2.49	152.85 ± 4.05
Other Bkgs	318.28 ± 18.42	39.61 ± 4.03	29.29 ± 3.98	19.97 ± 5.41	19.97 ± 2.30
Total Bkg	6537.93 ± 79.18	560.17 ± 14.82	354.86 ± 10.39	180.34 ± 5.14	181.83 ± 13.34
DATA	6541.00	578.00	329.00	184.00	183.00
signal(15)	2.45 ± 1.09	0.49 ± 0.49	0.49 ± 0.49	3.42 ± 1.29	28.37 ± 3.72
signal(30)	3.91 ± 1.38	2.93 ± 1.20	2.45 ± 1.09	2.93 ± 1.20	18.59 ± 3.01

Table 4.21: post-fit results in the merged regime of the 2-lepton channel.

Process	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
$t\bar{t} + light$	661.74 ± 37.94	79.81 ± 5.81	52.08 ± 5.30	26.63 ± 3.07	3.44 ± 0.43
$t\bar{t} + c\bar{c}$	149.09 ± 15.38	13.51 ± 2.98	9.34 ± 1.02	3.86 ± 1.03	1.00 ± 0.20
$t\bar{t} + b\bar{b}$	210.52 ± 26.93	29.30 ± 2.08	19.26 ± 4.69	10.69 ± 2.69	1.79 ± 0.43
$Z \rightarrow \ell\ell$	921.30 ± 23.12	349.62 ± 14.03	350.78 ± 15.99	237.35 ± 9.53	51.44 ± 1.79
other Bkgs	102.87 ± 7.48	22.80 ± 0.81	18.26 ± 3.14	15.90 ± 2.70	4.01 ± 0.47
Total Bkg	2045.51 ± 43.85	495.03 ± 12.15	449.69 ± 13.83	294.27 ± 9.24	61.45 ± 7.61
DATA	2057.00	510.00	429.00	292.00	57.00
signal(30)	1.96 ± 0.98	4.40 ± 1.47	9.78 ± 2.19	36.68 ± 4.24	51.84 ± 5.04
signal(60)	1.96 ± 0.98	3.42 ± 1.29	6.36 ± 1.76	48.91 ± 4.89	75.32 ± 6.07

Table 4.22: post-fit results in the resolved regime of the 2-lepton channel.

Process	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
$t\bar{t} + light$	94037.86 ± 62.99	14289.61 ± 24.22	4594.51 ± 13.61	1115.36 ± 6.67	167.73 ± 2.58	24.88 ± 0.99
$t\bar{t} + c\bar{c}$	35090.65 ± 39.02	4298.17 ± 13.57	1327.85 ± 7.54	324.05 ± 3.72	49.29 ± 1.45	6.21 ± 0.51
$t\bar{t} + b\bar{b}$	8518.89 ± 19.04	919.48 ± 6.18	300.86 ± 3.51	73.63 ± 1.72	11.29 ± 0.67	1.07 ± 0.20
$Z \rightarrow \nu\nu$	9914.42 ± 35.61	5149.08 ± 31.87	2388.50 ± 22.79	797.30 ± 13.71	172.61 ± 6.39	44.24 ± 3.30
$W \rightarrow \ell\nu$	23150.24 ± 103.74	8900.99 ± 86.75	2315.69 ± 46.60	457.09 ± 35.21	84.28 ± 6.47	20.82 ± 3.18
QCD	151998.19 ± 1100.32	30399.15 ± 544.81	7667.47 ± 269.33	1533.78 ± 119.23	154.65 ± 34.01	11.83 ± 8.24
other Bkgs	13892.58 ± 52.70	2768.87 ± 22.46	969.57 ± 12.87	278.94 ± 7.35	59.82 ± 3.57	15.23 ± 1.85
Total Bkg	336602.81 ± 1109.67	66725.34 ± 553.79	19564.45 ± 275.04	4580.14 ± 125.54	699.66 ± 35.52	124.28 ± 9.67
Data	336535.00	66647.00	19678.00	4597.00	707.00	131.00
signal(M15)	24.45 ± 4.00	21.10 ± 3.60	23.70 ± 3.93	41.62 ± 4.98	50.30 ± 5.49	47.44 ± 5.22
signal(M30)	23.04 ± 3.97	31.74 ± 4.38	34.60 ± 4.69	64.29 ± 6.28	43.63 ± 5.12	31.82 ± 4.22

Table 4.23: post-fit results in the merged regime of the 0-lepton channel.

Process	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
$t\bar{t} + light$	21743.02 ± 30.17	1737.21 ± 8.44	1805.52 ± 8.58	411.00 ± 4.06	41.38 ± 1.28	38.55 ± 1.23
$t\bar{t} + c\bar{c}$	12722.35 ± 23.48	827.18 ± 5.94	869.33 ± 6.09	186.25 ± 2.80	15.10 ± 0.80	15.85 ± 0.82
$t\bar{t} + b\bar{b}$	5520.91 ± 15.29	299.42 ± 3.51	367.50 ± 3.88	99.53 ± 2.01	7.95 ± 0.57	11.83 ± 0.69
$Z \rightarrow \nu\nu$	761.30 ± 10.22	155.06 ± 5.57	246.79 ± 7.26	114.06 ± 5.22	18.47 ± 2.29	32.48 ± 3.40
$W \rightarrow \ell\nu$	1232.71 ± 49.65	156.99 ± 8.48	223.60 ± 10.11	88.28 ± 6.16	10.46 ± 2.90	15.48 ± 2.79
QCD	22060.64 ± 414.24	1754.96 ± 128.57	1484.55 ± 100.76	351.48 ± 58.16	6.92 ± 4.89	5.83 ± 5.83
other Bkgs	2825.14 ± 22.61	227.95 ± 6.38	258.81 ± 6.87	61.73 ± 3.45	5.51 ± 1.07	8.55 ± 1.34
Total Bkg	66866.08 ± 419.97	5158.77 ± 129.59	5256.11 ± 102.37	1312.32 ± 59.06	105.78 ± 6.43	128.58 ± 7.60
Data	66880.00	5141.00	5175.00	1398.00	103.00	130.00
signal(M30)	21.06 ± 3.54	8.24 ± 2.34	43.28 ± 5.22	38.86 ± 4.78	21.72 ± 3.50	59.08 ± 5.67
signal(M60)	20.39 ± 3.58	5.21 ± 1.78	29.79 ± 4.27	64.22 ± 6.06	20.52 ± 3.45	80.90 ± 6.73

Table 4.24: post-fit results in the resolved regime of the 0-lepton channel.

4.9.6 Goodness of Fit

The above fit results prove fit quality for fitting to one binned dataset, but it is important to know that the fit quality is good for all possible binned toy datasets. For this, as a first step, we can generate and run the fit procedure for many toy datasets and check if all these fits have converged (status of RooFit minimization and covariance matrix quality). After that, there exist different possible fit diagnostics for binned ML fit.

The Saturated Model GOF: Computes a goodness-of-fit measure for binned fits based on the saturated model, as prescribed by the Statistics Committee (note). This quantity is similar to a chi-square, but can be computed for an arbitrary combination of binned channels with arbitrary constraints. A saturated model is a model where each data bin (in a binned fit) has an independent parameter to match the observed event count exactly. The likelihood of the saturated model is the best possible likelihood that any model can achieve. By comparing the likelihood of our model (L_{fit}) to the likelihood of the saturated model (L_{sat}), we can evaluate how much information is lost by using our chosen model.[66]

The test statistic is:

$$\lambda = -2 \log \left(\frac{L_{\text{fit}}}{L_{\text{sat}}} \right) = -2 (\log L_{\text{fit}} - \log L_{\text{sat}})$$

where:

- L_{fit} is the likelihood of the fitted model. - L_{sat} is the likelihood of the saturated model.

If the p-value of λ is low ($p < 0.05$), the fit is likely not a good description of the data. After running the test for 1000 binned toy datasets, we obtain the following distributions.

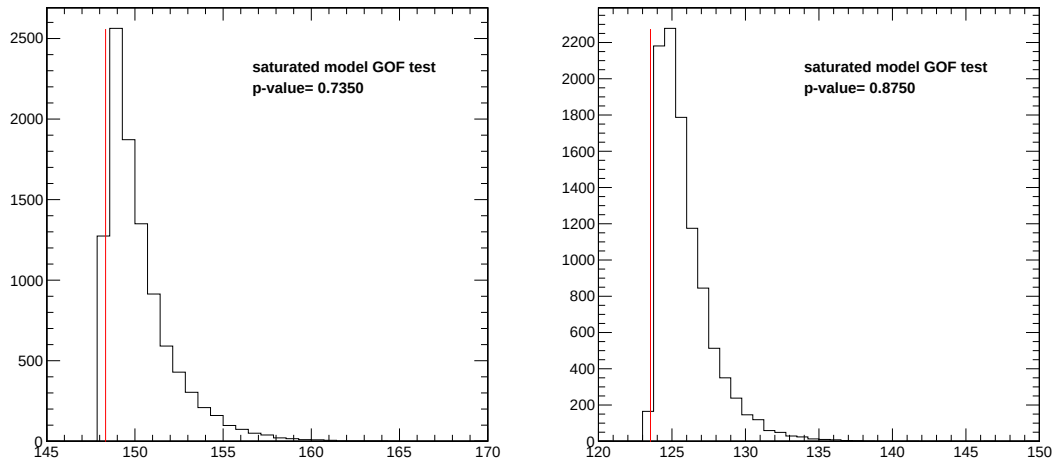


Figure 4.54: Saturated model distribution derived by applying saturated model GOF to 1000 toy datasets, for $m_a = 15$ GeV in the merged regime (left) and for $m_a = 60$ GeV in the resolved regime (right) of the 0-lepton channel.

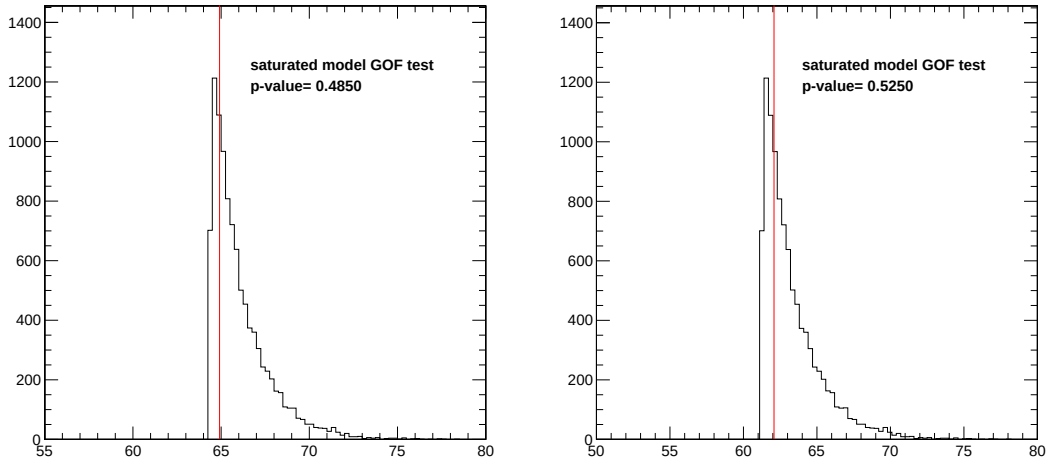


Figure 4.55: Saturated model distribution derived by applying saturated model GOF to 1000 toy datasets, for $m_a = 15$ GeV in the merged regime (left) and for $m_a = 60$ GeV in the resolved regime (right) of the 2-lepton channel.

Kolmogorov-Smirnov test: Compute a goodness-of-fit measure for binned fits using the Kolmogorov-Smirnov test. It is based on the largest difference between the cumulative distribution function and the empirical distribution function of any bin.[66]It measures the absolute difference:

$$D = \max |F_{\text{model}}(x) - F_{\text{data}}(x)|$$

If D is small, the model is a good fit.

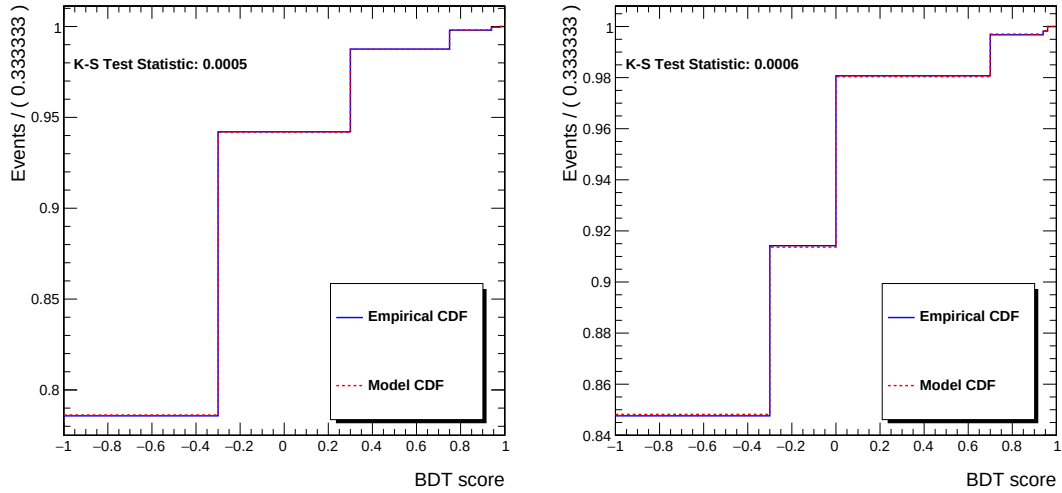


Figure 4.56: cumulative distribution function and the empirical distribution and the resulting K-S test value for $m_a = 15$ GeV in the merged regime (left) and for $m_a = 60$ GeV in the resolved regime (right) of the 0-lepton channel.

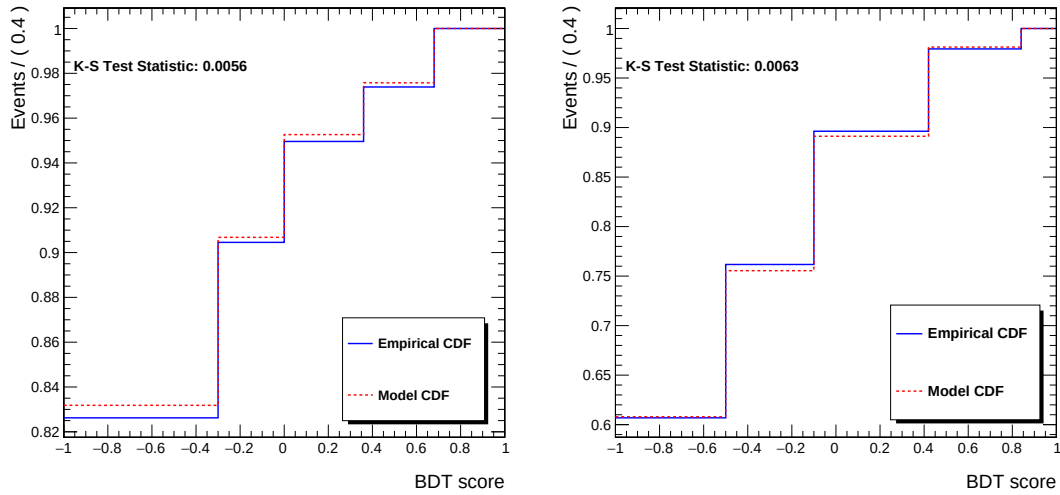


Figure 4.57: cumulative distribution function and the empirical distribution and the resulting K-S test value for $m_a = 15$ GeV in the merged regime (left) and for $m_a = 60$ GeV in the resolved regime (right) of the 2-lepton channel.

Also, after generating and fitting many toy datasets, we can plot the resulting NLL values. The distribution should have a global maximum and be symmetric with gaussian shape.

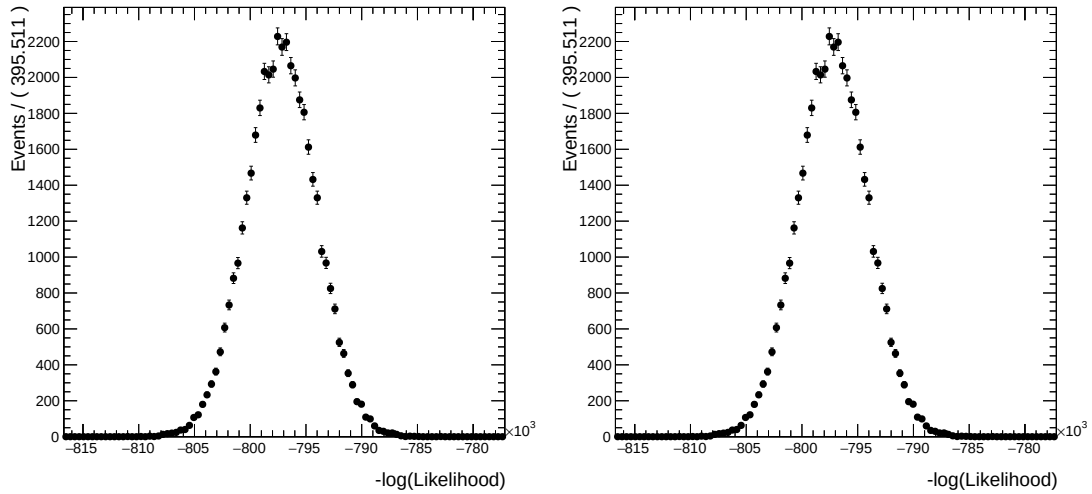


Figure 4.58: NLL distribution after generating and fitting 10000 binned toy datasets for $m_a = 15$ GeV in merged regime (left) and for $m_a = 60$ GeV in resolved regime (right) in 2-lepton channel.

