Measurement of the production of the standard Higgs boson associated to a pair of top and antitop quarks and search for a charged Higgs boson decaying into a pair of t and b quarks with the ATLAS detector at LHC

Xuan Yang

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

préparée dans le cadre d'une cotutelle *entre l'Université Grenoble Alpes* et Shandong University

Spécialité : **PHYSIQUE SUBATOMIQUE&ASTROPARTICULES**

Arrêté ministériel : le 6 janvier 2005 – 25 mai 2016

Présentée par

Xuan YANG

Thèse dirigée par **Johann Collot** et **Marine Kuna**
codirigée par **Cunfeng FENG**

préparée au sein des **Laboratoire de Physique Subatomique et de Cosmologie**

dans **l'école doctorale de physique de Grenoble et Institute of frontier and interdisciplinary science of Shandong University**

Mesure de la production du boson de Higgs standard en association avec une paire de quarks top et antitop et recherche d'un boson de Higgs chargé se désintégrant en une paire de quarks top et bottom à l'aide du détecteur ATLAS au LHC.

Thèse soutenue publiquement le **10 juin 2020** ,
devant le jury composé de :

Monsieur Shenjian CHEN
Professeur Nanjing University, Membre
Monsieur Zhijun LIANG
Professeur associé IHEP, Rapporteur
Madame Susan Shotkin-Gascon
Professeure IP2I Lyon, Université de Lyon, Rapporteure
Monsieur Lianliang MA
Professeur Shandong University, Membre
Monsieur Christophe Furget
Professeur LPSC, University Grenoble-Alpes, Président
Monsieur Pascal Pralavorio
Docteur CPPM Marseille, CNRS/IN2P3, Membre

Abstract

The Higgs field plays an important role in the Standard Model (SM). It is not only responsible for the mass of elementary particles but also the only scalar field in the SM. Hence the study of the Higgs boson becomes one of the most popular topics at LHC. It provides a great opportunity to probe the SM predictions and search for beyond SM evidence. New physics beyond the SM usually has a more complicated Higgs scenario and predict more than one Higgs boson, such as a charged Higgs boson. Searching for these new particles will give us a direct hint of new physics.

In this thesis, two main contributions are presented, searching for a heavy charged Higgs boson (H^+) decaying to $t\bar{b}$ and searching for $t\bar{t}H$ production xsin multilepton final states. The search for H^+ is performed with data collected by the ATLAS detector in 2015 and 2016, corresponding to an integrated luminosity of 36.1 fb^{-1}. There is no significant excess of H^+ signal over the SM backgrounds observed in the mass range from 200 GeV to 2000 GeV. The largest deviation from the SM hypothesis obtained from the fit is observed at 300 GeV, corresponding to a local p_0 value of 1.13 %.

The search for $t\bar{t}H$ production in multilepton final state is performed with data collected during 2015-2017, corresponding to an integrated luminosity of 79.9 fb^{-1}. The Yukawa coupling of the Higgs boson and top quark can be directly determined from the measurement of $t\bar{t}H$ production cross section. This analysis aims at observation of $t\bar{t}H$ production using only multilepton final states. There are six final states defined by the number of light leptons (electrons or muons) and hadronic taus, and 25 event categories to search for $t\bar{t}H$ signal and estimate fake leptons simultaneously. The studies in the final state with two same-sign leptons and one hadronic tau is discussed in detail. An excess of $t\bar{t}H$ events over other SM backgrounds is found with an observed (expected) significance of 1.8 σ (3.1 σ). The measured $t\bar{t}H$ production cross section is 294^{+182}_{-162} fb, which is consistent with the SM prediction.

Keywords: Experimental Particle physics, Standard Model, New physics, Higgs bosons, Top quark

Résumé

Introduction

Le modèle standard (SM) de la physique des particules représente notre meilleure compréhension des constituants élémentaires de la matière à l'heure actuelle. Le champ de Higgs joue un rôle unique et important dans le SM. C'est le seul champ scalaire fondamental et il est responsable de la génération de la masse des particules. Le boson de Higgs prédit par le SM a été observé par les collaborations ATLAS et CMS au LHC en 2012 [1, 2, 3]. Cependant, ceci n'est qu'un point de départ. Afin de sonder les propriétés du boson de Higgs, nous devons rechercher et mesurer le boson de Higgs dans différents modes de production et de désintégration. En particulier, la production de Higgs en association avec une paire de quarks supérieurs ($t\bar{t}H$) est intéressante car le couplage de Yukawa entre le boson de Higgs et le quark top peut être déterminé directement au niveau de l'arbre à partir de ce processus. La section efficace de $t\bar{t}H$ est très faible par rapport aux autres modes de production. Pour une collision proton-proton à $\sqrt{s} = 13$ TeV, elle ne représente qu'environ 1 % de la section efficace totale de production du boson de Higgs. Les collaborations ATLAS et CMS recherchent la production de $t\bar{t}H$ depuis le début de l'exploitation du LHC. Celle-ci été observée en 2017 par l'expérience ATLAS utilisant l'ensemble de données 2015-2016 et en combinant $H \to \gamma\gamma$ et $H \to b\bar{b}$ et la désintégration de Higgs en leptons [4]. CMS a également observé la production de $t\bar{t}H$ en combinant les données du run1 et du run2 collectées en 2016 [5].

L'une de mes principales contributions dans ma thèse est la recherche de $t\bar{t}H$dans l'état final multileptonique. Cet état final nécessite au moins deux leptons légers provenant à la fois de la désintégration du boson de Higgs et de la désintégration du quark top. Cette analyse cible principalement les modes de décroissance $H \to WW^*$, $H \to \tau\tau$ et $H \to ZZ^*$. Elle pâtit d'une précision statistique assez faible, d'une séparation difficile avec les bruits de fond irréductibles ($t\bar{t}W$) et d'une mauvaise estimation du bruit non-prompt. Dans cette analyse, ma principale contribution est:

- développement du logiciel d'analyse du groupe et production de ntuple
- optimisation de l'analyse classique par coupures et comptage dans le canal 2ℓSS+$1\tau_{\text{had}}$ (nécessitant deux leptons de même signe et un tau hadronique)
- Estimation de faux leptons dans le canal 2ℓSS+$1\tau_{\text{had}}$
- analyse statistique

Au-delà du SM, il est fort probable que le secteur de Higgs soit plus compliqué. Dans le SM, le champ scalaire se compose d'un seul doublet SU (2), ce qui est la solution la plus simple. Avec des structures de champs scalaires plus compliquées, de nouveaux de bosons de Higgs apparaissent. Par conséquent, la recherche des nouveaux bosons de Higgs est une sonde directe de la nouvelle physique au-delà du SM. La recherche d'une nouvelle physique est également une tâche importante en physique des particules, car nous savons déjà que le SM ne peut pas expliquer tous les résultats de l'expérience, par exemple la masse des neutrinos ou la matière noire. Dans ma thèse, j'ai également contribué à la recherche d'un boson de Higgs chargé se désintégrant en une paire de quarks top et bottom ($H^+ \to t\bar{b}$) dans la gamme de masse de 200 GeV à 2000 GeV en utilisant les données collectées en 2015 et 2016 par le détecteur ATLAS. Cette analyse est impactée par de

grands bruits de fond provenant du processus $t\bar{t}$, en particulier du fond irréductible $t\bar{t} + b\bar{b}$. Nous avons mis en œuvre une analyse multivariée pour extraire les événements H^+ des bruits de fond $t\bar{t}$. Nous avons également essayé de développer une méthode de reconstruction de masse par arbre de décision boosté (BDT) pour augmenter le pouvoir discriminant concernant les H^+ de faible masse. De nombreuses études ont également été réalisées pour modéliser les événements $t\bar{t} + b\bar{b}$ et comprendre la systématique. Mes contributions dans cette analyse ont été:

- comparaison données / MC dans l'état final dileptonique
- optimisation de la région du signal dans l'état final dileptonique
- études de BDT dans l'état final dileptonique
- reconstruction de masse à l'état final dileptonique

Cette thèse est organisée comme suit. Le chapitre 2 donne une introduction théorique. Le chapitre 3 présente le Grand collisionneur de hadrons et le détecteur ATLAS, ainsi que la reconstruction d'objets intéressants pour l'analyse. Le chapitre 4 décrit l'analyse de la recherche d'un boson de Higgs chargé, suivi de l'analyse multileptonique $t\bar{t}H$ dans le chapitre 5. Au final, une conclusion simple et des perspectives pour l'avenir sont données dans le chapitre 6.

Théorie

Jusqu'à présent, la théorie la plus aboutie pour décrire les éléments physiques les plus fondamentaux est le modèle standard (SM). Il décrit les particules fondamentales et leurs interactions à l'exception de la gravité. Ces particules sont classées en fermions et bosons en fonction de leur spin. Les fermions ont un spin demi-entier et constituent la matière conventionnelle dans la nature. D'un autre côté, les bosons ont un spin entier et sont les porteurs des interactions de base. Il y a trois générations dans les fermions, présentant des masses croissantes. Naturellement, seule la première génération, avec la masse la plus faible, est stable et toute la matière conventionnelle est essentiellement constituée de fermions de première génération. Les fermions de deuxième et troisième génération sont instables et se désintègrent en particules plus légères en très peu de temps. Les fermions peuvent être classés en leptons et quarks en fonction des interactions auxquelles ils participent. Les leptons ne subissent pas d'interaction forte contrairement aux quarks. L'une des caractéristiques les plus intéressantes des quarks est qu'ils portent des charges fractionnaires. Dans le SM, il existe deux types de bosons: les bosons de spin 1 (bosons vectoriels, également appelés bosons de jauge, qui sont les porteurs de force des interactions fondamentales; les bosons de spin 0, jusqu'à présent le seul boson de spin 0 connu est le boson de Higgs, qui est associé au champ de Higgs qui génère les masses de particules élémentaires. Les photons portent l'interaction électromagnétique, les bosons W et Z portent l'interaction faible et les gluons portent l'interaction forte.

Il existe trois interactions de base dans le SM: l'interaction électromagnétique, l'interaction faible et l'interaction forte. L'interaction gravitationnelle n'est actuellement pas modélisée au niveau microscopique par le SM. L'interaction électromagnétique a une portée infinie et a été très bien décrite à grande échelle par les équations de Maxwell. La théorie décrivant le champ électromagnétique à l'échelle microscopique est l'électrodynamique quantique (QED). Les interactions faible et électromagnétique sont unifiées dans la partie du SM appelée théorie électrofaible (EWT). L'interaction

forte ne se produit qu'entre les quarks que l'on trouve confinés dans les hadrons. Elle est décrite par la chromo-dynamique quantique (QCD).

Afin de générer la masse de particules, la symétrie de jauge de la théorie électrofaible doit être brisée tout en gardant le lagrangien parfaitement symétrique. Cela se fait via la rupture de symétrie spontanée (SSB). Dans le SM, un doublet de champs scalaires complexes (champ de Higgs) est introduit et brise la symétrie de jauge (mécanisme de Brout-Englert-Higgs). Il génère trois bosons de Goldstone sans masse qui sont absorbés par les bosons de jauge. Un boson scalaire massif (boson de Higgs) apparaît comme vestige de cette symétrie.

Il convient de noter que le mécanisme de Higgs dans le SM est la solution la plus simple. Il est également possible d'inclure des champs scalaires plus complexes, comme un doublet SU (2) supplémentaire (modèle à deux doublets de Higgs, 2HDM). La plupart des théories allant au-delà du modèle standard (BSM) nécessitent un scénario de Higgs plus compliqué. Par exemple, un type particulier de 2HDM est le scénario de Higgs pour le modèle standard supersymétrique minimal (MSSM). Dans les modèles 2HDM, il y a quatre bosons de Higgs massifs prévus en plus du boson de Higgs du SM: deux bosons de Higgs chargés, un bosons de Higgs neutre et un bosons de Higgs pseudo-scalaire. Le boson de Higgs chargé est une caractéristique importante de cette théorie. Il se désintègre généralement en une paire de quarks supérieurs et inférieurs. La recherche du boson de Higgs chargé est une sonde directe de recherche de nouveaux phénomènes physiques.

Le LHC et Le détecteur ATLAS

Avec une énergie dans le centre de masse nominale de 14 TeV, le grand collisionneur de hadrons le LHC est le collisionneur proton-proton le plus énergétique et une opportunité formidable d'étudier la physique subatomique. L'accélérateur du CERN (organisation européenne pour la recherche nucléaire) est situé sur la frontière franco-suisse dans un tunnel sous-terrain à une profondeur de 50 à 175 mètres.

Quatre détecteurs sont placés sur le LHC : A Toroidal LHC ApparatuS (ATLAS) et Compact Muon Solenoid (CMS) sont des expériences généralistes tandis que LHCb et ALICE sont destinés respectivement à la physique des hadrons beaux et l'étude des plasmas de quarks et de gluons. Le tunnel du LHC est un héritage du grand collisionneur électrons-positrons (LEP) en opération de 1989 à 2000. Le collisionneur a une circonférence de 27km et un gradient horizontal de 1.4 %. Comparativement aux leptons, les hadrons ont l'avantage de perdre très peu d'énergie par radiation synchrotron dans les accélérateurs circulaires grâce à un ratio de masse favorable. Les premières collisions proton-proton du LHC eurent lieu en Novembre 2009 avec une énergie dans le centre de masse de 900 GeV, puis ont augmenté à 2.37 TeV en quelques jours. La première prise de données à $\sqrt{s} = 7$ TeV a commencé en 2011 pour être augmentée à $\sqrt{s} = 8$ TeV en 2012. Pendant cette période appelée « Run 1 », le LHC a délivré une luminosité intégrée de 28 fb^{-1}. Après trois ans de mise à jour du collisionneur, la prise de donnée du Run 2 a eu lieu de 2015 à fin 2018 avec une énergie dans le centre de masse de 13 TeV délivrant une luminosité intégrée de 150 fb^{-1}.

Le détecteur ATLAS est l'une des expériences généralistes auprès du LHC. L'un de ses objectifs les plus importants a été la recherche du boson de Higgs et de nouvelle physique à l'échelle du Teraélectronvolt, ainsi que l'approfondissement de la compréhension des processus QCD et électrofaibles, ainsi que la physique de la saveur. Le détecteur est le plus grand jamais construit au

monde avec une longueur de 44 mètres, une hauteur de 25 mètres et une masse de plus de 7000 tonnes. Sa structure est en couche avec une géométrie cylindrique et une très bonne hermiticité. ATLAS comporte plusieurs sous détecteurs : en partant du centre (le plus proche du tube à vide du LHC) vers l'extérieur, on trouve un trajectographe interne (ID) pour mesurer l'impulsion des traces chargées, suivi par les calorimètres électromagnétiques et hadroniques pour mesurer l'énergie déposée par les particules pour finir par le spectromètre à muons (MS) qui permet de déclencher sur les muons et de reconstruire leurs traces.

Recherché la production de bosons de Higgs chargés

Après la découverte du boson de Higgs en 2012, la question est de savoir s'il s'agit de la seule particule scalaire fondamentale ou si celle-ci pourrait être la première manifestation d'un secteur de Higgs plus riche allant au-delà du modèle standard (BSM) . Presque tous les scénarios BSM Higgs contiennent plus d'un boson de Higgs parmi lesquels les bosons Higgs chargés sont généralement présents. Le boson de Higgs chargé est prédit par plusieurs modèles, parmi toutes les théories, le modèle à deux doublets de Higgs (2HDM) est très populaire. Un type particulier de 2HDM, à savoir le type II 2HDM, correspond au secteur de Higgs du modèle standard supersymétrique minimal (MSSM). Pour un boson de Higgs chargé de grande masse (plus lourd que le quark top), le principal mode de production au LHC est associé à un quark top, $gg \rightarrow tbH^+$ (appelé schéma à 4 saveurs ou 4FS) ou $gb \rightarrow tH^+$ (appelé schéma à 5 saveurs ou 5FS). La production et la désintégration de H^+ sont également contrôlées par deux paramètres: le rapport des valeurs moyennes dans le vide des deux doublets de Higgs (tan β) et l'angle de mélange entre les bosons de Higgs CP-pairs (α). Pour un boson de Higgs chargé de grande masse et $cos(\beta - \alpha) \approx 0$, la désintégration du H^+ est dominée par $H^+ \rightarrow tb$.

Dans cette analyse, j'ai recherché la production de bosons de Higgs chargés lourds à travers $gg \rightarrow tbH^+$ et H^+ se désintégrant en $H^+ \rightarrow tb$, entre 200 GeV et 2000 GeV H^+, en utilisant les données collectées par le détecteur ATLAS en 2015 et 2016, correspondant à une luminosité intégrale de 36,1 fb^{-1}. Les événements de signal et de fond sont estimés par des échantillons de simulation. Pour les événements du signal, deux canaux principaux ont été considérés, selon que le quark top se désintègre semi-leptoniquement ($t\ toWb \rightarrow lvb$) ou hadroniquement ($t \rightarrow Wb \rightarrow jjb$): canal lepton-plus-jets (lorsque l'un des quarks top se désintègre semi-leptoniquement et l'autre se désintègre de façon hadronique) et canal di-lepton (lorsque les deux quarks top se désintègrent semi-leptoniquement). Dans la suite, nous ne traiterons que le canal di-lepton.

La sélection du canal di-lepton repose essentiellement sur le fait qu'il y a exactement deux leptons chargés avec une charge opposée, puisque les deux leptons chargés proviennent des deux quarks top qui ont une charge opposée. La quantité de mouvement transverse du lepton dominant doit être supérieure à 27 GeV et pour l'autre lepton, son p_T doit être supérieure à 10 GeV. S'agissant des événements ee, le lepton sous-dominant doit satisfaire p_T plus grand que 15 GeV. Pour les événements de saveur identique (ee ou $\mu\mu$), la masse invariante de la paire de leptons ne doit pas être dans la fenêtre de masse Z (83 GeV à 99 GeV) et doit être supérieure à 15 GeV, pour supprimer les événements $Z + jets$ et Drell-Yan. De plus, les événements doivent contenir au moins 2 jets étiquetés comme des quarks b (70 % WP) avec p_T supérieur à 25 GeV.

Les événements passant la présélection seront divisés en régions de contrôle (CR) et régions de signal (SR) en fonction de la pureté du signal, de la multiplicité en jets et de la multiplicité en jets étiquetés quark b. Les régions de signal sont $3j3b$, $\leq 4j3b$ et $\leq 4j \leq 4b$ avec $3j2b$ et $\leq 4j2b$ comme régions de contrôle. Une région est identifiée CR ou SR en fonction de sa signification statistique ($S/\sqrt{B}$).

Le bruit de fond le plus important est le ttbar + jets, en particulier le ttbar accompagné de deux jets b supplémentaires. Il a les mêmes états finaux que le signal H^+ et il est difficile d'obtenir une bonne évaluation par simulation. Plusieurs pondérations basées sur des variables cinématiques sont appliquées aux événements ttbb pour obtenir une meilleure modélisation. Ce bruit est également laissé flottant dans l'ajustement final pour qu'il puisse mieux reproduire les données réelles. D'autres bruits de fond mieux connus sont directement tirés de la simulation.

Pour extraire le signal du bruit de fond, la méthode à Arbre de Décision Boostée (BDT) est implémentée dans l'analyse. Les BDT sont entraînés séparément sur tous les échantillons H^+ et toutes les SR contre les bruits de fond ttbar uniquement, car ce sont le processus dominants dans la plupart des cas. Les variables d'entrée comprennent les variables de forme d'événement, les variables cinématiques globales et les propriétés cinématiques des combinaisons d'objets du détecteur sélectionnés. Les variables d'entrée sont sélectionnées et optimisées pour chaque point de masse séparément. Concernant les variables corrélées hautement linéaires, seule celle ayant le meilleur pouvoir de séparation est conservée. Une dizaine de variables sont sélectionnées pour chaque point de masse.

Afin d'améliorer la séparation entre le signal H^+ et le bruit de fond ttbb, nous avons essayé de mettre en œuvre une reconstruction de masse de l'état final di-lepton. En principe, ceci n'est pas faisable car l'état final comporte deux neutrinos. Nous avons essayé de mettre en Œœuvre la méthode BDT associée à une correspondance d'identité donnée par la simulation. Les jets dans l'état final sont appariés aux partons simulés si leur ΔR <0,4. De cette façon, nous pouvons reconstruire l'événement et le considérer comme un signal pour dans le processus de reconstruction. D'un autre côté, nous pouvons supposer d'autres combinaisons entre les jets reconstruits et les partons vrais et les considérer comme du bruit dans le processus de reconstruction. Ensuite, un BDT est formé entre eux. Lors de son application dans l'analyse, chaque combinaison possible obtient un score de BDT et celle dont le score se rapproche le plus de 1 est considérée comme la reconstruction correcte. Cette méthode améliore légèrement nos résultats, mais elle n'est pas incluse dans le résultat final en raison de son intense consommation en calcul. Nous avons recherché le boson de Higgs chargé de 200 GeV à 2000 GeV et aucun excès significatif n'a été observé. La plus grande déviation de l'hypothèse SM obtenue à partir de l'ajustement est observée à 300 GeV, correspondant à une valeur locale p_0 de 1,13 %.

Recherché la production de $t\bar{t}H$

Les couplages entre le boson de Higgs et les autres particules du Modèle Standard (MS) sont bien prédits par la théorie, et leur mesure précise est un moyen de chercher de la nouvelle physique au-delà du MS. Dans le MS, le couplage des fermions au boson de Higgs est proportionnel à leur masse. Par conséquence, parmi toutes les particules élémentaires, le quark top est celle qui présente le plus fort de ces couplages ce qui en fait un paramètre clé du MS. Expérimentalement, le couplage

peut être déterminé de manière indirecte, à travers la boucle de top, par la mesure de la section efficace de production du boson de Higgs par fusion de gluons, ou par la production associée du boson de Higgs avec une paire de quarks top qui est un processus du premier ordre. Le couplage est attendu avec une intensité de l'ordre de l'unité et sensible à de potentiels effets de nouvelle physique qui peuvent modifier sa valeur prédite par le MS. La production ttH peut être observée à travers différentes topologies d'événements suivant le mode de désintégration des quarks top et du boson de Higgs. Les états finaux les plus pertinents incluent ceux où le boson de Higgs se désintègre en paire de quarks beaux, en deux photons, ou en paire de bosons W, Z, ou de quark tau, présentant un nombre suffisant de leptons dans leur état final, un ensemble de signatures appelées multileptons (ML). Ces canaux sont relativement limités par la statistique, mais peuvent atteindre une bonne sensibilité en utilisant les signatures claires de leptons pour supprimer les bruits de fonds des autres processus du MS. Différents canaux sont construits selon le nombre de leptons légers reconstruits (électrons, muons) et le nombre de tau se désintégrant en hadrons (τ^{had}). Le sujet de cette thèse est l'étude du canal $2\ell SS + \tau^{had}$ présentant deux électrons ou muons de même signe et un tau de décroissance hadronique. En dehors des leptons, au moins quatre jets dont un étiqueté beau sont demandés pour qu'un événement entre dans la sélection.

L'analyse présentée dans cette thèse utilise 80 fb^{-1} de données de collisions proton-proton enregistrées par le détecteur ATLAS avec $\sqrt{s} = 13$ TeV de 2015 à 2017. Le signal et le bruit de fond irréductible sont estimés à partir de la simulation avec des corrections issues des données. Les bruits de fonds réductibles issus de leptons non prompts ou d'erreur de reconstruction de la charge sont estimés à partir des données. Les leptons non prompts sont principalement issus des désintégrations semi-leptoniques de hadrons beaux, les électrons non prompts peuvent également être issus de conversions de photons. Les tau non prompt viennent principalement de jets mal identifiés.

Trois méthodes ont été développés pour estimer le bruit de fond des leptons non prompts à partir des données : Fake Factor (FF), la Méthode de la Matrice (MM) et le Template Fit (TF). Dans le résultats final, la méthode de TF est utilisée comme valeur centrale et FF comme incertitude systématique. Dans la méthode de Template Fit, des facteurs de normalisation sont assignés à chacune des différentes sources de leptons non-prompts, puis déterminés par un ajustement simultané de plusieurs régions de contrôle construites pour les contraindre. La forme de la contribution non prompt est prise des simulations. Les leptons tau non prompts sont estimés dans la région de contrôle avec deux leptons de signe opposé et un tau hadronique où ils sont dominants.

Deux méthodes sont développées pour extraire l'intensité du signal recherché. L'une de ces approches est une analyse multivariée de type arbre de décision ou Boosted Decision Tree (BDT). Compte-tenu de la faible statistique de données disponibles pour le canal $2\ell SS + \tau^{had}$, j'ai développé une méthode de catégorisation où la région de signal (SR) est divisée entre plusieurs catégories enrichies en signal grâce à des coupures sur des variables discriminantes. L'optimisation des variables et des coupures a été faite ainsi : une boucle sur les variables les plus discriminantes est effectuée pour trouver leur valeur de coupure optimale basée sur la significance. Les dix meilleurs variables sont conservées et combinées par deux, trois ou quatre, pour ne conserver que la meilleure combinaison basé sur les résultats d'un ajustement sur un set de données asimov. Une fois la combinaison

de variables choisies, les catégories avec peu de statistique et peu de signal sont fusionnées. Au final, la méthode de la catégorisation atteint une sensibilité similaire à celle du BDT.

Pour choisir la stratégie de l'analyse $2\ell SS + \tau^{had}$, j'ai effectué et comparé les analyses statistiques complètes combinant les différentes méthodes d'estimation des leptons non prompts (FF, MM, TF) ainsi que les différentes méthodes d'analyses (BDT, catégorisation, une catégories inclusive), et comparé les résultats. En l'absence de différence statistique significative parmi ces options, la méthode du Template Fit a été choisie pour être la plus robuste vis-à-vis des différentes sources de non prompt, ainsi que l'option de ne conserver qu'une catégorie pour augmenter la robustesse de l'analyse.

La valeur mesurée par ajustement aux données indique une intensité du signal normalisée à la valeur attendue par le SM de $\mu_{ttH} = 0.58^{+0.36}_{-0.33}$ pour une section efficace de $\sigma_{ttH} = 294^{+182}_{-162}$ fb ce qui est compatible avec la prédiction du MS ($\sigma^{SM}_{ttH} = 507^{+35}_{-50}$ fb). La significance observée (attendue) est de 1.8 (3.1) σ. Dans ce résultats, la mesure du contribution du bruit de fond irréductible ttW est supérieure aux prédictions théoriques actuelles du modèle standard. En combinant plusieurs états finaux ttH, dont ttH en multileptons, ttH(bb), ttH($\gamma\gamma$) and ttH(ZZ→4l), ATLAS a maintenant observé le mode de production ttH avec une significance observée (attendue) de 5.8 (4.9) σ.

Conclusion

La découverte du boson de Higgs en 2012 confirme le modèle standard de la physique des particules. Désormais, la recherche et les mesures de ses différents modes de production et de désintégration, ainsi que la recherche de nouvelles particules apparaissent comme les prochaines étapes. Dans cette thèse, des études ciblant l'observation de la production associée $t\bar{t}H$, ainsi que la recherche d'un boson de Higgs chargé de grande masse, le tout dans des canaux multileptoniques, sont présentées.

La recherche d'un boson de Higgs chargé de grande masse est présentée au chapitre 4. Les événements sont classés dans des régions de signal ou de contrôle selon les multiplicités de jet et de b-jet afin obtenir une sensibilité optimale. Des méthodes multivariées sont utilisées pour distinguer le signal H^+ des bruits de fond. Le bruit de fond dominant, $t\bar{t} + b\bar{b}$, est irréductible et assez mal modélisé par simulation. Par conséquent, les événements $t\bar{t}$ sont classés en $t\bar{t} + ljets$, $t\bar{t} + c\bar{c}$ et $t\bar{t} + b\bar{b}$ en fonction de la saveur des jets supplémentaires. Les contributions au bruit $t\bar{t} + c\bar{c}$ et $t\bar{t} + b\bar{b}$ sont laissés flottantes dans l'ajustement. Une pondération supplémentaire prenant SHERPA 2.1.1 comme référence est appliquée aux événements $t\bar{t} + b\bar{b}$ pour obtenir une meilleure modélisation. Une méthode de reconstruction de masse basée sur un arbre de décision boosté est développée pour améliorer la sélectivité du signal H^+ de faible masse. Cette méthode montre une amélioration de la sensibilité, mais n'a pas été utilisée dans l'analyse en raison du temps de traitement informatique supplémentaire jugé excessif. La recherche est effectuée sur 18 points de masse, de 200 GeV à 2000 GeV. Aucun excès significatif n'est observé ce qui permet de placer des limites supérieures sur les sections efficaces de production. L'espace des paramètres est également contraint.

La recherche de la production de $t\bar{t}H$ dans les états finaux multileptoniques est présentée au chapitre 5. Il existe plusieurs sous-canaux selon le nombre de leptons légers et de taus hadroniques. Une méthode par ajustement sur modèle permet d'estimer les faux leptons et la contribution $t\bar{t}W$.

Des méthodes multivariées sont utilisées pour extraire le signal $t\bar{t}H$. Ma contribution se concentre sur l'état final 2ℓSS+$1\tau_{\text{had}}$. Une méthode classique par coupures et comptage est développée et optimisée comme approche alternative. La sensibilité atteint essentiellement le même niveau que la méthode par arbre de décision boosté. Une méthode matricielle est également appliquée aux 2ℓSS+$1\tau_{\text{had}}$ afin de confirmer l'estimation des faux leptons. Une nouvelle étude des erreurs systématiques prenant en compte les écarts d'estimation des faux leptons est utilisée, dans le but d'améliorer leur règle de somme et pour prendre en compte l'effet de forme. Enfin, un ajustement simultané sur tous les sous-canaux est effectué pour extraire la force du signal. Le résultat s'avère être :

$$\hat{\mu} = 0.58^{+0.26}_{-0.25}\ (\text{stat.})^{+0.19}_{-0.15}\ (\text{exp.})^{+0.13}_{-0.11}\ (\text{bkg. th.})^{+0.08}_{-0.07}\ (\text{sig. th.}) = 0.58^{+0.36}_{-0.33}. \quad (1)$$

La section efficace du processus $t\bar{t}H$ mesurée est en accord avec la prédiction du modèle standard.

Le run2 du LHC s'est achevé fin 2018. Le total des données pp collectées pendant cette période correspond à une luminosité intégrée de 140 fb^{-1}. La prochaine période de prise de données (run3) commencera en 2021, puis le LHC passera dans sa phase à haute luminosité (HL-LHC) en 2026. J'ai également participé au développement du nouveau détecteur interne d'ATLAS (projet ITk), conçu pour remplacer l'actuel trajectomètre dans le cadre de mon travail d'intérêt collectif destiné à devenir auteur des publications d'ATLAS. Ma tâche a consisté à développer un outil de débogage afin de vérifier la compatibilité de la simulation du détecteur Geant4. HL-HLC collectera jusqu'à 3000 fb^{-1} de données de collisions proton-proton. Cela nous permettra de mesurer avec précision les propriétés du boson de Higgs et de rechercher certaines de ses désintégrations rares ($\mu\mu$, $Z\gamma$, $c\bar{c}$).

Toute propriété mesurée du boson de Higgs qui s'écarterait des prévisions du SM sera un indice d'une nouvelle physique. D'un autre côté, il y a encore de la place pour la recherche directe de nouvelles particules. Presque toutes les théories BSM incluent des bosons de Higgs supplémentaires, en particulier des bosons de Higgs chargés. L'observation de ces nouvelles particules scalaires serait une preuve directe d'une nouvelle physique. Nous pouvons également améliorer notre compréhension de la simulation et des erreurs systématiques pour obtenir de meilleurs résultats. Le développement de nouvelles méthodes d'analyse, telles que les techniques d'apprentissage automatique, est également un moyen d'augmenter la sensibilité à l'avenir. Tous ces efforts nous aideront à mieux comprendre notre monde.