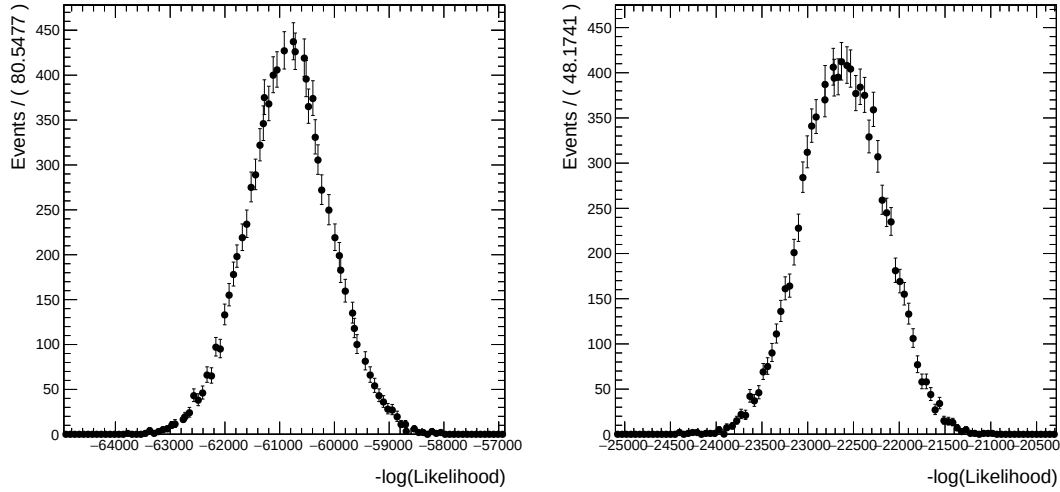


Figure 4.59: NLL distribution after generating and fitting 10000 binned toy datasets for $m_a = 15$ GeV in merged regime (left) and for $m_a = 60$ GeV in resolved regime (right) in 2-lepton channel.

4.10 Upper limits to the signal yield using Bayesian method

4.10.1 Introduction to Bayesian Statistics

Bayesian statistics provides a framework for incorporating prior knowledge into probability estimation. Unlike frequentist inference, which relies solely on observed data, Bayesian inference updates a prior belief using new evidence to form a posterior probability distribution. [67]

Given a model parameter θ and observed data D , Bayes' theorem states:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}. \quad (4.38)$$

Here:

- $P(\theta|D)$ is the posterior probability distribution,
- $P(D|\theta)$ is the likelihood function,
- $P(\theta)$ is the prior probability distribution,
- $P(D)$ is the evidence, serving as a normalization constant.

Bayesian and frequentist methods differ mainly in interpretation: Bayesian inference treats parameters as random variables with probability distributions, while frequentist inference estimates fixed parameters from data. Frequentist confidence intervals rely on the concept of long-run frequency, whereas Bayesian credible intervals directly provide a probability statement about the parameter given the data.[68]

4.10.2 Credible Intervals

A fundamental concept in Bayesian analysis is the credible interval (CI), which quantifies uncertainty in parameter estimation. Credible intervals are fairly straightforward to compute.[69] As the Bayesian inference returns a distribution of possible POI values (the posterior), the credible interval is just the range containing a particular percentage of probable values. For instance, the 95% credible interval is simply the portion of the posterior distribution that contains 95% of the values. There exist different types of CIs. Since we want to compute an upper limit for our POI, we choose an "upper bound type" interval, calculated from the CDF of the posterior distribution by integration from the lower possible posterior distribution value up to value corresponding to 95% of the posterior integral. So the $(1 - \alpha)\%$ credible interval for a parameter θ satisfies:

$$\int_{\theta_{\text{low}}}^{\theta_{\text{high}}} P(\theta|D)d\theta = 1 - \alpha. \quad (4.39)$$

This interval contains the true parameter value with probability $1 - \alpha$, given the prior and observed data. Unlike frequentist confidence intervals, Bayesian credible intervals do not rely on hypothetical repeated experiments but instead reflect direct probability statements about the parameter. The choice of prior $P(\theta)$ significantly influences the posterior distribution and, consequently, the extracted credible intervals. In our case, the POI values obtained from the fit results are our prior.[67],[68]

In many cases, analytical computation of the posterior distribution is intractable, necessitating numerical methods such as Markov Chain Monte Carlo (MCMC) to estimate credible intervals. [69], [70]

Markov Chain Monte Carlo (MCMC) and the Metropolis-Hastings Algorithm

The MCMC method is a powerful numerical technique used to sample from complex posterior distributions. Markov Chain Monte Carlo is a class of algorithms designed to sample from probability distributions that are difficult to sample from directly, particularly high-dimensional distributions. MCMC algorithms construct a Markov chain, a sequence of random samples where each sample depends only on the previous one (the "Markov property"). The chain is designed so that its stationary distribution is the target distribution we want to sample from. MCMC algorithms use a proposal distribution to generate candidate samples. The proposal distribution determines how the algorithm explores the sample space. A key step in MCMC is the acceptance or rejection of proposed samples. This step ensures that the samples converge to the target distribution. The acceptance probability is calculated based on the target distribution and the proposal distribution.[71],[72] Instead of evaluating the posterior directly, MCMC generates a sequence of correlated samples that approximate the true posterior distribution [67].

One widely used MCMC technique is the Metropolis-Hastings algorithm, which proceeds as follows:

1. Initialize with an arbitrary starting value θ_0 .
2. Propose a new candidate value θ^* from a proposal distribution $q(\theta^*|\theta_n)$.
3. Compute the acceptance ratio:

$$r = \frac{P(\theta^*|D)q(\theta_n|\theta^*)}{P(\theta_n|D)q(\theta^*|\theta_n)}. \quad (4.40)$$

4. Accept the new value with probability $\min(1, r)$. If accepted, set $\theta_{n+1} = \theta^*$; otherwise, retain the current value $\theta_{n+1} = \theta_n$.
5. Repeat for a large number of iterations to construct the Markov chain.

The generated samples approximate the posterior distribution, allowing computation of Bayesian credible intervals and extraction of upper limits. Some practical considerations for MCMC algorithm need to be taken into account are:

- **Burn-in steps:** The initial samples of the Markov chain may not accurately represent the target distribution. Therefore, it's common to discard the first part of the chain (the "burn-in" period) before using the samples for estimation. Using a number of burn-in steps is likely to result in a bias towards earlier stages of the chain before a reasonable convergence.
- **Proposal Distribution Tuning:** The choice of proposal distribution can significantly impact the efficiency of the MCMC algorithm. Ideally, the proposal distribution should be tuned to achieve a reasonable acceptance rate and to explore the sample space effectively.
- **Convergence Diagnostics:** It's important to assess whether the Markov chain has converged to the stationary distribution. Various diagnostic tools can be used, such as examining trace plots, and autocorrelation functions. Some trace plots representing the correlations between POI and the nuisance parameters during MC sampling are shown in the appendix.

In our case, since we have run the fit, we use the resulting correlation matrix to construct a proposal that includes knowledge of our sample space. This leads to a proposal acceptance rate of about 30% in the 0-lepton channel and 35% in the 2-lepton channel. The number of iterations is 3000, with 3000 burn-in steps. The figures 4.60 and 4.61 depict the posterior distributions of the expected signal yield, for $m_a = 60$ GeV in the resolved regime and for $m_a = 15$ GeV in the merged regime of the 0-lepton and of the 2-lepton channel respectively. The posterior distributions of expected yields for all possible "a" masses are shown in the appendix. The red line depicts the signal yield corresponding to the 95% credible interval, from which we will extract the 95% CL upper limit values on the signal strength, as it is described in the next section. It is notable that in the resolved regime, posterior distributions become wider for small "a" masses, and as "a" mass increases, they become narrower. The opposite happens in the merged regime; as "a" mass increases, posteriors become wider. As it will be shown later, this tension observed in the posteriors is also depicted on the upper limits.

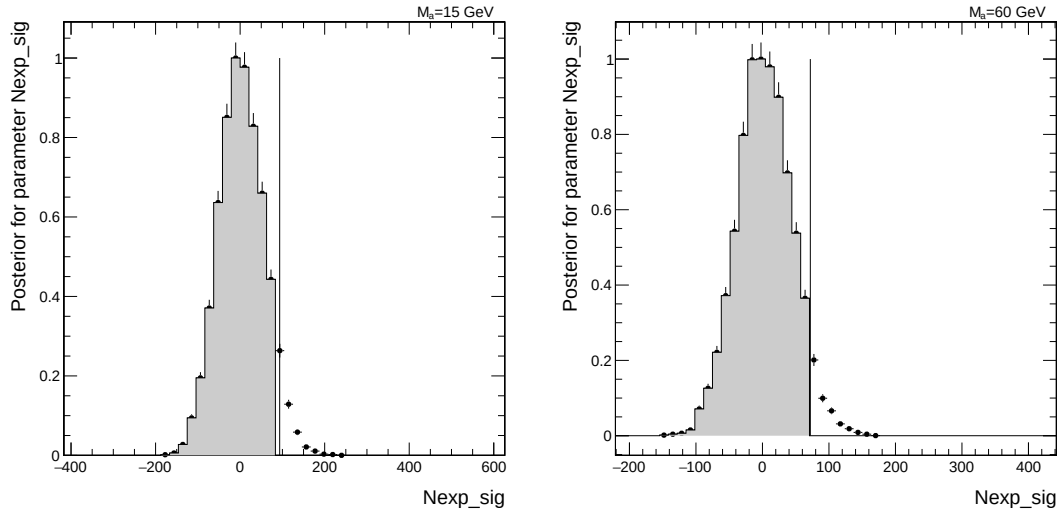


Figure 4.60: Posterior distribution of the expected signal yield for $m_a = 15 \text{ GeV}$ in the merged regime (left) and for $m_a = 60 \text{ GeV}$ in the resolved regime (right) of the 0-lepton channel.

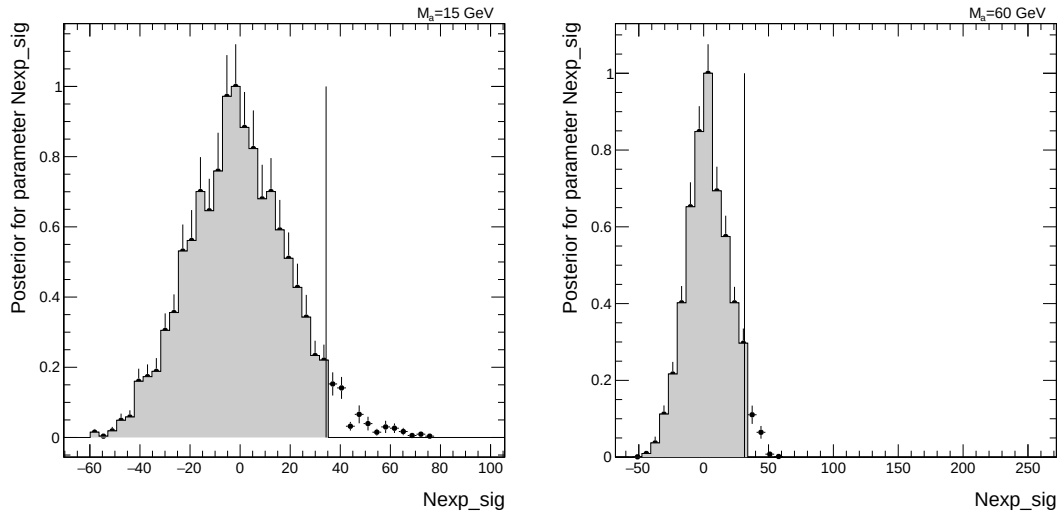


Figure 4.61: Posterior distribution for the expected signal yield for $m_a = 15 \text{ GeV}$ in the merged regime (left) and for $m_a = 60 \text{ GeV}$ in the resolved regime (right) of the 2-lepton channel.

4.10.3 Upper limits on the signal yield using Bayesian methods

To extract upper limits of signal yield, we integrate the posterior probability distribution up to a desired credibility level. Given an upper limit θ_{lim} at confidence level C , we solve:

$$\int_{\theta_{low}}^{\theta_{lim}} P(\theta|D)d\theta = C. \quad (4.41)$$

A common choice is $C = 0.95$, yielding a 95% credible upper limit on the signal yield. Then, dividing the obtained value by the theoretical N_{exp} of each "a" mass sample, we obtain the corre-

sponding upper limits of signal strength [70] .

Metropolis-Hastings algorithm is particularly useful when extracting upper limits in cases with low statistics or non-Gaussian likelihood distributions. By generating samples from the posterior distribution, this method enables robust estimation of credible intervals and upper limits on signal yields. Bayesian upper limits are estimated using RooStats MCMC calculator. The process involves:

1. Defining a prior distribution $\pi(\theta)$.
2. Computing the likelihood function $p(x|\theta)$.
3. Sampling from the posterior distribution using MCMC.
4. Determining θ_{up} such that the cumulative posterior probability reaches $1 - \alpha$.

Chapter 5

Results: Model independent Upper Limits

We finally derive the signal strength upper limits extraction for all possible "a" masses in each regime and channel. Expected upper limit means that instead of data, sum of MC background datasets is used to derive the upper limit. In other words, expected upper limit is the upper limit assuming no signal and only SM background. This expected upper limit is compared with the observed upper limit derived using actual data taken at CMS. However, as we have not used real data yet, only the expected upper limit is meaningful.

The expected limit is computed by generating many toy MC binned datasets (1000 MC toys) and computing the upper limit for each of them. We accept upper limit values only for MC toys where the fit procedure successfully converged, which happened in almost all cases. The median expected upper limit, as well as the 68% and 95% quantiles are computed and shown in the distributions of the limits. To compute them, we sorted the upper limit values for all MC toys in ascending order. For N toys, median is the value of $N^{th}/2$ toy, lower and upper borders of 1σ band are the values of $0.16N^{th}$ and $0.84N^{th}$ toy respectively, and lower and upper borders of 2σ band are the values of $0.025N^{th}$ and $0.975N^{th}$ respectively. The observed limits are indicative since observed data are MC pseudo-data, as discussed above. The results are presented in Figures 5.1 and 5.2. The median expected 95% CL upper limit values are shown in Tables 5.2 and 5.1 for all possible "a" masses in each regime and channel.

In the resolved regime, we observe that upper limits of the branching ratio generally decrease as the mass of the "a" boson increases. For lower M_a points ($M_a < 20$ GeV) significance is limited by signal acceptance, since (bb) pairs tend to be collimated. In the merged regime, sensitivity is recovered for M_a between 15 and 25 GeV especially in the 0-lepton channel, leading to a significant improvement in upper limits of the branching ratio. In the 2-lepton channel, sensitivity is recovered basically around $M_a = 15$ GeV, while for higher "a" masses, lower branching ratios are obtained in the resolved regime.

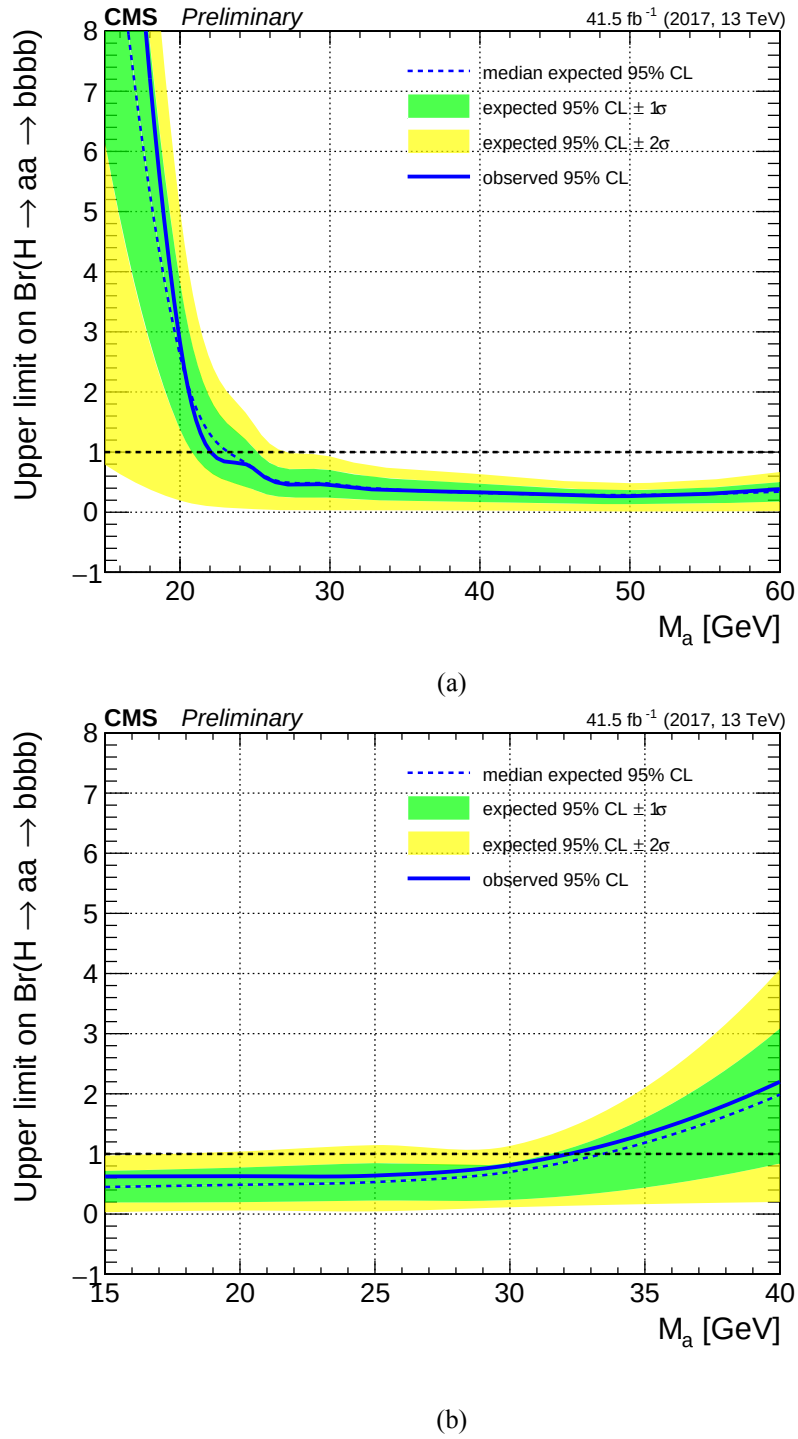


Figure 5.1: Model independent 95% CL upper limits on $\text{Br}(H \rightarrow aa \rightarrow bbbb)$ (a) in the resolved and (b) in the merged regime, 0-lepton channel. (linear scale)