Acknowledgements

The long journey of doctoral study seems to be coming to an end when I am writing the acknowledgements. This is already the sixth year. I made many happy memories with my supervisors, classmates and friends during this long period, but I also had hard and lost time. All these experiences have made me more mature and let me have confidence for my new life in the future.

Firstly, I would like to thank my supervisor, Prof.Cunfeng FENG of Shandong University, for his great guide and support of my research life. He led me into the field of experimental particle physics, taught me the basic skills and knowledge of research. I am fortunate to have such a kind supervisor. And I should also say thanks to my co-supervisor, M.Johann Collot of University Grenoble-Alpes in France, for his kind and meticulous help during my study based at Grenoble. I spent almost three years working at LPSC. He helped me to pass through the most difficult time when I just started living in a new country. And for sure he is very experienced in science, he gives me lots of useful suggestions. I would like to thank my third supervisor, Mme. Marine Kuna, who guided me to finish the ttH multilepton analysis. She is very optimistic, and always full of enthusiasm for work. She helped me a lot to improve my skills on research, language as well as presentation.

Thanks to all the great members of Charged Higgs and ttH multilepton analysis group. Physics analysis is really depending on team-work. I cannot finish my studies without collaborations from other people. Especially I would like to thank Tamara Vazquez Schroeder and Richard Polifka, who were the convenors of the ttH multilepton analysis. I will never forget the supports I got from them when I met hard times. And I should also thank Zhi Zheng and Zhi Li, who were also working on ttH multilepton. They gave me lots of brilliant ideas on the analysis methods.

Thanks to Ye Tian, Rui Yuan and Yanhong Yu, who are my collegues and friends in Shandong University. We spent almost six years together. Those happy memories we had seem to be just in front of me. I should also thank Feng Sun, who has been one of my best friends since undergraduate. He helped on programming very much.

Finally I should thank my parents and all the other family members who support me all the time. During my doctoral study, I have met lots of people. I cannot thank them one by one, but I will never forget the supports that I received.

Contents

Chapter 1

Introduction

The Standard Model (SM) represents our best understanding of the elementary constituents of matter at the moment. The Higgs field plays a unique and important role in the SM. It is the only fundamental scalar field and it is responsible for the particle mass generation. The Higgs boson predicted by the SM has been observed by the ATLAS and CMS collaborations at LHC in 2012 [1, 2, 3]. However this is just the starting point, in order to probe the properties of the Higgs boson, we have to search and measure the Higgs boson in different production and decay modes. In particular, the Higgs production in association with a pair of top quarks ($t\bar{t}H$) is interesting since the Yukawa coupling between the Higgs boson and the top quark can be determined directly at tree level from this process. The cross section of $t\bar{t}H$ is very small compared to other production modes. For proton-proton collision at $\sqrt{s} = 13$ TeV, it only accounts for about 1 % of the total Higgs production cross section. The ATLAS and CMS collaborations have been looking for $t\bar{t}H$ production since the beginning of the LHC operation and it was observed in 2017 by the ATLAS experiment using 2015-2016 dataset and combining $H \to b\bar{b}$, $H \to \gamma\gamma$ and Higgs decays to leptons [4]. CMS also observed $t\bar{t}H$ production by combining run1 data and run2 data collected in 2016 [5]. One of my main contributions in my thesis is the search for $t\bar{t}H$ in the multilepton final state. Multilepton final state requires at least two light leptons coming from both Higgs decay and top quark decay. This analysis mainly targets at $H \to WW^*$, $H \to \tau\tau$ and $H \to ZZ^*$ decay modes. It suffers from low statistics, from a difficult distinction to irreducible backgrounds ($t\bar{t}W$) and a complicated estimate of non-prompt backgrounds. My main contributions were:

- Group analysis framework development and ntuple production
- Cut-and-count analysis optimization in 2ℓSS+$1\tau_{\text{had}}$ channel (requiring two same-sign leptons and one hadronic tau)
- Fake lepton estimate in 2ℓSS+$1\tau_{\text{had}}$ channel
- Statistical analysis

Looking beyond the SM, there is also the possibility that the Higgs sector can be more complicated. In the SM, the scalar field consists of only one SU(2) doublet, which is the simplest solution. With more complicated scalar field structures, there are more bosons predicted in addition to the observed Higgs boson. Therefore looking for the new Higgs bosons is a direct probe of new physics beyond the SM. Searching for new physics is also an important task in particle physics, since we already know that the SM cannot explain all the experiment results, for example the neutrino mass

and dark matter. I also contributed to the search for a charged Higgs boson decaying to bottom and top quarks ($H^+ \to t\bar{b}$) in my thesis. This analysis was developed to search for a heavy charged Higgs boson in the 200 GeV to 2000 GeV mass range using data collected in 2015 and 2016 by the ATLAS detector. It suffers from the large backgrounds induced by $t\bar{t}$ process, especially the irreducible $t\bar{t} + b\bar{b}$ background. We took the advantage of multivariate analysis to extract the H^+ events from $t\bar{t}$ backgrounds. We also tried to develop a mass reconstruction method to increase the discriminant power for low mass H^+. Lot of studies were performed to model the $t\bar{t} + b\bar{b}$ events and understand the systematics. My contributions in this analysis were:

- Data/MC comparison in dilepton final state
- signal region optimization in dilepton final state
- BDT studies in dilepton final state
- Mass reconstruction in dilepton final state

This thesis is organized as follows. Chapter 2 gives an introduction to the theoretical basis. Chapter 3 introduces the Large Hadron Collider and the ATLAS detector, as well as the objects reconstruction. Chapter 4 describes the analysis deployed to search for a charged Higgs boson, followed by the $t\bar{t}H$ multilepton analysis in chapter 5. In the end, a simple conclusion and outlook for the future is given in chapter 6.

Chapter 2

Theory of scalar fields

The fundamental elements and structure of matter have been one of the most attracting questions of the nature since the beginning of the history of human being. Many great physicists have thought about this puzzle and proposed various theories. For a long time, atoms were regarded as the fundamental element of matter. But as the development of experimental techniques, electron was observed and then confirmed as being part of an atom around the beginning of the 20th century. Following this, our understanding of nature has progressed rapidly. So far, we have built a working theory, namely the Standard Model (SM), that describes the "fundamental" elements of matter and their interactions very well. It has been tested by many experiments and so far its predictions agree with observations of the microcosm. But observations in the microcosm (dark matter, dark energy, baryon asymmetry, etc) let us believe that it is not a perfect theory and not the "final" theory we are looking for. Several new theories beyond the SM have been proposed and LHC provides a perfect environment to test them.

In this chapter, some theory basics will be discussed. The Standard Model will be discussed in section 2.1, and the theory beyond Standard Model which is related to my thesis will be discussed in section 2.2.

2.1 The Standard Model and Higgs Mechanism

2.1.1 Introduction

The SM describes the elementary particles and their interactions with the exception of gravity. All the particles of the SM are presented in figure 2.1. These particles are categorized into fermions and bosons according to their spin. Fermions have half-integer spin and they make up the conventional matter in nature. On the other hand, bosons have integer spin and they are the carriers of the basic interactions.

There are three generations in the fermions, with increasing masses. Naturally only the first generation, with the lowest mass, is stable and all the conventional matter consists of the first generation fermions. The second and third generation fermions are unstable and decay into lighter particles in a very short time. Fermions can be further categorized into leptons and quarks according to the interactions they participate in.

Leptons do not undergo strong interaction. Classified by charges, there are two sets of leptons: charged leptons (e, μ, τ) and neutral leptons (neutrinos) (ν_e, ν_μ, ν_τ). The three generations are also called flavors, so that each flavor of charged lepton has a corresponding neutrino. The mass of

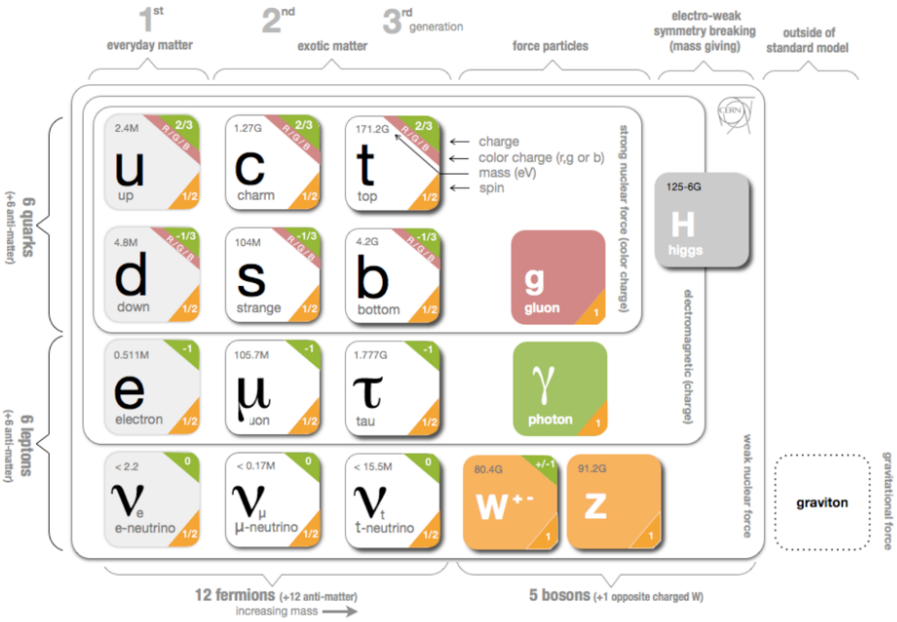


FIGURE 2.1: Elementary particles predicted in the Standard Model.

neutrinos was considered to be zero in the past, but results from neutrino experiments showed that they do have very small masses (sub eV scale). Electron is very light but stable, while μ is much heavier but still has a long lifetime compared to other unstable particles. Thus μ usually can travel through a long distance before it decays into electron, so that we can detect it in the detectors. τ lepton is very heavy (GeV scale) thus it decays almost immediately after it is produced. The basic information of leptons is summarized in table 2.1.

	Mass (MeV)	Q(e)
e	0.51	-1
μ	105.65	-1
τ	1776.84	-1
ν_e	< 0.2 eV	0
ν_μ	< 0.2 eV	0
ν_τ	< 0.2 eV	0

TABLE 2.1: Leptons in the SM with corresponding mass and electric charge [6].

Quarks participate in strong interaction. One of the most interesting realities of quarks is that they carry fractional charges. According to their charges, there are two sets of quarks: with 2/3 |e| (up (u), charm (c), top (t)), with -1/3 |e| (down (d), strange (s), bottom (b)). The type of quarks is also referred to as flavor. Another different intrinsic property they have is the color charge, which is necessary to describe the strong interaction. Each flavor of quarks can have three different colors: red (R), green (G), blue (B), which is analogous to the primary colors in nature. Because of the phenomenon known as color confinement, all observable particles should have white color (eg. consists of three colors (RGB) or consists of color-anticolor ($R\bar{R}$)), quarks can never be directly found in isolation. In principle, only compound states of quarks (hadrons) can be observed. They

consist of baryons (mode of three quarks) and mesons (mode of a pair of quark and anti-quark). The basic information of quarks is summarized in table 2.2.

	Mass (MeV)	Q(e)
u	$2.16^{+0.49}_{-0.26}$	2/3
d	$4.67^{+0.48}_{-0.17}$	-1/3
c	1270 ± 20	2/3
s	93^{+11}_{-5}	-1/3
t	172900 ± 400	2/3
b	4180^{+30}_{-20}	-1/3

TABLE 2.2: Quarks in the SM with corresponding mass and electric charge [6].

In the SM, there are two boson types: spin 1 bosons (vector bosons, also called gauge bosons, which are the force carriers of the fundamental interactions; spin 0 bosons, so far the only known spin 0 boson is the Higgs boson, which is associated to the Higgs field that generates the elementary particle masses. Photons carry the electromagnetic interaction, W and Z bosons carry the weak interaction and gluons carry the strong interaction. Since quarks have color charges, to carry the strong interaction, gluons are bi-color charged, with one unit of color and one unit of a different anti-color. Therefore there are eight (2^3) different gluons in nature. All the massive particles interact with the Higgs boson to get mass (only photons and gluons are considered to be massless in the SM).

Force (interaction) is very normal in our daily life. The fundamental interactions are categorized into four different types in the SM:

- Electromagnetic interaction: all electric charged particles participate in
- Weak interaction: all particles participate in
- Strong interaction: only quarks participate in
- Gravity interaction: not included in the SM yet, will not be discussed here

Some basic information about the fundamental interactions are listed in table 2.3. Electromagnetic interaction has an infinite range and has been described by Maxwell's equations very well at large scale. The theory describing the electromagnetic field at micro scale is Quantum electrodynamics (QED). The weak interaction was first observed from the β decay of nuclei, and then it was combined with the electromagnetic interaction since they are indistinguishable at high energy. A theory called Electroweak theory (EWT) has been developed to describe both of them. Note that only in weak interaction quarks and leptons can change their flavor by emitting or absorbing a W boson therefore also change their charge meanwhile. Also weak interaction is the only one that can violate C, P and CP symmetries (CPT symmetry is considered to be conserved also in weak interaction). The strong interaction only happens between quarks and bind them into hadrons. It is described by Quantum Chromo Dynamics (QCD) theory. In this theory, the strength of interactions becomes weaker as the distance scale decreases (the corresponding energy scale increases), known as asymptotic freedom. Therefore perturbation calculation can only be performed at high energy scale where the interaction is weak enough. The calculation at low energy scale (e.g. proton) remains a difficult task. The tree level interactions of SM particles are presented in figure 2.2.

Interaction	Relative strength	Range (meter)
Strong	1	10^{-15}
Electromagnetic	1/137	∞
Weak	10^{-6}	10^{-18}
Gravity	10^{-39}	∞

TABLE 2.3: The four fundamental interactions in nature.

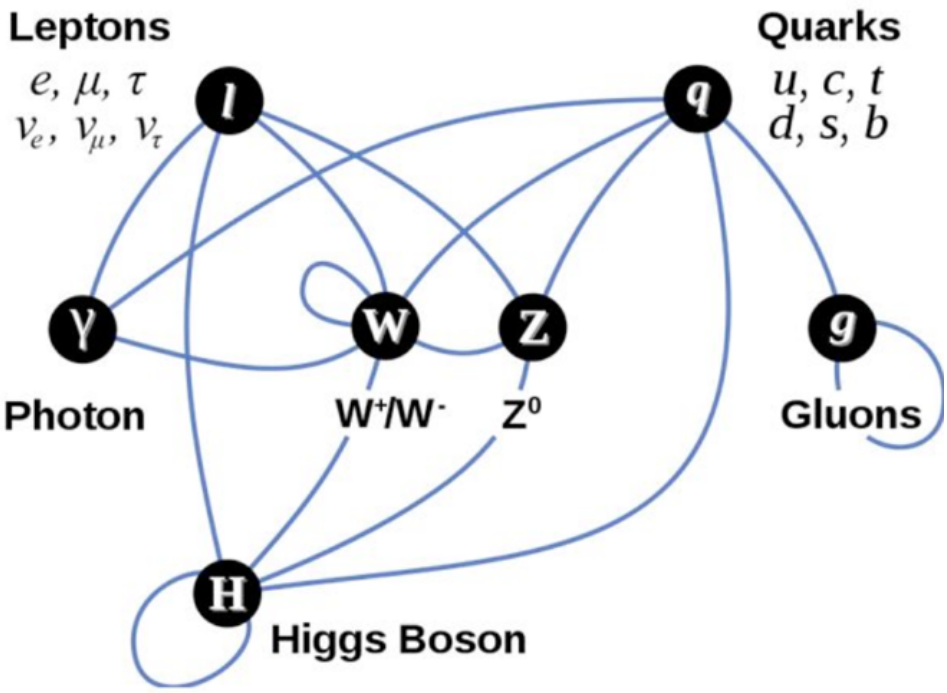


FIGURE 2.2: The tree level couplings in the SM.

2.1.2 Theory of the Standard Model

The SM is a quantum field theory. Field is the basic language to describe modern particle physics. Every particle in the SM corresponds to a field. In this field theory, different configurations of the fields can result in the same observable quantities. The transformation from one configuration to another is called gauge transformation. The invariance while transforming one configuration to another is called gauge invariance. From theory, any invariance during the transforming of field corresponds to a symmetry, therefore the gauge invariance is also called gauge symmetry. A theory that includes the gauge symmetry is considered to be a gauge theory [7]. The SM is built based on Gauge symmetries. The Lagrangian of fields in the SM is left invariant under gauge transformations.

2.1.2.1 Quantum electrodynamics

In the QED theory, the particle is described by a local field $\phi(x)$, where x is the space-time coordinate. Then the Lagrangian which describes its properties and interactions with fields is defined as:

$$\mathcal{L}(\phi, \partial_\mu \phi). \tag{2.1}$$

Applying the principle of the least action, we can achieve the Euler-Lagrange equation:

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)}\right) - \frac{\partial \mathcal{L}}{\partial \phi} = 0. \tag{2.2}$$

A free fermion with mass m can be described by a Lagrangian of a Dirac field:

$$\mathcal{L}_{Dirac} = \bar{\psi}(x)(i\gamma^\mu \partial_\mu - m)\psi(x), \tag{2.3}$$

where $\psi(x)$ is the spinor field representing the fermions, $\bar{\psi}(x) = \psi(x)^\dagger \gamma^0$ is the adjoint and γ^μ are the Dirac matrices. We can get the Dirac motion equation for a free fermion by applying the Euler-Lagrange equation 2.2 to the Lagrangian 2.3:

$$(i\gamma^\mu \partial_\mu - m)\psi(x) = 0. \tag{2.4}$$

It is obvious that the Dirac Lagrange 2.3 remains invariant under the global transformation of U(1) group:

$$\psi(x) \to U\psi(x) = e^{i\theta}\psi(x), \tag{2.5}$$

where θ is an arbitrary real constant. But the Lagrangian is not invariant under a local gauge transformation, that means θ is a function of the space-time coordinate x^μ. Thus we need to replace the derivative ∂_μ by the covariant derivative D_μ in order to keep it invariant under the local gauge transformation. This will also introduce a gauge vector field A_μ:

$$D_\mu \psi(x) \to e^{i\theta(x)} D_\mu \psi(x), \tag{2.6}$$

$$D_\mu \equiv \partial_\mu + iqA_\mu(x). \tag{2.7}$$

By doing the same thing in the Dirac Lagrangian 2.3, we can obtain the invariant Lagrangian:

$$\begin{aligned}\mathcal{L} &= \bar{\psi}(x)(i\gamma^\mu D_\mu - m)\psi(x) \\ &= \mathcal{L}_{Dirac} - q\bar{\psi}(x)\gamma^\mu A_\mu \psi(x).\end{aligned} \tag{2.8}$$

The second term represents the interaction between a fermion with electric charge q and the photon field A_μ. Notice that the photon also has kinetic energy. Thus the full QED Lagrangian should be:

$$\mathcal{L}_{QED} = \bar{\psi}(x)(i\gamma^\mu \partial_\mu - m)\psi(x) - q\bar{\psi}(x)\gamma^\mu A_\mu \psi(x) - \frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x), \tag{2.9}$$

where the last term represents for the kietic energy of photon field, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic field strength tensor.

It is obvious that if a mass term of photon with a form $\frac{1}{2}m^2 A_\mu A^\mu$ is included in the Lagrangian, the invariance under local gauge transformations will be violated. The electromagnetic interaction is represented by U(1) group since it has invariance under U(1) gauge transformation.

2.1.2.2 Quantum chromodynamics

As stated before, each quark can carry one of the three color states: red, green and blue. Thus quarks are color-triplets under SU(3) group and can be written as:

$$q_f = \begin{pmatrix} q_f^R \\ q_f^G \\ q_f^B \end{pmatrix} \tag{2.10}$$

and the adjoint:

$$\bar{q}_f = (\bar{q}_f^{\bar{R}}, \bar{q}_f^{\bar{G}}, \bar{q}_f^{\bar{B}}), \tag{2.11}$$

where f is the flavor of the quark. Then again the Lagrangian of free quarks with mass m can be written as:

$$\mathcal{L}_{free} = \sum_f \bar{q}_f(i\gamma^\mu \partial_\mu - m)q_f. \tag{2.12}$$

The Lagrangian should be invariant under SU(3) transformations in the color space:

$$q(x) \to Uq(x) = e^{i\frac{\lambda_a}{2}\theta_a}q(x), \tag{2.13}$$

where U is a 3×3 Unitary matrices. The generators of SU(3) color group, $\frac{\lambda_a}{2}$ (a = 1,...,8), are the Gell-Mann matrices, θ_a is a set of arbitrary parameters. To achieve the local gauge invariance (replacing θ_a by $\theta_a(x)$), as was done in QED, ∂_μ has to be replaced by the covariant derivative:

$$D_\mu \equiv \partial_\mu + ig_s\frac{\lambda_a}{2}G_\mu^a(x), \tag{2.14}$$

where g_s is the strong coupling constant, and $G_\mu^a(x)$ is the the vector field of gluons which are the propagators of strong interaction. Thus the kinetic energy term of the gluon field, $\frac{1}{4}G_{\mu\nu}^a G_a^{\mu\nu}$, should

also be included, where the color field tensor $G^a_{\mu\nu}$ is

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f_{abc} G^b_\mu G^c_\nu, \tag{2.15}$$

where f_{abc} are the structure constants of the SU(3) color group. Then the full QCD Lagrangian can be written as:

$$\mathcal{L}_{QCD} = \sum_f \bar{q}_f (i\gamma^\mu \partial_\mu - m) q_f - g_s \sum_f (\bar{q}_f \gamma^\mu \frac{\lambda_a}{2} q_f) G^a_\mu - \frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a. \tag{2.16}$$

Again, to achieve the local gauge invariance, there is no mass term for the gluon field.

2.1.2.3 Weak interactions

From the experimental facts, the field of weak interaction should be a doublet (i.e. in each generation of fermions, there are two particles, u and d, e and ν_e, etc). The simplest group with doublet representation is SU(2). Then following the procedure in QED and QCD, the Lagrangian of weak interaction should be invariant under SU(2) transformations. First the notations of a single family of quarks can be written as:

$$\psi_1(x) = \begin{pmatrix} u \\ d \end{pmatrix}_L, \qquad \psi_2(x) = u_R, \qquad \psi_3(x) = d_R \tag{2.17}$$

These notations are also valid for leptons:

$$\psi_1(x) = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L, \qquad \psi_2(x) = \nu_{eR}, \qquad \psi_3(x) = e^-_R \tag{2.18}$$

Here L is referring to left-handed field while R is referring to right-handed fields. Then the free Lagrangian of massless fermions can be given as:

$$\mathcal{L}_{free} = i\bar{u}(x)\gamma^\mu \partial_\mu u(x) + i\bar{d}(x)\gamma^\mu \partial_\mu d(x) = \sum_{j=1}^{3} i\bar{\psi}_j(x)\gamma^\mu \partial_\mu \psi_j(x). \tag{2.19}$$

It is obvious to see that the left-handed fields, $\psi_1(x)$ are invariant under global SU(2) transformations as following:

$$U_L \equiv e^{i\frac{\sigma_i}{2}\alpha^i}, \quad (i = 1, 2, 3) \tag{2.20}$$

The right-handed fields, $\psi_2(x), \psi_3(x)$, are singlets under SU(2). However $\psi_1(x), \psi_2(x), \psi_3(x)$ are also invariant under U(1) transformations. The transformations can be written as:

$$\begin{aligned} \psi_1(x) &\to \psi_1'(x) \equiv e^{iy_1\beta} U_L \psi_1(x), \\ \psi_2(x) &\to \psi_2'(x) \equiv e^{iy_2\beta} \psi_2(x), \\ \psi_3(x) &\to \psi_3'(x) \equiv e^{iy_3\beta} \psi_3(x), \end{aligned} \tag{2.21}$$

Thus the electromagnetic interaction and weak interaction could be combined, and then the Lagrangian is required to be invariant under the transformation group:

$$G \equiv SU(2)_L \otimes U(1)_Y, \tag{2.22}$$

where Y is the hypercharge. Notice that again there is no mass term included in equation 2.19 since the mass term would mix the left- and right-handed fields and then spoil the symmetry. Then the Lagrangian is also required to be invariant under local $SU(2)_L \otimes U(1)_Y$ gauge transformations with $\alpha^i = \alpha^i(x)$ and $\beta = \beta(x)$. This could also be done by replacing ∂_μ by the covariant derivative D_μ:

$$\begin{aligned} D_\mu \psi_1(x) &\equiv [\partial_\mu - ig\tilde{W}_\mu(x) - ig' y_1 B_\mu(x)]\psi_1(x), \\ D_\mu \psi_2(x) &\equiv [\partial_\mu - ig' y_2 B_\mu(x)]\psi_2(x), \\ D_\mu \psi_3(x) &\equiv [\partial_\mu - ig' y_3 B_\mu(x)]\psi_3(x). \end{aligned} \tag{2.23}$$

where $\tilde{W}_\mu \equiv \frac{\sigma_i}{2} W^i_\mu(x)$ denotes the $SU(2)_L$ matrix field. Then the Lagrangian

$$\mathcal{L} = \sum_{j=1}^{3} i\bar{\psi}_j(x)\gamma^\mu D_\mu \psi_j(x), \tag{2.24}$$

is invariant under the local G transformations. In order to build the kinetic terms for the gauge field, the corresponding strength tensor is introduced:

$$B_{\mu\nu} \equiv \partial_\mu B_\nu - \partial_\nu B_\mu, \tag{2.25}$$

$$W^i_{\mu\nu} \equiv \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g\epsilon^{ijk} W^j_\mu W^k_\nu. \tag{2.26}$$

$B_{\mu\nu}$ remains invariant under local G transformations while $W^i_{\mu\nu}$ transforms covariantly. Therefore the kinetic term can be written as:

$$\mathcal{L}_{kinetic} = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W^i_{\mu\nu} W^{\mu\nu}_i. \tag{2.27}$$

From the above discussion we can see that the gauge symmetry forbids to include a mass term for the gauge fields (gauge bosons), as well as the fermions since the mass term will connect the left- and right-handed fields and break the gauge symmetry. The experiment results, however, show that gauge bosons ($W^\pm$, Z^0) and fermions are for sure massive. In the following section the mechanism that generates mass for the fields will be introduced.

2.1.2.4 Higgs mechanism

So far, we have developed Lagrangians with very good properties, which can describe fermions, interactions between fermions and gauge fields and the self-interaction of gauge fields. However, the key problem is that the gauge bosons are massless, which is true for photons but not for the $W^\pm$ and Z^0. The experimental results show that they are quite heavy. In order to generate the mass, the gauge symmetry has to be broken while keeping the Lagrangian fully symmetric. This could be done by Spontaneous Symmetry Breaking (SSB).

Let us start with a complex scalar field $\phi(x)$, with Lagrangian:

$$\mathcal{L} = \partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi), \qquad V(\phi) = \mu^2 \phi^\dagger \phi + h(\phi^\dagger \phi)^2. \tag{2.28}$$

Obviously $\mathcal{L}$ is invariant under the global phase transformation of scalar field:

$$\phi(x) \to \phi'(x) \equiv e^{i\theta}\phi(x). \tag{2.29}$$

The parameter, h, should be larger than 0 in order to have a ground state. About the first term in $V(\phi)$, there are two possibilities for μ^2:

1. $\mu^2 > 0$: The potential has only the trivial minimum at $\phi = 0$. It describes a scalar particle with mass μ and quartic coupling h.

2. $\mu^2 < 0$: The minimum can be obtained at:

$$|\phi_0| = \sqrt{\frac{-\mu^2}{2h}} \equiv \frac{v}{\sqrt{2}} > 0, \qquad V(\phi_0) = -\frac{h}{4}v^4. \tag{2.30}$$

   Considering the U(1) transformations on the Lagrangian, there are infinite degenerate states of minimum energy, $\phi_0(x) = \frac{v}{\sqrt{2}} exp(i\theta)$.The symmetry is spontaneously broken when we choose a particlar number for θ, i.e. $\theta = 0$, as the ground state solution. The excitations over can be parametrized as:

$$\phi(x) \equiv \frac{1}{\sqrt{2}}[v + \varphi_1(x) + i\varphi_2(x)], \tag{2.31}$$

   where $\varphi_1(x)$ and $\varphi_2(x)$ are real fields, then the potential can be written as:

$$V(\phi) = V(\phi_0) - \mu^2\varphi_1^2 + hv\varphi_1(\varphi_1^2 + \varphi_2^2) + \frac{h}{4}(\varphi_1^2 + \varphi_2^2)^2. \tag{2.32}$$

   Then it is obvious that φ_1 describes a massive state with mass $m_{\varphi_1}^2 = -2\mu^2$ while φ_2 is massless.

From this discussion, we can see that the first case is just a usual single ground state. But the second case breaks the symmetry spontaneously and generates another massless particle. The appearance of this massless particle can be understood as following: the field φ_2 describes excitations around a flat direction in the potential, for example the states with the same energy as ground state. Since these excitations do not need any energy, they obviously correspond to a massless particle. In fact the generation of massless particle in association with SSB mechanism is a general result known as the Goldstone theorem. The remaining issue is that, these massless particles still do not exist in the nature. The solution is the Brout-Englert-Higgs mechanism. Considering a doublet of complex scalar fields:

$$\Phi(x) \equiv \begin{pmatrix} \phi^{(+)}(x) \\ \phi^{(0)}(x) \end{pmatrix}. \tag{2.33}$$

Then the Lagrangian of Goldstone model can be written the same as:

$$\mathcal{L}_S = (D_\mu \Phi)^\dagger D^\mu \Phi - \mu^2 \Phi^\dagger \Phi - h(\Phi^\dagger \Phi)^2, \qquad (\mu^2 < 0, h > 0), \tag{2.34}$$

where the covariant derivative has been applied:

$$D^{\mu}\Phi = [\partial^{\mu} - ig\tilde{W}^{\mu} - ig' y_{\Phi} B^{\mu}]\Phi. \tag{2.35}$$

It is obviously invariant under $SU(2)_L \otimes U(1)_Y$ transformations. The potential is similar to what has been discussed in SSB mechanism, there are infinite degenerate states with minimum energy satisfying:

$$|\langle 0|\phi^{(0)}|0\rangle| = \sqrt{\frac{-\mu^2}{2h}} \equiv \frac{v}{\sqrt{2}}. \tag{2.36}$$

Once a particular ground state is chosen, the $SU(2)_L \otimes U(1)_Y$ symmetry gets spontaneously broken to the electromagnetic subgroup $U(1)_{QED}$ which is still a true symmetry of the vacuum. Another three massless bosons should be generated according to the Goldstone theorem. If the scalar doublet is parametrized in the general form:

$$\Phi(x) = e^{i\frac{\sigma_j}{2}\theta^j(x)} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}, \tag{2.37}$$

there are four real fields, $\theta^j(x)$ and H(x). In the local $SU(2)_L$ invariance of the Lagrangian, the dependence on $\theta^j(x)$ can be rotated away. Therefore, the unphysical massless excitations are excluded by the additional ingredient from the gauge symmetry (unitary). Also notice that the covariant derivative 2.35 couples the scalar doublet to the gauge bosons. If considering the physical gauge, $\theta^j(x) = 0$, the kinetic term of the scalar Lagrangian can be written as:

$$(D_{\mu}\Phi)^{\dagger} D^{\mu}\Phi \xrightarrow{\theta^j=0} \frac{1}{2}\partial_{\mu}H\partial^{\mu}H + (v+H)^2\{\frac{g^2}{4}W_{\mu}^{\dagger}W^{\mu} + \frac{g^2}{8\cos^2\theta_W}Z^{\mu}Z_{\mu}\}, \tag{2.38}$$

where

$$\begin{aligned} W_{\mu} &= \frac{W_{\mu}^1 + iW_{\mu}^2}{\sqrt{2}}, \\ Z_{\mu} &= \frac{g'W_{\mu}^3 - gB_{\mu}}{\sqrt{g^2 + g'^2}}, \\ \cos\theta_W &= \frac{g}{\sqrt{g^2 + g'^2}}. \end{aligned} \tag{2.39}$$

Clearly, the vacuum expectation value of the neutral scalar, $\frac{v}{\sqrt{2}}$, generates a quadratic term for the $W^{\pm}$ and Z^0 bosons. In this way these bosons acquire masses:

$$m_W = m_Z \cos\theta_W = \frac{1}{2}vg. \tag{2.40}$$

Reviewing what has been done so far, we have found a way to give masses to the gauge bosons. We add a doublet of scalar, Φ, to the $SU(2)_L \otimes U(1)_Y$ model. The total Lagrangian remains invariant under local gauge transformations, but the SSB breaks down part of the symmetry and generated three massless Goldstone bosons. These Goldstone bosons can be however eliminated from the Lagrangian due to the underlying local gauge symmetry and then generate masses for the gauge bosons. But the photon is still massless since $U(1)_{QED}$ remains unbroken symmetry. From the

equation 2.38, it should be also noticed that there is remaining scalar particle H, namely the Higgs boson.

Rewriting the scalar Lagrangian 2.34 in terms of the physical fields (unitary gauge) as following:

$$\mathcal{L}_S = \frac{1}{4}hv^4 + \mathcal{L}_H + \mathcal{L}_{HG^2}, \tag{2.41}$$

where

$$\mathcal{L}_H = \frac{1}{2}\partial_\mu H \partial^\mu H - \frac{1}{2}M_H^2 H^2 - \frac{M_H^2}{2v}H^3 - \frac{M_H^2}{8v^2}H^4, \tag{2.42}$$

$$\mathcal{L}_{HG^2} = M_W^2 W_\mu^\dagger W^\mu \{1 + \frac{2}{v}H + \frac{H^2}{v^2}\} + \frac{1}{2}M_Z^2 Z_\mu Z^\mu \{1 + \frac{2}{v}H + \frac{H^2}{v^2}\}. \tag{2.43}$$

and the Higgs mass is given by:

$$M_H = \sqrt{-2\mu^2} = \sqrt{2h}v. \tag{2.44}$$

Thus we can see that the coupling between the Higgs boson and gauge bosons are proportional to the squared gauge boson masses. All the couplings can be determined by M_H, M_W, M_Z and the vacuum expectation value v which is related to the fermion constant. The masses of W and Z bosons are measured from experiments and then v can be further determined from the muon life time. But still there is another free paramter, h, in equation 2.44, therefore the Higgs mass can not be predicted from theory.

So far the problem of the masses of gauge bosons have been solved, but the fermions are still remaining massless. A fermionic mass term $\mathcal{L}_m = -m\bar{\psi}\psi = -m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$ is forbidden since it would break the gauge symmetry. Considering that we have introduced a doublet of scalar field, then we can write a gauge invariant fermion-scalar coupling (Yukawa couping) as following (only for the first generation of fermions):

$$\mathcal{L}_Y = c_1(\bar{u}, \bar{d}) \begin{pmatrix} \phi^{(+)} \\ \phi^{(0)} \end{pmatrix} d_R + c_2(\bar{u}, \bar{d}) \begin{pmatrix} \phi^{(0)*} \\ -\phi^{(-)} \end{pmatrix} u_R + c_3(\bar{\nu_e}, \bar{e}) \begin{pmatrix} \phi^{(+)} \\ \phi^{(0)} \end{pmatrix} e_R + h.c. \tag{2.45}$$

where the second term involves the $\mathcal{C}$-conjugate scalar field $\phi^c \equiv i\sigma_2\phi^*$. The Yukawa Lagrangian becomes simpler in the unitary gauge after SSB:

$$\mathcal{L}_Y = \frac{1}{\sqrt{2}}(v + H)\{c_1\bar{d}d + c_2\bar{u}u + c_3\bar{e}e\}. \tag{2.46}$$

Therefore the fermions can get mass from the Yukawa coupling to the Higgs boson:

$$m_d = -c_1\frac{v}{\sqrt{2}}, \qquad m_u = -c_2\frac{v}{\sqrt{2}}, \qquad m_e = -c_3\frac{v}{\sqrt{2}} \tag{2.47}$$

Notice that, the masses of fermions however are still arbitrary since the free parameters c_i are unkown. But the fermion Yukawa coupling strengths are proportional to the masses:

$$\mathcal{L}_Y = -(1 + \frac{H}{v})\{m_d\bar{d}d + m_u\bar{u}u + m_e\bar{e}e\} \tag{2.48}$$

2.1.3 SM Higgs boson production and decay at LHC

The LHC is a high energy proton-proton collider, the Higgs boson can be produced by the hard scattering among quarks and gluons inside the protons. The main production mechanism at LHC are [8, 9]:

- gluon fusion (ggF), $gg \to H$ (a)
- vector boson fusion (VBF), $qq \to qqH$ (b)
- in association to a gauge boson (VH), $q\bar{q} \to VH$ (c)
- in association to a pair of top quarks (ttH), $gg \to t\bar{t}H$ (d)

The feynman diagrams for these production are presented in figure 2.3. There are also several other production mechanism, such as in association to a pair of bottom quark (bbH) or in association to a single top quark (tH). They are also valuable to probe certain parameter space, but are hard to search due to a very tiny cross section or huge backgrounds. The cross sections for the production of the Higgs boson is a function of $\sqrt{s}$, the center-of-mass enery, for the pp collisions, as shown in figure 2.4. The dominant contributions are from ggF and VBF processes, followed by VH and ttH. The contribution from $b\bar{b}H$ is comparable to $t\bar{t}H$, but the background is much higher due to large QCD jets in pp collisions. In particular, the ttH production is of interest in my thesis since it provides a possibility to probe the coupling of the Higgs boson to the top quark (λ_t) directly at tree level. λ_t can also be probed through indirect measurements, i.e. a top quark loop is included in the ggF process, as well as in the Higgs decay to photons. Its value is expected to of the order of unity as $v = 246$ GeV and thus is sensitive to new physics effects. The search for $t\bar{t}H$ production in multilepton final states are described in chapter 5.

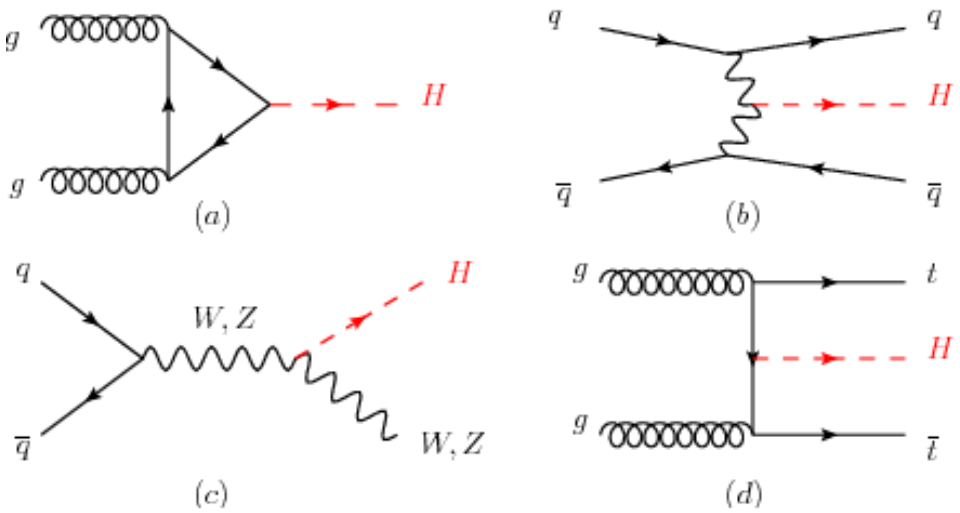


FIGURE 2.3: The four major Higgs boson production processes at LHC.

The lifetime of the Higgs boson is extremely short. The branching ratio of Higgs decay modes is related to the involved particle mass. The branching ratios are summarized in figure 2.5 as a function of the Higgs mass. The dominant decay modes are $H \to b\bar{b}$, $H \to WW^*$, $H \to gg$ and $H \to \tau^+\tau^-$. There are also some rare decay modes, such as $H \to \gamma\gamma$, $H \to \mu^+\mu^-$, $H \to Z\gamma$.

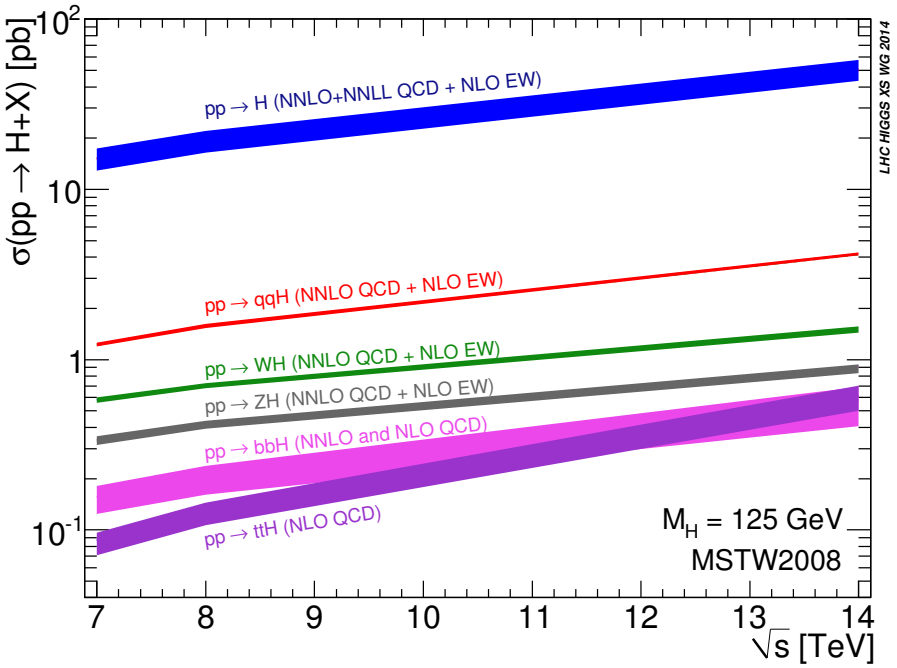


FIGURE 2.4: Standard Model Higgs boson production total cross sections. Uncertainties are shown as band.

Notice that despite the fact that the Higgs boson does not couple to massless particles at tree level, it can decay to massless particles through a loop of heavy particle (W, Z, t). Thus they can provide indirect information on the Higgs coupling to those heavy particles. The measured cross sections of major production modes scaled to the SM prediction are presented in figure 2.6 broken into different Higgs decay channels.

The Higgs boson was first observed by both ATLAS and CMS experiment in 2012 [1, 2, 3]. The main discovery channels were $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$, which had very tiny branching ratios but very clean backgrounds. Utill now, all major production modes and major decay channels have been observed. The state-of-art Higgs boson mass measurement is 2.7:

$$M_H = 125.9 \pm 0.24\text{GeV} \tag{2.49}$$

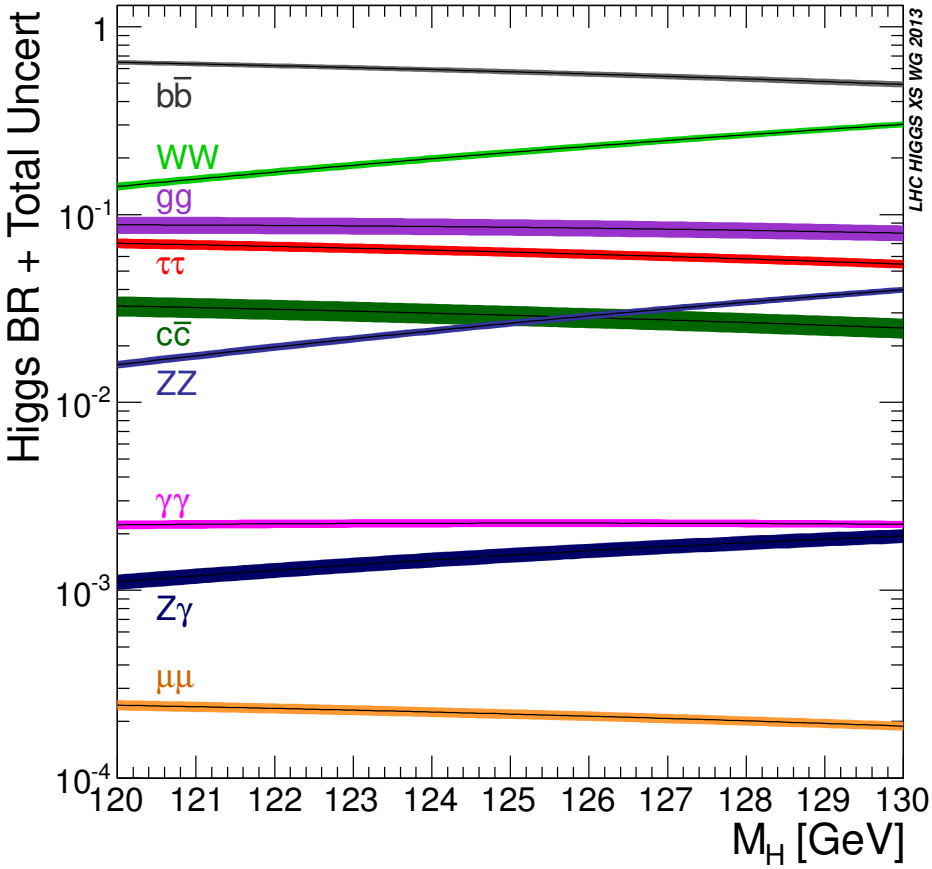


FIGURE 2.5: Standard Model Higgs boson decay branching ratios and total uncertainties.

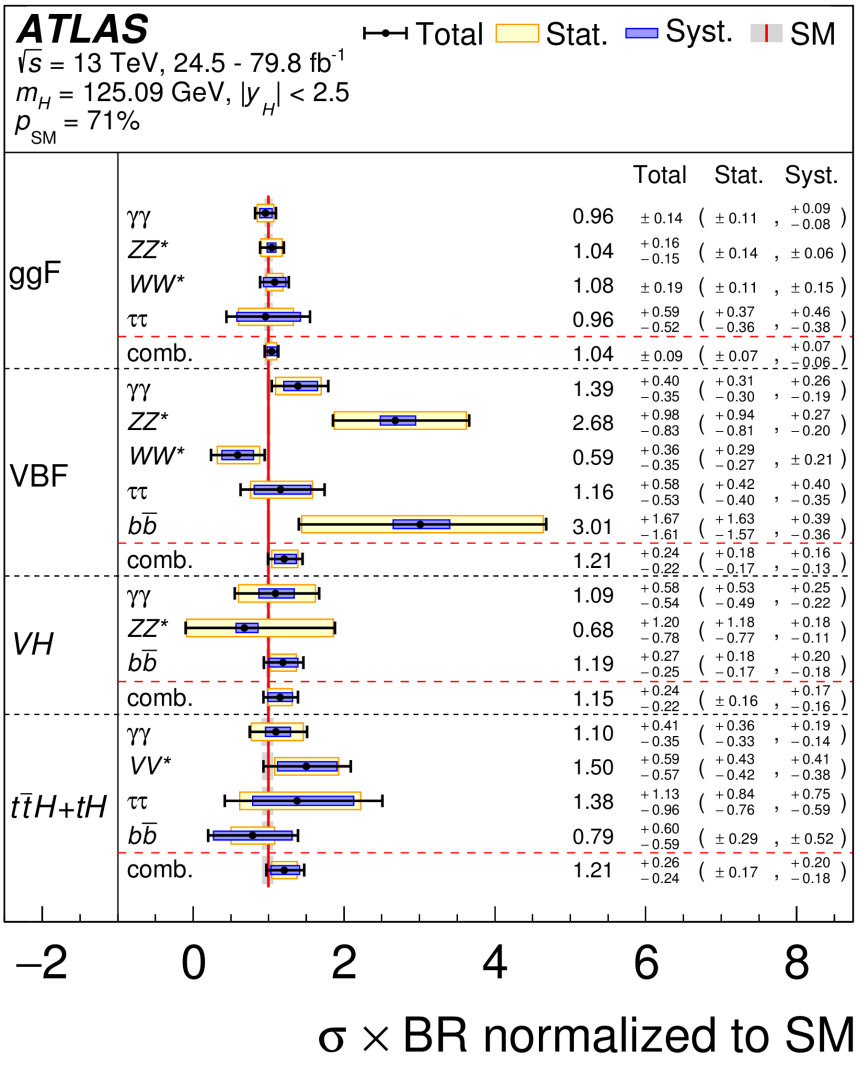


FIGURE 2.6: Cross sections times branching fraction for ggF, VBF, VH and ttH+tH production in each relevant decay mode, normalized to their SM predictions [10].

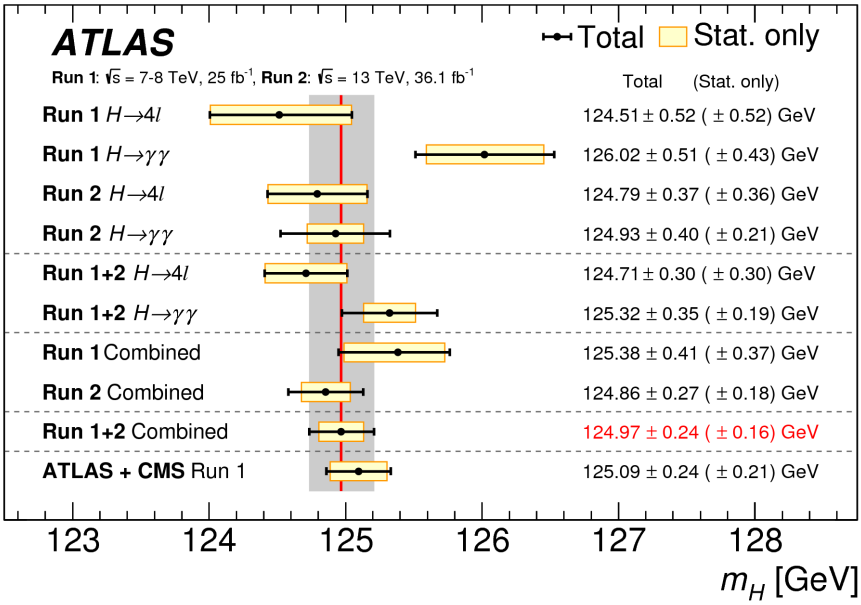


FIGURE 2.7: Summary of the Higgs boson mass measurements from ATLAS and CMS, compared with the combined Run 1 measurement [11].

2.2 Higgs sector beyond the Standard Model

2.2.1 Introduction

Even though the SM has achieved tremendous success, many hints show that it cannot be a complete theory to explain everything in the Universe. Some famous remaining puzzles are:

- **Hierarchy Problem.** The energy scales in the SM cross a huge range, from EW scale (125 GeV) to the Plank scale (10^{19} GeV). More technically, the question is equivalent to why the mass of the Higgs boson is so lighter than the Plank mass. One solution is that there is a cancellation between bare Higgs mass and quadratic radiative corrections, which requires extremely fine-tuning that sounds not natural.
- **Dark matter.** Cosmology experiments show that the universe consists of 5 % normal matter, 26 % dark matter and 69 % dark energy. But all SM particles are excluded from dark matter candidates which means there must be new particles that are not included in the SM.
- **Neutrino mass.** Neutrinos are regarded as massless particles in the SM. But recent neutrino experiments prove that neutrinos are massive due to neutrino oscillations. Adding neutrino mass terms is not easy in the SM as right-handed neutrinos do not have interactions.
- **Matter-antimatter asymmetry.** Particles and anti-particles are considered to be symmetric in the SM, which means there should be the same amount of matter and anti-matter in the universe. But the observations show that our universe is mostly made up of matter. In order to have enough matter to build the observed universe, stronger CP violation than what has been measured in the SM is required.
- **Gravitation** Gravitation is not described in the SM.

Many new physics theories beyond the SM (BSM) have been proposed to explain those puzzles. The most popular one is supersymmetry (SUSY), where each particle (fermions and bosons) has a superpartner, as illustrated in figure 2.8.

Usually a more complicated Higgs scenario is required in BSM theories. As discussed in section 2.1.2.4, only the simplest possible scalar structure, one SU(2) doublet, is assumed. But more complicated scalar structures are also possible to generate the masses. One critical evidence of the scalar structure is the parameter ρ defined as (at tree level):

$$\rho = \frac{\sum_{i=1}^{n}[I_i(I_i+1) - \frac{1}{4}Y_i^2]v_i}{\sum_{i=1}^{n}\frac{1}{2}Y_i^2 v_i}, \tag{2.50}$$

where n is the number of scalar multiplets ϕ_i, I_i is weak isospin, Y_i is weak hypercharge, v_i is the vacuum expectation value of the neutral components. Experimentally, ρ is measured to be very close to one. Both SU(2) singlets with $Y = 0$ and SU(2) doublets with $Y = \pm 1$ can give $\rho = 1$. So by adding scalars, more and more complicated models can be achieved. Thus the simplest extention of the SM consists of simply adding one other scalar doublet to the Higgs sector. This is the so-called two-Higgs-doublet-model (2HDM) [12, 13]. It is also the model that was used for my thesis

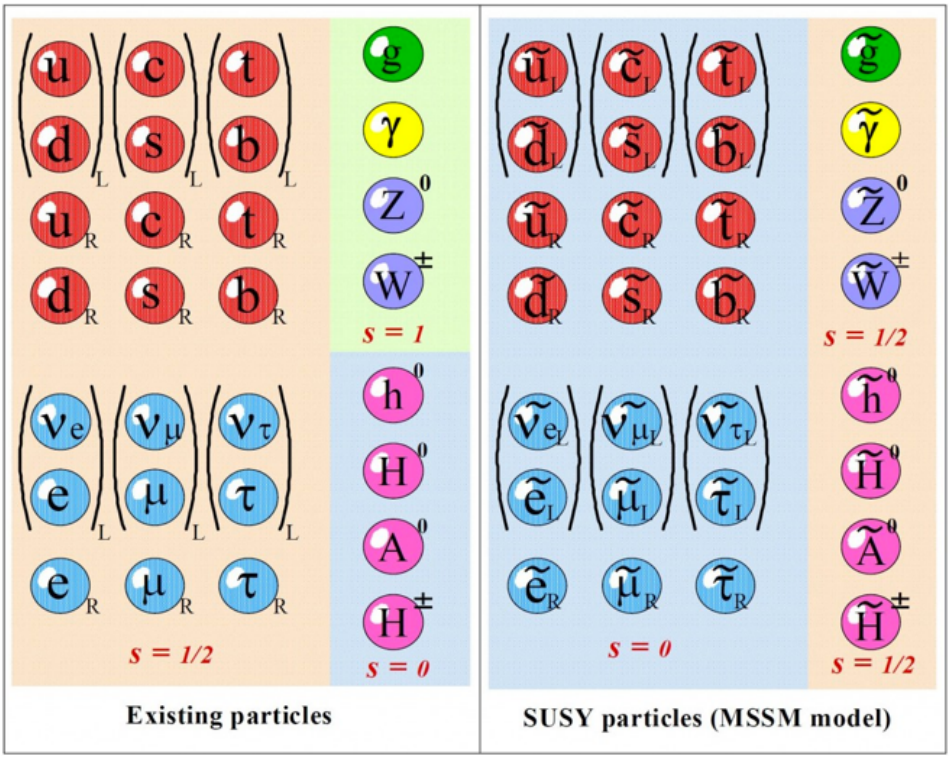


FIGURE 2.8: Particles predicted by the Minimal Supersymmetric Standard Model (MSSM). In the left block there are SM particles with four new Higgs bosons predicted by MSSM. Their superpartners are shown in the right block.

work described in chapter 4. This model covers the popular supersymmetric theories, such as the Minimal Supersymmetric Standard Model (MSSM), to provide a way to solve the asymmetry between matter and antimatter due to additional sources of CP violation.

2.2.2 Two Higgs doublet model

Generally speaking, there are various vacuum structures of 2HDMs. The most general scalar potential contains 14 parameters and can have CP-conserving, CP-violating and charge-violating minima. To simplify the discussion, several assumptions are made: CP is conserved in the Higgs sector; CP is not spontaneously broken; discrete symmetries eliminate all quartic terms in either of the doublets from the potential; all possible real quadratic coefficients are considered, including a term which softly breaks these symmetries. Then the most general scalar potential for two doublets, Φ_1 and Φ_2 with hypercharge +1 can be written as:

$$\begin{aligned} V = m_{11}^2\Phi_1^\dagger\Phi_1 + m_{22}^2\Phi_2^\dagger\Phi_2 - m_{12}^2(\Phi_1^\dagger\Phi_2 + \Phi_2^\dagger\Phi_1) + \frac{\lambda_1}{2}(\Phi_1^\dagger\Phi_1)^2 + \frac{\lambda_2}{2}(\Phi_2^\dagger\Phi_2)^2 \\ + \lambda_3\Phi_1^\dagger\Phi_1\Phi_2^\dagger\Phi_2 + \lambda_4\Phi_1^\dagger\Phi_2\Phi_2^\dagger\Phi_1 + \frac{\lambda_5}{2}[(\Phi_1^\dagger\Phi_2)^2 + (\Phi_2^\dagger\Phi_1)^2], \end{aligned} \tag{2.51}$$

where all their parameters are real. The minimization of the above potential is:

$$\langle\Phi_1\rangle_0 = \begin{pmatrix} 0 \\ \frac{v_1}{\sqrt{2}} \end{pmatrix}, \quad \langle\Phi_2\rangle_0 = \begin{pmatrix} 0 \\ \frac{v_2}{\sqrt{2}} \end{pmatrix}. \tag{2.52}$$

where the two complex scalar doublets are:

$$\Phi_a = \begin{pmatrix} \phi_a^+ \\ (v_a + \rho_a + i\eta_a)/\sqrt{2} \end{pmatrix}, a = 1, 2. \tag{2.53}$$

There are in total eight fields (ϕ^+, v, ρ, η) in the 2HDM. As the Higgs mechanism in the SM, three of them are "eaten" by the gauge bosons to give masses to the W and Z bosons. The remaining five fields are physical scalar fields, including one charged scalar, two neutral scalars and one pseudoscalar. With the minimization given by equation 2.52, the mass terms for the charged scalar are given by:

$$\mathcal{L}_{\phi^\pm mass} = [m_{12}^2 - (\lambda_4 + \lambda_5)v_1 v_2](\phi_1^-, \phi_2^-) \begin{pmatrix} \frac{v_2}{v_1} & -1 \\ -1 & \frac{v_1}{v_2} \end{pmatrix} \begin{pmatrix} \phi_1^+ \\ \phi_2^+ \end{pmatrix}. \tag{2.54}$$

There is a zero eigenvalue corresponding to the charged Goldstone boson $G^\pm$ that is eaten by the $W^\pm$. The mass of the remaining charged Higgs boson is $m_+^2 = [m_{12}^2/(v_1 v_2) - \lambda_4 - \lambda_5](v_1^2 + v_2^2)$. The mass terms for the pseudoscalar are given by:

$$\mathcal{L}_{\eta mass} = \frac{m_A^2}{v_1^2 + v_2^2}(\eta_1, \eta_2) \begin{pmatrix} v_2^2 & -v_1 v_2 \\ -v_1 v_2 & v_1^2 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}. \tag{2.55}$$

There is a pseudoscalar Goldstone boson that is eaten by Z^0, and a remaining pseudoscalar Higgs boson with mass $m_A^2 = [m_{12}^2/(v_1 v_2) - 2\lambda_5](v_1^2 + v_2^2)$. The mass terms for the scalars are written as:

$$\mathcal{L}_{\rho mass} = -(\rho_1, \rho_2) \begin{pmatrix} m_{12}^2 \frac{v_2}{v_1} + \lambda_1 v_1^2 & -m_{12}^2 + \lambda_{345} v_1 v_2 \\ -m_{12}^2 + \lambda_{345} v_1 v_2 & m_{12}^2 \frac{v_1}{v_2} + \lambda_2 v_2^2 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix}, \tag{2.56}$$

where $\lambda_{345} = \lambda_3 + \lambda_4 + \lambda_5$. Notice that the mass matrix of the scalars can be diagonalized by performing a rotation with an angle α. The mass matrices of the charged scalar and the pseudoscalar can also be diagonalized by performing a rotation with an angle β, where $\tan\beta \equiv \frac{v_2}{v_1}$ indicates the ratio of VEV of two doublets. These two parameters, α and β, determine the interactions of the Higgs fields with the vector bosons and the fermions, therefore they are the key points when discussing phenomenology.

An important feature of 2HDM is whether allowing flavor-changing neutral currents (FCNC) at tree level. The mass matrix of a quark with electric charge of $-1/3$ in 2HDM is:

$$M_{ij} = y_{ij}^1 \frac{v_1}{\sqrt{2}} + y_{ij}^2 \frac{v_2}{\sqrt{2}}. \tag{2.57}$$

In the SM, diagonalization of the mass matrix will automatically diagonalize the Yukawa couplings, therefore there is no tree level FCNC. However in 2HDMs, generally y^1 and y^2 will not be diagonalized simultaneously, thus the Yukawa couplings will not be flavor diagonal which allows FCNC at tree level. Here only natural flavor conservated models are considered since FCNC at tree level would cause many phenomenological problems. The conservation is achieved by a discrete or continuous symmetry.

Considering the quark sector of the 2HDM, there are two possibilities: all quarks couple to only

one of the Higgs doublets, namely type I 2HDM; the right-handed quarks with electric charge of 2/3 couple to one Higgs doublet and right-handed quarks with electric charge of -1/3 (d_R) couple to the other. By convention, the right-handed quarks with electric charge of 2/3 (u_R) always couple to Φ_2. The type I 2HDM is enforced with a $\Phi_1 \to -\Phi_1$ discrete symmetry, whereas the type II 2HDM is enforced with a $\Phi_1 \to -\Phi_1, d_R^i \to -d_R^i$ discrete symmetry. In these two models, it is conventionally assumed that the right-handed leptons satisfy the same discrete symmetry as the d_R^i and then the leptons couple to the same Higgs doublet as the quarks with electric charge of -1/3. This is however not guaranteed by any law, therefore there are two other possiblilities for the leptons: the right-handed quarks couple to Φ_2 and the right-handed leptons couple to Φ_1, namely the lepton-specific model; the couplings of quarks are the same as in type II 2HDM but the right-handed leptons couple to Φ_2. These four models are summarized in table 2.4.

Model	u_R^i	d_R^i	e_R^i
Type I	Φ_2	Φ_2	Φ_2
Type II	Φ_2	Φ_1	Φ_1
Lepton-specific	Φ_2	Φ_2	Φ_1
Flipped	Φ_2	Φ_1	Φ_2

TABLE 2.4: Models with natural flavor conservation. The superscript i is a generation index. The u_R^i always couple to Φ_2 by convention.