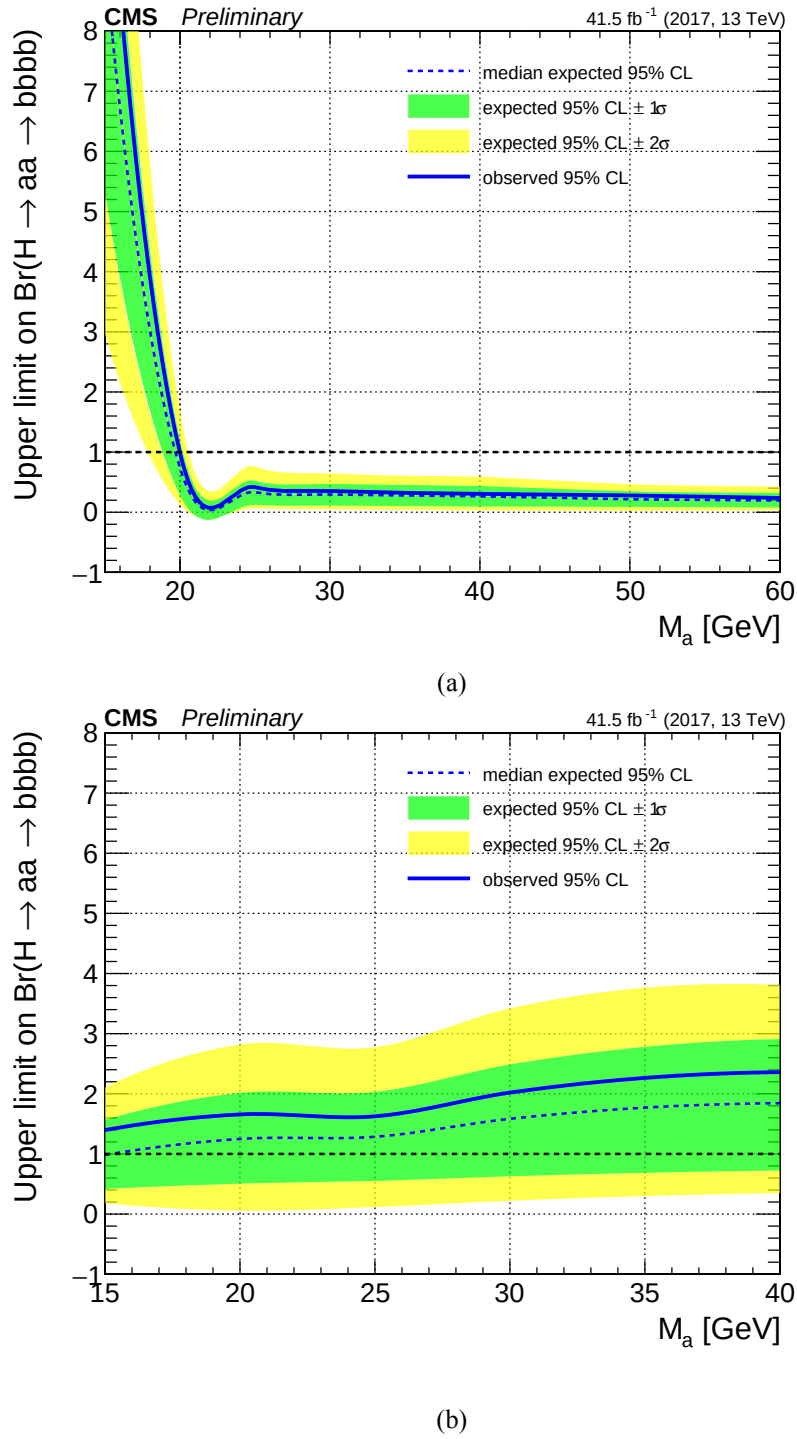


Figure 5.2: Model independent 95% CL upper limits on $\text{Br}(H \rightarrow aa \rightarrow bbbb)$ (a) in the resolved and (b) in the merged regime, 2-lepton channel. (linear scale)

m_a point	95% UL of Expected Signal Strength	
	Resolved Regime	Merged Regime
15 GeV	9.19	0.98
20 GeV	0.73	1.25
25 GeV	0.33	1.28
30 GeV	0.29	1.58
40 GeV	0.26	1.84
50 GeV	0.22	—
60 GeV	0.20	—

Table 5.1: The 95% CL upper limit of the expected signal strength in the 2-lepton channel.

m_a point	95% UL of Expected Signal Strength	
	Resolved Regime	Merged Regime
15 GeV	11.76	0.45
20 GeV	2.62	0.50
25 GeV	0.69	0.56
30 GeV	0.48	0.72
40 GeV	0.34	2.07
50 GeV	0.25	—
60 GeV	0.33	—

Table 5.2: The 95% CL upper limit of the expected signal strength in the 0-lepton channel.

5.1 Outlook

$H \rightarrow 4b$ decay in VH associated production is going to be further investigated. An effort to combine 0-lepton, 1-lepton, and 2-lepton channels will be made. As we show, the expected limits presented for the 0-lepton channel seem particularly competitive results. A possibility to improve existing 2-lepton upper limits remains open. Furthermore, we will add the rest of Run 2 MC data and hopefully the Run 3 MC data very soon to increase our statistics, as well as to add the real data in order to extract the observed upper limits as well. We are considering contributing to the development of a ParticleNet boosted double-b tagger for AK8 jets, customized for $a \rightarrow bb$ decay. This method is targeting the sensitivity improvement of the already defined merged regime. For the lowest "a"-boson masses, between 12-15 GeV, (bb) pairs become collimated enough to be reconstructed inside the cone of a single AK4 jet. Thus, if AK4 substructure could be investigated in order to define a double b-tagger for AK4 jets, this would help in the investigation of this very boosted regime. Some other more minor steps that need to be taken in the context of this analysis

are to develop a data-driven method for QCD estimation in the 0-lepton channel, to calculate data-MC scale factors, and include systematic uncertainties in detail.

Appendix A

CMS MC Samples

Sample Name	Xsec (pb)	KFactor
DYJetsToLL_M-50.TuneCPS.13TeV-madgraphMLM-pythia8/	4895	1.18
QCD_HT300to500.TuneCP5 13TeV-madgraph-pythia8/	322600.0	1
QCD_HT500to700.TuneCP5 13TeV-madgraph-pythia8/	29980.0	1.0
QCD_HT700to1000.TuneCP5 13TeV-madgraph-pythia8/	6334.0	1.0
QCD_HT1000to1500.TuneCPS.13TeV-madgraph-pythia8/	1088.0	1.0
QCD_HT1500to2000.TuneCPS.13TeV-madgraph-pythia8/	99.11	1.0
QCD_HT2000toInf.TuneCPS.13TeV-madgraph-pythia8/	20.23	1.0
TTToHadronic.TuneCPS.PSweights.13TeV-powheg-pythia8/	377.96	1.0
TTTo2L2Nu.TuneCPS.PSweights.13TeV-powheg-pythia8/	88.29	1.0
TTToSemiLeptonic.TuneCPS.13TeV-powheg-pythia8/	365.34	1.0
WJetsToLNu_HT-100to200.TuneCPS.13TeV-madgraphMLM-pythia8/	1395.0	1.21
WJetsToLNu_HT-1200to2500.TuneCPS.13TeV-madgraphMLM-pythia8/	1.074	1.21
WJetsToLNu_HT-200To400.TuneCPS.13TeV-madgraphMLM-pythia8/	407.9	1.21
WJetsToLNu_HT-2500ToInf.TuneCPS.13TeV-madgraphMLM-pythia8/	0.03216	1.21
WJetsToLNu_HT-400To600.TuneCPS.13TeV-madgraphMLM-pythia8/	57.48	1.21
WJetsToLNu_HT-600To800.TuneCPS.13TeV-madgraphMLM-pythia8/	12.87	1.21
WJetsToLNu_HT-800To1200.TuneCPS.13TeV-madgraphMLM-pythia8/	5.366	1.21

WJetsToLNu.TuneCPS.13TeV-madgraphMLM-pythia8/	5294.0	1.0
ST_s-channel-4f-leptonDecays.TuneCP5.PSweights.13TeV-amcatnlo-pythia8/	6334.0	1.0
ST_t-channel_antitop-4f-InclusiveDecays.TuneCP5.PSweights.13TeV-powheg-pythia8/	3.74	1.0
ST_t-channel_top-4f-InclusiveDecays.TuneCP5.PSweights.13TeV-powheg-pythia8/	80.95	1.0
ST_tW-antitop-5f-InclusiveDecays.TuneCP5.PSweights.13TeV-powheg-pythia8/	136.02	1.0
ST_tW-top-5f-InclusiveDecays.TuneCP5.PSweights.13TeV-powheg-pythia8/	35.85	1.0
WW.TuneCPS.13TeV-pythia8/	115.3	1.0
WZ.TuneCPS.13TeV-pythia8/	48.1	1.0
WminusH.HToBB.WToLNu.M125.13TeV-powheg-pythia8/	0.10899	1.0
WplusH.HToBB.WToLNu.M125.13TeV-powheg-pythia8/	0.17202	1.0
ZH.HToBB.ZToNuNu.M125.13TeV-powheg-pythia8/	0.09322	1.0
ZJetsToNuNu.HT-100to200.13TeV-madgraph/	302.8	1.23
ZJetsToNuNu.HT-200To400.13TeV-madgraph/	92.59	1.23
ZJetsToNuNu.HT-2500ToInf.13TeV-madgraph/	0.005146	1.23
ZJetsToNuNu.HT-400To600.13TeV-madgraph/	13.18	1.23
ZJetsToNuNu.HT-600To800.13TeV-madgraph/	3.257	1.23
ZJetsToNuNu.HT-800To1200.13TeV-madgraph/	1.496	1.23
ZZ.TuneCPS.13TeV-pythia8/	16.523	1.0

Appendix B

Expected signal yield in all the "a" mass samples

In tables B.1 and B.2, the expected event yield in the resolved and in the merged regime is shown for all "a" mass samples, of the 0-lepton and of the 2-lepton channel respectively.

	a mass point							
Regime	M(12)	M(15)	M(20)	M(25)	M(30)	M(40)	M(50)	M(60)
Merged	238	208	212	205	230	175	188	161
Resolved	56	41	66	126	192	237	278	224

Table B.1: Expected signal yield in the merged and in the resolved regime of the 0-lepton channel, for all "a" mass points.

	a mass point							
Regime	M(12)	M(15)	M(20)	M(25)	M(30)	M(40)	M(50)	M(60)
Merged	31	35.2	31.4	34.2	30.8	27.4	19.5	17.6
Resolved	22.3	19.5	60.4	95	104.5	110	127.8	136

Table B.2: Expected signal yield in the merged and in the resolved regime of the 2-lepton channel, for all "a" mass points.

Appendix C

AK8 Jet Characteristics

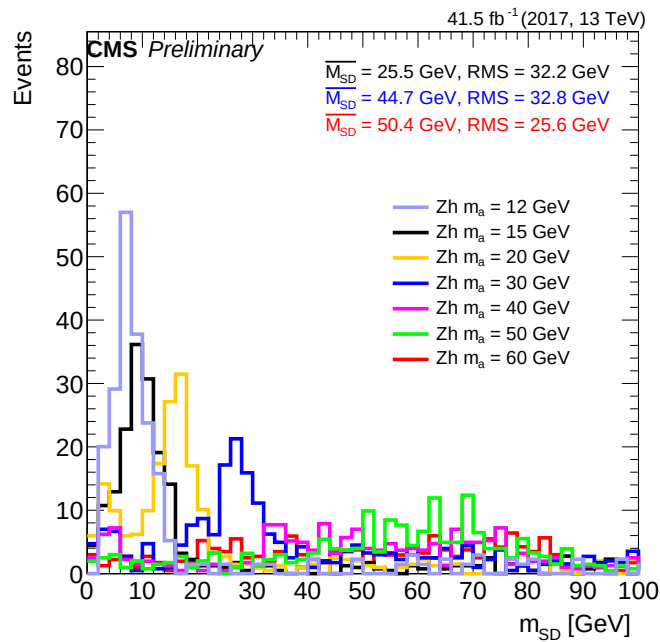


Figure C.1: fat jet m_{sd} for many possible "a" masses. The peaks of distributions are closer to truth value of m_a for low "a" masses, as "a" boson is fully reconstructed by a fat jet. However they are shifted towards lower values than the m_a truth value (mass regression application is planned in the future.)

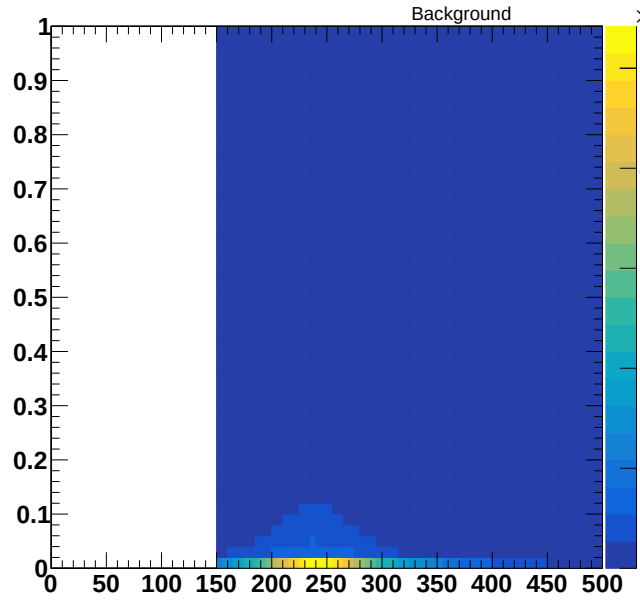


Figure C.2: In this figure we see that above the cutoff fat jet $p_T = 150 \text{ GeV}$ the discriminant does not seem to be correlated with the p_T . The distribution is shown for the total background to include more statistics.

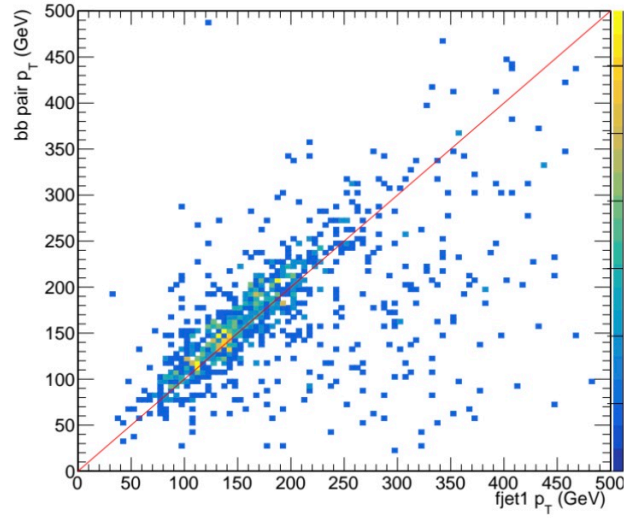


Figure C.3: In this figure we see in the same plane, the p_T of a fat jet that is matched to a bb pair coming from the same "a" boson in generator level and the p_T of that (bb) pair.

Appendix D

Wh contribution to the 0-lepton channel

Let's have a deeper look at how the Wh process results in such a large contribution to total signal yield in the 0-lepton channel, as shown in Table 4.10. In tables D.1 and D.2 the expected events, surviving after each selection cut, are shown for the Zh process and for the Wh process respectively.

Process	Raw	Exactly 0 leptons	$MET > 170$ GeV	Merged: $N_{fjet} \geq 1$ and $N_b \geq 1$	Resolved: $N_b \geq 3$
Zh $m_a = 15$ GeV	35577	34448 $eff = 93.9\%$	633 $eff = 1.72\%$	116.5 $eff = 0.31\%$	22.7 $eff = 0.04\%$
Zh $m_a = 30$ GeV	36688	34414 $eff = 93.8\%$	662 $eff = 1.8\%$	123.7 $eff = 0.33\%$	123.2 $eff = 0.33\%$
Zh $m_a = 60$ GeV	36685	34428 $eff = 93.8\%$	630 $eff = 1.7\%$	99.3 $eff = 0.27\%$	147 $eff = 0.4\%$

Table D.1: Event flow table of the Zh process in the 0-lepton channel

Process	Raw	Exactly 0 leptons	$MET > 170$ GeV	Merged: $N_{fjet} \geq 1$ and $N_b \geq 1$	Resolved: $N_b \geq 3$
Wh $m_a = 15$ GeV	56848	49689 $eff = 87.4\%$	419 $eff = 0.73\%$	91.7 $eff = 0.16\%$	15.7 $eff = 0.03\%$
Wh $m_a = 30$ GeV	56847	49553 $eff = 87.1\%$	426 $eff = 0.75\%$	105.3 $eff = 0.18\%$	68.9 $eff = 0.12\%$
Wh $m_a = 60$ GeV	56843	49675 $eff = 87.3\%$	371 $eff = 0.65\%$	60.6 $eff = 0.1\%$	74.3 $eff = 0.13\%$

Table D.2: Event flow table of the Wh process in the 0-lepton channel

In order to visualize the effect of each cut on the surviving events, we plot each variable on which we apply a cut exactly before the step of the event flow where this cut is applied. So, beginning with step-1 of the event cut flow, where we demand exactly 0 leptons, we plot the lepton multiplicity before any selection cut, in Zh and Wh separately. Then we plot the MET in step-1 of the cut flow, where we have demanded 0 leptons. Finally, in step-2, where we applied the MET cut at 170 GeV, we plot the multiplicity of b-tagged jets and the multiplicity of fat jets and cross-cleaned b-tagged jets.

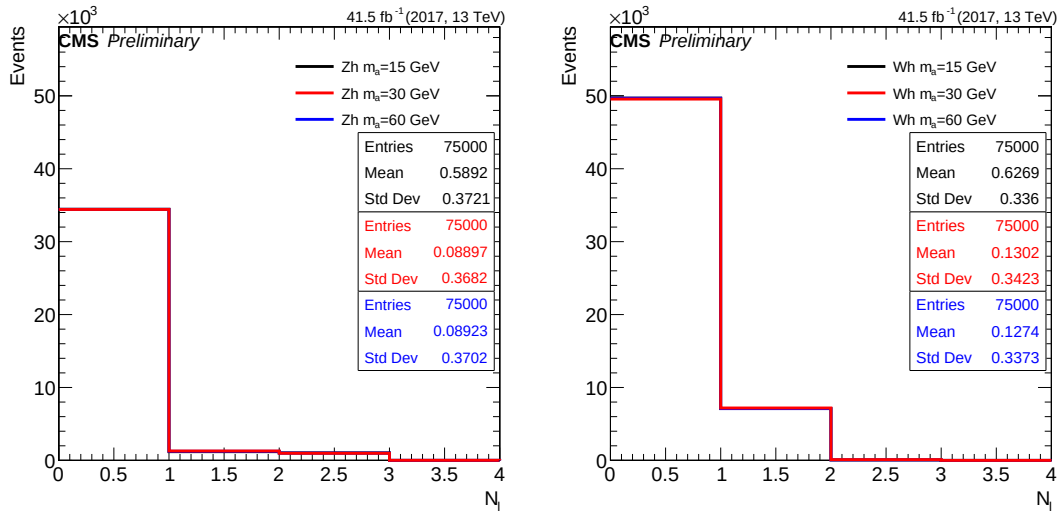


Figure D.1: Leptons multiplicity in Zh (left) and Wh (right) process before any selection cut.