From the definition of $\tan\beta$, the VEV can be written as $v_1 = v\cos\beta$ and $v_2 = v\sin\beta$. Then the neutral Goldstone boson is $G^0 = \eta_1\cos\beta + \eta_2\sin\beta$. The physical pseudoscar, A, is the linear combination of η orthogonal to G^0:

$$A = \eta_1\sin\beta - \eta_2\cos\beta. \tag{2.58}$$

The physical scalars are a lighter h and a heavier H, which are orthogonal to each other:

$$h = \rho_1\sin\alpha - \rho_2\cos\alpha, \tag{2.59}$$

$$H = -\rho_1\cos\alpha - \rho_2\sin\alpha, \tag{2.60}$$

The Higgs boson in the SM is:

$$\begin{aligned} H_{SM} &= \rho_1\cos\beta + \rho_2\sin\beta \\ &= h\sin(\alpha-\beta) - H\cos(\alpha-\beta). \end{aligned} \tag{2.61}$$

Considering the charged scalar, the most general Yukawa couplings to fermions can be written as:

$$\mathcal{L}_{H^\pm} = -H^+\left(\frac{\sqrt{2}V_{ud}}{v}\bar{u}(m_u X P_L + m_d Y P_R)d + \frac{\sqrt{2}m_\ell}{v}Z\bar{\nu}_L\ell_R\right) + h.c. \tag{2.62}$$

where V_{ud} is the element of the CKM matrix corresponding to the charge 2/3 quark u and the charge -1/3 quark d. The values of X, Y and Z depend on the models and are summarized in table 2.5. Just as the case in the SM, the mass of Higgs bosons in the 2HDM cannot be predicted by theory.

	Type I	Type II	Lepton-specific	Flipped
X	$\cot\beta$	$\cot\beta$	$\cot\beta$	$\cot\beta$
Y	$\cot\beta$	$-\tan\beta$	$\cot\beta$	$-\tan\beta$
Z	$\cot\beta$	$-\tan\beta$	$-\tan\beta$	$\cot\beta$

TABLE 2.5: The parameters in equation 2.62 for the natural flavor conservation models.

The couplings of the charged Higgs boson to fermions is related to the $\tan\beta$ and masses. Therefore it couples mostly to the third generation quarks, as well as tau leptons. The production of the charged Higgs boson depends on whether its mass is lower or higher than the top mass. For low mass charged Higgs boson, the dominant production mode is radiating from top quark $t \to bH^+$; on the contrary, for high mass the dominant production mode is in association with a top quark. For the decay of charge Higgs boson, it is related to the $\tan\beta$, the branch ratio of decaying to $\tau\nu$ increases for larger $\tan\beta$. But decay to tb is always dominant for high mass charged Higgs boson. This justifies that most of the searches for high-mass charged Higgs are performed in the $H^+ \to tb$ channel.

Chapter 3

LHC & the ATLAS Detector

The Large Hadron Collider (LHC) is the largest collider providing the highest proton-proton collision energy. It is located at the European Organization for Nuclear Research (CERN) near the France-Switzerland border, at a underground depth between about 50 m to 175 m. The proton beam energy can be up to 7 TeV with center-of-mass energy $\sqrt{s} = 14$ TeV, which provides a great opportunity to study subatomic physics. There are four detectors located at different places at LHC, among which A Toroidal LHC ApparatuS (ATLAS) and Compact Muon Solenoid (CMS) are general purpose detectors while LHCb and ALICE are designed for b-hadron physics and Quark-Gluon-Plasma investigation, respectively. In this chapter, the design of LHC is presented in section 3.1, the ATLAS detector is described in section 3.2 as well as the future upgrade of the ATLAS detector in section 3.3 where I finished my qualification task to become an ATLAS author (developing debugging tools for detector simulation) and finally the physical object reconstructions in section 3.4.

3.1 The Large Hadron Collider

The tunnel of LHC is inherited from the Large Electron-Positron Collider (LEP) which operated from 1984 to 1989. The tunnel has a circumference of 26.7 km and features a gradient horizontal of 1.4 %. Comparing to a circular electron-positron collider, a hadron collider can significantly benefit from a larger radius and does not suffer from synchrotron radiation since the mass of protons is around 2000 times heavier than that of electrons.

The acceleration of injected particles is through a succession of machines, which is shown in figure 3.1 [14]. The protons are produced by a duoplasmatron source through stripping orbiting electrons from hydrogen atoms, and then injected to the linear accelerator called Linac2 where they are accelerated to 50 MeV. After Linac2, the beam is firstly fed into the Proton Synchrotron (PS) Booster, which has a radius of 25 meters and is accelerated to 1.4 GeV. Secondly the beam is fed into another PS, where it gets accelerated to 26 GeV and finally into the Super PS(SPS) with a circumference of 6.9 kilometers where it gets accelerated to 450 GeV. The particles now have got enough energy to be fed into the LHC and the injection is done via two transfer lines where the beam circulate in opposite directions.

The acceleartion of particles in the main ring is done by electric fields produced in superconducting radio-frequency (RF) cavities operating at 400 MHz with a 5 MV/m gradient to reach the designed energy of 14 TeV. The acceleration continues for about 20 minutes. Particles circulating at the nominal energy (7 TeV) lose about 7 keV per turn due to the synchrotron radiation. This loss

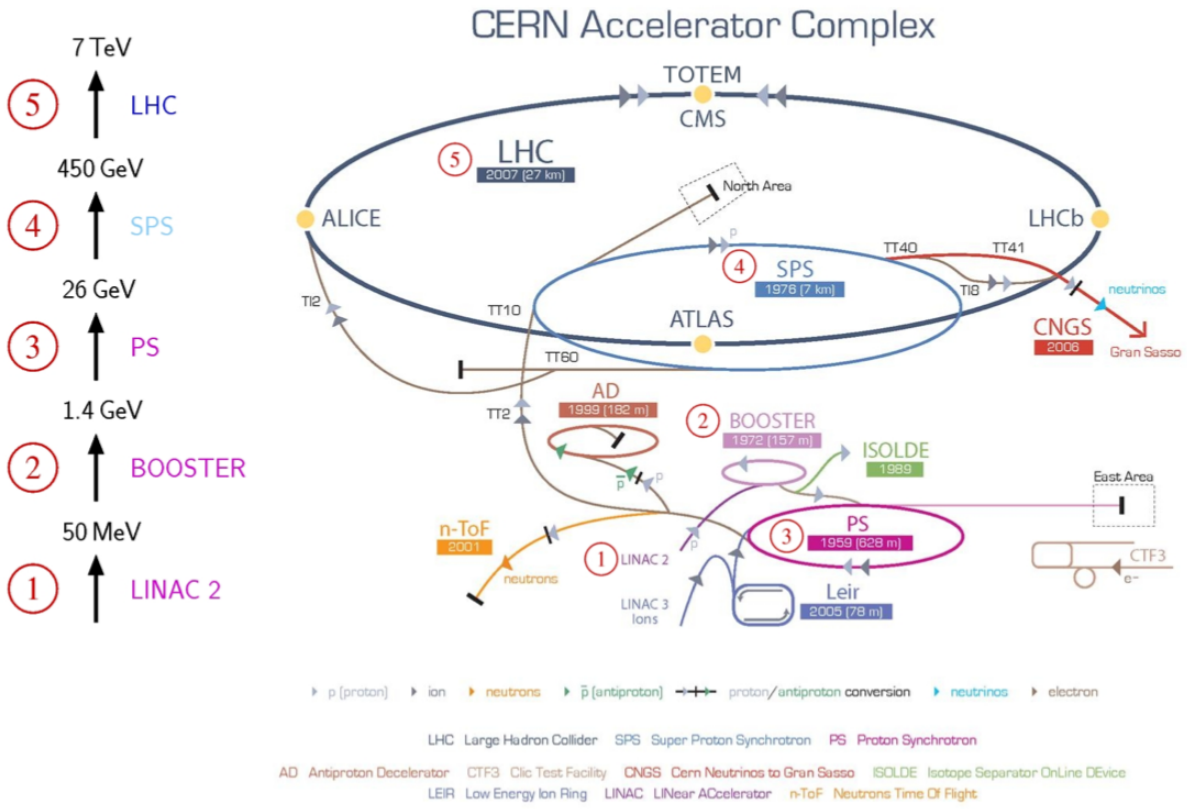


FIGURE 3.1: The accelerator chain at CERN.

is continuously corrected by the RF cavities. The chain of machines can also accelerate lead and xenon ions, which are injected through a Low Energy Ion Ring and then transfered through PS and SPS. So far, LHC has not yet reached its designed energy. It was running at $\sqrt{s} = 7$ and 8 TeV in 2011 and 2012, and 13 TeV since it restarted in 2015.

Another important parameter of colliders other than the beam energy is the instantaneous luminosity. It indicates the potential number of collisions per second. For a particular process, the number of collisions per second is given by:

$$N_{event} = \mathcal{L}_{inst} \times \sigma_{event} \tag{3.1}$$

where $\mathcal{L}_{inst}$ is the instantaneous luminosity and σ_{event} is the cross-section of a the process. The instantaneous luminosity is defined as:

$$\mathcal{L}_{inst} = \frac{\mu n_b f_r}{\sigma_{inel}} \equiv \frac{f_r N^2}{4\beta^* \epsilon} F \tag{3.2}$$

where μ is the average number of inelastic interactions per bunch crossing, n_b is the number of colliding bunch pairs, f_r is the machine revolution frequency and σ_{inel} is the pp inelastic cross-section. It depends on the collider parameters, as following:

- The number of protons in the bunch (N)
- The collision transverse area ($4\beta^*\epsilon$)
- The frequency of collisions (f_r)
- Geometric luminosity reduction factor (F), caused by the crossing angle

For ATLAS and CMS, the designed luminosity of pp collisions is $10^{34} cm^{-2} s^{-1}$. The experiments can benefit a lot from an increase of luminosity, with the caveat that they suffer from additional pile-up events as shown in figure 3.2 [15]. There are two kinds of pile-ups, in-time pile-up refers to multiple proton pairs interacting in a bunch crossing, and out-time pile-up refers to the collisions from surrounding bunch crossing in the read-out time window of the given bunch crossing. Pile-up events can increase the multiplicity of tracks and vertices in an event as well as the overall energy. Thus the reconstruction of physical objects such as vertices, tracks and jets will be more challenging. In general pile-up events are modelled by minimum bias events based on Monte Carlo simulation. To account for different pile-up environments in different data-taking years, there are different pile-up profiles used in the Monte Carlo simulation. The most often used MC campaign are listed below:

- MC15a: use pile-up profile for 2015 data
- MC15c: use a combined profile for 2015+2016 data
- MC16a: use a combined profile for 2015+2016 data
- MC16d: use a profile for 2017 data
- MC16e: use a profile for 2018 data

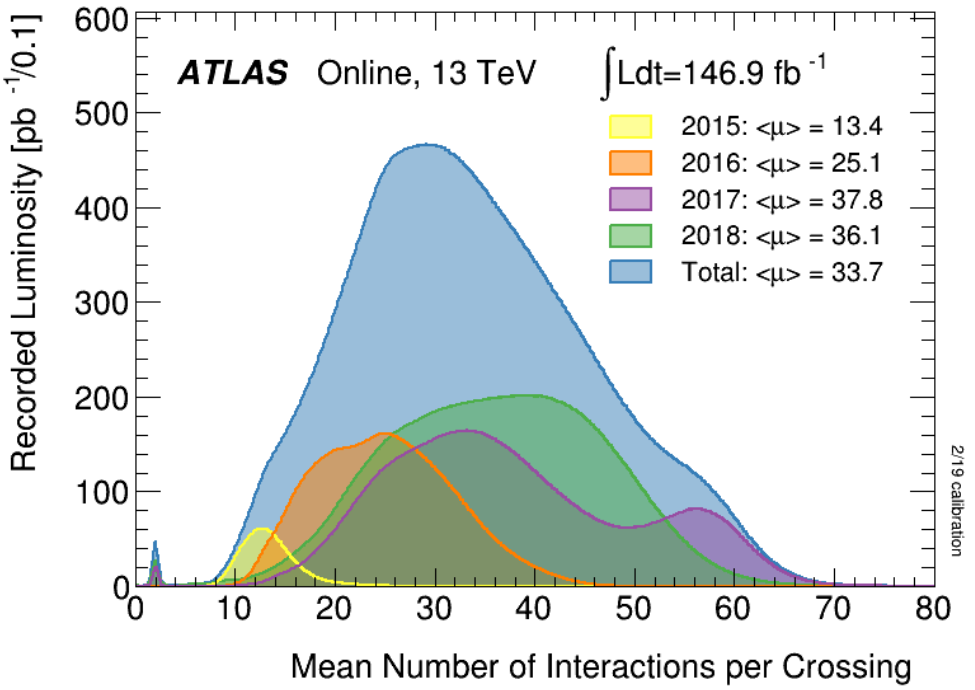


FIGURE 3.2: Number of interactions per bunch-crossing in Run2 data. More interactions usually lead to larger influence of pile-up.

The first pp collision at LHC occurred in November 2009, at a center-of-mass energy of 900 GeV and then increased to 2.37 TeV in a few days. The first normal data taking at 7 TeV started in 2011 and then increased to 8 TeV in 2012, this period is called "Run 1". The total integrated luminosity delivered by LHC during Run 1 is about 28 fb^{-1}. After 3 years Phase-0 upgrade (Long shutdown 1), the data taking restarted in 2015 with the center-of-mass energy increased to 13 TeV , this period is called "Run 2". Run 2 ended at the end of 2018 with a total delivered luminosity of 150 fb^{-1} . The evolution of integrated luminosity as a function of time is presented in figure 3.3 [15].

The next data taking period, Run 3, will start in 2021 with the center-of-mass energy increased to 14 TeV and after that, the LHC will go to long shutdown 3 and then restart as "High Luminosity LHC (HL-LHC)" in 2026. The detectors, ATLAS and CMS, will also be upgraded to meet the requirements of HL-LHC and provide better performance. The expected integrated luminosity for HL-LHC is about 3000 fb^{-1}. The schedule of ATLAS experiment can be found in figure 3.4.

3.2 The ATLAS Detector

The ATLAS detector is a general purpose particle detector placed at the LHC. The most important objects of the ATLAS detector are to search for the Higgs boson and new physics processes at TeV scale, but it can also probe other physical goals, such as QCD, electroweak interactions and flavor physics. To achieve this, it is the largest detector in the world with a length of 44 m and height of 25 m and weight about 7000 tons. It has a layered structure with cylindrical geometry and almost whole hermetic coverage. The ATLAS detector consists of several subdetectors, each subdetector has specific functionalities in the reconstruction of events. From the center (closed to

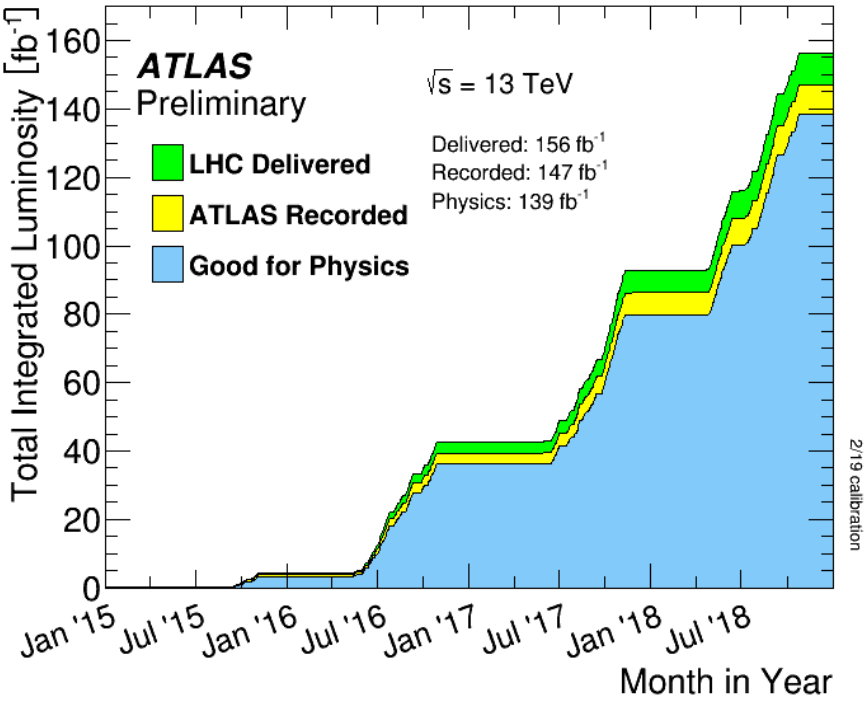


FIGURE 3.3: Delivered luminosity as a function of time in ATLAS Run 2.

FIGURE 3.4: The timeline of ATLAS experiment.

the LHC beam pipe) to the outermost part, first is the inner detector (ID) responsible for the precise measurement of momenta of charged particles, followed by the electromagnetic and hadronic calorimeters measuring the energy deposition of particles and finally the muon spectrometer (MS) to provide tracking and triggering of muons. The overall layout of the ATLAS detector is shown in figure 3.5 [16].

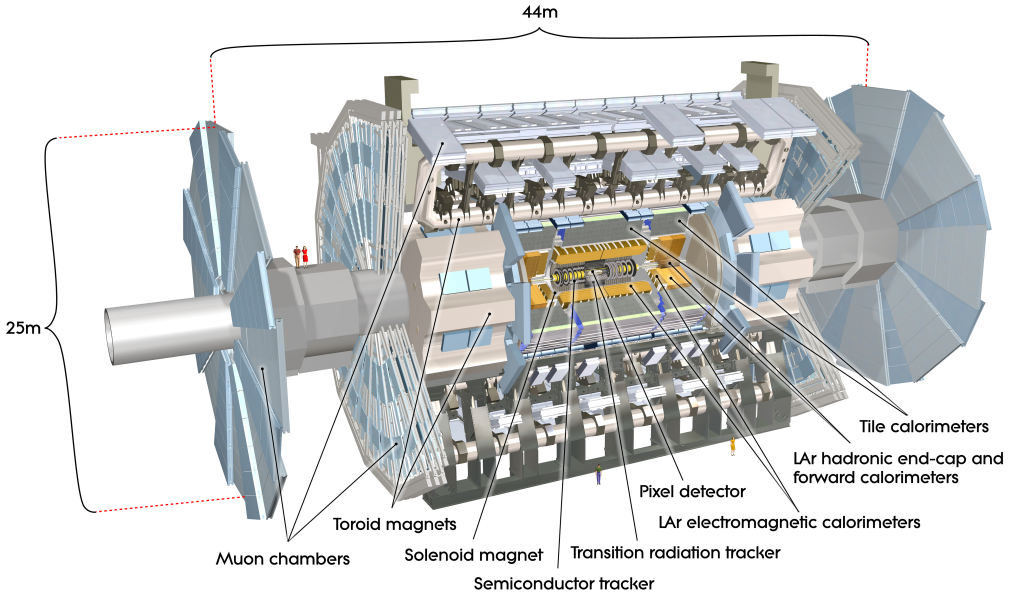


FIGURE 3.5: The overview of the ATLAS detector.

3.2.1 The ATLAS coordinate system

ATLAS defines a coordinate system to describe tracks of particles in the detector. The origin of the coordinate system is set to be at the interaction point (IP) where the beams collide with each other. The x-axis points from IP to the center of LHC ring, while the y-axis points upwards from IP. The z-axis is along the beam direction which is transverse to the x-y plane and directed towards the LHCb detector to gain a right-handed coordinate system.

It is also convenient to define cylindrical coordinats since the ATLAS detector is cylindrical shaped. In this coordinate system, the z-axis remains the same, while the polar angle θ is measured between the z-axis and x-y plane and the azimuthal angle ϕ is measured from the x-axis around the beam in the x-y plane. In practice, the pseudorapidity η is usually used instead of the polar angle θ, defined as:

$$\eta = -\ln(\tan\frac{\theta}{2}) \tag{3.3}$$

It only depends on the polar angle θ, but it should be noted that pseudorapidity is a relativistic approximation of the rapidity y, defined as:

$$y = \frac{1}{2}\ln(\frac{E+p_z}{E-p_z}) \tag{3.4}$$

There are also other two commonly used variables in the ATLAS experiment. One is the transverse momentum, p_T, which is defined in the x-y plane, $p_T = |\mathbf{p}| \sin\theta$. The other one is the angular distance between two objects defined in the $\eta - \phi$ space, as following:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{3.5}$$

3.2.2 Inner tracking detector

The inner detector (ID), shown in figure 3.6 [17], is closest to the beam pipe and immersed in the solenoid, with a length of about 7 meters and a radius of about 1.15 meters. It covers an angular range of $|\eta| < 2.5$ and is designed to provide precise measurement of both primary and secondary vertices of charged tracks, as well as the momenta of charged tracks. The charged tracks are curved by the strong magnetic field (2T at the solenoid center) provided by the surrounding solenoid and then the charge is measured from the direction of curvature while the momentum is measured from the radius of curvature.

The ID is composed of three subdetectors, the pixel detector, the semiconductor tracker (SCT) and the transition radiation tracker (TRT), from inside to outside.

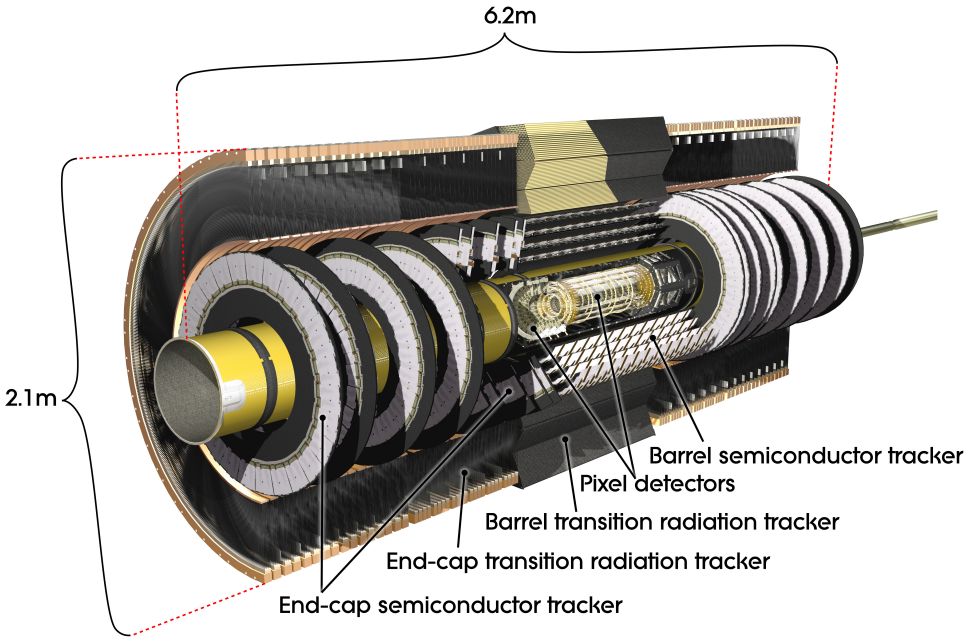


FIGURE 3.6: An overview of the ATLAS inner detector.

3.2.2.1 Pixel detector

The pixel detector consists of four cylindrical layers surrounding the beam pipe (barrel) and six disk layers in the endcap (three at each end of the barrel). The innermost layer is the insertable b-layer (IBL), which was installed in the first long shutdown and functioning since run-II started. It is positioned at a radius of 33 mm and it aims at providing improvements on b-tagging and secondary vertex reconstruction. The first, second and third layers of the barrel locate at 50.5 mm,

88.5 mm and 122.5 mm, respectively. The disk layers are located at z positions of 495 mm, 580 mm and 650 mm. The silicon sensors are pixelated, with a size of 50 μm in ϕ and 400 μm in r/z. They are assembled into about 40000 groups, together with readout electronics. The total number of readout channels is about 80.4 million.

3.2.2.2 Semiconductor tracker (SCT)

The semiconductor tracker is made of silicon strips with a size of 80 μm$\times$ 12 cm. It also has a barrel section and two endcaps. The barrel section consists of four layers, located at radii of 299 mm, 371 mm, 443 mm and 514 mm. The endcap section consists of nine disks on each end at z-positions of 853.8 mm, 934 mm, 1091.5 mm, 1299.9 mm, 1399.7 mm, 1771.4 mm, 2115.2 mm, 2505 mm and 2720.2 mm. The hits in the SCT can provide information to reconstruct the track momentum with 30 % better precision up to p_T = 500 GeV. Each SCT module is made of two layers of silicon strips with a small stereo angle of 40 mrad to each other. This helps improving the space resolution along the longest strip dimension. The resolution of z-position is around 1 mm and the resolution of two separate tracks within 200 μm. The total number of readout channels in SCT is about 6.3 million.

3.2.2.3 Transition Radiation Tracker (TRT)

The transition radiation tracker is an assembly of drift tubes with an inner radius of 56 cm and outer radius of 108 cm. It is designed to help the reconstruction of tracks and vertices, as well as electron identification. It also consists of one barrel section and two endcaps at each end. The barrel section is 114 cm long and made of more than 50000 axially arranged drift tubes (straws). Each endcap section consists of more than 122000 radially arranged straws, divided into 20 indepent wheels. The drift tubes are filled with a mixed gas of mainly xenon and carbon dioxide. The total number of TRT readout channels is about 351000.

3.2.3 Calorimeter

The ATLAS Calorimeter (Figure 3.7 [18]) is separated into one barrel region (coverage $|\eta| < 1.475$) and two endcap regions, just as the inner detector. The barrel region is divided into the LAr electromagnetic barrel and the Tile barrel sub-systems. Each of the endcap region contains the LAr electromagnetic endcap (EMEC), the LAr hadronic endcap (HEC) and the LAr forward (FCal) subsystems. The FCal is located near the inner detector while the HEC is situated behind the EMEC. The calorimeter system is mainly designed to measure the energy of particles. Calorimeters provide a much better and faster reconstruction of energy for very high energy particles than the inner detector. It is also essential for reconstruction of photons which are the key features of many important physical processes. Generally speaking, different particles behave differently in the calorimeter depending on their properties of interaction with materials. For instance, electrons develop an electromagnetic shower in the calorimeter while muons go through the whole calorimeter with minimum energy deposit.

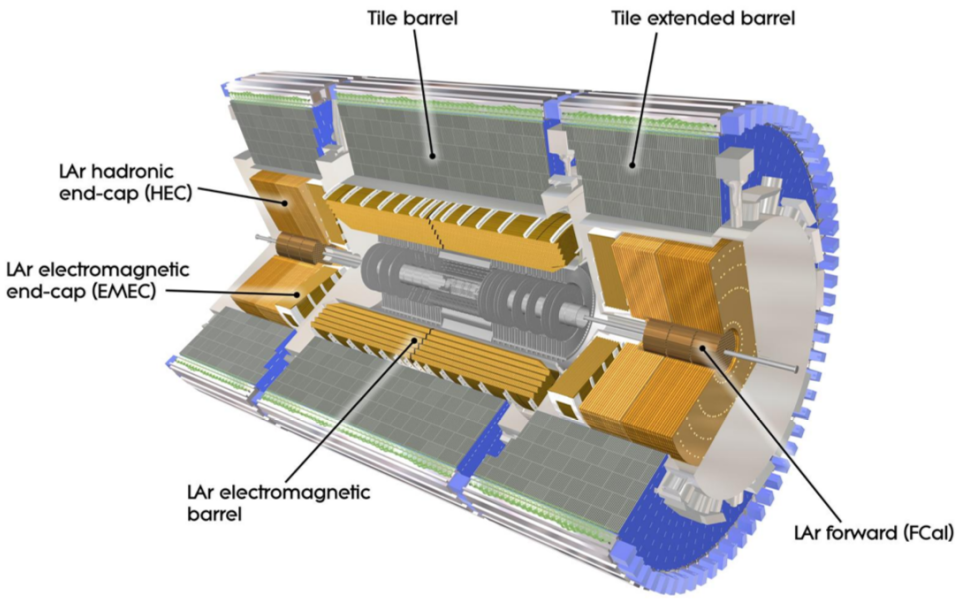


FIGURE 3.7: An overview of the ATLAS calorimeter system.

3.2.3.1 Electromagnetic calorimeter

The electromagnetic (EM) calorimeter (ECal) consists of a barrel section with a coverage of $|\eta| < 1.475$ and two endcap sections covering $1.375 < |\eta| < 3.2$. The barrel section comprises two half-barrels divided by a gap at $z = 0$. Each half-barrel has a length of 3.2 m, inner diameter of 2.8 m and outer diameter of 4 m, with a weight of 57 tons. The endcap section comprises two coaxial wheels in which the inner wheels cover the ranges $1.375 < |\eta| < 2.5$ and the outer wheels cover the ranges $2.5 < |\eta| < 3.2$. The width of a wheel is 63 cm and the weight is 27 tons.

The ECal is a sampling calorimeter. It uses Liquid Argon (LAr) as active medium and lead as absorber, with accordion-shaped electrodes made of Kapton and copper. The accordion shape allows for a full range coverage without any cracks and ensures that each particle travels through approximately the same amount of materials. High energy photons and electrons produce a cascade of electromagnetic particles (an EM shower) containing many low energy electrons and positrons. These charged products ionize the Argon atoms and produce free charges (electrons). Due to the applied high voltage, these free electrons drift to the readout electrodes to induce a signal. These electrodes are placed between the absorbers and surrounded by copper plates with a potential of 2000 V, leading to a drift time of 450 ns. The signal strength on the readout is approximately proportional to the number of drifting electrons and determines the energy measurement and its resolution. In general, the smaller leakage of the EM shower the better the energy resolution is. Therefore a presampler is placed within $|\eta| < 1.8$ to correct for the inhomogeneous energy losses of electrons and photons in the inner detector. It consists of a thin LAr layer only (without lead) with a width of 1.1 cm in the barrel and 0.5 cm in the endcap.

3.2.3.2 Hadronic calorimeter

The hadronic calorimeter (HCal) measures the energy and direction of hadrons (pion, proton, neutron, etc). These hadrons usually originate from the fragmentation and hadronization of quarks and gluons and gather in a small cone which is called "jet". Jets can pass through ID and ECal without significant loss of energy. The HCal is placed just outside the ECal and also comprises one barrel section and two endcap regions. The barrel is divided into a central section which covers $|\eta| < 1.0$ and two extended barrels which cover $0.8 < |\eta| < 1.7$. The inner radius is 2.28 m and extends to an outer radius of 4.25 m, making a total thickness of 9.7 interaction lengths (λ) in the x-y plane. In the barrel section, the HCal uses steel as absorber and plastic scintillator sampling sheets, namely "tiles", as active medium. Hadrons hit the absorber and interact with the nuclei, producing a hadronic shower. Charged particles in the shower produce fluorescence photons when they pass through the scintillators. The photons are collected by the fibres at the end of each tile and carried to the photomultiplier tubes to convert them to an electric signal. The energy is determined from this signal.

The Hadronic End-cap Calorimeter (HEC) uses liquid argon as ECal, but uses copper plates as absorber. Each endcap region comprises two wheels placed behind the EMEC. The HEC covers $1.5 < |\eta| < 3.2$, with a small overlap of the tile calorimeter.

3.2.4 Muon spectrometer

The Muon spectrometer (MS) is the outermost part of the ATLAS detector, as illustrated in figure 3.8 [19]. It is designed to provide accurate measurement of the muon momentum indepently from the inner detector, as well as providing a muon trigger. The MS consists of two types of precision detectors ($|\eta| < 2.7$) and two types of triggering detectors ($|\eta| < 2.4$).

The tracks of muons are bent by the the large superconducting air-core toroid magnet. In the range of $|\eta| < 1.4$, the magnetic field is provided by the large barrel toroid. In $1.6 < |\eta| < 2.7$, the magnetic field is provided by two smaller endcap magnets which are inserted into the ends of the barrel toroid. In $1.4 < |\eta| < 1.6$, namely the transition region, the magnetic field is involved by a combination of barrel and endcap magnets.

3.2.4.1 Muon precision detectors

The measurement of momentum in the barrel region is performed by the Monitored Drift Tube chambers (MDTs). The chambers consist of three to eight layers of drift tubes, operating at an absolute pressure of 3 bar. The resolution is 80 μm per tube in the bounding plane, or 35 μm per chamber on average. The Cathode Strip Chambers (CSC) provide the measurement in the forward region ($2.0 < |\eta| < 2.7$). The CSCs are multiwire proportional chambers with cathode planes segmented into strips in orthogonal directions. The resolution is 40 μm in the bending plane and 5 mm in the transverse plane.

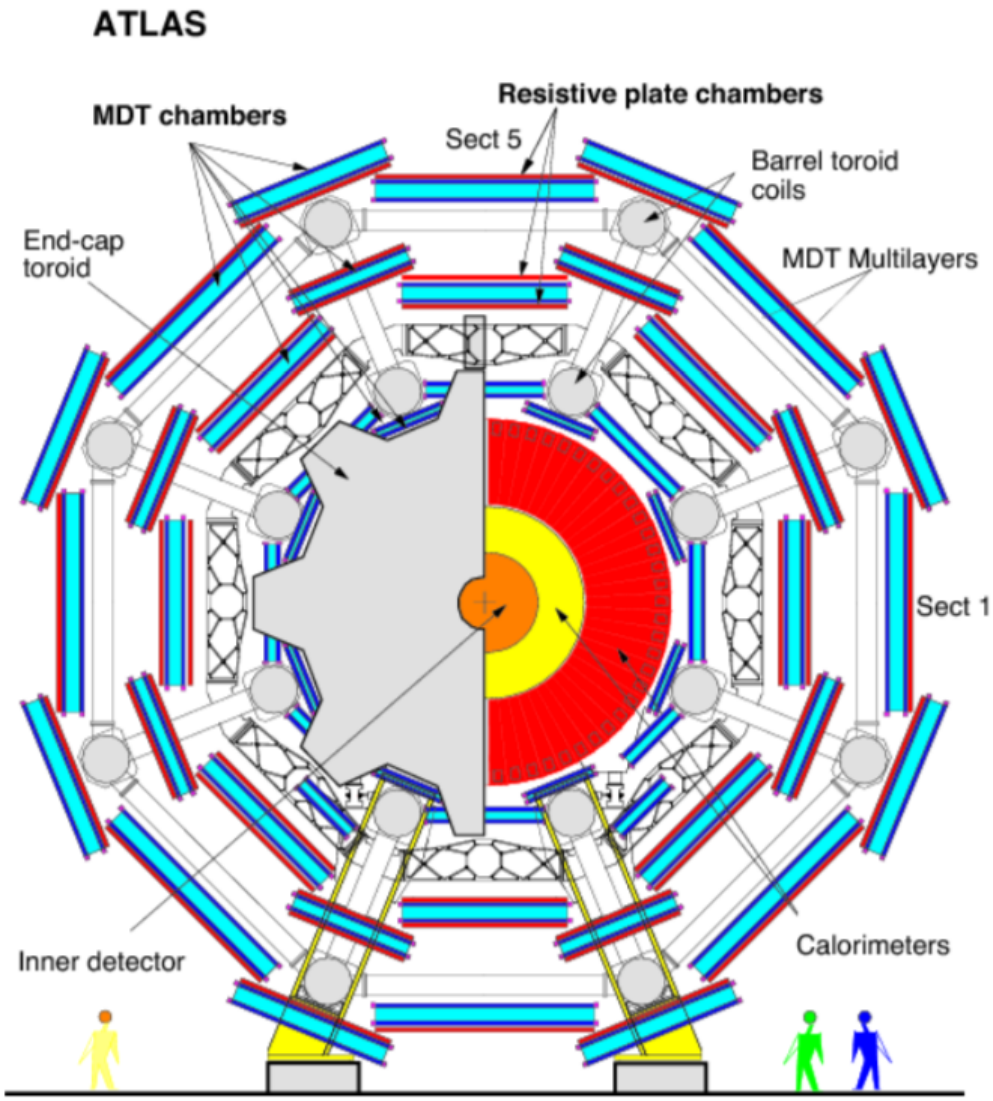


FIGURE 3.8: Schematic view of the Muon spectrometer projected to x-y plane.

3.2.4.2 Muon trigger detectors

The trigger detectors can provide a fast triggering within a few tens of nanoseconds after the passage of muons to deliver the track information. In the barrel region, the triggering is provided by the Resistive Plate Chambers (RPC) while in the endcap regions it is provided by the Thin Gap Chambers (TGC). There is a threefold purpose for these trigger detectors: provide bunch-crossing information, provide well-defined p_T threshold and measure the muon coordinate in the direction orthogonal to the bending plane.

3.2.5 Trigger

The designed luminosity of LHC is 10^{34} cm^{-2}s^{-1} and the bunch crossing rate is 40 MHz, thus it is not possible to read and store the information of all collisions. For the ATLAS physics program, only a small fraction of events are of interest, this also leads to the requirement of a trigger system. Given the limitation of offline processing capacity, the maximum event rate that can be stored is 400 Hz which means the event rate has to be reduced by a factor of 100k. In order to reach this goal, the ATLAS has a three level trigger system, level 1 (L1), level 2 (L2) and event filter [20]. A schematic layout of the ATLAS run2 trigger system is shown in figure 3.9. The L1 trigger is based on hardware, by using customized electronics. The L2 trigger and event filter are also called high level trigger (HLT) which are both based on software. Each subsequent level refines the decisions made in the previous levels and applies additional selection where necessary.

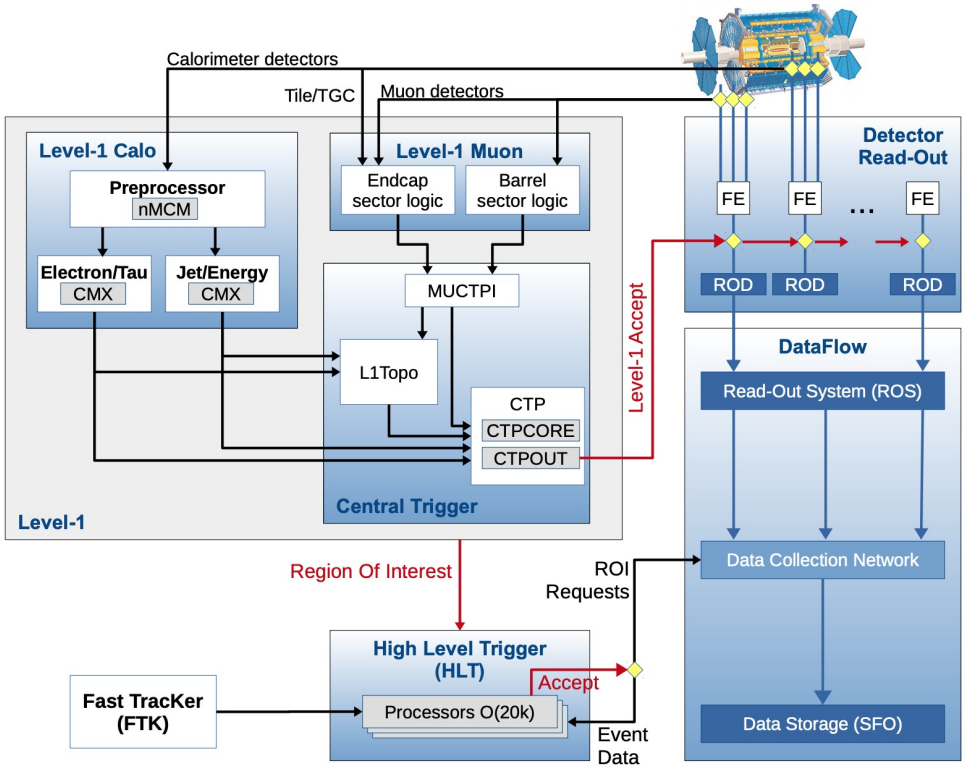


FIGURE 3.9: Schematic layout of the ATLAS trigger and data acquisition system in Run-2 [21].

The L1 trigger searches for objects with high transverse momentum, such as muons, electrons, photons, jets as well as missing transverse energy (MET). It uses reduced-granularity information

from detectors: RPC and TGC for muons and all of the calorimeter sus-systems for EM clusters, jets, τ and MET. The maximum L1 accept rate is 75 kHz which is limited by capability of the read-out systems. The decision taken by L1 trigger must reach the front-end electronics within 2.5 μs after the corresponding bunch-crossing. Results from the muon and calorimeter trigger are further processed by the central trigger processor, where a trigger menu made of combinations of trigger selections is implemented. Events passing L1 trigger are then sent to next stage, high level triggers.

The L2 selections use all available detector data and reduce the event rate to about 3.5 kHz in about 40 ms for each event on average. The final stage is the event filter, which reduces the event rate to roughly 400 Hz. The process is done by using offline analysis softwares in about four seconds for each event.

3.2.6 Magnetic systems

The ATLAS has a unique hybrid magnet system that consists of four large superconducting magnets, one solenoid and three toroids. This system is 22 m in diameter and 26 m in length, with a stored energy of 1.6 GJ in total. The solenoid is parallel to the z-axis and provide an axial magnetic field with a maximum strength of 2 T. It is also designed to have minimum radiative thickness in front of the ECal. The toroids consist of one barrel section and two endcap sections, which can provide a toroidal magnetic field of roughly 0.5 T and 1.0 T for the muon spectrometers in the barrel and endcap regions, respectively. A simple illustration of the magnet geometry can be found in figure 3.10 [22].

3.2.7 Simulation of ATLAS detectors

The ATLAS collaboration provides a full set of toolkits to simulate the performance of the ATLAS detector [23]. The event simulation [24] can convert the events generated by the Monte Carlo generators to an output which has an identical format to the one deliverd by the ATLAS detector, allowing for a direct comparison between the real data and theoretical predictions. The simulation of detectors is done by using the GEANT4 package [25]. This toolkit can basically simulate the entire ATLAS detectors including geometry, materials and subsystems (electronics, trigger, etc), as well as the response of detectors to the passing particles. All these informations are processed by the same procedure as the real data taking, converting signals to voltages and finally digitized signals (digits). The outputs of digitization is called RAW data and can be used for object reconstructions.

This is so called "full simulation", which can provide a precise description of the interaction of particles with the detector but is very time-consuming. This leads to the fast simulation strategies, which can quickly mimic the response from GEANT4 simulation and thus allows a fast production of simulated events for physics analysis. The framework of fast simulation is called ATLFASTII [26], which consists of two components: the Fast Calorimeter Simulation (FastCaloSim) for calorimeter simulation and the FAST ATLAS Tracking system for ID and MS simulations. Any of these sub-detectors can be simulated with the nominal GEANT4 indepently to others, providing the flexibility to meet the demands of different physics analysis.

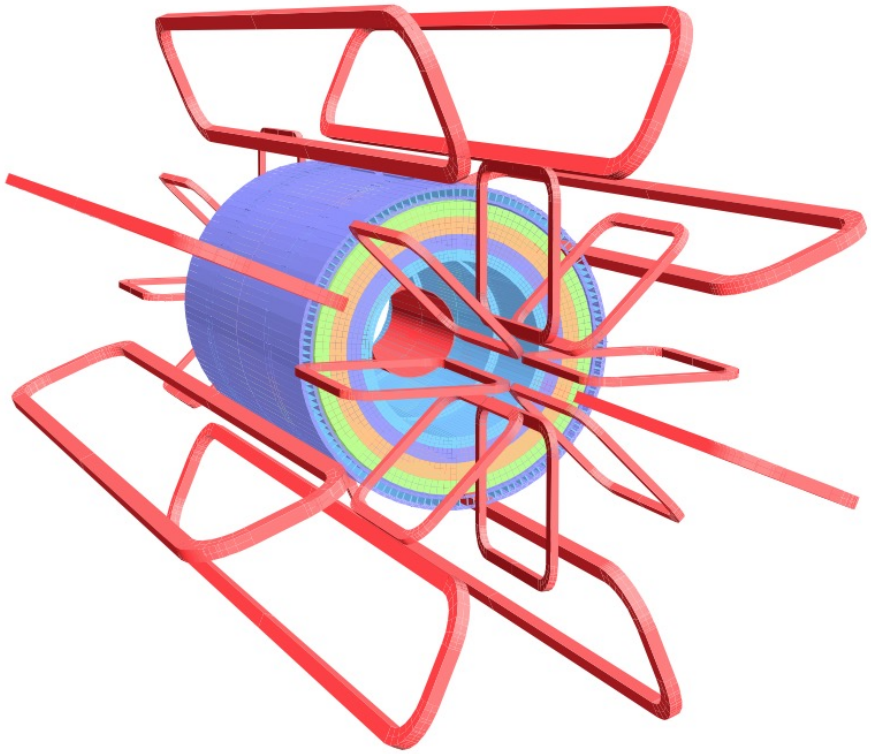


FIGURE 3.10: The geometry of ATLAS magnetic systems. The solenoid (shown in red) lies inside the calorimeter volume. The toroids (red square) are outside the colorimeter volume.

3.3 Upgrade of Inner Detector towards High-Luminosity LHC (HL-LHC) : the Inner Tracker project (ITk)

So far LHC has been the most important scientific tool devoted to study the fundamental nature at the energy frontier during the past two decades. LHC will continue operating in the next 20 years and will be upgraded several times which will be resulted in an integrated luminosity of 3000 $fb^{-1}$for the proton-proton collisions [27]. This section will focus on the upgrade of the inner detector where I worked on one of the most important subdetectors to be replaced in HL-LHC. The upgrades to achieve the accelerator performance will be mostly realised during two long shutdowns:

- Long Shutdown 1 (LS1) was completed in 2015 to prepare the accelerator for operation at 13 TeV.
- Long Shutdown 2 (LS2) will take place in 2019/2020 to further improve the performance of LHC, accompanied by significant detector upgrades (Phase-I)
- Long Shutdown 3 (LS3) will start at the end of 2023 to include major performance upgrades of the accelerator for the high-luminosity LHC (HL-LHC) which also requires replacements of several major detector components (Phase-II).

The expected luminosity of HL-LHC is $5 \times 10^{34}cm^{-2}s^{-1}$ ($7.5 \times 10^{34}cm^{-2}s^{-1}$) with an average number of interactions per bunch-crossing $\langle\mu\rangle$ = 140(200). This leads to an even higher pile-up comparing to current LHC and the integrated radiation dose is also epected to increase by approximately a factor of ten. Thus the HL-LHC will present an extremely challenging environment for the ATLAS detector, well beyond that for which it was designed. The ATLAS collaboration decided that it will replace the Inner Detector with a new all-silicon tracker [28] to maintain tracking performance in this high-occupancy and high radiation dose environment.

Several new technologies are introduced to make sure that the tracker can work properly in this extreme environment. Meanwhile the material is also reduced to improve the performance at forward region. A new track trigger is implemented due to the new read-out scheme, which could improve the ATLAS online data selection capability. The new tracker consists of two parts: a greatly enlarged Pixel detector which has roughly twice the radius and four times the length of the current one; a segmented strip detector which has more than three time the silicon area comparing the current SCT in order to cover the full radius of the solenoid inner bore.

The design of the new tracker is basically driven by the performance required in track reconstruction efficiency, track resolution, vertex reconstruction and large eta region.

- ***Track reconstruction efficiency*** The track reconstruction efficiency for muons with $p_T > 1$ GeV is required to be 99 % within $|\eta| < 2.7$. For particles that interact more strongly with the detector material, such as charged hadrons and electrons, it must be possible to reconstruct those originating from within a transverse window of ± 2 mm around the interaction region. It should have a good two-track separation resolution to measure the tracks in the core of high-energy jet with high efficiency. It also should be possible to reconstruct electrons from

photon conversions up to a radius of about 30 cm, which allows the study of the material distribution within the Pixel system.

- ***Track resolution*** In order to avoid contamination from falsely identified high momentum tracks, the fraction of tracks with an estimated transverse momentum $|p_T^{Reco} - p_T^{true}|/p_T^{true} > 0.5$ should not exceed a rate of 1 % over the entire detector coverage for particles following the p_T-spectrum of the underlying min bias events. The rate of charge misidentification should also be smaller than 0.5 % for muons with p_T = 500 GeV and not exceed 10 % at p_T = 2 TeV. The track extrapolation to the calorimeter must have spatial resolution better than that of the fitted energy-deposit cluster centroid.

- ***Vertex reconstruction*** With 200 pile-up events, the mean separation of primary vertices is typically less than 1 mm. Therefore it is not possible to reconstruct all vertices in a triggered event individually. However, it is important that high transverse momentum objects (muons, electrons and tracks in high transverse energy jets) coming from a common vertex can all be correctly associated to the same vertex with good efficiency. Vertex reconstruction in such an environment poses stringent requirements on the tracking resolution close to the interaction point and imposes the need to minimise the amount of material of the inner layers of the Pixel detector.

- ***Large eta region*** A wide range of physics analyses of particular relevance to the HL-LHC physics program, for example vector boson fusion (VBF), require track reconstruction beyond the $\eta = \pm 2.7$ of the current Inner Detector. Forward jet reconstruction, missing transverse energy resolution and pile-up jet rejection all benefit significantly from extending tracking coverage out to $\eta = \pm 4.0$. This extended coverage is achieved through adding further Pixel rings.

The layout of the ITk Detector is optimised by focusing on tracking performance, cost effectiveness, ease of construction and installation, as well as the ability to operate the detector up to the end of the HL-LHC operation. Two options (figure 3.11) to extend the Pixel Detector coverage up to η = 4 have been investigated, named Inclined and Extended. The Inclined layout can minimise the materials transersed by particles in the very forward region and then provide better parameter resolution, as well as providing more hits to reduce the complexity of the Pixel Detector end-cap ring system and cover the same η range as Extended layout with less silicon surface. In the central region of the ITk Detector, sensors are arranged in cylinders around the beam axis, with (starting from inside) five pixel layers followed by two short-strip layers of paired stereo modules then two long-strip layers of paired stereo modules. The forward regions will be covered by six strip disks and a number of pixel rings leading to one or more hits depending on the ring layer and η position. An overview of the ITk layout can be found in figure 3.12 and figure 3.13.

It is also important to reduce the material seen by the particles, since the scattering of particles inducted in the detector material is an importance source of noise in the track reconstruction. Thus it is important for us to understand and minimise the material in the inner detector. By re-designing the readout system and the cooling system, the material in the large η region is extremely reduced, as shown in figure 3.14.

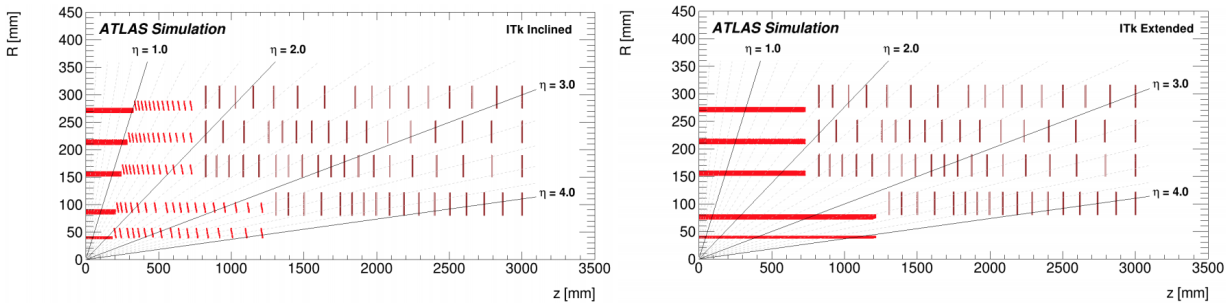


FIGURE 3.11: Diagrams showing simulated energy deposits in active layers for the two candidate layouts zoomed in on the Pixel barrel. The Inclined layout shows in left and Extended layout shows in right. The Inclined layout provides better resolution while the Extended layout has simpler stave support.

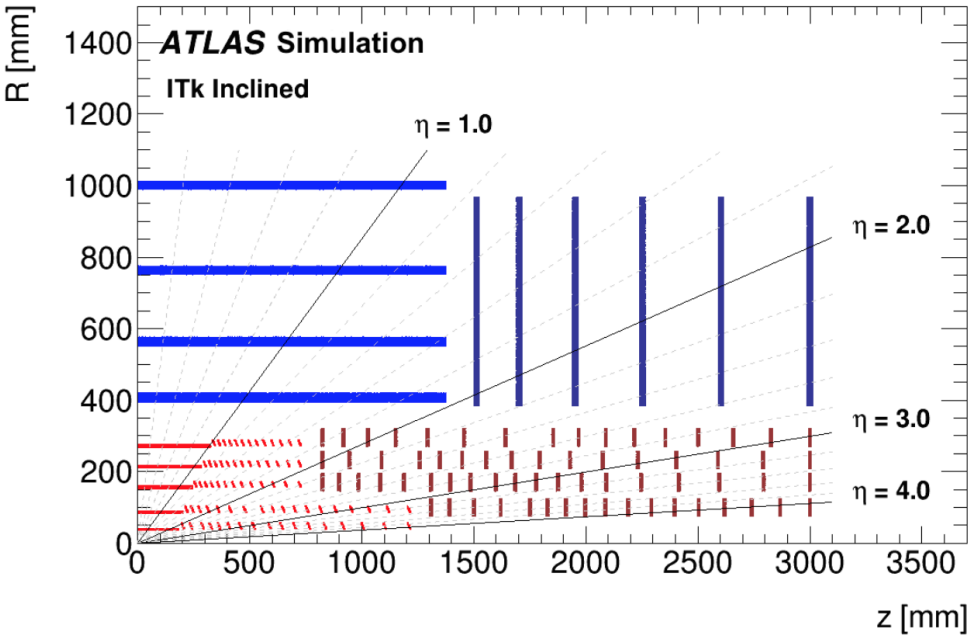


FIGURE 3.12: Schematic layout of the ITk. Here only one quadrant and only active detector elements are shown. The horizontal axis is the axis along the beam line with zero being the interaction point. The vertical axis is the radius measured from the IP. The outer radius is set by the bore of the solenoid.

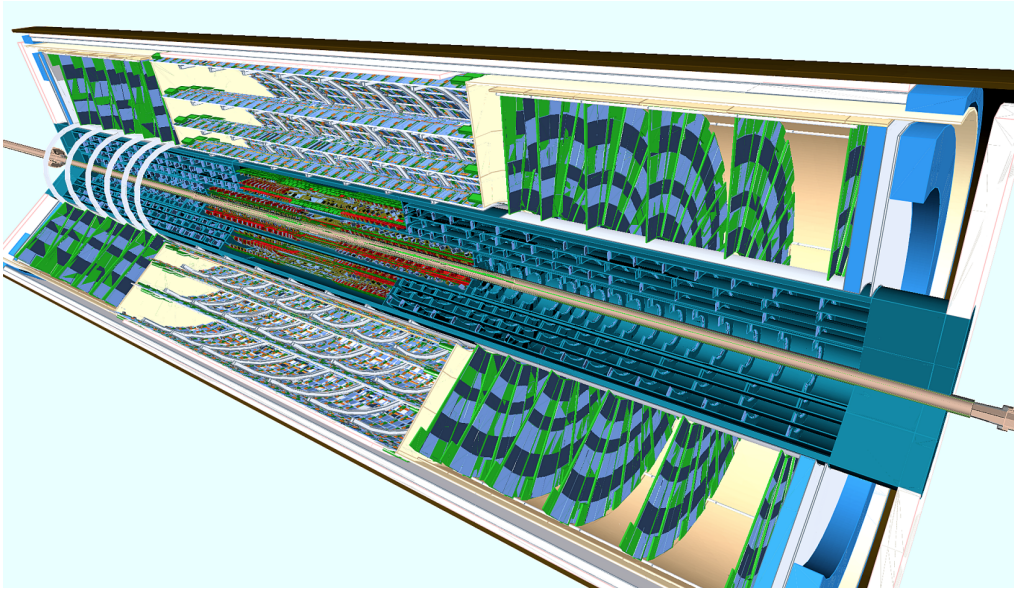
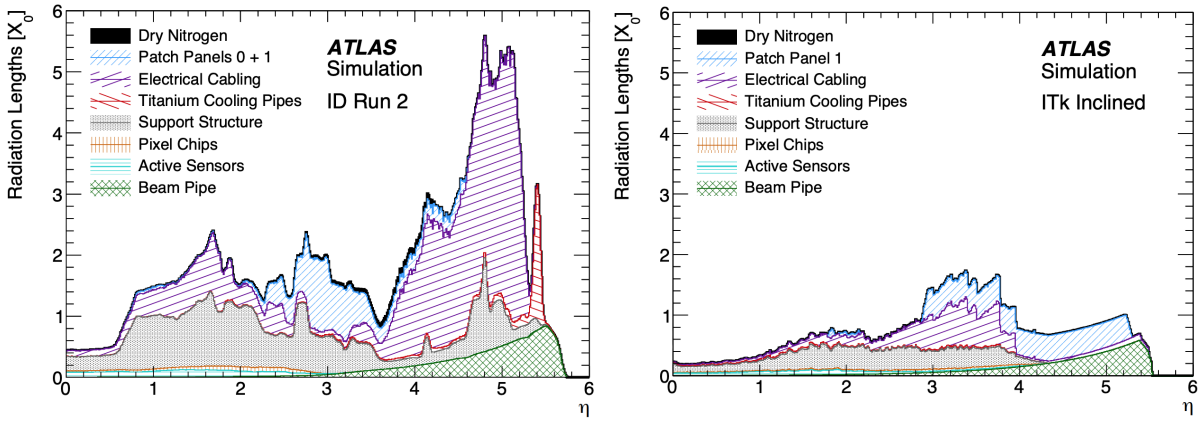


FIGURE 3.13: A visualisation of the ITk as implemented in the simulation framework. The beam pipe is shown in brown. The Pixel detector is closest to the beam pipe and the Strip detector is outside Pixel.



(A) Run2 ID (B) ITk

FIGURE 3.14: Radiation length X_0 versus η for the current ATLAS ID (Pixel, SCT, TRT) in the left and ITk (Pixel, Strip) in the right.

In order to estimate and optimize the performance of ITk, the ATLAS simulation tools, Athena, is used and the new detector geometry is implemented by C++ code. The parameters, such as size and mass of the various detector components, are provided by engineers. Therefore there is always a possibility that mistakes could be implemented when importing these numbers to the simulation program. I developed a tool, called "MassDebugger" [29], to spot potential bugs in the implemented geometry. This tool can read and print all the volumes implemented in GEANT4, and convert it to human-readable format. It is convenient to check the mass of each component by using this tool and it can also provide a tree-structure of the Pixel Detector components (figure 3.15).

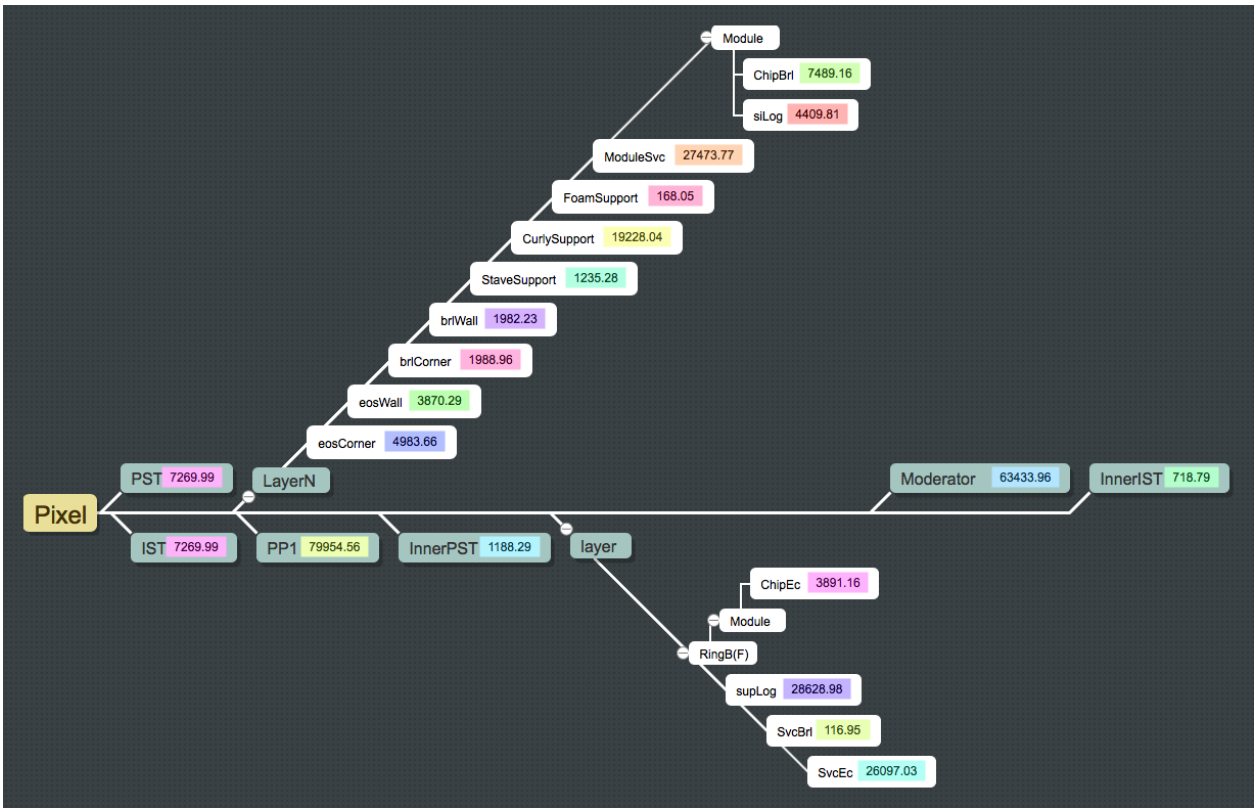


FIGURE 3.15: A diagram of the tree-structure of the Pixel Detector in ITk (layout version step2). The relationship and mass (shown in the colorful boxes) are read from the mass debugging tool developed by me.

3.4 Object reconstruction

The particles passing through the ATLAS detector are reconstructed as physical objects by dedicated algorithms, using the information collected from each sub-detector. The reconstruction algorithms are developed from simulation and then corrected by real data. They are evolving with time as more and more data is collected. The often use objects, electrons (3.4.1 [30]), muons (3.4.3 [31]), taus (τ) (3.4.4 [32]), photons (γ) (3.4.2 [33]), jets (3.4.5 [34, 35]) as well as missing transverse energy (MET) (3.4.6 [36]), are described in this section.

3.4.1 Electrons

Electrons play an important role in many ATLAS physics analysis, because they are primary signature for lots of processes, as well as suppressing the large backgrounds from QCD processes. Due to the fact that particles can interact with detector material through electromagnetic processes, the reconstruction and identification of electrons are big challenges. There are also fake electrons (non-prompt electrons) that originate from photon conversions and hadron decays.

3.4.1.1 Reconstruction

Electrons can lose a significant amount of energy due to bremsstrahlung when interacting with the material. Then the radiated photon can produce an electron-positron pair and the converted electron and positron can again interact with the detector material. These photons, electrons and positrons are usually emitted in a very collimated cone and are normally reconstructed as part of the same electromagnetic cluster. These interactions can also happen inside the inner detector, generating multiple tracks in the inner detector. As a result, it is possible to produce and match multiple tracks to the same electromagnetic cluster, all originating from the same primary electron. The reconstruction of electron candidates within the kinematic region encompassed by the high-granularity electromagnetic calorimeter and the inner detector is based on three fundamental components characterising the signature of electrons: localised clusters of energy deposits found within the electromagnetic calorimeter, charged-particle tracks identified in the inner detector and close matching in η space of the tracks to the clusters to form the final electron candidates. Figure 3.16 shows a schematic illustration of the elements that enter into the reconstruction and identification of an electron.