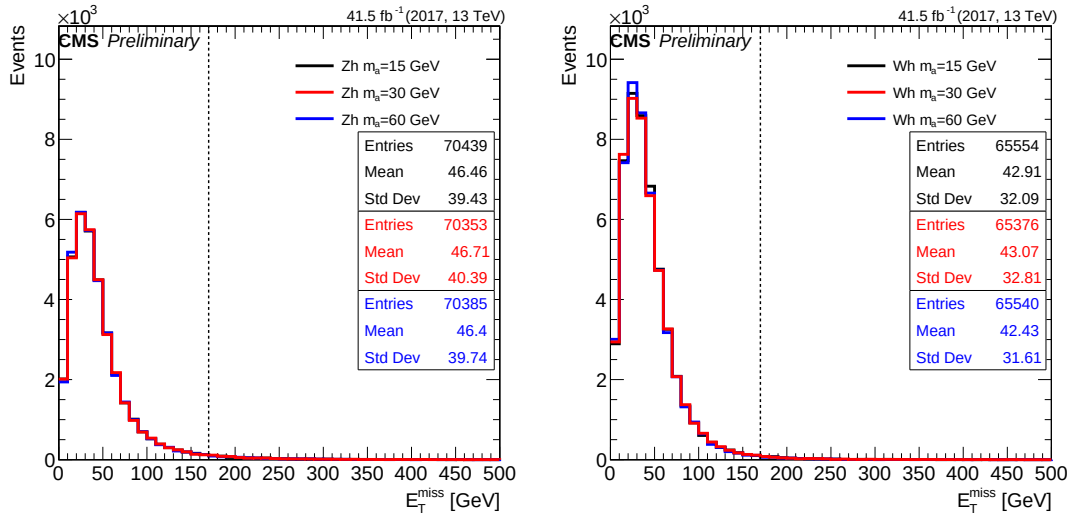


Figure D.2: MET distribution in Zh (left) and Wh (right) process after "exactly 0 leptons" cut (step 1 of the event cut flow).

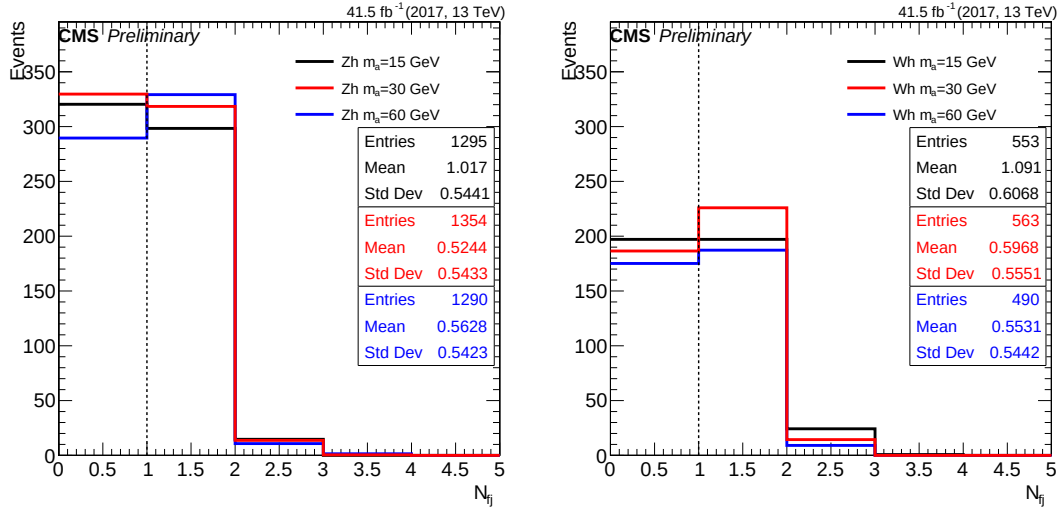


Figure D.3: Multiplicity of fat jets in Zh (left) and Wh (right) process after MET cut at 170 GeV (step 2 of the event cut flow).

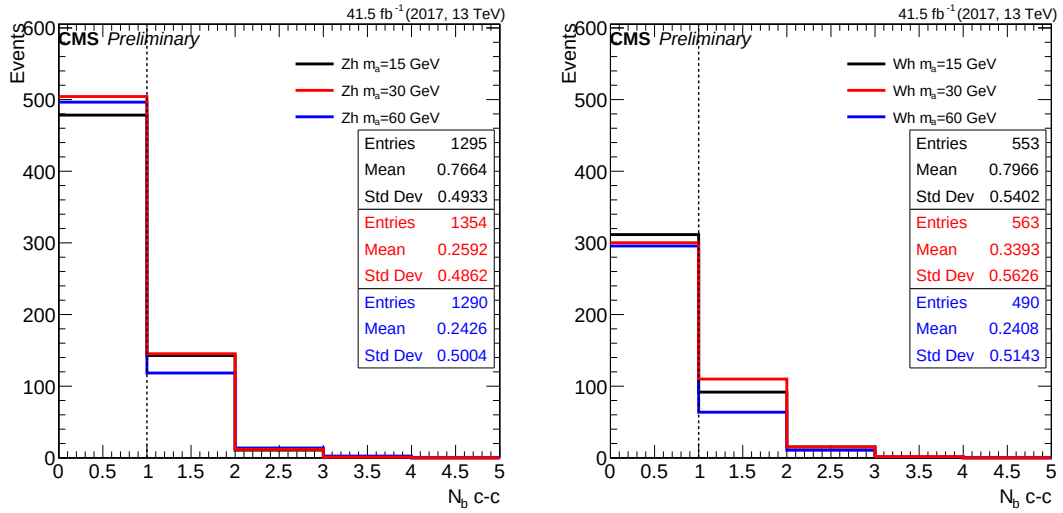


Figure D.4: Multiplicity of cross-cleaned b-tagged jets in Zh (left) and Wh (right) process after MET cut at 170 GeV (step 2 of the event cut flow).

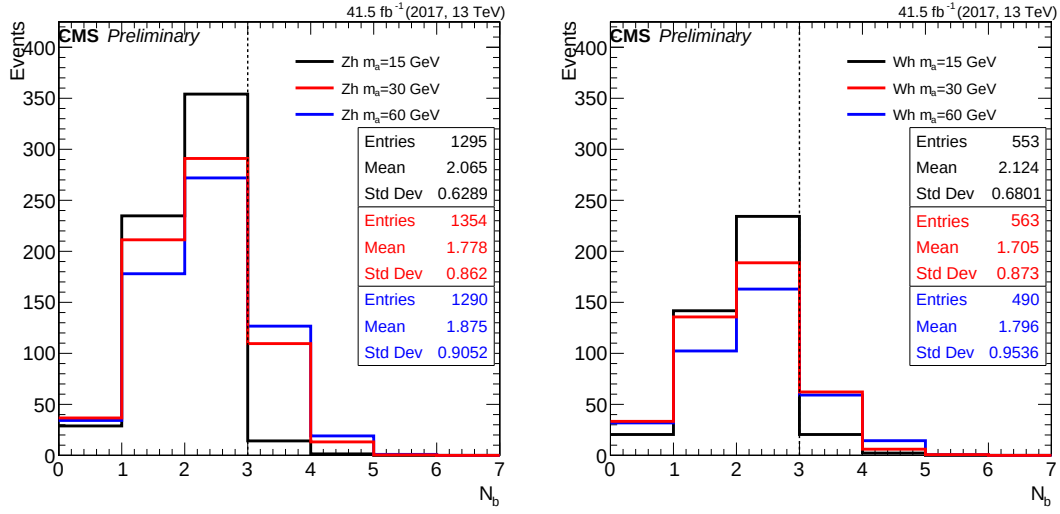


Figure D.5: Multiplicity of b tagged jets in Zh (left) and Wh (right) process after MET cut at 170 GeV (step 2 of the event cut flow).

In Figures C.6, C.7 we show the multiplicity of fat jets and cross-cleaned b-tagged jets, respectively, in the merged regime. And finally, in Figure C.8 we show the multiplicity of b-tagged jets in the resolved regime.

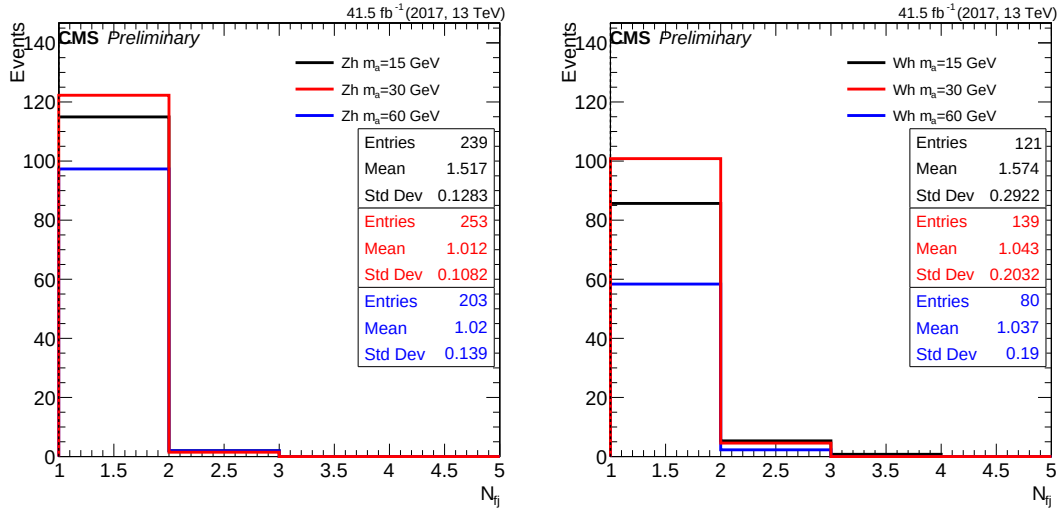


Figure D.6: Multiplicity of fat jets in Zh (left) and Wh (right) process in the merged regime (step 3 of the event cut flow).

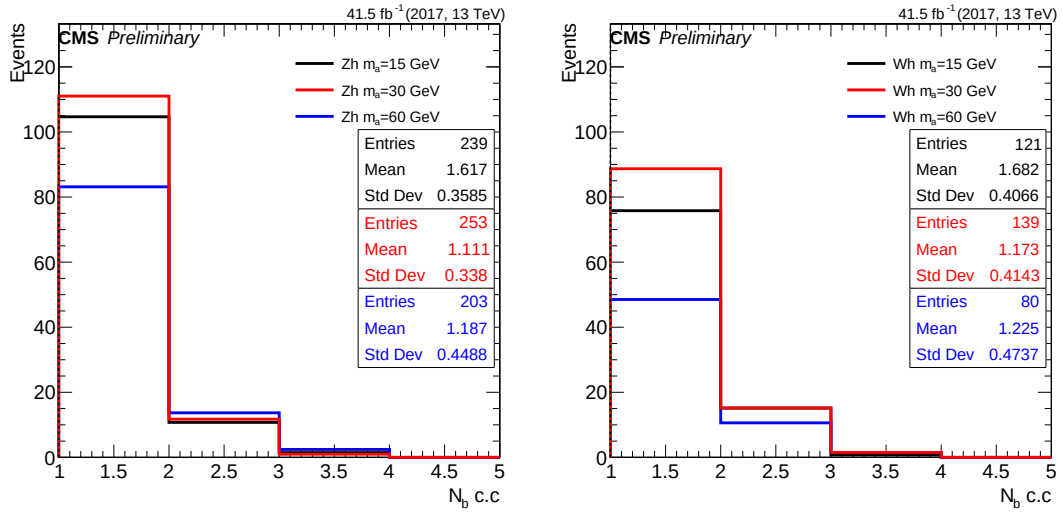


Figure D.7: Multiplicity of cross-cleaned b-tagged jets in Zh (left) and Wh (right) process in the merged regime (step 3 of the event cut flow).

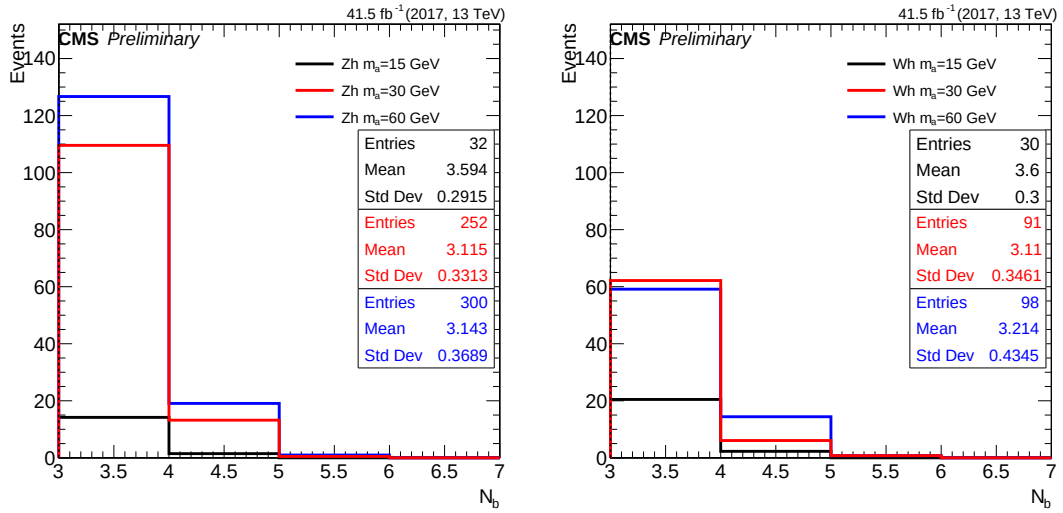


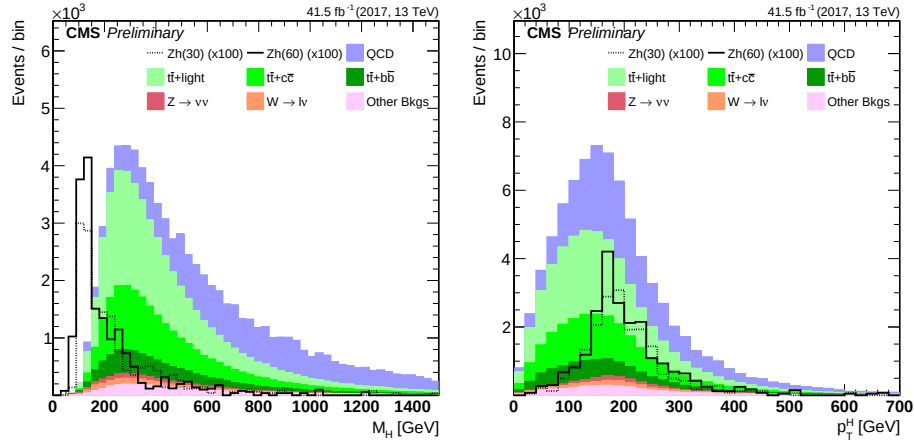
Figure D.8: Multiplicity of b-tagged jets in Zh (left) and Wh (right) process in the resolved regime (step 3 of the event cut flow).

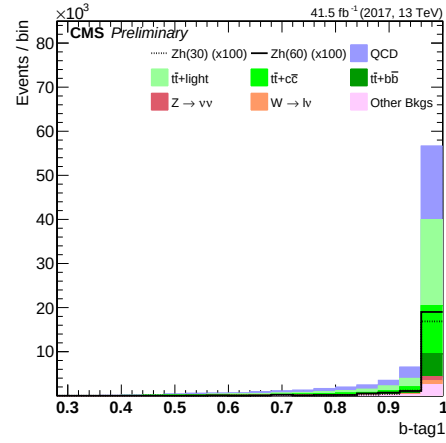
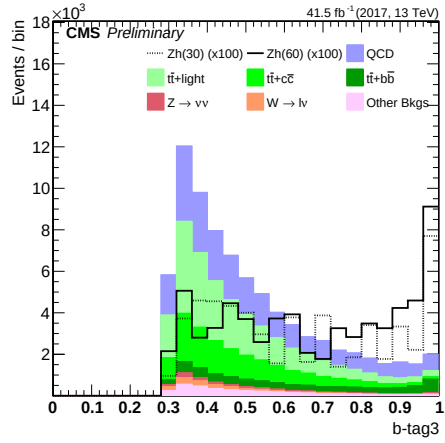
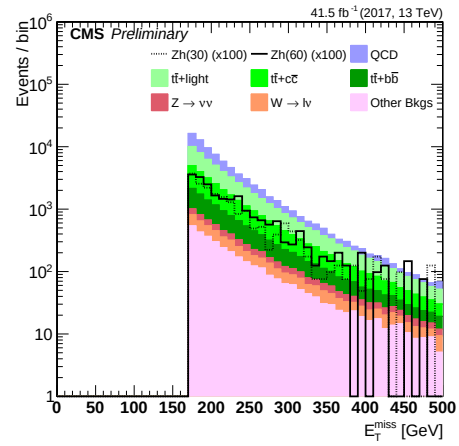
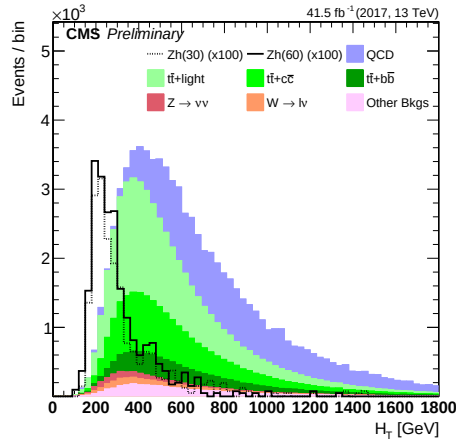
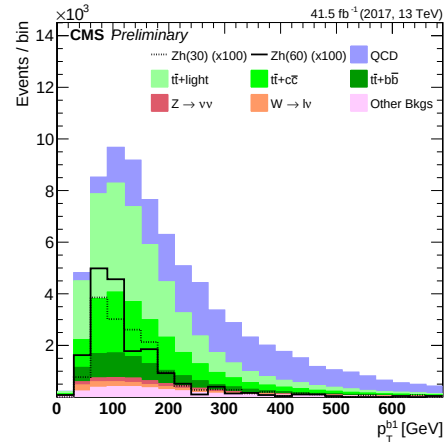
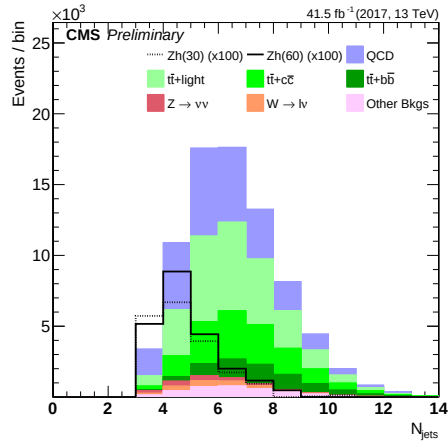
Appendix E