The reconstruction is performed in three steps, seed-cluster reconstrution, track reconstruction and electron-candidate reconstrution. The ECal is divided into a grid of 200×256 elements (towers) in the $\eta \times \phi$ space and the seed is built by searching a local energy deposit in the towers. The basic building block for track reconstruction is a hit in one of the inner detector tracking layers. Charged-particle reconstruction in the inner detector begins by assembling clusters from these hits. Three-dimensional measurements (space-points) are created from these clusters. Track seeds are then formed from sets of space-points in the silicon-detector layers. Then the track candidate is obtained from a χ^2 fit of the hits. Finally the electron-reconstruction is completed by matching the track-candidate to the candidate calorimeter seed cluster and the determination of the final cluster size.

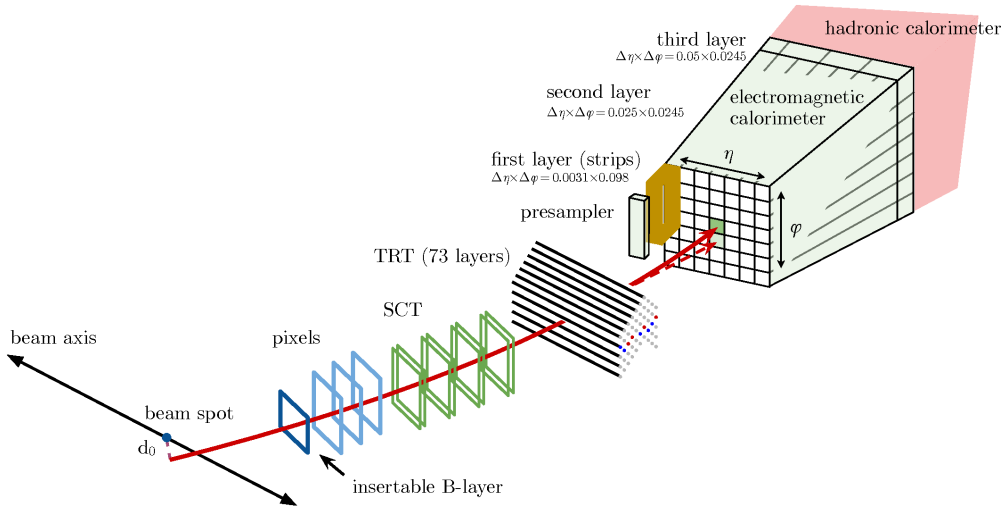


FIGURE 3.16: A schematic illustration of the path of an electron through the detector. The red trajectory shows the hypothetical path of an electron, which first traverses the tracking system (pixel detectors, then silicon-strip detectors and lastly the TRT) and then enters the electromagnetic calorimeter. The dashed red trajectory indicates the path of a photon produced by the interaction of the electron with the material in the tracking system [30].

3.4.1.2 identification

Prompt electrons entering the central region of the detector ($|\eta| < 2.47$) are selected using a likelihood-based (LH) identification. The inputs to the LH include measurements from the tracking system, the calorimeter system, and quantities that combine both tracking and calorimeter information. The electron LH is based on the products for signal, LS , and for background, LB, of n probability density functions (pdfs), P:

$$L_{S(B)}(\mathbf{x}) = \prod_{i=1}^{n} P_{S(B),i}(x_i) \tag{3.6}$$

where $\mathbf{x}$ is the vector of the various quantities used in the identification. $P_{SB,i}(x_i)$ is the value of the signal(background) pdf for quantity i at value x_i. The signal is prompt electrons, while the background is the combination of jets that mimic the signature of prompt electrons, electrons from photon conversions in the detector material, and non-prompt electrons from the decay of hadrons containing heavy flavours.

The electron LH identification is based on a discriminant d_L formed for each electron candidate:

$$d_L = \frac{L_S}{L_S + L_B} \tag{3.7}$$

The discriminant d_L has a sharp peak at unity (zero) for signal (background). This sharp peak makes it inconvenient to select operating points as it would require extremely fine binning. An inverse sigmoid function is used to transform the distribution:

$$d'_L = -\tau^{-1} ln(d_L^{-1} - 1) \tag{3.8}$$

where parameter τ is fixed to 15. An example of the distribution of a transformed discriminant is shown in figure 3.17 for prompt electrons from Z boson decays and for background.

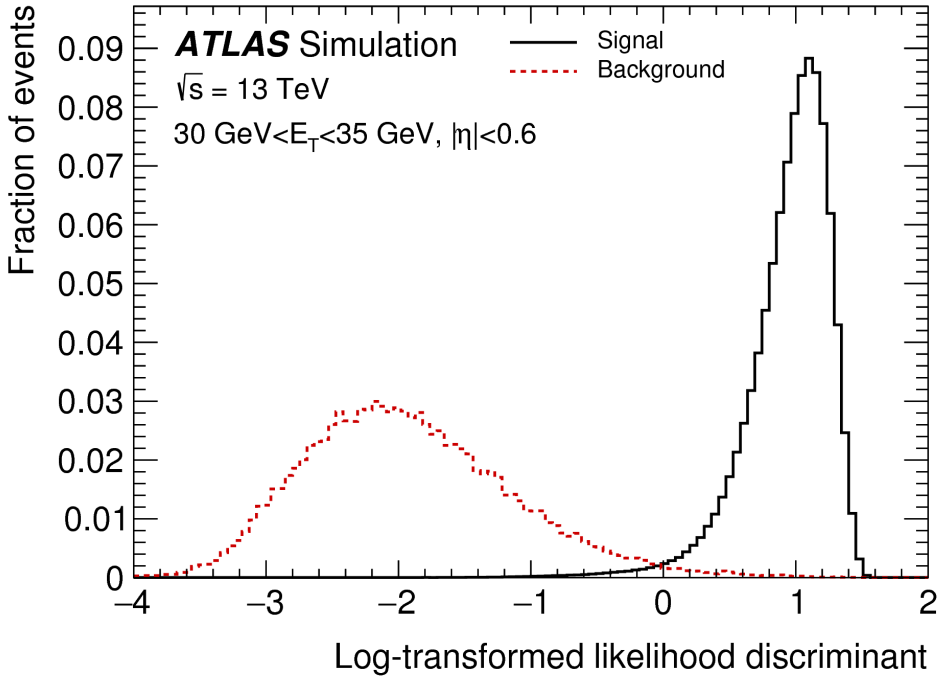
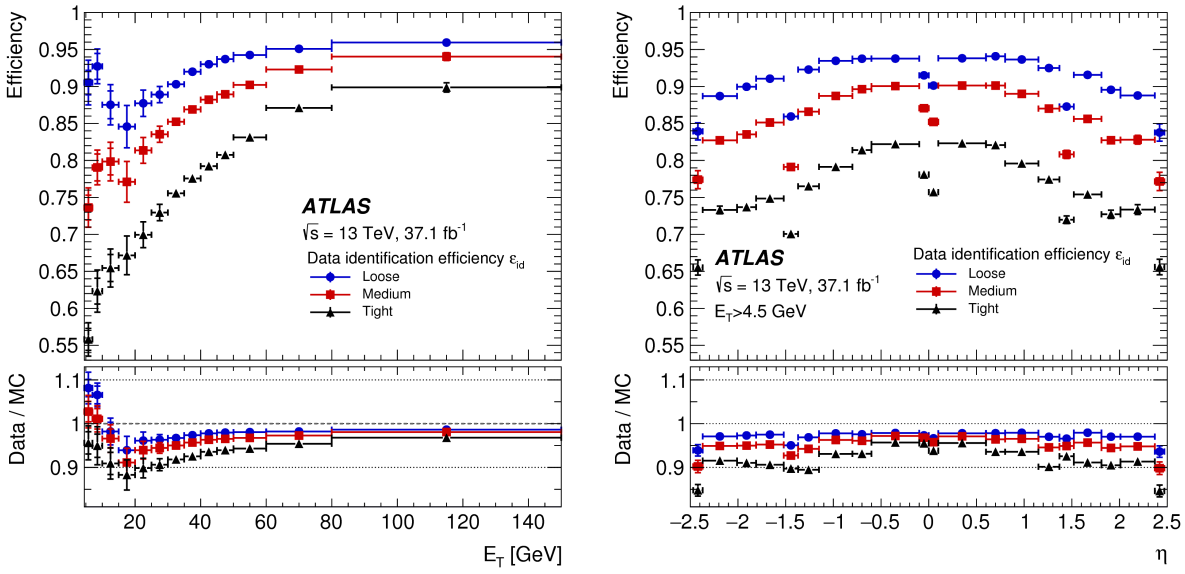


FIGURE 3.17: The transformed LH-based identification discriminant d_L' for reconstructed electron candidates with good quality tracks with 30 GeV $< E_T <$ 35 GeV and $|\eta| < 0.6$. The black histogram is for prompt electrons in a $Z \rightarrow ee$ simulated sample and the red histogram is background in a generic two-to-two process simulation sample. The histograms are normalised to unit area [30].

To cover the various required prompt-electron signal efficiencies and corresponding background rejection factors needed by the physics analyses carried out within the ATLAS Collaboration, four fixed values of the LH discriminant are used to define four operating points, VeryLoose, Loose, Medium, and Tight. The lower efficiencies of the Medium and Tight operating points compared to Loose result in an increased rejection of background. The identification is optimised in bins of cluster η and bins of E_T. The selected bins in cluster $|\eta|$ are based on calorimeter geometry, detector acceptances and the variation of the material in the inner detector. The pdfs of the various electron-identification quantities vary with particle energy, which motivates the bins in E_T. The performance of LH identification is shown in figure 3.18.

In addition to the identification, most analyses require electrons to fulfil isolation requirements, to further discriminate between signal and background. The isolation variables quantify the energy of the particles produced around the electron candidate and allow to disentangle prompt electrons from other, non-isolated electron candidates such as electrons originating from converted photons produced in hadron decays, electrons from heavy flavour hadron decays, and light hadrons misidentified as electrons. Two discriminating variables have been designed for this purpose:

- a calorimetric isolation energy, $E_T^{cone0.2}$, defined as the sum of transverse energies of topological clusters calibrated at the electromagnetic scale, within a cone of $\Delta R = 0.2$ around the candidate electron cluster.



(A) (B)

FIGURE 3.18: Measured LH electron-identification efficiencies in $Z \to ee$ events for the Loose (blue circle), Medium (red square), and Tight (black triangle) operating points as a function of E_T (3.18a) and η (3.18b). The data efficiencies are obtained by applying data-to-simulation efficiency ratios that are measured in $J/\psi \to ee$ and $Z \to ee$ events to the $Z \to ee$ simulation. For both plots, the bottom panel shows the data-to-simulation ratios [30].

- a track isolation, $p_T^{varcone0.2}$, defined as the sum of transverse momenta of all tracks, satisfying quality requirements, within a cone of $\Delta R = \min(0.2, 10\ \text{GeV}/E_T)$ around the candidate electron track and originating from the reconstructed primary vertex of the hard collision, excluding the electron associated tracks.

A variety of selection requirements on the quantities $E_T^{cone0.2}/E_T$ and $p_T^{varcone0.2}/E_T$ have been defined to select isolated electron candidates. The resulting operating points are divided into two classes:

- efficiency targeted operating points: varying requirements are used in order to obtain a given isolation efficiency ε_{iso} which can be either constant or as a function of E_T.
- fixed requirement operating points: in this case the upper thresholds on the isolation variables are constant.

3.4.2 Photons (γ)

The reconstruction of photons are performed separately for converted and unconverted photons. Converted photons are characterized by the presence of at least one track originating from a vertex inside the inner detector matched to the EM cluster while not required for unconverted photons. A conversion vertex in the inner detector is identified as the originating from a vertex inside the tracker matched to the of tracks of an EM cluster. A track is added to the converted photon collection if it is matched in η and ϕ to a reconstructed EM cluster. The matching requires that the impact parameter associated to the track is inside a (η, ϕ) window with a radius of 0.05 from the cluster

center, after extrapolation from its last measurement point to the second EM layer. The track is extrapolated to the position corresponding to the expected maximum energy deposit for EM showers. If a EM cluster cannot be associated to a track, then it is considered to be an unconverted photon candidate.

Almost all converted photons are also considered as electrons. The presence of a track associated to the conversion vertex is used to identify between these two categories. If the track associated to the EM cluster coincides with a track originating from a conversion vertex, then the electron is regarded as a converted photon object. If the track cannot be associated to a conversion vertex, then the object is identified as an electron. In case that an initially reconstructed electron with a matched track only has TRT information, the candidate is considered to be a converted photon even if there is a conversion vertex associated to the EM cluster. There is also a recovery procedure for the unconverted photons: if the reconstructed electron has the best track candidate with only TRT information or the converted photon candidate condition is not passed and the track p_T is smaller than 2 GeV, the candidate is considered to be an unconverted photon. The calibration of photons is similar to that of electrons.

Photon identification with high signal efficiency and high background rejection is required for transverse momenta from 10 GeV to the TeV scale to distinguish prompt photons from background photons. The identification is done by cutting on several discriminating variables characterising the lateral and longitudinal shower development in the electromagnetic calorimeter and the shower leakage fraction in the hadronic calorimeter. Prompt photons produce narrower energy deposits in the electromagnetic calorimeter and have smaller leakage to the hadronic calorimeter comparing to background photons from jets. There are two selections, `Loose` and `Tight`, defined in ATLAS. The `Loose` selection is only based on shower shapes in the second layer of the electromagnetic calorimeter and on the energy deposited in the hadronic calorimeter, with a efficiency of 97 % at E_T = 20 GeV and 99 % at $E_T > 40$ GeV for both converted and unconverted photons. The `Tight` selection adds information from finely segmented strip layer of the calorimeter and optimized separately for unconverted and converted photons, with an efficiency of 85 % at $E_T > 40$ GeV.

3.4.3 Muons

3.4.3.1 Reconstruction

Muon reconstruction is first performed independently in the ID and MS. The information from individual subdetectors is then combined to form the muon tracks that are used in physics analyses. Muon reconstruction in the MS starts with a search for hit patterns inside each muon chamber to form segments. The MDT segments are reconstructed by performing a straight-line fit to the hits found in each layer. The RPC or TGC hits measure the coordinate orthogonal to the bending plane. Segments in the CSC detectors are built using a separate combinatorial search in the η and ϕ detector planes. Muon track candidates are then built by fitting together hits from segments in different layers. The algorithm used for this task performs a segment-seeded combinatorial search that starts by using the segments generated in the middle layers of the detector as seeds where more trigger hits are available. The search is then extended to use the segments from the outer and inner layers as seeds. The segments are selected using criteria based on hit multiplicity and fit quality

and are matched using their relative positions and angles. At least two matching segments are required to build a track, except in the barrel–endcap transition region where a single high-quality segment with η and ϕ information can be used to build a track.

Since the same segment can be initially used to build several track candidates, an overlap removal algorithm selects the best assignment to a single track, or allows for the segment to be shared between two tracks. The hits associated with each track candidate are fitted using a global χ^2 fit. A track candidate is accepted if the χ^2 of the fit satisfies the selection criteria.

3.4.3.2 Identification

Muon identification is performed by applying quality requirements that suppress background, mainly from pion and kaon decays, while selecting prompt muons with high efficiency and/or guaranteeing a robust momentum measurement. Several variables offering good discrimination between prompt muons and background muon candidates are studied in simulated $t\bar{t}$ events. Muons from W decays are categorized as signal muons while muon candidates from light-hadron decays are categorized as background. Four muon identification selections (`Medium`, `Loose`, `Tight`, and `High-pT`) are provided to address the specific needs of different physics analyses. `Medium`, `Loose` and `Tight` are inclusive categories in that muons identified with tighter requirements are also included in the looser categories. The reconstruction efficiencies for signal and background obtained from $t\bar{t}$ simulated events are listed in table 3.1, separating low ($4 < p_T < 20$ GeV) and high ($20 < p_T < 100$ GeV) transverse momentum muon candidates.

	Low		High	
Selection	ϵ_{μ}^{MC} [%]	$\epsilon_{Hadrons}^{MC}$ [%]	ϵ_{μ}^{MC} [%]	$\epsilon_{Hadrons}^{MC}$ [%]
Loose	96.7	0.53	98.1	0.76
Medium	95.5	0.38	96.1	0.17
Tight	89.9	0.19	91.8	0.11
High-pt	78.1	0.26	80.4	0.13

TABLE 3.1: Efficiency for prompt muons from W decays and hadrons decaying in flight and misidentified as prompt muons computed using a $t\bar{t}$ MC sample, separating low ($4 < p_T < 20$ GeV) and high ($20 < p_T < 100$ GeV) transverse momentum muon candidates with $|\eta| < 2.5$. No isolation requirement is applied in the selection.

3.4.4 The (τ) lepton

The τ lepton is extremely massive comparing to electron and muon, thus it has a very short lifetime and decays either leptonically or hadronically before it reaches the ATLAS detector. In ATLAS, only hadronically decaying τs (labelled as τ_{had}) are considered for reconstruction, while the leptonically decay τs are reconstructed as the corresponding lepton. The hadronic decay products contain one or three charged pions in 72 % and 22 % of all cases, respectively, and always with a τ neutrino. In 68 % of all hadronic decays, at least one associated neutral pion is produced as well [6]. These charged and neutral hadrons are the signal of τ_{had} in the detector. τ candidates are seeded by jets formed using the procedure described in section 3.4.5.

A dedicated energy calibration is applied to τ candidates in order to correct the energy deposition measured in the detector to the average value of the energy carried by the measured decay products at the generator level. There are two calibrations available, known as the baseline calibration and the boosted regression tree (BRT) based calibration.

The tau identification algorithm is designed to reject backgrounds from quark and gluon initiated jets. The method is based on a Boost Decision Tree (BDT) (Introduced in 4.3.4.1). For τ associated with one or three tracks, the BDTs are trained separately with simulated $Z/\gamma^* \rightarrow \tau\tau$ for signal and di-jet events selected from data for background. Three working points labelled as `loose`, `medium` and `tight` are provided, corresponding to different tau identification efficiency values with the efficiency designed to be independent of p_T. The target efficiencies are 0.6, 0.55 and 0.45 for the generated one-track loose, medium and tight working points, and 0.5, 0.4 and 0.3 for the corresponding generated three-track target efficiencies. The input variables to the BDT are corrected such that the mean of their distribution for signal samples is constant as a function of pile-up. This ensures that the efficiency for each working point does not depend strongly on the pile-up conditions.

3.4.5 Jets

3.4.5.1 Reconstruction

Jets usually originate from the hadronization of quarks and gluons which are abundantly produced in a high energy proton-proton collision. They are also a common feature in many physics analysis. In both of the analysis presented in this thesis, the search for H^+ and ttH to leptons, jet multiplicities are used to define signal regions and control regions.

The reconstruction of jets uses the anti-k_t algorithm with the radius parameter R=0.4, with the positive-energy topological clusters (topoclusters) of calorimeter cell energies as input. These 3D clusters are built from topologically connected calorimeter cells, starting with a seed cell which contains a significance (signal vs. noise) above 4 σ. The noise is defined as the sum in quadrature of electronic and pile-up noise. Cells near the seed with a significance higher than 2 σ are added iteratively to build the cluster. There is also a splitting step in the topocluster algorithm: All cells in a cluster are searched for local maxima in terms of energy content, and the local maxima are then used as seeds for a new iteration of topological clustering, which will split the original cluster into more topoclusters. A topocluster by definition has the same energy as the sum of all included cells and has no mass. The reconstructed direction is the same as an unit vector originating from the center of the ATLAS coordinate system pointing to the energy-weighted topocluster barycenter. In order to suppress pile-up contributions, a vertex finding algorithm, Jet Vertex Fraction (JVF), is developed. JVF is defined as the fraction of track p_T assigned to a jet that originates from the hard scatter vertex. Jets with small JVF value are likely to have most of their tracks originating from pile-up vertices.

Four qualities of jets are defined with different levels of fake jet rejection: `Looser`, `Loose`, `Medium`, `Tight`. The Loose selection has the highest jet efficiency while the Tight selection has the highest background rejection. The definitions of the selections are based on the reconstruced energy at the

cell level, the jet energy deposited in the direction of the shower developement and the number of reconstructed tracks that are matched to the jet.

3.4.5.2 Calibration and Systematics

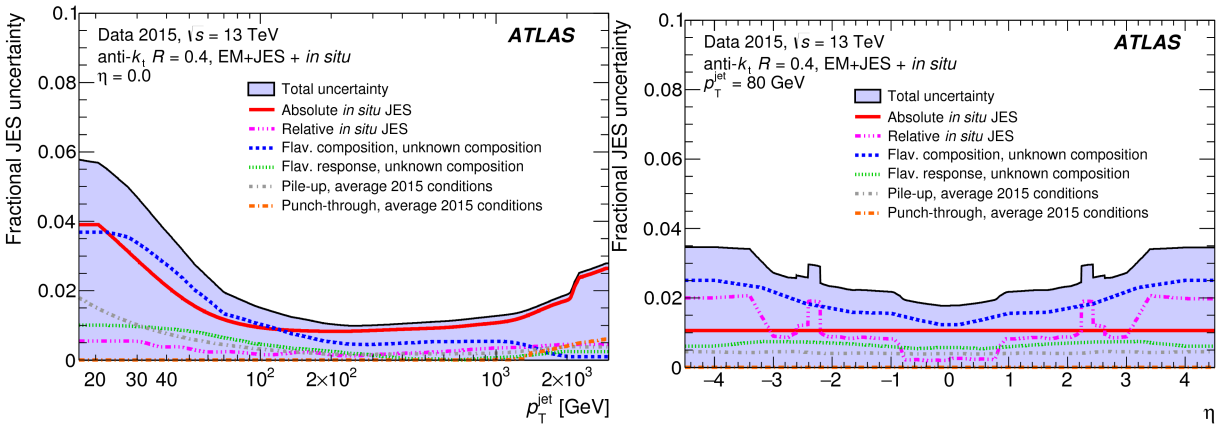
The calibration of jets consists of several different parts, including pile-up corrections, jet energy scale, η corrections, global sequential corrections and in situ corrections. First a jet is corrected to point back to the identified hard-scatter vertex. Next the pile-up effect is removed using an area-based subtraction procedure and residual corrections. The jet energy is then calibrated by applying a p_T and η-dependent correction derived from the nominal 2015 MC simulation, biases in the η position of the jets, caused by the different calorimeter technologies in different η regions, are also corrected. Further corrections are applied to the jets that reduce the dependence of the jet energy measurement on the longitudinal and transverse structure of the jets and also correct for jets that are not fully contained in the calorimeter. Finally, for jets in the data, an additional correction is applied that changes their calibration to its correct value based on in situ studies.

The systematic uncertainties on jets are mostly coming from jet energy scale (JES). The uncertainties are basically derived from in situ corrections. The final calibration includes a set of 80 JES uncertainty terms (3.19) propagated from the individual calibrations and studies, in which 67 of these uncertainties come from the Z/γ jet and multijet balance in situ calibrations accounting for assumptions made in the event topology, MC simulation, statistics, and propagated uncertainties of the electron, muon, and photon energy scales. The other 13 uncertainties are derived from other sources:

- Four pile-up uncertainties account for potential MC mismodeling of N_{PV}, μ, ρ and the residual p_T dependence
- Three η-intercalibration uncertainties account for potential physics mismodeling, statistical uncertainties, and the method non-closure in the $2.0 < |\eta| < 2.6$ region.
- Three additional uncertainties account for differences in the jet response and simulated jet composition of light-quark, b-quark, and gluon-initiated jets.
- Three other uncertainties account for the the GSC punch-through correction, AFII modeling (only applied for fast simulation samples) and high p_T jets.

3.4.5.3 b-jets tagging

The identification of jets originating from b quarks is an important tool since many processes of interests feature the presence of b quarks, for instance, top quarks almost always decay to a b quark and a W boson. It would be helpful to distinguish new phenomena, as well as suppress light flavor jets backgrounds. Figure 3.20 shows an illustration of the production of a b-jet. Comparing to charm hadron and light flavor hadrons, b hadron can fly longer distance before decaying in the inner detector. Thus it is possible to identify jets originating from b hadrons from other jets by finding a displaced secondary vertex. There are three main algorithms for b-tagging commonly used in the ATLAS experiment.



(A) (B)

FIGURE 3.19: Combined uncertainty on the JES of fully calibrated jets as a function of (3.19a) jet p_T at $\eta = 0$ and (3.19b) η at $p_T = 80$ GeV. Systematic uncertainty components include pile-up, punch-through, and uncertainties propagated from the Z/γ-jet and MJB (absolute in situ JES) and η-intercalibration (relative in situ JES). The flavor composition and response uncertainties assume a 50 % quark and 50 % gluon composition with a 100 % uncertainty (unknown composition) [34].

- Based on impact parameters (IP2D, IP3D)
- Inclusive secondary vertex reconstrution algorithms (SV)
- Decay chain reconstrution algorithms (JetFitter)

These algorithms provide complementary information and are combined to a single multivariate discriminant (MV). In ATLAS run2, the classifier is a Boosted Decision Tree (BDT), with an admixture of charm and light jets as backgrounds in order to improve the charm jet rejection. Figure 3.21 shows the performance of the MV2 (MV2c10) algorithm with a 10 % charm jet admixture. In run 2 the performance increases about 10 %, mainly due to the new Insertable B-Layer (IBL) in the inner detector.

Depending on the b jet efficiency, which is defined as the rate of a true b jet passing the b-tagging algorithm, there are four working points (WP) calibrated for analysis, listed in table 3.2. The WPs are set by applying cuts on the MV2c10 output value. Additionally, a particular working point, pseudo continuous b-tagging (PCBT), which divides the MV2c10 weight into five bins instead of two bins (tagged or not tagged), is developed. In PCBT, every jet is assigned a value following the rules:

1. not tagged (failed at 85 % WP)
2. tagged at 85 % WP but failed at 77 % WP
3. tagged at 77 % WP but failed at 70 % WP
4. tagged at 70 % WP but failed at 60 % WP
5. tagged at 60 % WP

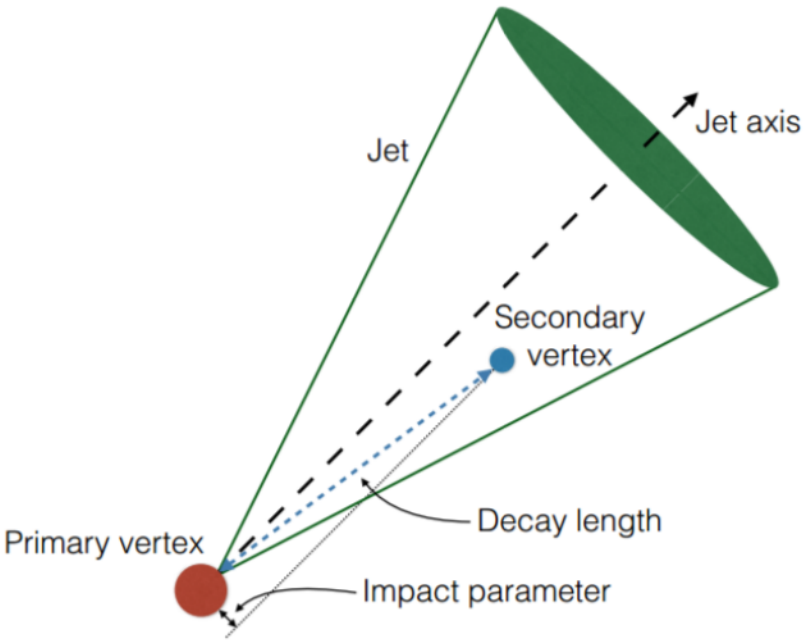


FIGURE 3.20: Illustration of the production of a b-tagged jet.

WP (%)	cut value	b-tagging eff. [%]	c-jets rejection	tau rejection	light jets rejection
60	0.94	61.14	22	150	1204
70	0.83	70.84	8	39	313
77	0.64	77.53	4	16	113
85	0.11	85.23	2	6	28

TABLE 3.2: b-tagging benchmarks of MV2c10 tagger. Rejection means that among given number of jets one jet would pass the tagger. For example c-jet rejection of 150 for 60 WP means that among 150 c-jets 1 jet would pass the MV2c10 tagger at 60 WP.

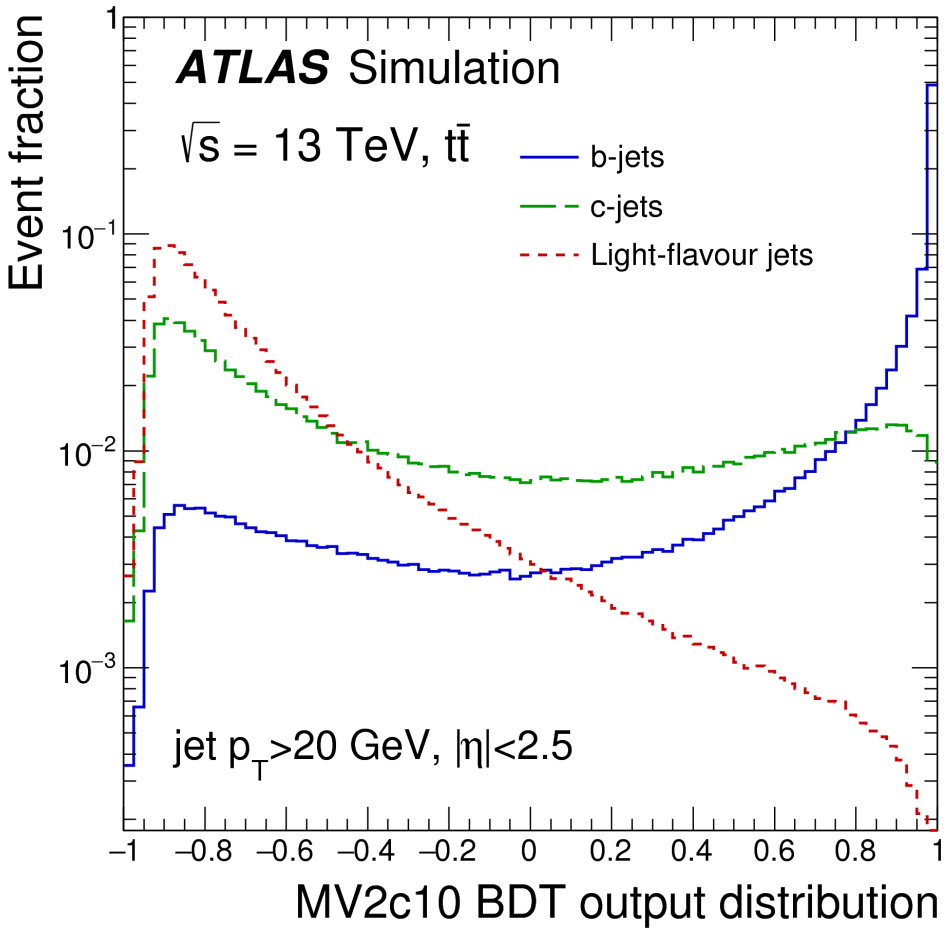


FIGURE 3.21: The MV2c10 output for b-jets (solid line), c-jets (dashed line) and light jets (dotted line) in simulated $t\bar{t}$ events [35].

3.4.6 Missing transverse energy

Missing transverse engergy (MET) represents the undetectable particles, such as neutrinos, dark matter particles. It indicates the conservation of momentum in the transverse plane. In ATLAS, MET is defined as the negative vector-sum of the transverse momenta of all reconstructed and calibrated physics objects (hard term), and the unclustered energy (soft term), as indicated in equation 3.9. The latter one is estimated from tracks with low p_T which are associated to the primary vertex but not assigned to any hard object. The hard term takes into account all the corrections and calibrations applied to the objects.

$$E_{x(y)}^{miss} = E_{x(y)}^{miss,e} + E_{x(y)}^{miss,\gamma} + E_{x(y)}^{miss,\tau} + E_{x(y)}^{miss,\mu} + E_{x(y)}^{miss,jets} + E_{x(y)}^{miss,soft} \tag{3.9}$$

The reconstruction of MET suffers from the pile-up events, especially for the soft term since the pile-up correction for the hard term is propagated to the MET. Therefore, a track-based technique, the Soft Term Vertex Fraction (STVF), is developed to correct the contribution from the soft term. The STVF is defined as the ratio of the scalar sum of p_T of all soft tracks associated with the primary vertex to the scalar sum of p_T of all soft tracks in the event. The soft term then is reweighted by STVF per event. The systematic uncertainties on MET reconstruction are established from $Z \rightarrow ll$ and $W \rightarrow lv$ events, as the difference between data and simulations of the MET distribution. An example of the performance of the MET reconstruction can be found in figure 3.22.

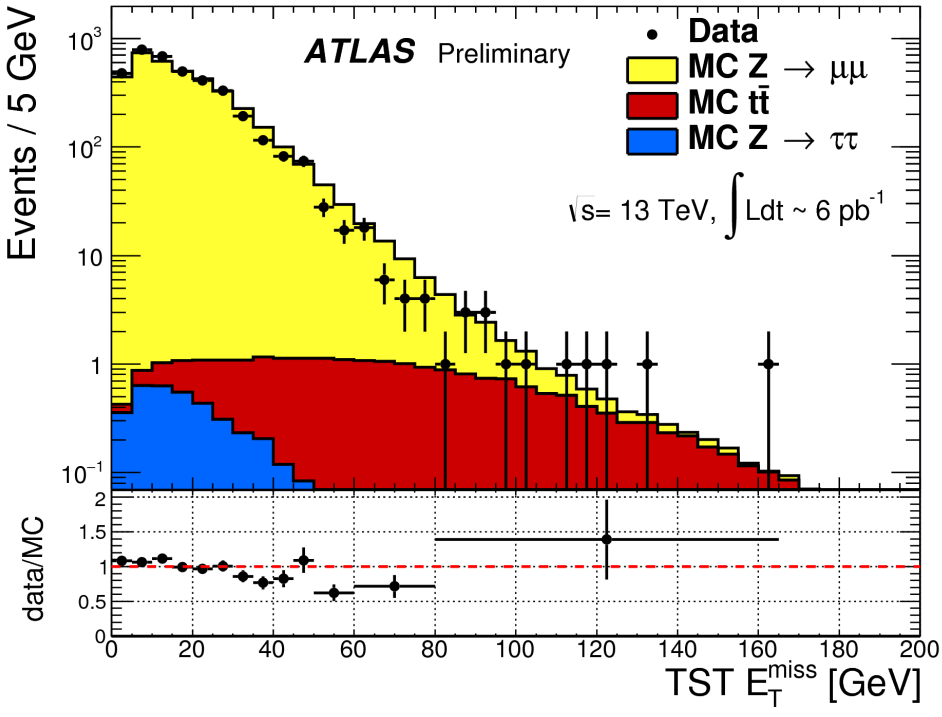


FIGURE 3.22: The distribution of TST (track-based soft term) E_T^{miss}, with 2015 data in $Z \rightarrow \mu\mu$ events. The simulation samples are superimposed and normalized to data [36].

Chapter 4

Search for a heavy charged Higgs boson decaying to top and bottom quarks

4.1 Introduction

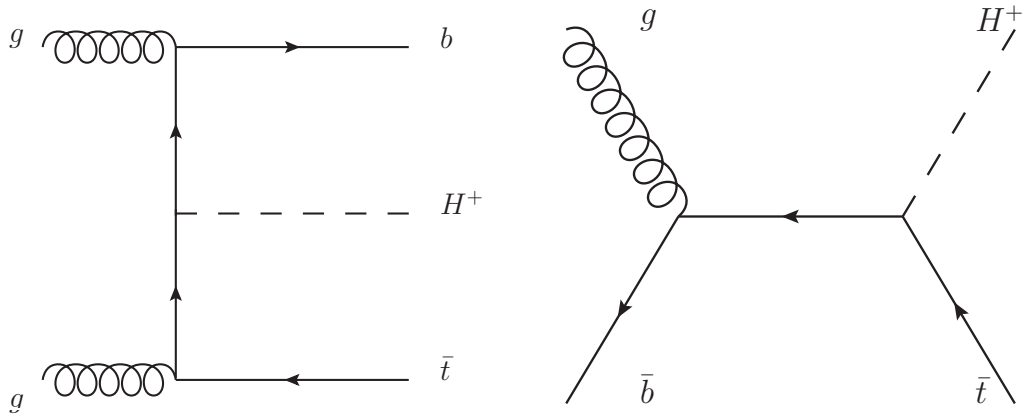
After the discovery of the Higgs boson [1, 2, 3], the question of whether it is the only fundamental scalar particle or it could be the first state of a beyond standard model (BSM) Higgs sector is still remaining. Almost all BSM Higgs scenarios contain more than one Higgs boson among which the charged Higgs bosons are usually present, H^+ or H^-. The detailed discussion on the theory is presented in Chapter 1. From now on, the charged Higgs is always denoted as H^+ for simplicity, since H^- is just its charge-conjugate.

The charged Higgs boson is predicted by several models, among all theories the two-Higgs-Doublet model (2HDM) [37, 13, 38, 12] is very popular. A particular type of 2HDM, namely type-II 2HDM, is corresponding to the Higgs sector of the Minimal Supersymmetric Standard Model (MSSM). In this analysis we mainly focus on this particular model, but we also test other possible models in the results.

The production and decay mechanism of the charged Higgs boson partly depends on its mass. For a light H^+ ($m_{H^+} < m_{top}$), the production of H^+ occurs mainly through the decay of top quark ($t \to bH^+$). For a heavy H^+($m_{H^+} > m_{top} + m_{bottom}$, where m_{top} and m_{bottom} are the masses of top and bottom quarks respectively), the leading production mode in LHC is in association with a top quark, $gg \to tbH^+$ (referred to as the 4-flavor scheme, or 4FS) or $gb \to tH^+$ (referred to as the 5-flavor scheme, or 5FS). In this analysis, the 4FS mode is selected to produce the signal Monte Carlo samples following the suggestion from theorists. The production modes of H^+ are summarized in figure 4.1.

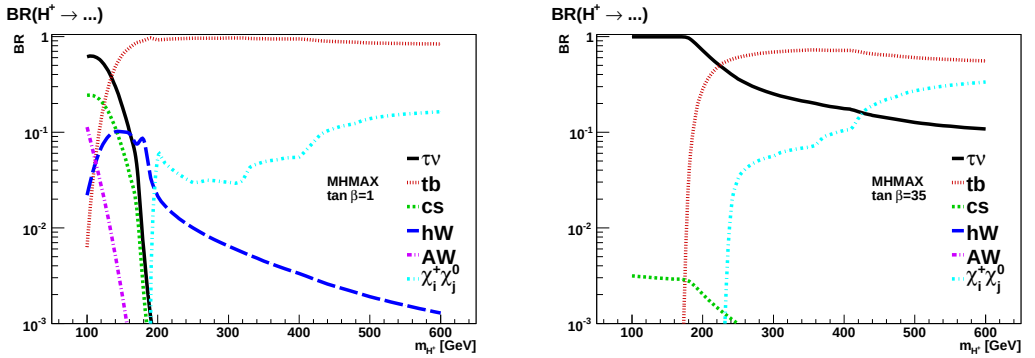
In the 2HDM, the production and decay of H^+ are also controlled by two parameters: the ratio of the vacuum expectation values of the two Higgs doublets (tan β) and the mixing angle between the CP-even Higgs bosons (α). For a heavy charged Higgs boson and $cos(\beta - \alpha) \approx 0$, the decay mode of H^+ is dominated by $H^+ \to tb$ [39, 8]. For certain scenarios the decay $H^+ \to \tau\nu$ also becomes important. Two examples of H^+decay branch ratio are given in figure 4.2 [40]. For high mass and low $tan\ \beta$ values, $H^+ \to tb$ dominates.

My contribution during my Ph.D study in this analysis is mainly in the di-lepton channel, including sensitivity study, mass reconstruction of H^+ boson and Data-MC comparison which will be presented in the following sections.



(A) four-flavor scheme (4FS) (B) five-flavor scheme (5FS)

FIGURE 4.1: Production mode of a heavy charged Higgs boson ($m_{H^+} > m_{top} + m_{bottom}$), for 4FS 4.1a and 5FS 4.1b respectively.



(A) H^+decay branch ratio with $\tan\beta = 1$ (B) H^+decay branch ratio with $\tan\beta = 35$

FIGURE 4.2: Branch ratio of H^+ decay modes as a function of H^+ mass, with $\tan\beta = 1$ (4.2a) and $\tan\beta = 35$ (4.2b). Plots are taken from Ref [41].

4.2 Overview of previous results

In ATLAS run1, the same search for heavy H^+ decaying to tb was performed on a dataset of 20.3 fb^{-1} integral luminosity at $\sqrt{s} = 8$ TeV [42]. The mass range where H^+ was searched was 200 GeV to 600 GeV and the signal was simulated with the 5FS diagram. The result in figure 4.3 shows that data approximately agrees with the SM prediction, but a broad excess of data compared to MC is observed for all mass points except 600 GeV point. The excess arrives at a maximum of 2.4σ at 250 GeV.

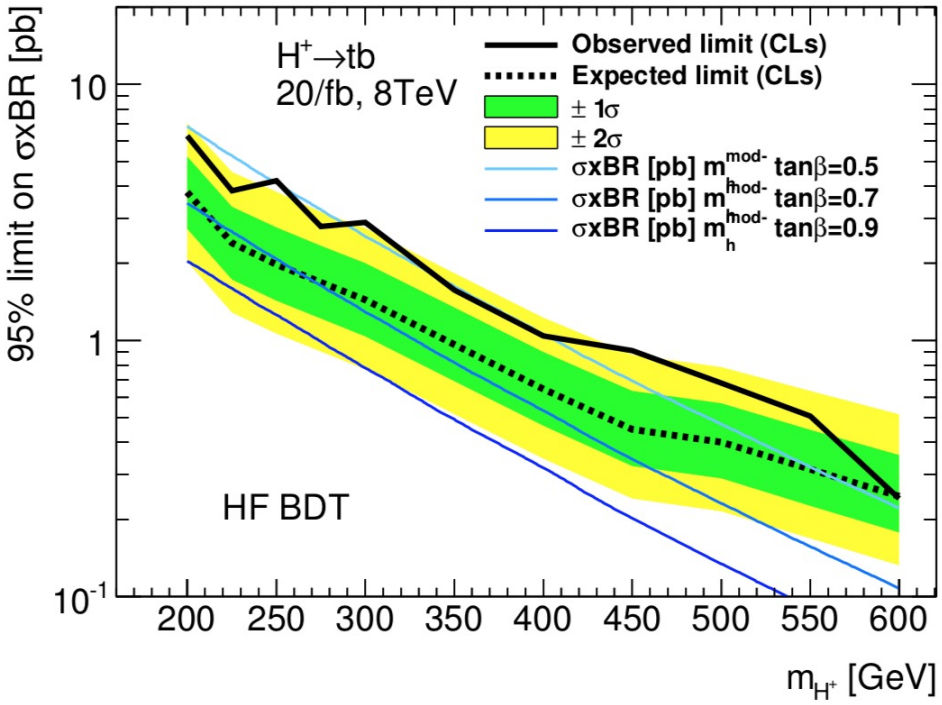


FIGURE 4.3: The observed and expected limit of run-I analysis with an integral luminosity of 20.3 fb^{-1} dataset. A broad excess is observed for all mass points except 600 GeV.

In ATLAS run2, a conference note result was released in ICHEP 2016, using 13.2 fb^{-1} data at $\sqrt{s} = 13$ TeV [43]. The result in figure 4.4 shows no beyond 2σ significant excess in the range from 300 GeV to 1000 GeV. In this analysis three points remain to be improved:

- The di-lepton channel is not included in the result, due to a lack of statistics in the signal MC samples. This limitation forces to merge signal regions (SR) and control regions (CR) to only one SR and one CR and resulted in a large anti-correlation between H^+ signal strength and $t\bar{t}$+heavy flavor normalization factors which could not be addressed in time for the publication.

- The low mass region (200 GeV- 300 GeV) is dropped because the discrimination here between signal and background is too poor and thus makes the fitted signal strength unstable.

- The high mass region (over 1000 GeV) is reduced to single SR with only two bins and no CR. This leads to an increase of sensitivity to systematic variations and lowers the overall sensitivity to the signal in these very pure regions.

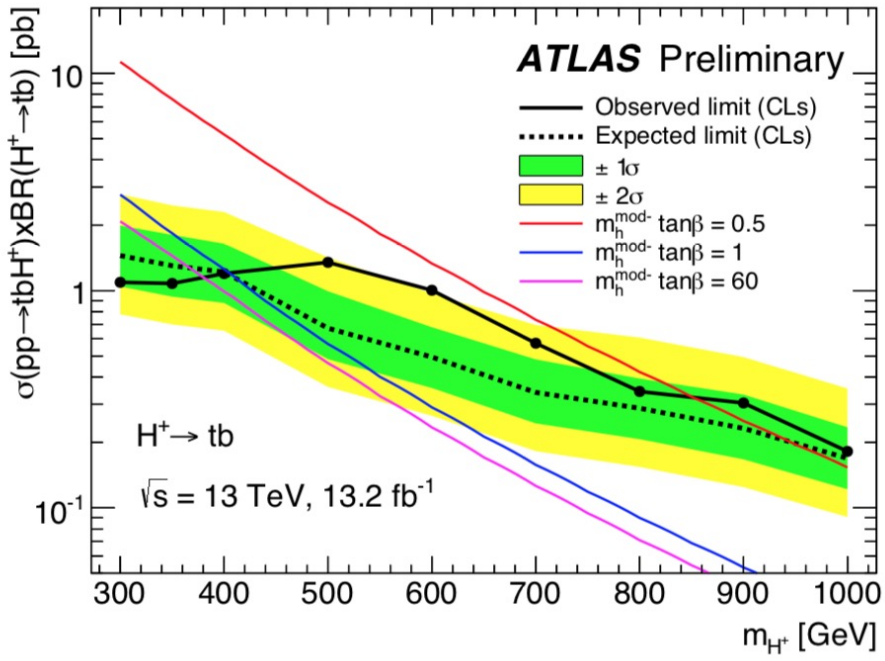


FIGURE 4.4: The observed and expected limit of ICHEP 2016 conference note, using an integral luminosity of 13.2 fb^{-1} dataset. The mass range is reduced to 300 GeV up to 1000 GeV only.

Several new technologies and methods are introduced to the current analysis to improve the performance on the points discussed above and will be presented in the following sections.

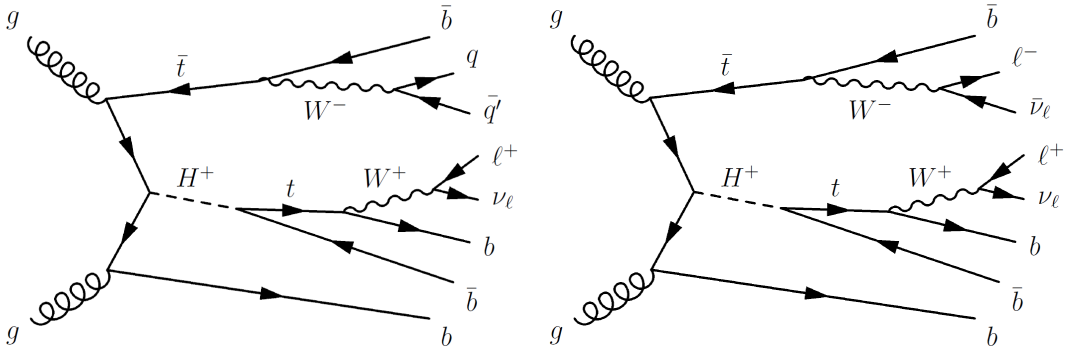
4.3 Analysis strategy

In this analysis search for the production of heavy charged Higgs bosons through $gg \rightarrow tbH^+$ and H^+decays through $H^+ \rightarrow tb$, in the 200 GeV to 2000 GeV H^+ mass range. For signal events, there are two main channels considered depending on the top decaying semi-leptonically ($t \rightarrow Wb \rightarrow lvb$) or hadronically ($t \rightarrow Wb \rightarrow jjb$): lepton-plus-jets channel (when one of the top decays semi-leptonically and the other one decay hadronically) and di-lepton channel (when both of the top quarks decay semi-leptonically), as also shown in figure 4.5.

In the following sections, the discussion will be focused on di-lepton channels to which I most contributed. The lepton-plus-jets channel will be mentioned when necessary.

4.3.1 Data, Signal and background modelling

ATLAS pp collision data and Monte Carlo (MC) simulation samples including H^+ signal and SM background used in the analysis are described in this section. All the MC samples are generated assuming that the mass of top quark is 172.5 GeV and the mass of higgs boson is 125 GeV. The



(A) lepton-plus-jets final state (B) di-lepton final state

FIGURE 4.5: Feynman diagrams of lepton-plus-jets final state (A) and di-lepton final state (B) respectively.

detector simulation is done with GEANT4 package, through a full simulation (FULLSIM) or a fast simulation of the calorimeter response (ATLFASTII). Data and simulation samples are processed equally through reconstrution software and the object identification efficiency, energy scales and energy resolutions of simulation samples are corrected to match data.

4.3.1.1 Data samples

The data used in this analysis were recorded in 2015 and 2016 with $\sqrt{s} = 13$ TeV pp collisions, corresponding to an integrated luminosity of 36.1 fb^{-1}. Only runs with stable beams and for which all detector components were functional are used.

4.3.1.2 Signal samples

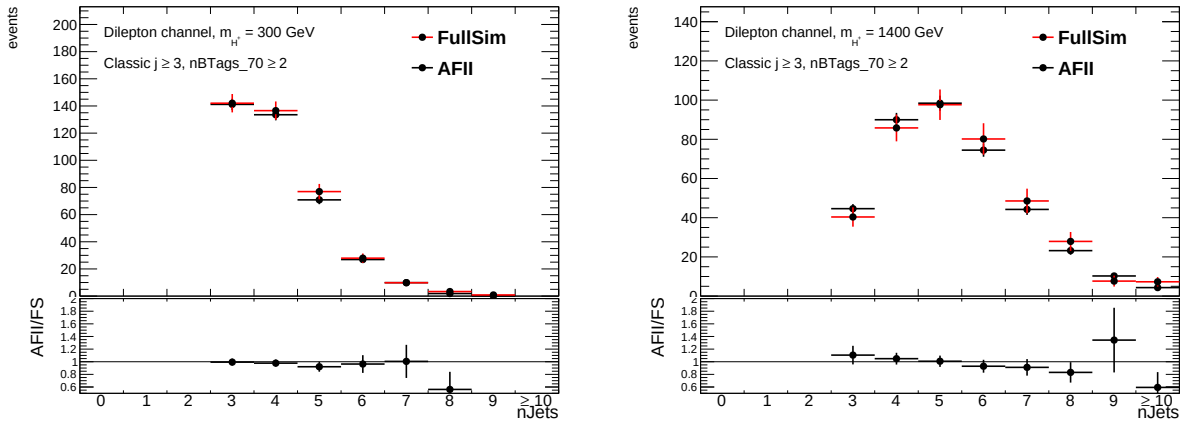
The H^+ signal is generated with MADGRAPH5_AMCATNLO (MG5_AMC) + PYTHIA8.2 [44, 45], which is a 4FS NLO generator. The parton distribution function (PDF) is NNPDF2.3 [46]. The width of H^+ is set to 0, and the static QCD scales, μ_R, μ_F, are set to $m_{H^+}/3$. 18 different H^+ mass points from 200 GeV to 2000 GeV are generated with the ATLFASTII (AFII) simulation. The 300 GeV and 1400 GeV samples are also available as FULLSIM (FS) to compare with ATLFASTII samples. The generated samples are summarized in table 4.1.

To increase the statistics in di-lepton final state, signal samples with di-lepton filter are also produced, which means the events must contain two charged leptons from W bosons decays which come from top quarks ($t \rightarrow Wb \rightarrow blv$). For this purpose, the `TTbarWToLeptonFilter` is used, which has a filter efficiency of 10 %. `TTbarWToLeptonFilter` is a event filter applied at the generator level which selects events with at least one W boson from the top quark leptonic decay.

Since this analysis highly depends on an accurate description of signal samples, it is important to understand the performance of ATLFASTII simulation. The acceptance of FULLSIM samples is very often higher than ATLFASTII. The difference is less than 1 % at 1400 GeV while around 4 % for 300 GeV. Figure 4.6 presents the comparison of number-of-jets between ATLFASTII and FULLSIM samples after the dilepton final state pre-selection (see section 4.3.2 for details). The agreement is good and no obvious trends can be seen.

DSID	H^+ mass [GeV]	Size (nominal)	Size (dilep filter)	$\sigma(\tan\beta = 1)$ [pb]	$\sigma(\tan\beta = 60)$ [pb]
341541	200	5.0M	2.00M	3.3642	3.1218
341542	225	1.5M	0.60M	2.6823	2.4761
341543	250	1.5M	0.60M	2.4642	1.9838
341544	275	1.0M	0.40M	1.7517	1.5993
341545	300	1.0M	0.40M	1.4224	1.2931
341546	350	0.8M	0.32M	0.9626	0.8697
341547	400	0.8M	0.32M	0.6626	0.5915
341548	500	0.7M	0.28M	0.3300	0.2927
341549	600	0.6M	0.24M	0.1749	0.1534
341550	700	0.6M	0.24M	0.0969	0.0844
341551	800	0.6M	0.24M	0.0559	0.0482
341552	900	0.6M	0.24M	0.0333	0.0286
341553	1000	0.7M	0.28M	0.0204	0.0175
341554	1200	0.9M	0.36M	0.0082	0.0069
341555	1400	1.2M	0.48M	0.0036	0.0030
341556	1600	1.2M	0.48M	0.0016	0.0014
341557	1800	2.0M	0.80M	0.0008	0.0006
341558	2000	2.0M	0.80M	0.0004	0.0003

TABLE 4.1: Generated H^+ samples. All samples are generated with ATLFASTII (AFII) and available as MC15c. The last two columns show the expected production cross-section, but in the analysis the cross-section is set to 1 pb for easily scaling between different theory models. The mass points 300 GeV and 1400 GeV are also simulated in FULLSIM (FS).



(A) 300 GeV (B) 1400 GeV

FIGURE 4.6: Comparing the number-of-jets after dilepton pre-selection between ATLFASTII and FULLSIM for 300 GeV(4.6a) and 1400 GeV(4.6b).

4.3.1.3 Background samples

The event generators and configurations for the background MC samples are summarized in table 4.2.

Physics process	Generator	Parton shower generator	Cross-section normalization	PDF set	Tune
$t\bar{t}$+ jets	POWHEG-BOX v2	PYTHIA 8.210	NNLO+NNLL	NNPDF3.0NLO	A14
$t\bar{t}b\bar{b}$	SHERPA 2.1.1	SHERPA 2.1.1	NLO for $t\bar{t}b\bar{b}$	CT10F4	SHERPA default
$t\bar{t}V$	MG5_AMC	PYTHIA 8.210	NLO	NNPDF3.0	A14
$t\bar{t}H$	MG5_AMC	PYTHIA 8.210	NLO	NNPDF3.0NLO	A14
Single top, Wt	POWHEG-BOX v1	PYTHIA 6.428	aNNLO	CT10	Perugia 2012
Single top, t-channel	POWHEG-BOX v1	PYTHIA 6.428	aNNLO	CT10F4	Perugia 2012
W+jets	SHERPA 2.2.1	SHERPA 2.2.1	NNLO	NNPDF3.0NNLO	SHERPA default
Z+jets	SHERPA 2.2.1	SHERPA 2.2.1	NNLO	NNPDF3.0NNLO	SHERPA default

TABLE 4.2: Nominal simulated background event samples. The generator, parton shower generator and cross-section used for normalization are shown together with the applied PDF set and tune. The $t\bar{t}b\bar{b}$ event sample generated using SHERPA 2.1.1 is used to reweight the events from the $t\bar{t}+ \geq 1b$ process in the $t\bar{t}$+ jets sample.

The dominant SM background, $t\bar{t}$ + jet process is generated with the POWHEG-BOX v2 NLO (Next-leading-order) generator [47, 48, 49, 50] with the NNPDF3.0 parton distribution function (PDF) set [51]. The `hdamp` parameter, which controls the transverse momentum of the first additional emission beyond the Born configuration, is set to 1.5 times the top quark mass [52]. Parton shower and hadronisation are modelled by PYTHIA8.2 [53] with the A14 tune. The sample only contains "non-all hadronic" W boson decays. There is also a special sample dedicated to dilepton final state, which is a subset of the "non-all hadronic" sample. The $t\bar{t}$ + jets sample is normalized to the TOP++2.0 [54] theoretical cross section of 832^{+46}_{-51} pb, calculated at next-to-next-to-leading order (NNLO) in QCD including resummation of next-to-next-to693 leading logarithmic (NNLL) soft gluon terms [55, 56, 57, 58, 59].

The $t\bar{t}$ + jet is generated inclusively with all possible flavors in the extra jets. The decay of bottom and charm hadrons is simulated with the EVTGEN v1.2.0 [60]. The sample is categorized to $t\bar{t}$ + light jets and $t\bar{t}$ + heavy flavor (HF) jets in terms of the flavor of extra jets. The $t\bar{t}$ + HF jets is further categorized to $t\bar{t}+ \geq 1c$ and $t\bar{t}+ \geq 1b$ depending on whether the additional jets originate from hadrons containing b or c quarks. The categorization is done in the following way:

1. If at least one jet is matched to ($\Delta R < 0.3$) a b-hadron with $p_T > 5$ GeV not originating from top quark, then the event is categorized as $t\bar{t}+ \geq 1b$.

2. If at least one jet is matched to a c quark but not b quark and this quark does not originate from top quark, the event is categorized as $t\bar{t}+ \geq 1c$.

3. $t\bar{t}$ + jet which does not fall into the above two categories are categorized as $t\bar{t}$ + light events.

The $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ categories are further subdivided into:

- $t\bar{t} + b$ or $t\bar{t} + c$: when exactly one jet is matched to exactly one b or c quark
- $t\bar{t} + b\bar{b}$ or $t\bar{t} + c\bar{c}$: when two jets are each matched to two b or two c quarks

- $t\bar{t}+B$ or $t\bar{t}+C$: when one single jet is matched to a $b\bar{b}$ or $c\bar{c}$ pair
- $t\bar{t}+3b$: when the event does not fall into the above categories
- $t\bar{t}+b$ (MPI/FSR): the additional HF jets can only be matched to b-hadrons from multi-parton interactions (MPI) and final-state gluon radiation (FSR)

To better model the $t\bar{t}+\geq 1b$ background, several reweightings to SHERPA+OPENLOOPS [61, 62, 63], which is a 4FS generator, are applied to the $t\bar{t}+\geq 1b$ events from nominal sample. First of all, a reweighting is applied to correct the relative normalization in each subcategory but keep the overall normalization unchanged, as shown in figure 4.7. Then several reweightings of kinematic variables are applied sequentially. First the p_T of the $t\bar{t}$ system is reweighted and then the p_T of top quarks. The last reweighting is based on the type of $t\bar{t}+\geq 1b$ events. If there is only one extra HF jet, then the p_T of this jet is used in the reweighting. If there are more than one extra HF jets, first the ΔR between the HF jets is reweighted and then the p_T of dijet system.

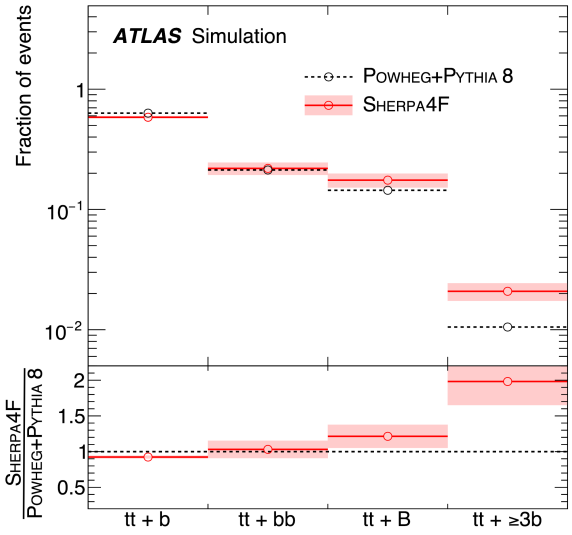


FIGURE 4.7: The relative predict fractions of different subcategoris of $t\bar{t}+\geq 1b$ before any event selection. The prediction of nominal POWHEG+PYTHIA 8 is compared to SHERPA4F. The nominal sample is reweighted to match the shape of SHERPA4F. The plot is taken from [64]

$W/Z+$jets events are generated using SHERPA 2.2.1 [62]. In the $W/Z+$jets samples, matrix elements are calculated using the COMIX [65] and OPENLOOPS matrix element generators and then merged with the SHERPA parton shower [66] using the ME+PS@NLO prescription [67]. he NNPDF3.0NNLO PDF set was used together with a dedicated parton shower tune developed by the SHERPA authors. The $W/Z+$jets events are normalized to NNLO cross-sections [68, 69, 70, 71, 72].

The Wt single top quark background is generated using the POWHEG-BOX v1 generators with the CT10 PDF set. The overlaps between $t\bar{t}$ and Wt final states are removed using the "diagram removal" scheme [73]. The t-channel single top quark events are generated using the POWHEG-BOX v1 generator with the 4FS for the NLO matrix element calculations and the fixed 4FS PDF set CT10F4. The top quark are decayed with MADSPIN [74], which preserves the spin correlations. The

samples are interfaced to PYTHIA 6.428 [75] with the PERUGIA 20112 UE tune [76]. The Wt and t-channel single top quark samples are normalized to the approximate NNLO (aNNLO) theoretical cross-section [77, 78, 79].

Samples of $t\bar{t}V$ ($V = W, Z$) events are generated at NLO in the matrix elements calculation using MG5_AMC with the NNPDF3.0NLO PDF set interfaced to PYTHIA 8.210 with the A14 UE tune. The $t\bar{t}H$ process is modelled using MG5_AMC with NLO matrix elements, NNPDF3.0NLO PDF set. The factorisation and renormalization scales are set to $\mu_F = \mu_R = m_T/2$, where m_T is defined as the scalar sum of the transverse masses $m_T = \sqrt{p_T^2 + m^2}$ of all final-state particles. The events are interfaced to PYTHIA 8.210 with the A14 UE tune.