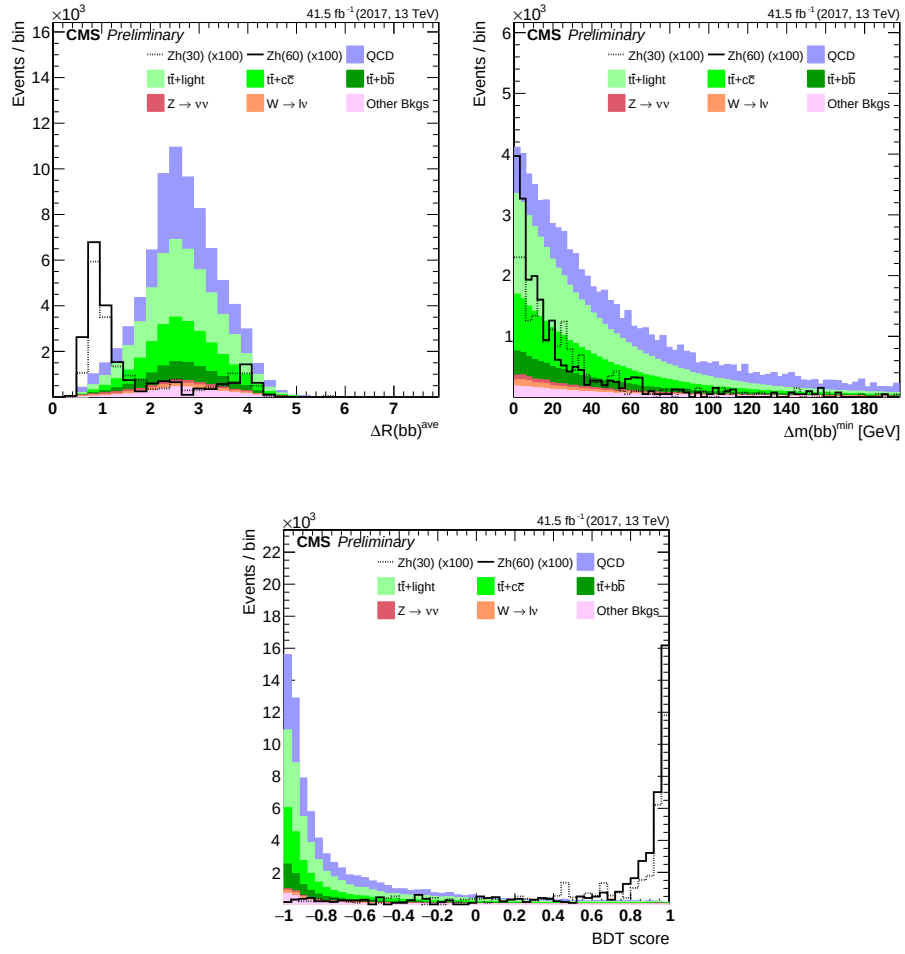
BDT input variable distributions

There follows a total presentation of the BDT input variables and their distributions after the application of all selection cuts in each regime of each channel. In the following histograms, distributions of the signal of two possible 'a' masses ($m_a = 15$ GeV and $m_a = 30$ GeV in the merged regime and $m_a = 30$ GeV and $m_a = 60$ GeV in the resolved regime) are superimposed with the stacked distributions of all backgrounds that are present in each channel. The signal distributions are scaled for visibility purposes. The output BDT score distribution is also the last figure shown in each regime of each channel.

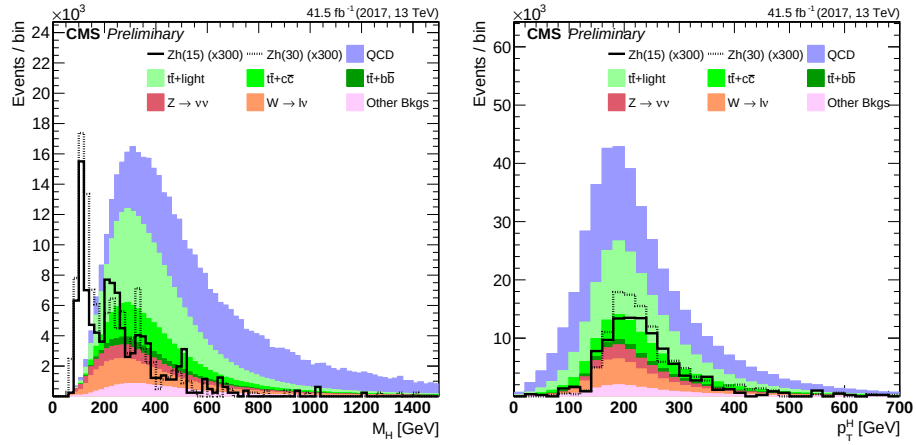
E.0.1 The resolved regime of the 0-lepton channel

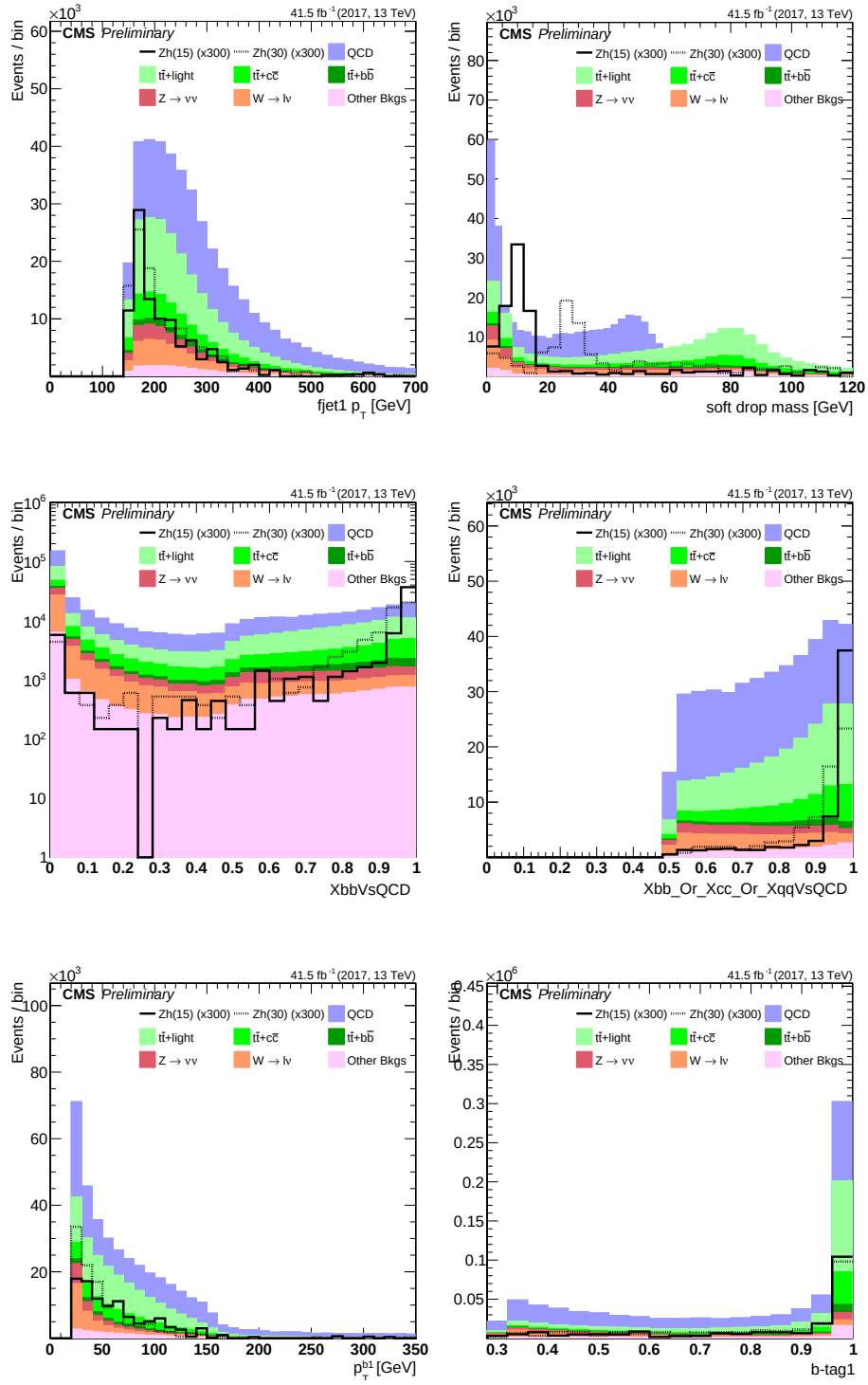


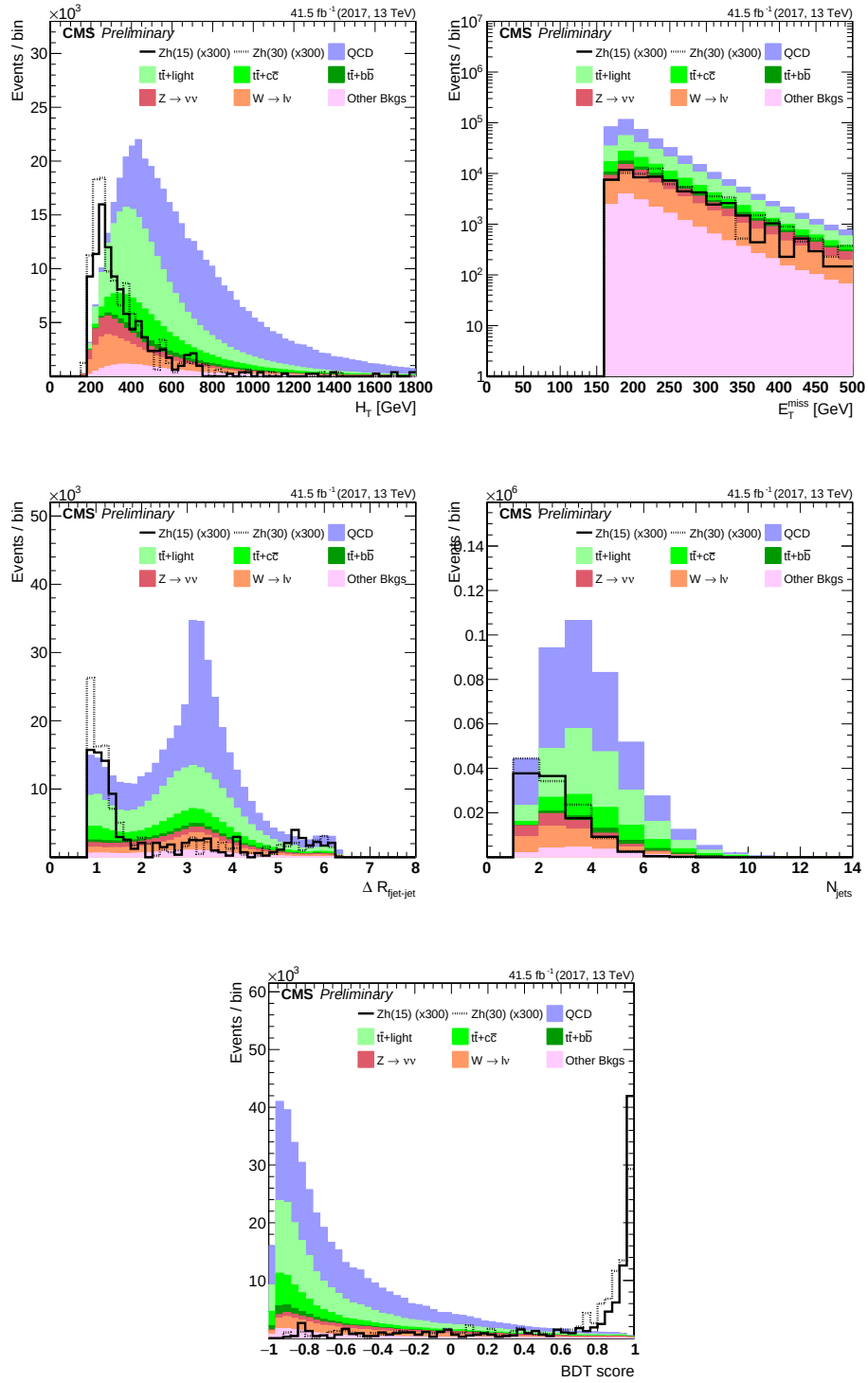




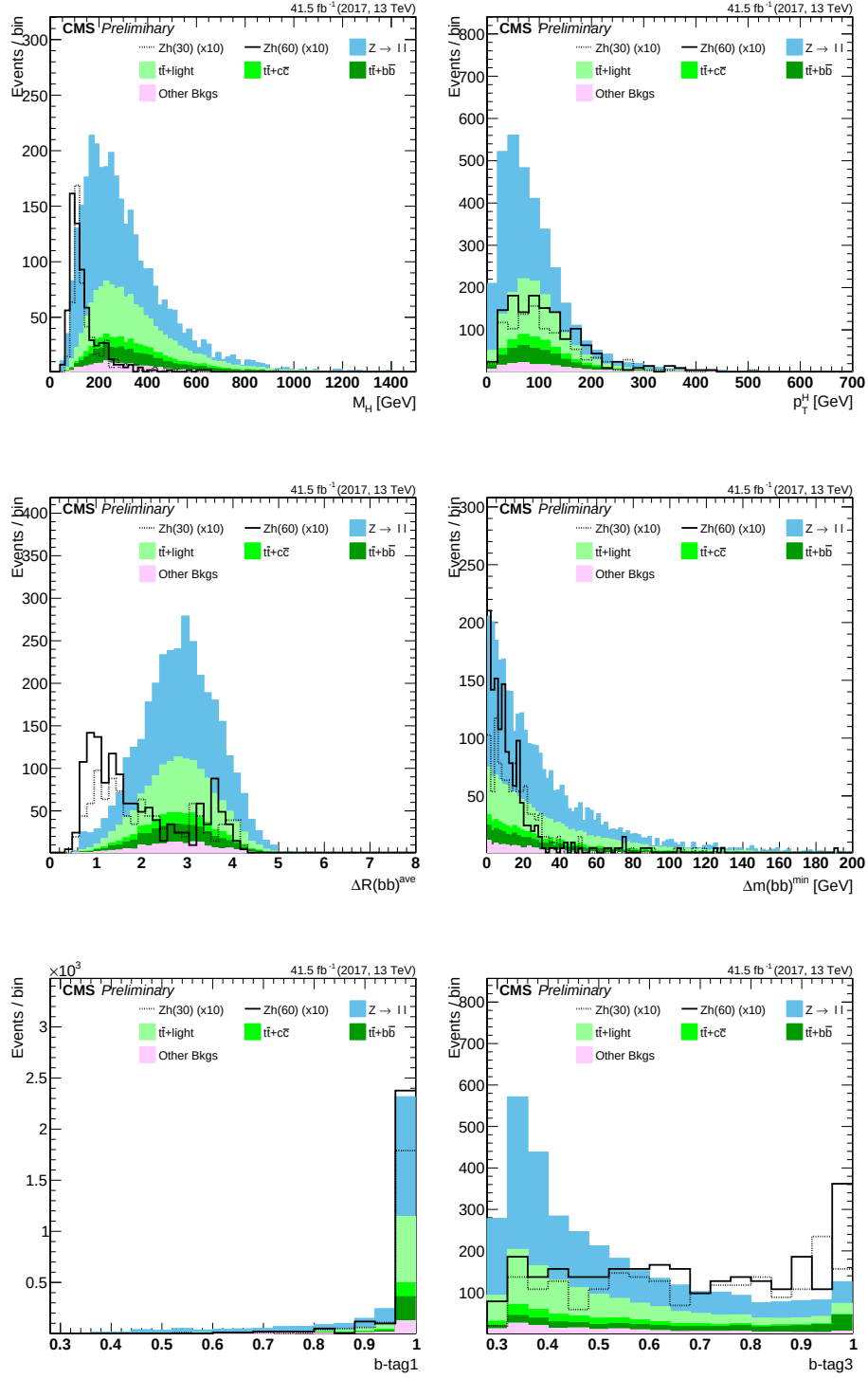
E.0.2 The merged regime of the 0-lepton channel

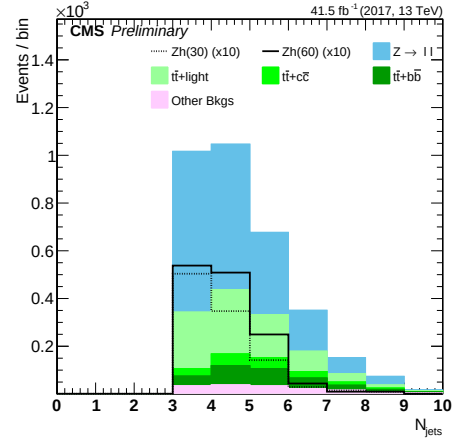
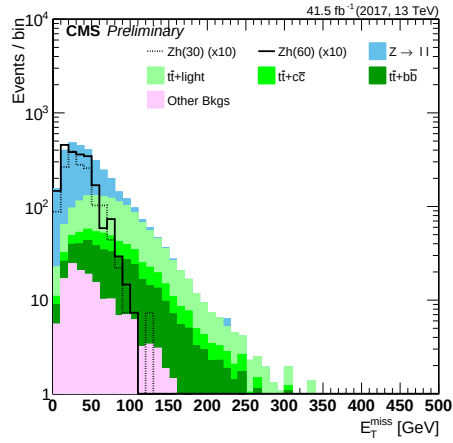
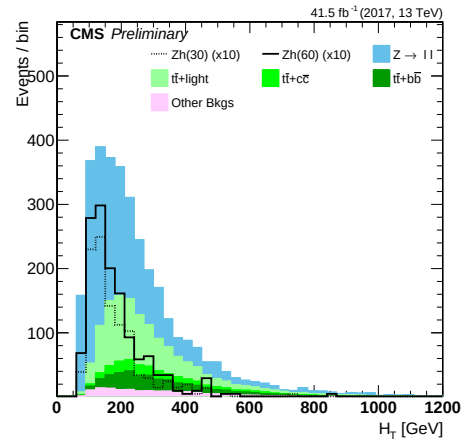
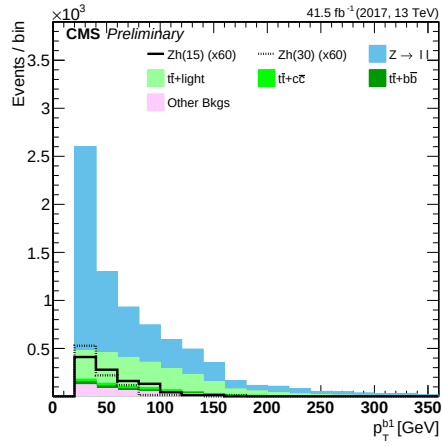
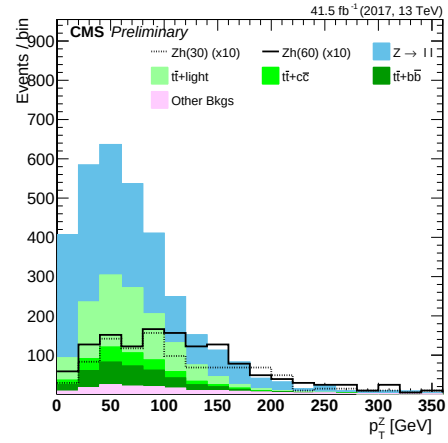
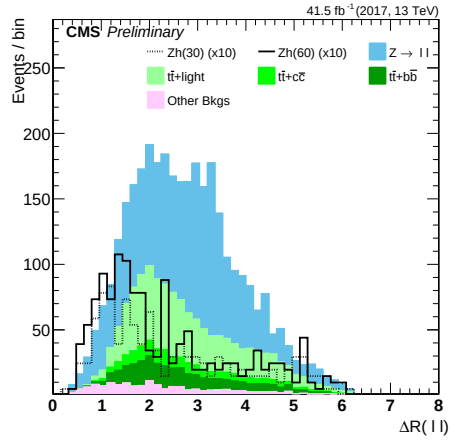


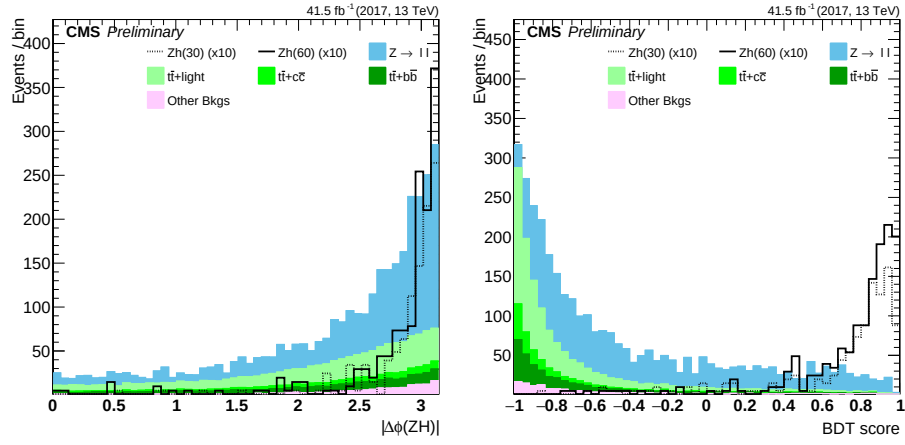




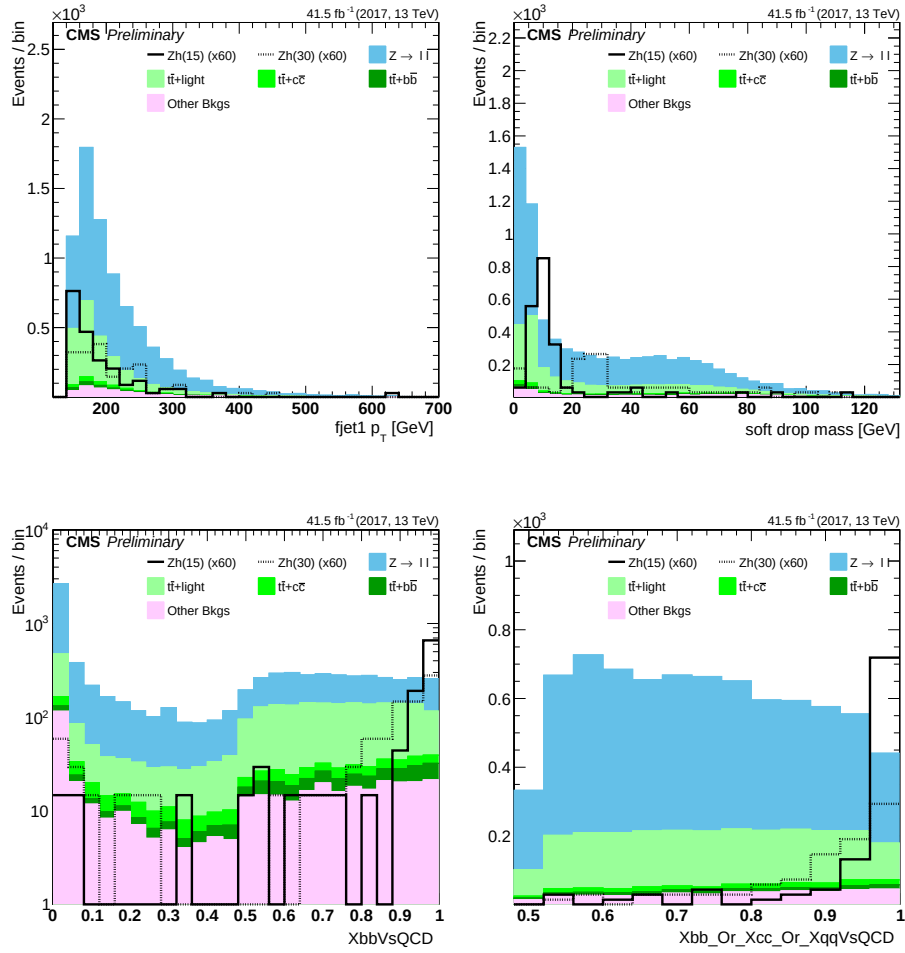
E.0.3 The resolved regime of the 2-lepton channel

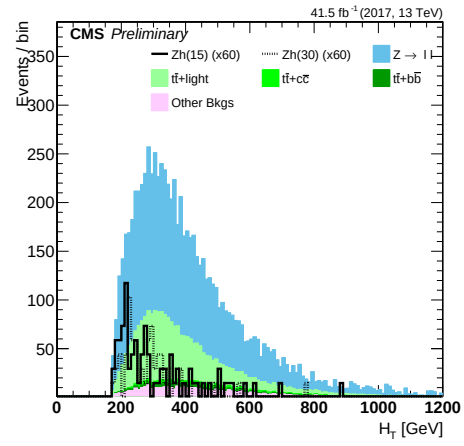
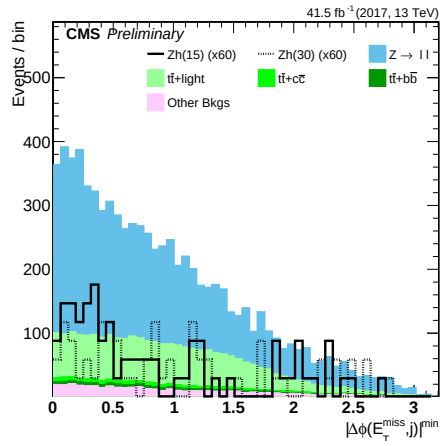
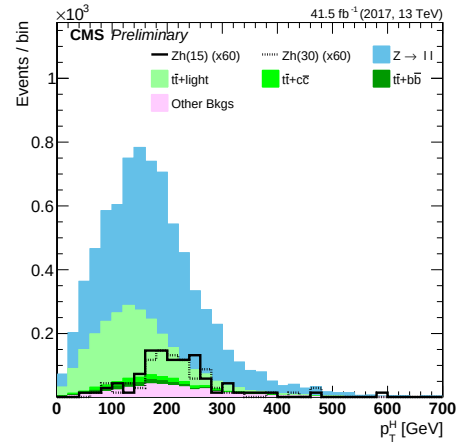
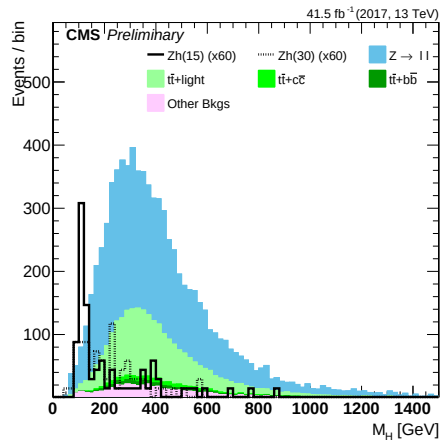


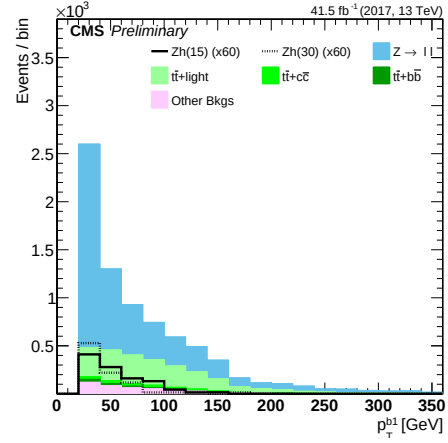
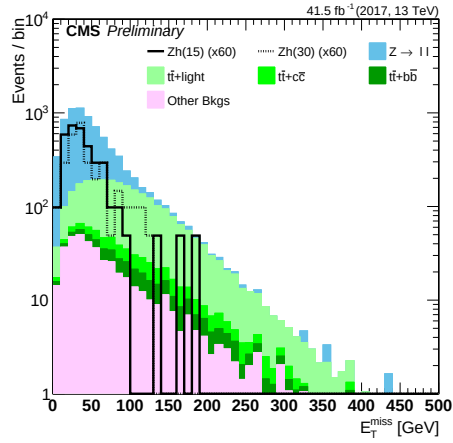
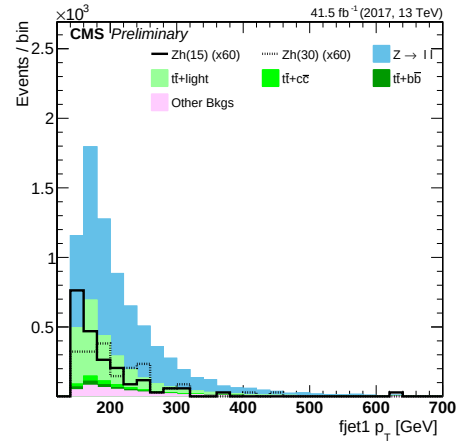
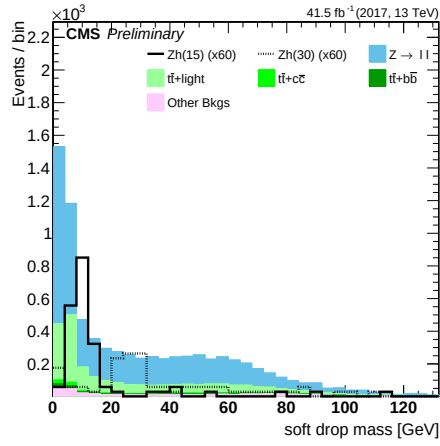
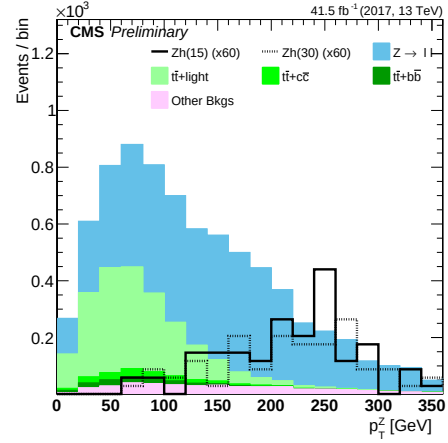
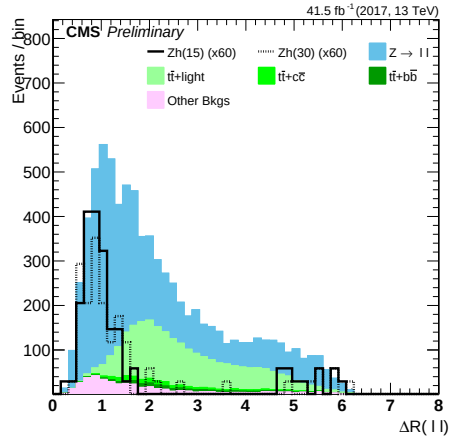


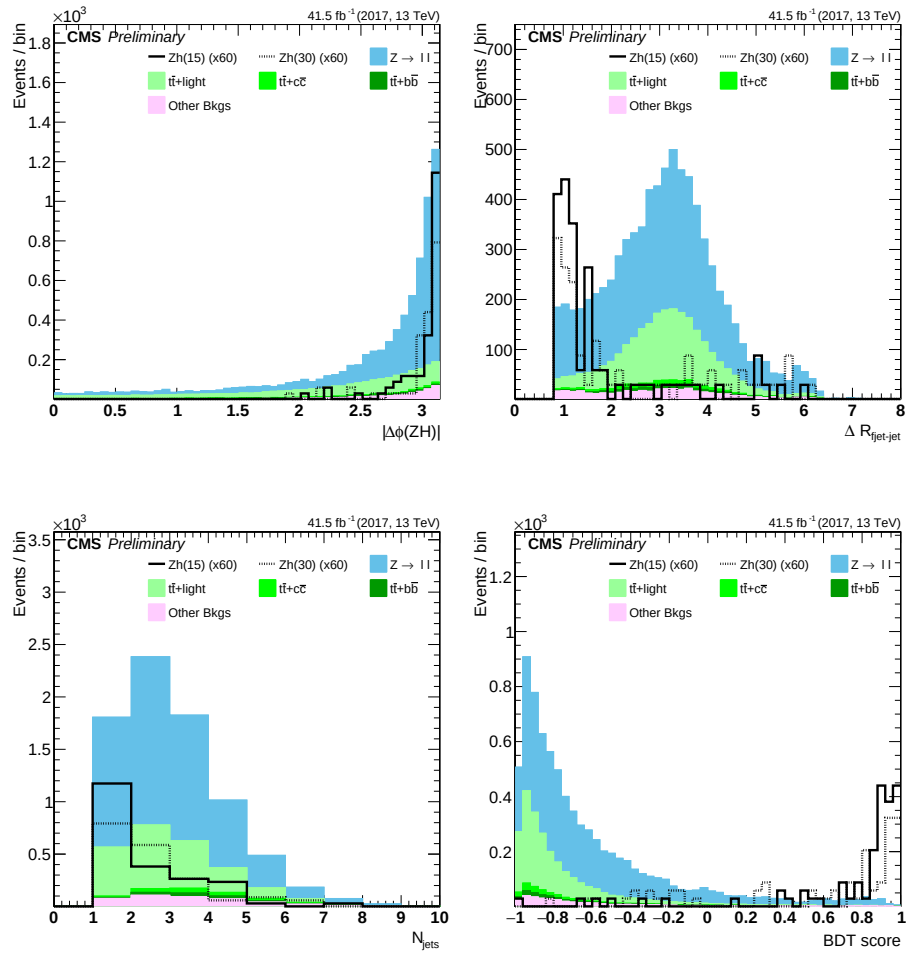


E.0.4 The merged regime of the 2-lepton channel









Appendix F

Nuisance parameters and their uncertainties

Process	Range	σ [pb]	Uncertainty [%]		
			Xsec	Systematic	Total
$t\bar{t} + \text{LF}$	$(0, 2N_{t\bar{t}+\text{LF}}^{\text{exp}})$	834	6%	32%	33%
$t\bar{t} + \text{bb}$	$(-0.8N_{t\bar{t}+\text{bb}}^{\text{exp}}, 2.8N_{t\bar{t}+\text{bb}}^{\text{exp}})$	834	6%	58%	58%
$t\bar{t} + \text{cc}$	$(-1.25N_{t\bar{t}+\text{cc}}^{\text{exp}}, 3.25N_{t\bar{t}+\text{cc}}^{\text{exp}})$	834	6%	75%	75%
W+jets	$(0, 2N_{\text{W+jets}}^{\text{exp}})$	65309	4%	30%	30%
QCD	$(-2N_{\text{QCD}}^{\text{exp}}, 4N_{\text{QCD}}^{\text{exp}})$	360121	-	-	50%
DY	$(0, 2N_{\text{DY}}^{\text{exp}})$	4895	2%	30%	30%
Other bkgs	$(-0.8N_{\text{EW}}^{\text{exp}}, 2.8N_{\text{EW}}^{\text{exp}})$	335.78	50%	30%	60%

Table F.1: Ranges and uncertainties applied to each nuisance parameter (the expected yield of each background process) in the fit procedure.

Appendix G

Posterior distributions of the expected signal yield

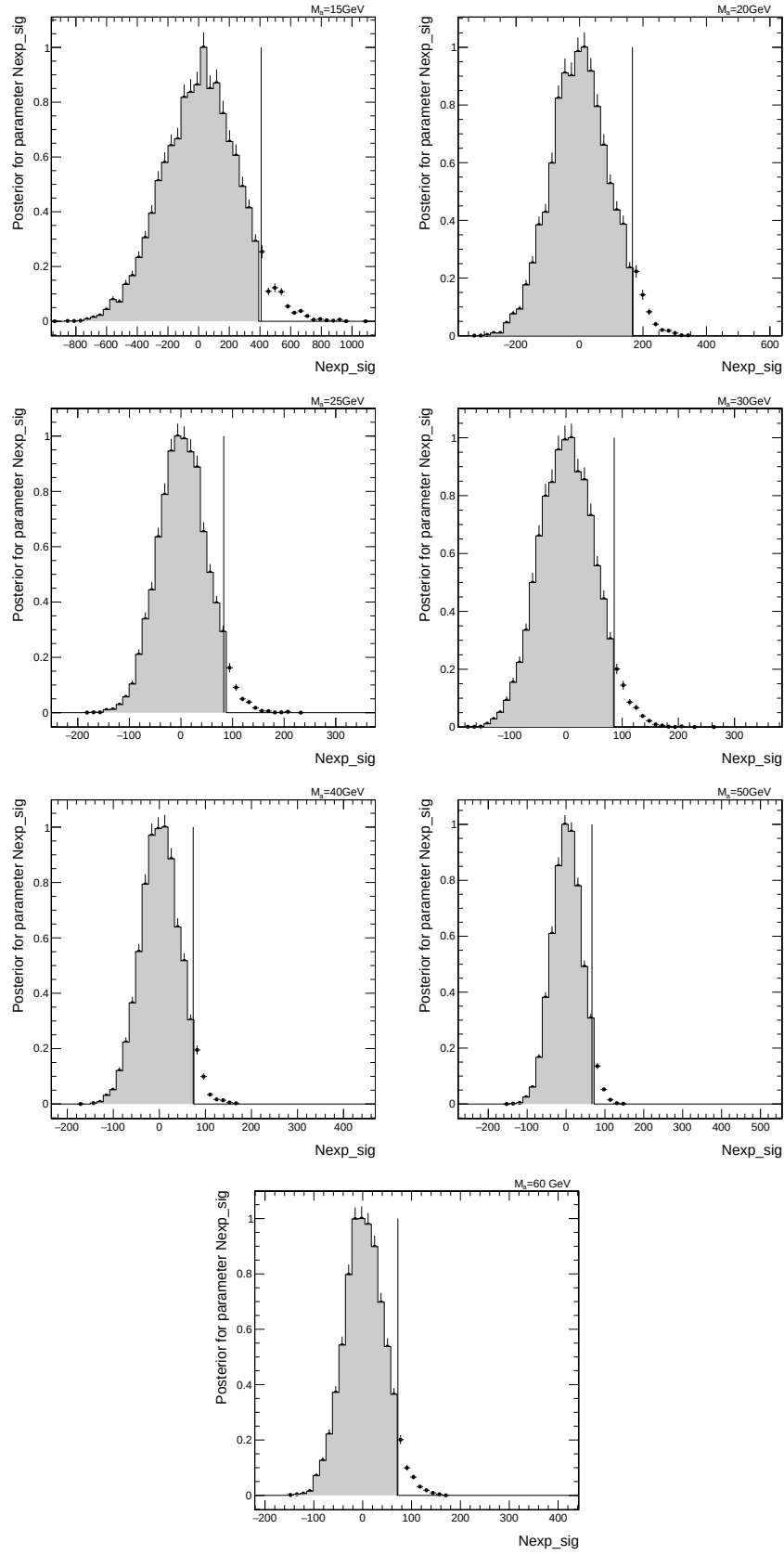


Figure G.1: Posterior distribution for the expected signal yield for m_a in the range [15-60] GeV in the resolved regime of the 0-lepton channel .

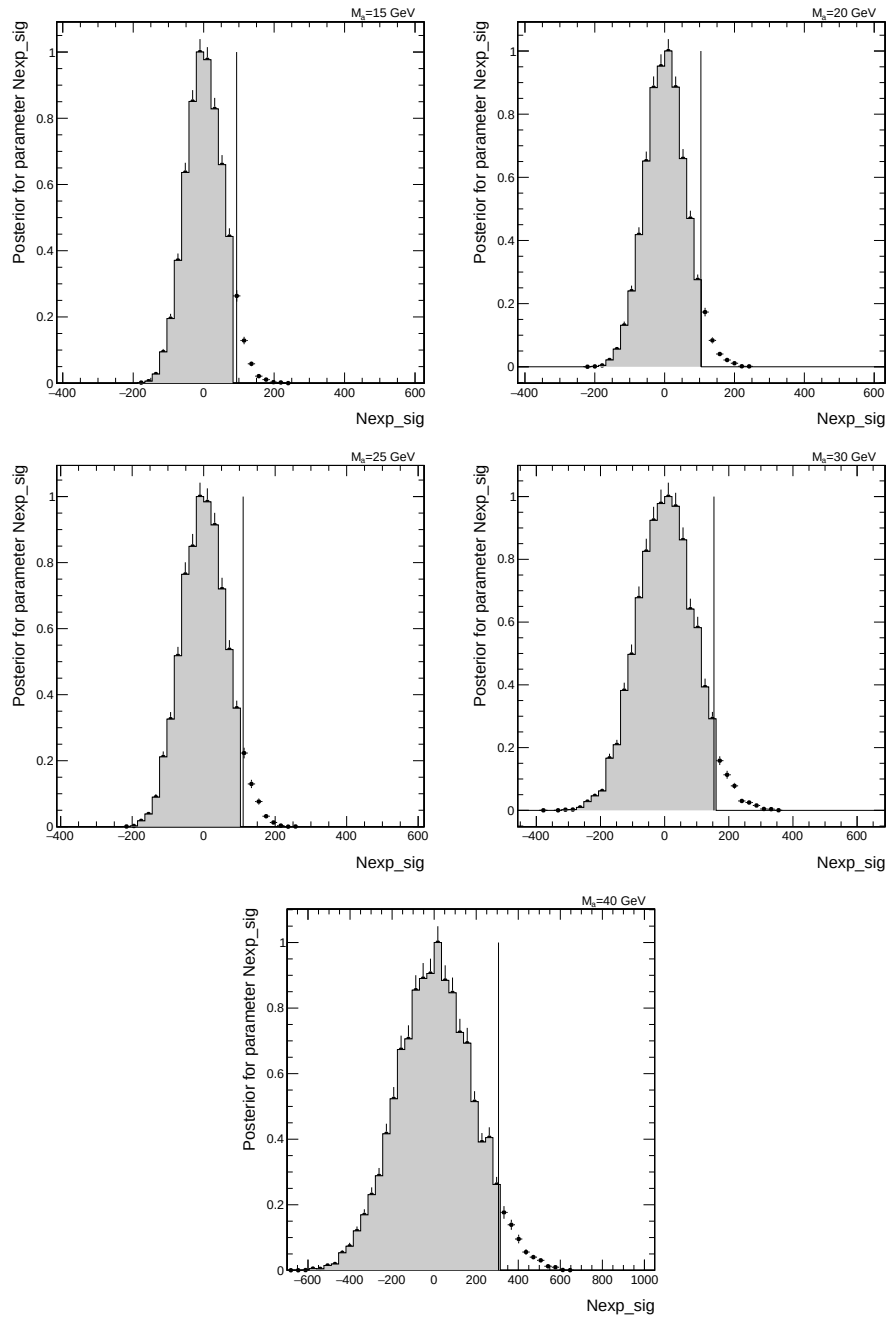


Figure G.2: Posterior distribution for the expected signal yield for m_a in the range [15-40] GeV in merged regime of the 0-lepton channel.

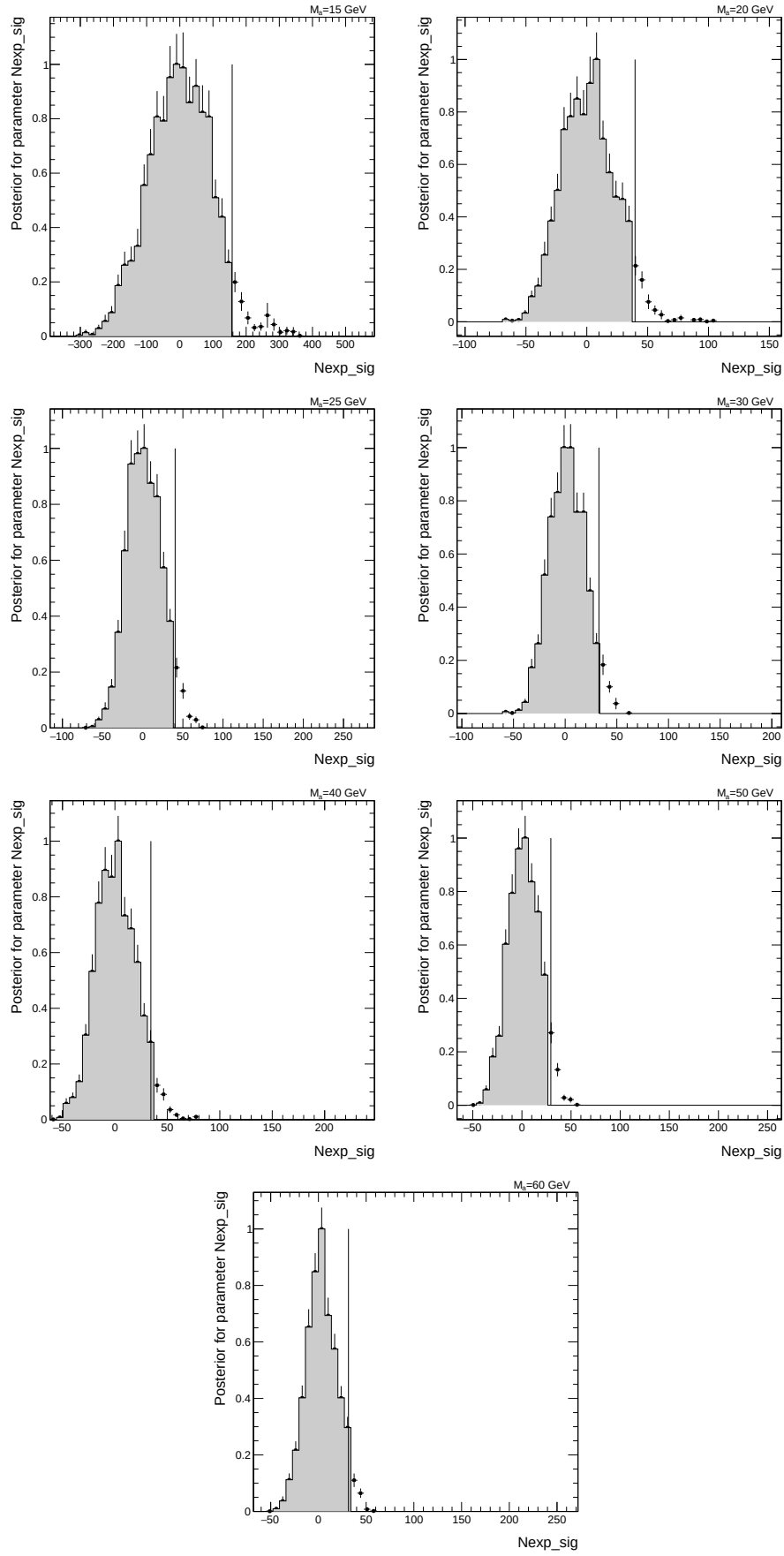


Figure G.3: Posterior distribution for the expected signal yield for m_a in the range [15-60] GeV in the resolved regime of the 2-lepton channel.

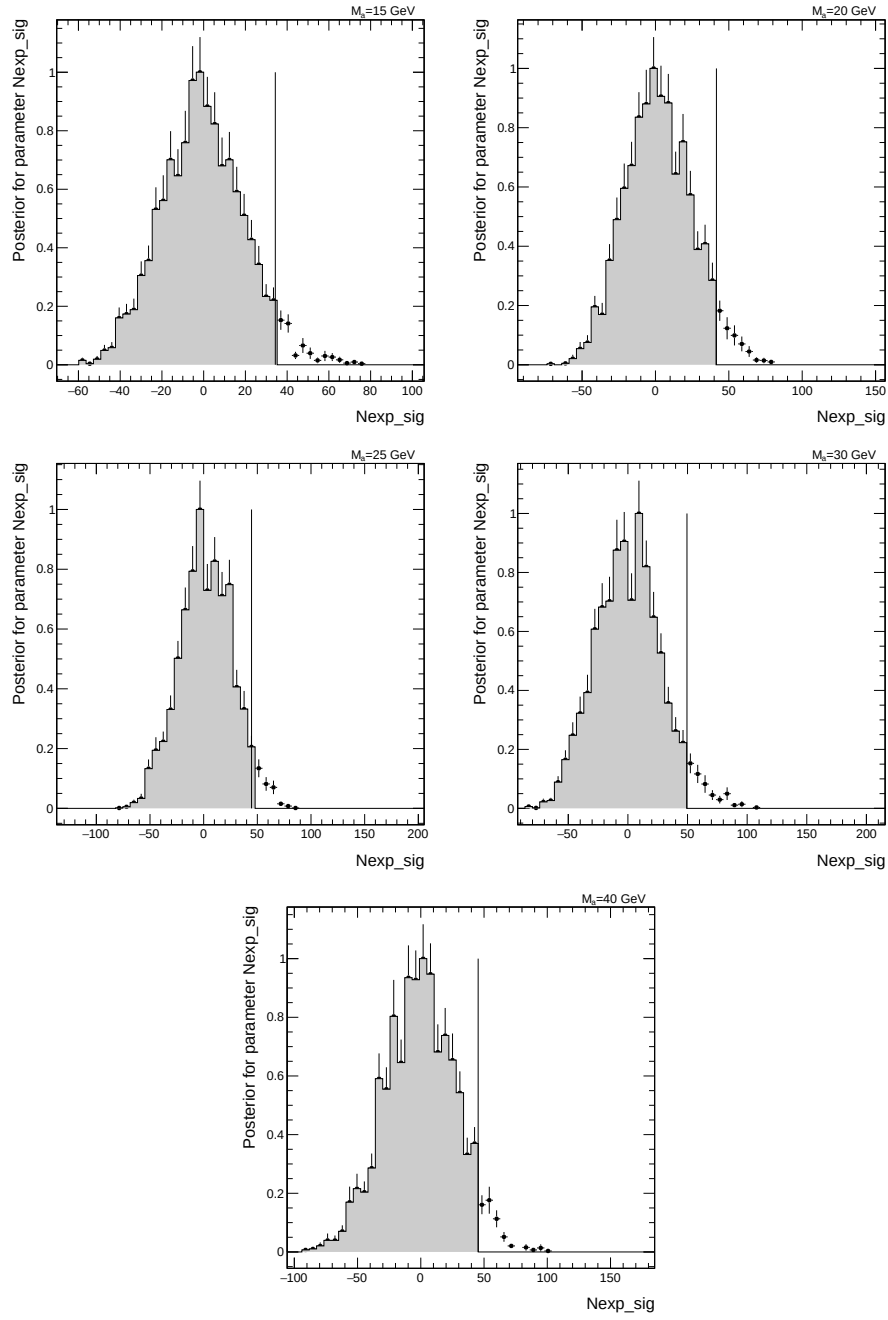


Figure G.4: Posterior distribution for the expected signal yield for m_a in the range [15-40] GeV in the merged regime of the 2-lepton channel.

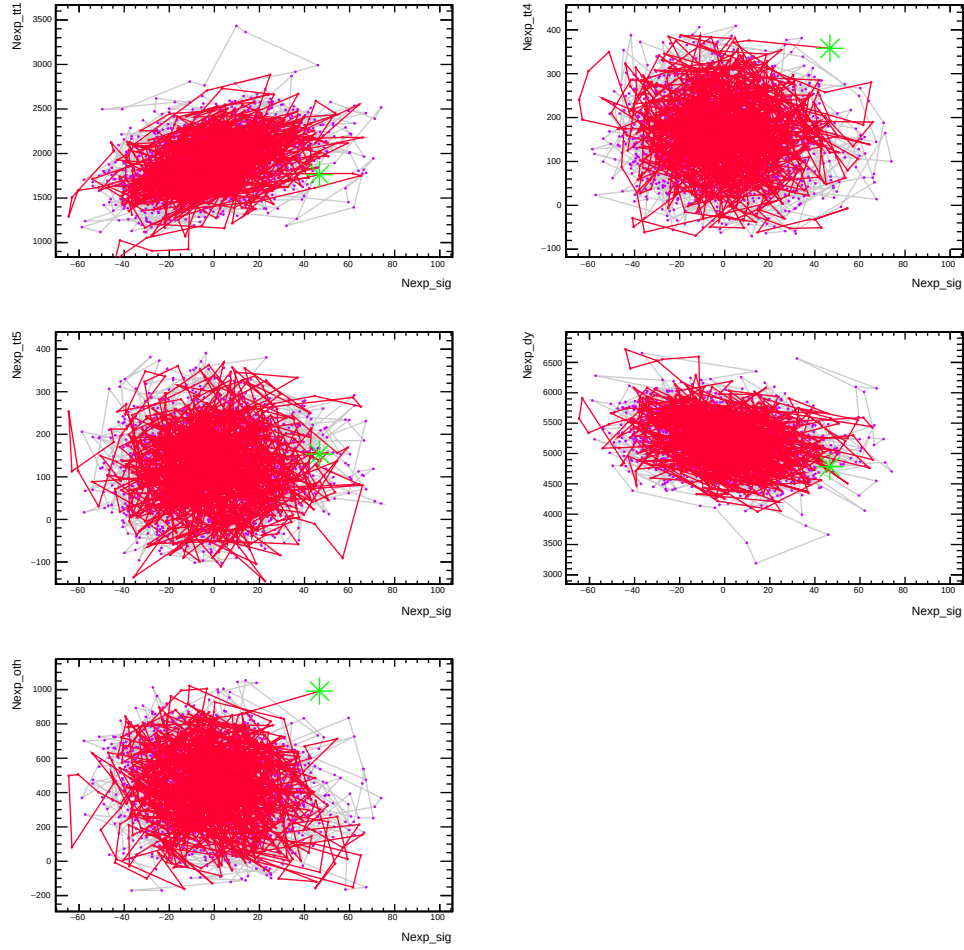
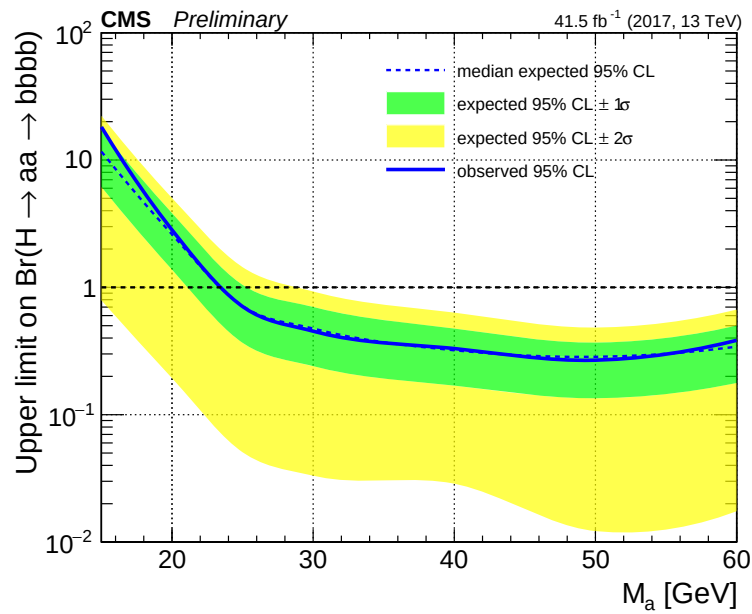


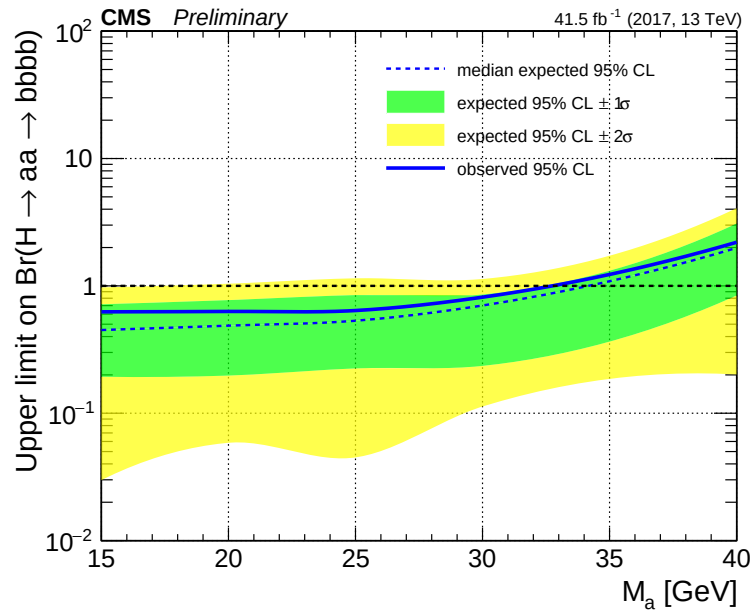
Figure G.5: Trace plot showing the correlations between POI and each nuisance parameter during the sampling procedure of MCMC algorithm, in the resolved regime, 2-lepton channel

Appendix H

Upper limits in logarithmic scale

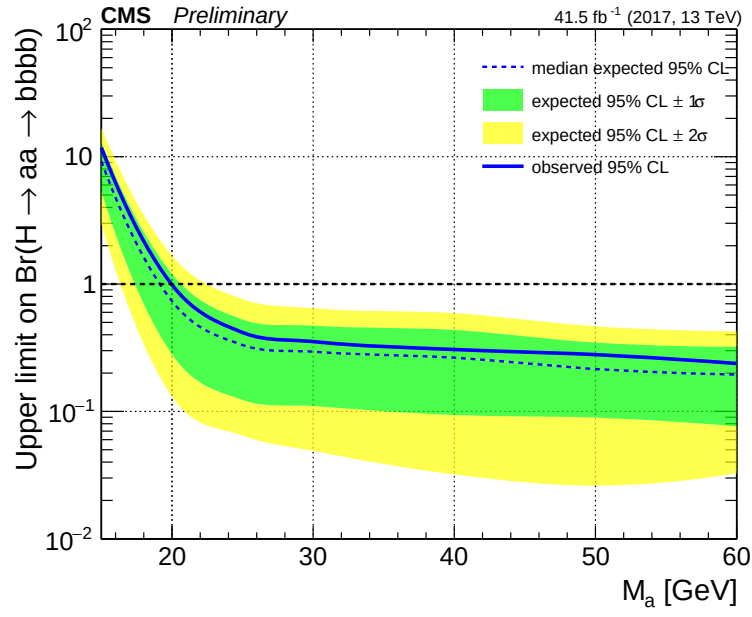


(a)

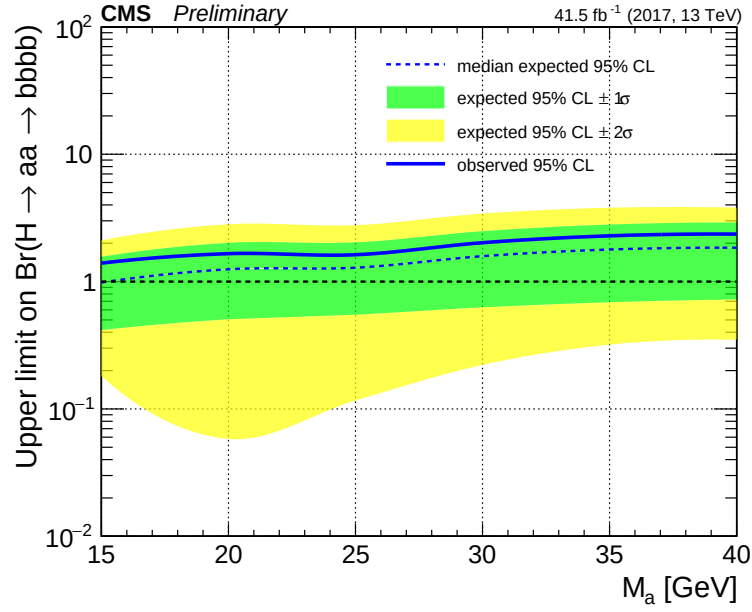


(b)

Figure H.1: Model independent 95% CL upper limits on $Br(H \rightarrow aa \rightarrow bbbb)$ (a) in the resolved and (b) in the merged regime of the 0-lepton channel. (logarithmic scale)



(a)



(b)

Figure H.2: Model independent 95% CL upper limits on $Br(H \rightarrow aa \rightarrow bbbb)$ (a) in the resolved and (b) in the merged regime of the 2-lepton channel. (logarithmic scale)

Bibliography

- [1] A. Djouadi, “The anatomy of electroweak symmetry breaking,” p. 1–216, Feb. 2008. [Online]. Available: <http://dx.doi.org/10.1016/j.physrep.2007.10.004>
- [2] CERN, “Origins of the Brout-Englert-Higgs Mechanism,” accessed: March 5, 2025. [Online]. Available: <https://home.cern/science/physics/origins-brout-englert-higgs-mechanism>
- [3] D. Caterina and T. Dan, “Searching for Dark Matter with the ATLAS Detector,” 2019. [Online]. Available: <https://atlas.cern/updates/feature/dark-matter>
- [4] M. Thomson, Modern Particle Physics. Cambridge University Press, 2013.
- [5] F. Halzen and A. D. Martin, Quarks and Leptons: An Introductory Course in Modern Particle Physics. Wiley, 1984.
- [6] D. Griffiths, “Introduction to Elementary Particles,” Cham, pp. 93–115, 2022. [Online]. Available: https://doi.org/10.1007/978-3-031-05625-3_5
- [7] F. Englert and R. Brout, “Broken Symmetry and the Mass of Gauge Vector Mesons,” Phys. Rev. Lett., vol. 13, p. 321, 1964. [Online]. Available: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.13.321>
- [8] D. N. Salerno, “Search for the Standard Model Higgs Boson Produced in Association with Top Quarks in the Fully Hadronic Final State at the CMS Experiment,” 2018. [Online]. Available: https://cds.cern.ch/record/2313961/files/TS2018_003_2.pdf
- [9] A. Collaboration, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC,” Phys. Lett. B, vol. 716, p. 1, 2012.
- [10] C. Collaboration, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC,” Phys. Lett. B, vol. 716, p. 30, 2012.
- [11] M. Dittmaier and S. Schumacher, “The Higgs boson in the Standard Model - From LEP to LHC: Expectations, Searches, and Discovery of a Candidate,” Progress in Particle and Nuclear Physics, vol. 70, pp. 4–7, 2013. [Online]. Available: <https://arxiv.org/abs/1211.4828>
- [12] V. A. Bednyakov, N. D. Giokaris, and A. V. Bednyakov, “On the Higgs mass generation mechanism in the Standard Model,” Physics of Particles and Nuclei, vol. 39, no. 1, p. 13–36, Jan. 2008. [Online]. Available: <http://dx.doi.org/10.1134/S1063779608010024>

- [13] Potter et al., “Handbook of LHC Higgs Cross Sections: 3. Higgs Properties: Report of the LHC Higgs Cross Section Working Group,” 2013. [Online]. Available: <http://cds.cern.ch/record/1559921>
- [14] S. L. Wu, “Brief history for the search and discovery of the Higgs particle – A personal perspective,” *Modern Physics Letters A*, vol. 29, no. 09, p. 1330027, Mar. 2014. [Online]. Available: <http://dx.doi.org/10.1142/S0217732313300279>
- [15] S. Navas et al., “Review of particle physics,” *Phys. Rev. D*, vol. 110, no. 3, p. 030001, 2024.
- [16] Cepeda et al., “Exotic Higgs Decays,” *Annual Review of Nuclear and Particle Science*, vol. 72, no. 1, p. 119–149, 2022. [Online]. Available: <http://dx.doi.org/10.1146/annurev-nucl-102319-024147>
- [17] Khachatryan et al., “Search for light bosons in decays of the 125 GeV Higgs boson in proton-proton collisions at $\sqrt{s}=8$ TeV,” *Journal of High Energy Physics*, vol. 2017, no. 10, 2017. [Online]. Available: [http://dx.doi.org/10.1007/JHEP10\(2017\)076](http://dx.doi.org/10.1007/JHEP10(2017)076)
- [18] Curtin et al., “Exotic decays of the 125 GeV Higgs boson,” *Physical Review D*, vol. 90, no. 7, 2014. [Online]. Available: <http://dx.doi.org/10.1103/PhysRevD.90.075004>
- [19] A. Breskin and R. Voss, *The CERN Large Hadron Collider: Accelerator and Experiments*. CERN Geneva, 2009.
- [20] CERN, “Facts and figures about LHC,” accessed: March 2025. [Online]. Available: <https://home.cern/resources/faqs/facts-and-figures-about-lhc>
- [21] —, “Linear Accelerator 4,” 2024. [Online]. Available: <https://home.cern/science/accelerators/linear-accelerator-4>
- [22] S. Beranek, “Model independent search for new particles in two-dimensional mass space using events with missing energy, two jets and two leptons, with the CMS detector,” Ph.D. dissertation, University of Aachen, 2017.
- [23] L. Rossi, “The High Luminosity LHC Project,” in *THE FUTURE OF OUR PHYSICS INCLUDING NEW FRONTIERS: Proceedings of the International School*. World Scientific, 2017, pp. 371–391.
- [24] C. Collaboration, “Description and performance of track and primary-vertex reconstruction with the CMS tracker,” *Journal of Instrumentation (JINST)*, vol. 9, p. P10009, 2014.
- [25] L. C. C. Collaboration), “Performance and operation of the CMS Phase 1 pixel detector,” *Nuclear Instruments and Methods in Physics Research Section A*, vol. 936, pp. 688–690, 2019.
- [26] A. Benaglia, “The CMS ECAL performance with examples,” 2014.
- [27] A. M. (CMS), “The CMS Electromagnetic Calorimeter: lessons learned during LHC Run 1, overview and future projections,” in *Proceedings of the TIPP 2014 Conference*, vol. TIPP2014, 2014, p. 029.

- [28] C. Collaboration, “Performance of the CMS hadron calorimeter with cosmic ray muons and LHC beam data,” *Journal of Instrumentation*, vol. 5, no. 03, p. T03012, 2010.
- [29] —, “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV,” *Journal of Instrumentation (JINST)*, vol. 13, p. P06015, 2018.
- [30] —, “CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger,” CERN, Technical Report, 2002.
- [31] W. Contributors, “Parton (particle physics),” *Wikipedia, The Free Encyclopedia*, accessed: March 2025.
- [32] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and Collider Physics*. Cambridge University Press, 2003.
- [33] J. M. Butterworth, G. Dissertori, and G. P. Salam, “Hard Processes in Proton-Proton Collisions at the Large Hadron Collider,” *Annual Review of Nuclear and Particle Science*, vol. 62, pp. 387–405, 2012.
- [34] “Hard scattering cross sections and parton distribution functions at the LHC,” 2013. [Online]. Available: <https://api.semanticscholar.org/CorpusID:120516526>
- [35] R. Placakyte, “Parton Distribution Functions,” 2011.
- [36] Y. V. Kovchegov and E. Levin, *Quantum Chromodynamics at High Energy*. Cambridge University Press, 2012.
- [37] R. D. Field, “The underlying event in hard scattering processes,” 2002.
- [38] T. Sjöstrand, S. Mrenna, and P. Skands, “PYTHIA 6.4 Physics and Manual,” *Journal of High Energy Physics (JHEP)*, vol. 2006, no. 05, p. 026, 2006.
- [39] S. Navas et al., “MONTE CARLO SIMULATIONS,” in *Review of Particle Physics*. Phys. Rev. D, 2024, vol. 110, no. 3, p. 030001.
- [40] A. M. S. et al. (CMS), “Particle-flow reconstruction and global event description with the CMS detector,” *Journal of Instrumentation*, vol. 12, no. 10, p. P10003, 2017.
- [41] S. K. Dooling, “Differential Cross Section Measurement of Drell-Yan Production and Associated Jets with the CMS Experiment at the LHC,” Ph.D. dissertation, U. Hamburg, Dept. Phys., 2014.
- [42] C. Collaboration, “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV,” *Journal of Instrumentation*, vol. 13, no. 05, p. P05011, 2018.
- [43] C. O. Data, “Cms physics objects,” 2024. [Online]. Available: <https://cms-opendata-workshop.github.io/workshop2024-lesson-physics-objects/04-jetmet.html>
- [44] M. D. Schwartz, “TASI Lectures on Collider Physics,” 2017. [Online]. Available: <https://arxiv.org/abs/1709.04533>

- [45] C. Collaboration, “Standard model cross sections for CMS at 13 TeV,” twiki page. [Online]. Available: <https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSectionsat13TeV>
- [46] G. Karapostoli et al., “Search for exotic decays of the Higgs boson to a pair of new light bosons in final states with four bottom quarks,” CERN, Tech. Rep. CMS AN-18-154, 2024, archive Hash: 7bbaa23-D, Archive Date: January 7, 2024.
- [47] P. D. Group, “Monte Carlo Particle Numbering Scheme,” 2024, accessed: March 5, 2025. [Online]. Available: <https://pdg.lbl.gov/2024/reviews/rpp2024-rev-monte-carlo-numbering.pdf>
- [48] Hayrapetyan et al., “Search for the decay of the Higgs boson to a pair of light pseudoscalar bosons in the final state with four bottom quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV,” Journal of High Energy Physics, no. 6, 2024. [Online]. Available: [http://dx.doi.org/10.1007/JHEP06\(2024\)097](http://dx.doi.org/10.1007/JHEP06(2024)097)
- [49] Aaboud et al., “Search for the Higgs boson produced in association with a vector boson and decaying into two spin-zero particles in the $H \rightarrow aa \rightarrow 4b$ channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector,” Journal of High Energy Physics, vol. 2018, no. 10, 2018. [Online]. Available: [http://dx.doi.org/10.1007/JHEP10\(2018\)031](http://dx.doi.org/10.1007/JHEP10(2018)031)
- [50] T. Sjöstrand, S. Ask, J. R. Christiansen et al., “An introduction to PYTHIA 8.2,” Comput. Phys. Commun., vol. 191, p. 159, 2015.
- [51] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with parton shower simulations: the POWHEG method,” Journal of High Energy Physics (JHEP), vol. 11, p. 070, 2007.
- [52] C. Collaboration, “Simplified template cross section measurements in VH, H to bb with full Run 2 data set,” CERN, Tech. Rep. CMS AN-2019/229, 2019, version 14, updated April 1, 2022.
- [53] M. Paterno and /Fermilab, “Calculating efficiencies and their uncertainties,” Fermi National Accelerator Lab. (FNAL), Batavia, IL (United States), Tech. Rep., 2004. [Online]. Available: <https://www.osti.gov/biblio/15017262>
- [54] A. Zabi, J. W. Berryhill, E. Perez, and A. D. Tapper, “The Phase-2 Upgrade of the CMS Level-1 Trigger,” 2020.
- [55] M. Cacciari, G. P. Salam, and G. Soyez, “The anti-kt jet clustering algorithm,” Journal of High Energy Physics, vol. 2008, no. 04, p. 063, 2008.
- [56] C. Collaboration, “Pileup mitigation at CMS in 13 TeV data,” Journal of Instrumentation, vol. 15, p. P09018, 2020.
- [57] D. Bertolini, P. Harris, M. Low, and N. Tran, “Pileup Per Particle Identification,” Journal of High Energy Physics, vol. 2014, no. 10, p. 059, 2014.
- [58] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, “Soft Drop,” Journal of High Energy Physics, vol. 2014, no. 05, p. 146, 2014.

- [59] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, “Towards an understanding of jet substructure,” Journal of High Energy Physics, vol. 2013, no. 09, p. 029, 2013.
- [60] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, and A. Stakia, “Jet Flavour Classification Using DeepJet,” Journal of Instrumentation, vol. 15, no. 12, p. P12012, 2020.
- [61] C. Collaboration, “Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques,” CMS DP 2020/002, 2020.
- [62] H. Qu and L. Gouskos, “Jet Tagging via Particle Clouds,” Physical Review D, vol. 101, p. 056019, 2020.
- [63] A. H. et al. (TMVA), “TMVA - Toolkit for Multivariate Data Analysis,” 2007.
- [64] W. Verkerke and D. P. Kirkby, “The RooFit toolkit for data modeling,” in eConf, vol. C0303241, 2003, p. MOLT007.
- [65] E. Gross and O. Vitells, “Wald’s Theorem and the “Asimov” Data Set,” ATLAS Statistics Forum, Dec. 2, 2009, 2009.
- [66] COMBINE, “Statistical Tests.” [Online]. Available: https://cms-analysis.github.io/HiggsAnalysisCombinedLimit/latest/what_combine_does/introduction/#statistical-tests
- [67] Navas et al., “Statistics,” in Review of Particle Physics. Phys. Rev. D, 2024, vol. 110, no. 3, p. 030001.
- [68] C. Röver, C. Messenger, and R. Prix, “Bayesian versus frequentist upper limits,” 2011. [Online]. Available: <https://arxiv.org/abs/1103.2987>
- [69] L. M. (CERN), “Statistical Software Tools RooFit/RooStats.” [Online]. Available: <https://agenda.infn.it/event/5719/contributions/60039/attachments/43297/51371/RooStatsVietri2013.pdf>
- [70] I. Narsky, “Estimation of upper limits using a Poisson statistic,” Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 450, no. 2–3, p. 444–455, 2000. [Online]. Available: [http://dx.doi.org/10.1016/S0168-9002\(00\)00305-3](http://dx.doi.org/10.1016/S0168-9002(00)00305-3)
- [71] COMBINE, “Combine Common statistical methods.” [Online]. Available: <https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/part3/commonstatsmethods/#bayesian-limits-and-credible-regions>
- [72] RooStatsWorkBook, “Tools for Statistical Tests and Inference.” [Online]. Available: <https://roostatsworkbook.readthedocs.io/en/latest/statisticaltests.html#binned-and-unbinned-models>