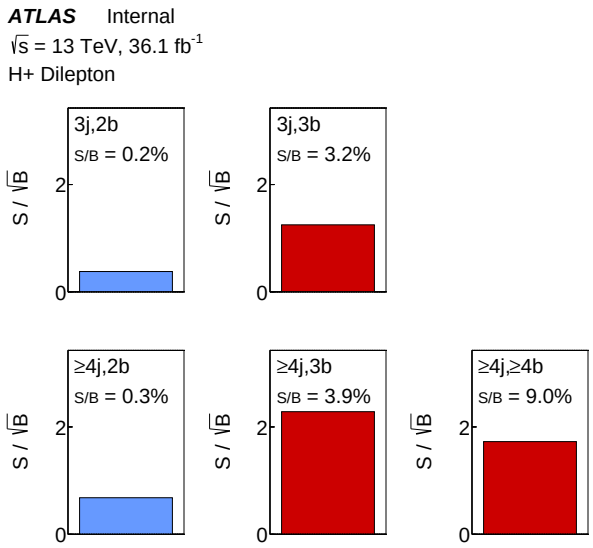
The minor $tH + X$ backgrounds, consisting of the production of a single top quark in association with a Higgs boson and jets ($tHjb$), and the production of a single top quark, a W boson and a Higgs boson (WtH), are treated as one background. The $tHjb$ process is simulated with MG5_AMC interfaced to PYTHIA 8.210 and the CT10 PDF set, and WtH is modelled with MG5_AMC interfaced to Herwig++ [80] using the CTEQ6L1 PDF set [81]. Additional minor SM backgrounds (diboson production, single top s-channel, tZ, tWZ, $4t$, $ttWW$) are also simulated, even though they contribute less than 1% in any analysis region.

4.3.2 Selections

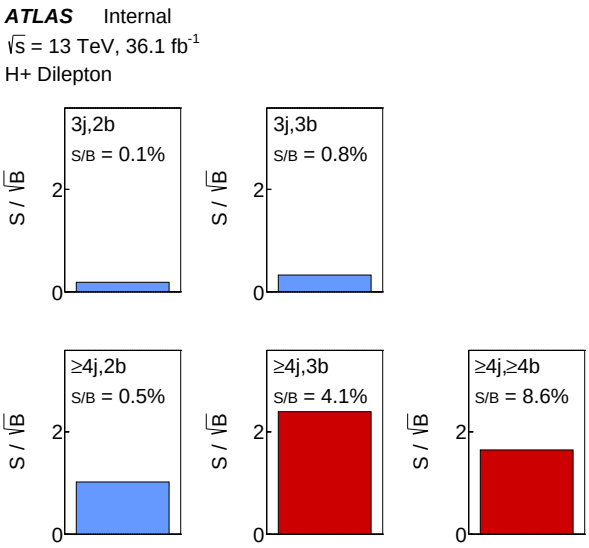
The selection of di-lepton channel is basically driven by the feature that there are exactly two charged leptons (ee, $\mu\mu$ or $e\mu$) with opposite charge, since the two charged leptons originate from the top quarks which have opposite charge. Therefore, same sign lepton pairs can only come from non-prompt leptons. The transverse momentum of the leading lepton must be larger than 27 GeV and for the other lepton its p_T must be larger than 10 GeV. And in case of ee events, the subleading lepton is requested to have p_T larger than 15 GeV. For the same flavor events (ee or $\mu\mu$), the invariant mass of the lepton pair must not be in the Z mass window (83 GeV to 99 GeV) and should be larger than 15 GeV, to suppress $Z + jets$ and Drell-Yan events. Additionally, the events must contain at least 2 b-tagged jets with p_T larger than 25 GeV.

Events passing the pre-selection will be split into control regions and signal regions depending on the signal purity based on jet multiplicity and b-tagged jet multiplicity. Since in the truth there are at least four b quarks, the signal events will tend to have more b-tagged jets. For the dilepton channel, signal regions are $3j3b$, $\geq 4j3b$ and $\geq 4j \geq 4b$ with $3j2b$ and $\geq 4j2b$ as control regions, as shown in figure 4.8. The decision of a region to be a CR or SR is taken based on a scan of significance ($S/\sqrt{B}$) for various H^+ mass (the cross-section is assumed to be 1 pb) against $t\bar{t}$ only backgrounds. Figure 4.9 shows the results of scanning, where the samples are normalized to a luminosity of 30 fb^{-1}.

It is obvious that the significance of different SRs highly depends on the H^+ mass. All SRs tend to be less sensitive at higher H^+ mass because for high mass H^+ the decay products are likely to be boosted to result in a large-R jet but not several small-R jets. Especially for $3j3b$ region, it becomes extremely non-sensitive at high mass, thus it is treated as a CR for $m_{H^+} \leq 1000$ GeV. The pre-fit background composition in each region can be found in figure 4.10.

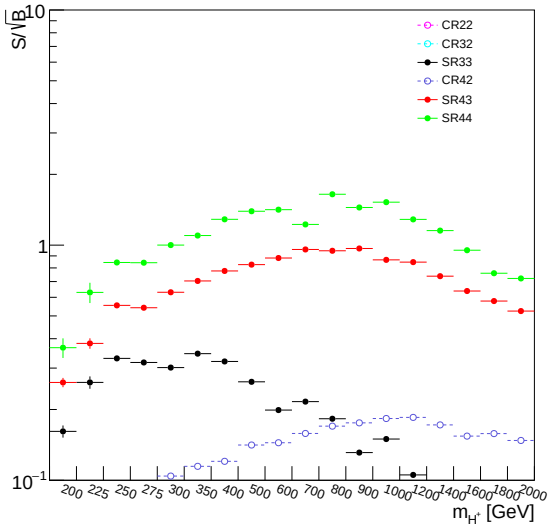


(A) Signal purity in each regions split by jet and b-tagged jet-multiplicity of 400 GeV H^+

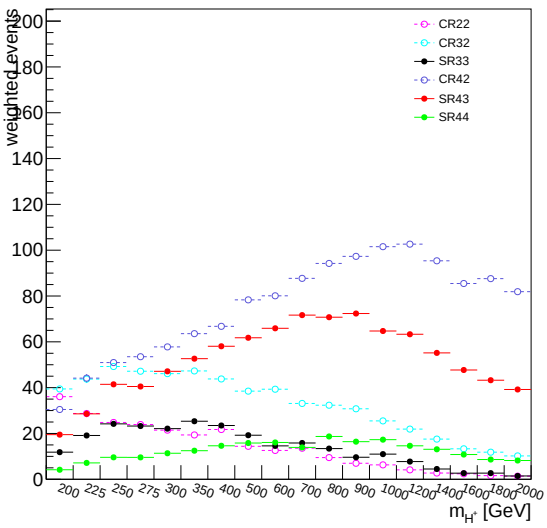


(B) Signal purity in each regions split by jet and b-tagged jet-multiplicity of 1200 GeV H^+

FIGURE 4.8: Signal purity in each region. Signal regions are denoted as red and control regions are denoted as blue.



(A) Significance of SRs as a function of H^+ mass



(B) Signal yields in SRs as a function of H^+ mass

FIGURE 4.9: Scanning of various signal regions, looking at significance ($S/\sqrt{B}$) in 4.9a and signal yields in 4.9b, as a function of H^+ mass. The decision of SR/CR is based on this scanning. The notion in the legend is defined as: CR/SR represent whether it is a CR or SR; the first number indicates the jet multiplicity; the second number indicates the b-jet multiplicity. For example, CR42 means a control region requireing at least 4 jets in which 2 jets are b-tagged.

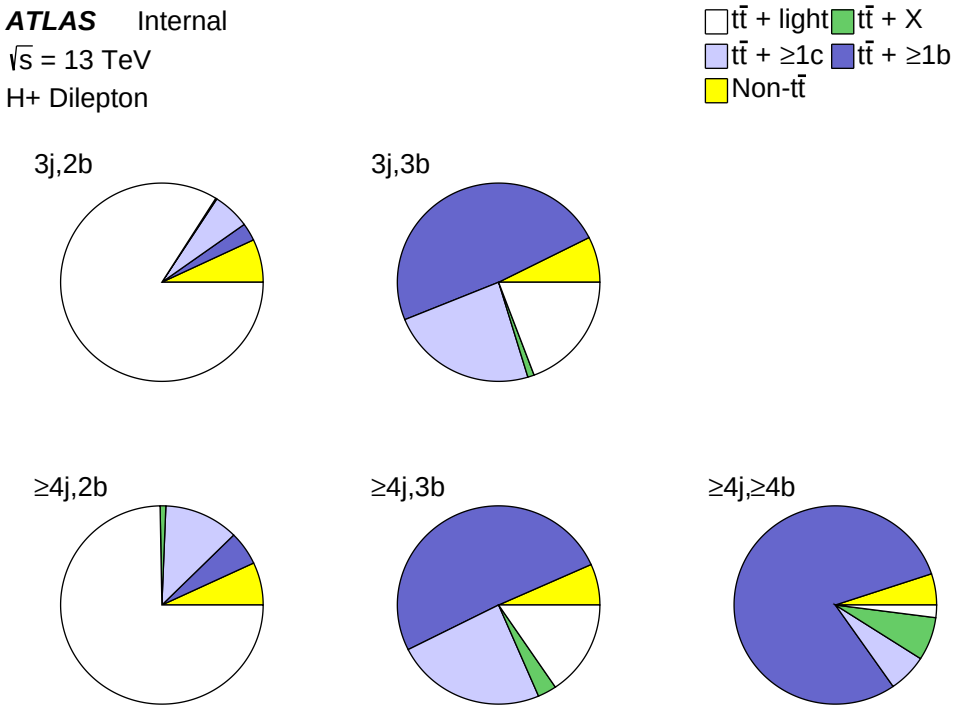


FIGURE 4.10: Prefit background composition in various CRs and SRs.

4.3.3 Background estimation

Most of backgrounds in dilepton final states are estimated from MC samples as described in section 4.3.1. Background processes in this analysis can be categorized into two cases:

- Reducible background: Some processes do not give the same final states as H^+ events, but due to the imperfect detector and reconstruction algorithm, there is always a chance for them to get into the SR phase space. These backgrounds can be reduced by improving particle identification and event selection. For instance, Z + jets, Diboson are reducible backgrounds in this analysis because they do not have b-jets.

- Irreducible background: Some background processes can give exactly the same final states, such as $t\bar{t}$ + HF jets. For instance, if a $t\bar{t}$ pair produced in association with a pair of $b\bar{b}$ from an additional gluon, then the final state of this process is $t\bar{t}b\bar{b}$, which is exactly the same as the H^+ signal $\bar{t}bH^+ \rightarrow \bar{t}bt\bar{b}$. C jets also have a higher probability to be mis-identified as b jets than light jets, which will lead $t\bar{t}c\bar{c}$ to fall into SR in the end. There is no simple way to reduce the contribution of $t\bar{t}$ + HF jets in the SR.

The normalization of Z + HF jets is scaled by 1.3 with an uncertainty of 35 % which is evaluated from a Z-enriched region. The regions is defined as $|m_{ll} - 91\ \text{GeV}| < 8\ \text{GeV}$ for same flavor lepton pairs (ee or $\mu\mu$) which is orthogonal to the nominal pre-selection. Then the ratio of data and MC prediction is taken as the scale factor, which is calculated separately for 0 HF jets, 1 HF jets and 2 HF jets. For 0 HF jets category, the scale factor is found to be 1.0 and in the other 2 categories the scale factor is found to be 1.3.

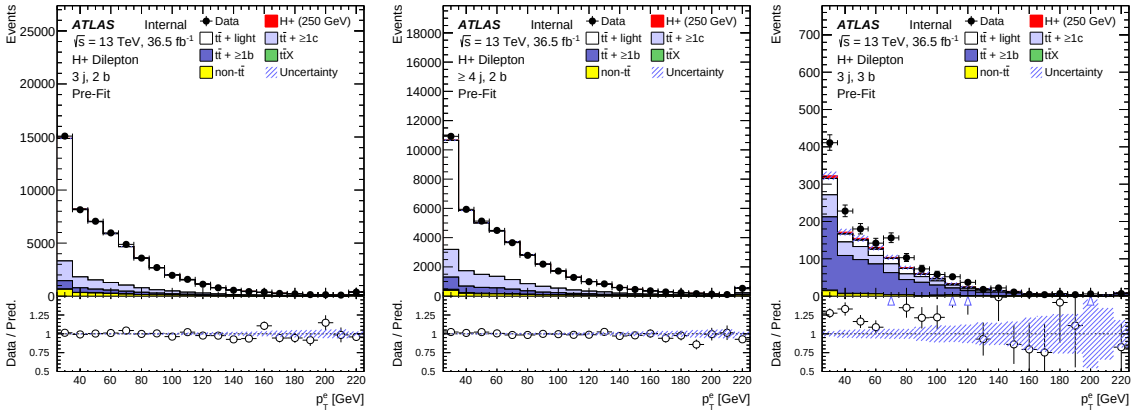
There is also the possibility that events with only one prompt lepton can pass the selection criteria, namely non-prompt background. Most of the non-prompt lepton in dilepton final states come from jets mis-identified as electrons. The contribution of non-prompt background is estimated using data-driven methods. The estimate is taken from MC simulation and the normalization factor is determined by comparing data and simulations in a same-sign dilepton CR. The final contribution of non-prompt events is less than 1 % of total backgrounds which is negligible.

Figure 4.11 and 4.12 show the control plots for several basic object properties in SRs and CRs. In the control regions ($3j2b$ and $\geq 4j2b$) the agreement between data and MC is very good, while a large discrepancy can be seen in the SRs ($3j3b$, $\geq 4j3b$ and $\geq 4j \geq 4b$). The irreducible backgrounds, $t\bar{t}$ + HF jets are likely to fall into SRs because of a higher b-jet multiplicity. But $t\bar{t}$ + light jets are more likely to go into CRs since it has a relatively lower b-jet multiplicity. Thus it is difficult to control $t\bar{t}$ + HF jets using CRs since the composition is very different. Another difficulty is that it is hard to get a precise prediction from theory which means MC is not fully reliable. In practice, $t\bar{t}$ + HF jets is made free-floating in SRs and a simultaneous fit is done on both signal strength and $t\bar{t}$ + HF jets normalization factors to get control of this background.

4.3.4 Multivariate analysis

4.3.4.1 Brief concepts of BDT

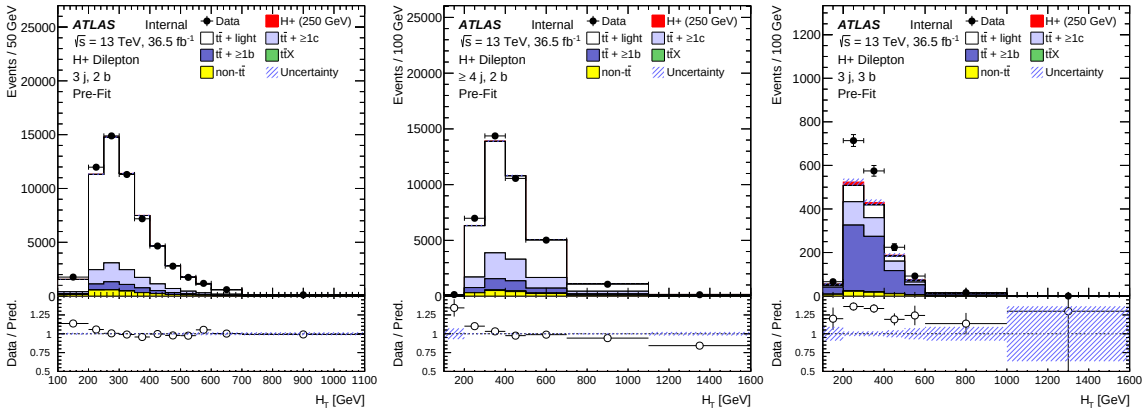
The multivariate analysis method has been widely employed in the ATLAS collaboration. Boosted Decision Tree (BDT) is one of the most popular method which is used to separate signal from



(A) $3j2b$ (B) $\geq 4j2b$ (C) $3j3b$

(D) $\geq 4j3b$ (E) $\geq 4j \geq 4b$

FIGURE 4.11: Control plots of electron p_T in all CRs and SRs. Processes which contain top quarks at the exception of $t\bar{t}$ such as $t\bar{t}V$, $t\bar{t}WW$, $t\bar{t}H$ are denoted as $t\bar{t}+X$. Processes without top quarks are denoted as non-$t\bar{t}$. 250 GeV $H^+$is taken as the signal. The modelling in CRs is quite good while not so satisfying in SRs.



(A) $3j2b$ (B) $\geq\ 4j2b$ (C) $3j3b$

(D) $\geq\ 4j3b$ (E) $\geq\ 4j\ \geq\ 4b$

FIGURE 4.12: Control plots of HT in all CRs and SRs. HT is the scalar sum of p_T of all visible objects (jets, leptons). 250 GeV $H^+$is taken as the signal. The modelling in CRs is quite good while not so satisfying in SRs.

backgrounds. In this analysis, the TMVA toolkit [82] is used for the training and application. In this thesis, only simple concepts are introduced, more detailed instructions can be found in the TMVA manual. A *decision tree* is a binary tree structure as illustrated in figure 4.13. Starting from the root note, a split of events into "Signal-like" and "Background-like" is done by searching for the best separation variables and corresponding cuts. This process is repeated several times until reaching the stop criterion. In the end, the events in the right are more like signal while in the left are more like background. The stop criterion is usually leaf size (number of events in a node) or depth (max depth is a given hyperparameter). The *boosting* expends a single tree to many trees which form a *forest*. Details about the boosting algorithm can be found in the TMVA manual. The output of an event is decided by all the trees by applying different weights. Boosting can stablize the response of the decision tree originating from the fluctuations of the training samples, as well as enhancing the performance with respect to a single decision tree.

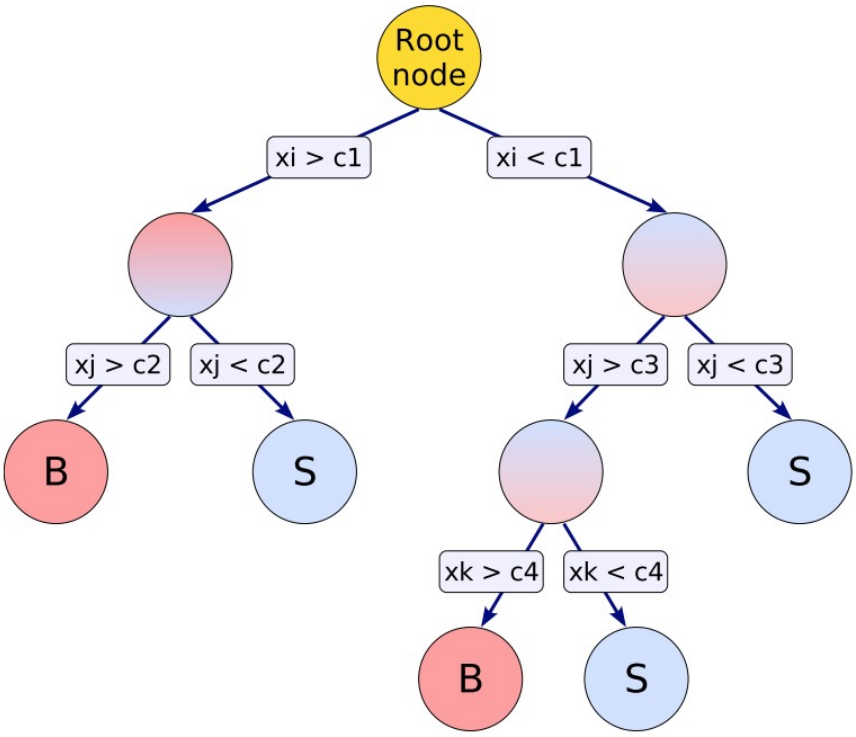


FIGURE 4.13: A sketch illustrating the decision tree in BDT. The sketch is taken from (reference here)

There are several important hyperparameters in the BDT training, which are simply optimized by testing several options to improve the performance of the BDT.

- NTrees: number of trees in the forest
- MaxDepth: maximum depth of the decision tree allowed (simply saying, one split increase the depth by 1)
- MinNodeSize: minimum number of events required in a leaf node. If the number is less than this parameter then the splitting stops.

- nCuts: number of cuts applied in a variable range to find optimal cut point.
- BoostType: boosting type for the trees in the forest.
- Shrinkage: Learning rate for GradBoost algorithm.
- UseBaggedBoost: Use only a random subsample of all events for growing the trees in each iteration

One of the major issues that need to be concerned during the training of BDT is called "over-training", which happens when the decision tree is over fitted to the training data so that it can not predict new data very well. The main reasons for over-training are:

- Lack of statistics of training sample
- The model is too complicated for the given sample (statistics is fixed)
- Bad sampling of training sample

In practice, the whole sample set is split into two parts, one for training and one for testing. Then the BDT output of testing sample is compared to the output of training sample. If there is no over-training, the two outputs should have basically the same shape after normalized to unit.

4.3.4.2 Classification BDT in di-lepton final states

A Boosted Decision Tree (BDT) is used as default discriminant to separate H^+ signal and backgrounds in SRs. BDTs are trained separately for all H^+ samples and all SRs against $t\bar{t}$ backgrounds only since they are the dominant process in most cases. The settings of the BDT training is summarized in table 4.3.

BDT parameters	Value
BoostType	Grad
UseBaggedBoost	True
BaggedSampleFraction	0.8
nTrees	400 (SR 3j3b, $\geq 4j3b$), 200 (SR $\geq 4j \geq 4b$)
Shrinkage	0.3
MaxDepth	3
MinNodeSize	5 %

TABLE 4.3: The BDT parameters used in the training. *nTrees* is lowered to 200 in SR $\geq 4j \geq 4b$ due to the lack of statistics.

The number of variables used in the training is around 10 and optimized for each mass point separately. All the variables initially considered in the BDT optimisation fall into the following categories:

- Event shape variables (transverse sphericity, transverse thrust and transverse thrust minor)
- Global kinematic variables, i.e. minimum ΔR of any jet pair.

- Kinematic properties of combinations of selected detector objects, i.e. invariant mass of leading jet and subleading jet M_{j1j2}.

Event shape variables reflect the feature of the whole event topology. They are useful when the signal event has particular features. Since in hadron colliders the momentum along beam axis is usually boosted, these variables are often defined in terms of transverse momentum. Those variables are also tested in the BDT, including *transverse sphericity*, *transverse thrust* and *transverse thrust minor*. *Sphericity* is most widely used for discriminating symmetric multi-jet topologies. In a hadron collider, the *Sphericity* is usually defined in the transverse plane, in terms of the eigenvalues ($\lambda_1,\ \lambda_2$, with $\lambda_1 > \lambda_2$) of the following matrix:

$$M_{xy} = \sum_i \frac{1}{|p_{T,i}|^2} \begin{bmatrix} p_{x,i}^2 & p_{x,i}p_{y,i} \\ p_{x,i}p_{y,i} & p_{y,i}^2 \end{bmatrix} \tag{4.1}$$

Then $S_\perp \equiv \frac{2\lambda_2}{\lambda_1+\lambda_2}$. It tends to be 1 for the events with circular symmetry in the transverse plane, while is 0 for pencil-like events.

The *transverse thrust* is defined as:

$$T_\perp = \max_{\hat{n}} \frac{\sum_i |\vec{p}_{T,i} \cdot \hat{n}|}{\sum_i |\vec{p}_{T,i}|} \tag{4.2}$$

where the sum is performed over transverse momentum of all charged particles in the event. The thrust axis $\hat{n}_T$ is the unit vector $\hat{n}$ in the equation 4.2 which maximize the ratio. The *transverse thrust* equals to 1 if the event is perfectly pencil-like, while equals to $\langle|\cos\phi|\rangle = 2/\pi$ for perfectly circular symmetric events in the transverse plane, where ϕ is the azimuthal angle between the thrust axis and each particle.

The plane defined by thrust axis $\hat{n}_T$ and the beam axis $\hat{z}$ is called *event plane*. Then the *transverse thrust minor* is defined as:

$$T_M = \frac{\sum_i |\vec{p}_{T,i} \cdot \hat{n}_m|}{\sum_i |\vec{p}_{T,i}|}, \hat{n}_m = \hat{n}_T \times \hat{z} \tag{4.3}$$

It measures the energy flow out of the event plane and equals to 0 for pencil-like events while $2/\pi$ for isotropic events. All these variables are tested together with other kinematic variables and are dropped since they do not rank high in the variable importance.

The detector objects in the analysis model appear as 4-vectors (energy, p_T, η, ϕ) of jets, charged leptons (e,μ) and missing transverse energy (MET) (to ref to obj definition). The jets and charged leptons are labelled by their ranking of p_T. To validate the variables and BDT training in signal depleted regions (3j2b and $\geq 4j2b$), no more than 4 jets and 2 b-tagged jets are used in the variable optimization. The total number of possible combinations of objects in each region is

$$\sum_{k=1}^{N} \frac{N!}{k!(N-k)!}$$

where N is the total number of objects and k is the number of objects to build the variable. For each combination, p_T, η, energy and invariant mass are stored as variables, as well as those variables taking the difference between objects such as Δp_T, $\Delta\phi$, $\Delta\eta$, ΔR, ΔE and ΔM (difference of invariant

mass) of the combinations. To avoid double-counting of jets and b-tagged jets, the jet objects are considered separately from b-tagged jets (denoted as b_j) and un-tagged jets (denoted as u_i), while jets are denoted as j_i indepently from b-tagging. Thus only variables such as $\Delta R(b_i, u_j)$, $\Delta M(b_i, b_j)$ are considered while variables such as $\Delta\eta(b_i, j_j)$ are ignored.

All the variables are ranked by a combination of TMVA separation and ROC curve integral, an example is shown in figure 4.14. The TMVA separation is defined as:

$$\langle S^2 \rangle = \frac{1}{2} \sum_{i=1}^{N_{bins}} \frac{(S_i - B_i)^2}{S_i + B_i} \Delta x_i \tag{4.4}$$

where S_i and B_i are the yields of signal and background in *bin i*, Δx_i is the bin width of *bin i*. A higher $\langle S^2 \rangle$ means a better separation power.

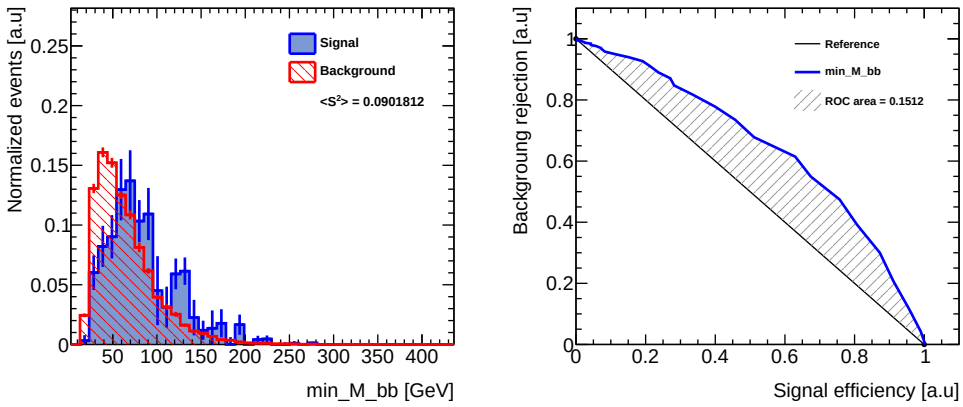
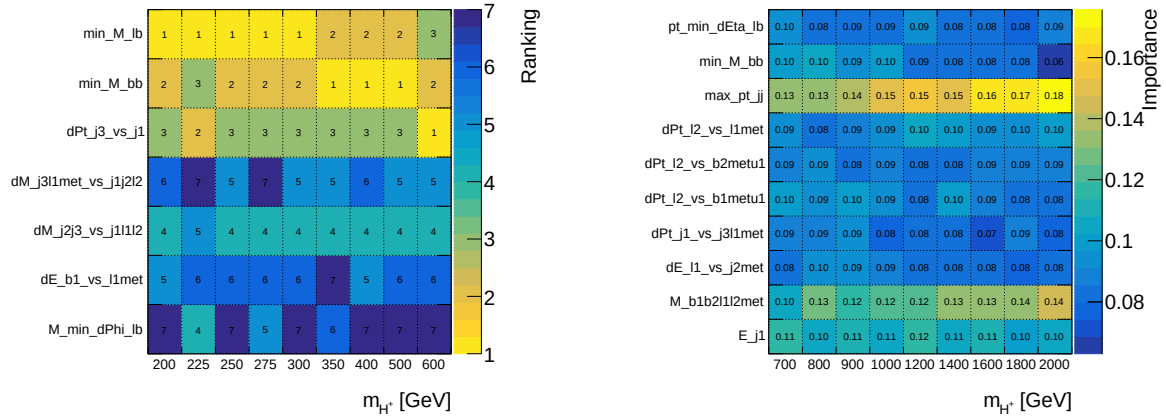


FIGURE 4.14: Example of BDT input variable min_M_bb ($m^{min}(b-pair)$) in $\geq 4j \geq 4b$ region for 400 GeV H^+. The left plot shows the separation of the variable and the right plot shows the integral of ROC curve. ROC curve is defined as a funtion of background rejection versus signal efficiency. A larger integral under the ROC curve means a better variable.

In most cases, kinematic variables can have a very high linear correlation between each other. To avoid the correlation and statistical fluctuations of variable ranking from mass point to mass point, the variables are grouped together if their correlation coefficient is larger than 0.5. The importance of the variables in each group is traced among the mass spectrum and the most sensitive variable in each group is used in the training for each mass point. Best variables in each group is selected for training. An example of the trace of variable importance is shown in figure 4.15. The BDT input variables and their explanations are listed in table 4.4, 4.5, 4.6 and 4.7 and the validation plots of variables are shown in figure 4.16 and 4.17:

The BDT is trained using half of the total events and the other half of events are used for testing. To use all the statistics of signal sample, the BDT is trained using "cross-training", a BDT is trained on odd events (the event number is odd) and then tested and applied to even events (the event number is even), and vice versa. In figure 4.18, the BDT outputs are shown for 400 GeV and 1200 GeV as examples for low mass and high mass H^+ respectively.



(A) Ranking of BDT input variables in a group as a function of H^+ mass in $\geq$ 4j3b SR in low mass range.

(B) Importance of BDT input variables in a group as a function of H^+ mass in $\geq$ 4j3b SR in high mass range.

FIGURE 4.15: Examples of variable trace among the mass spectrum. 4.15a shows in low mass range (200 GeV to 600 GeV) and 4.15b shows in high mass range (700 GeV to 2000 GeV).

variables \ SR	3j3b	$\geq$ 4j3b	$\geq 4j \geq 4b$
1	$m((j,b)^{p_T^{max}})$	$m((l,b)^{\Delta\phi^{min}})$	$p_T(b_2+l_1+l_2+E_T^{miss})$
2	$\Delta E(j_3,l_2)$	$\Delta E(b_1,l_1+E_T^{miss})$	$m^{min}(l,b)$
3	$E(j_3)$	$\Delta m(j_2+j_3,j_1+l_1+l_2)$	$m^{min}(b\text{-pair})$
4	$\Delta m(j_2+l_1,j_1+j_3+l_2+E_T^{miss})$	$\Delta m(l_1+j_3+E_T^{miss},j_1+j_2+l_2)$	$\Delta R(l_2,j_2+j_3+l_1+E_T^{miss})$
5	$\Delta R(j_2,j_1+l_2+E_T^{miss})$	$\Delta p_T(j_1,j_3)$	$\Delta p_T(j_1,j_3)$
6	$p_T(b_1)$	$m^{min}(b\text{-pair})$	H_T^{all}
7	$p_T((l,b)^{\Delta\eta^{max}})$	$m^{min}(l,b)$	–

TABLE 4.4: List of variables used in the BDT, for the various SR and H^+ masses below or including 600 GeV, ranked by importance.

variables \ SR	3j3b	$\geq$ 4j3b	$\geq 4j \geq 4b$
1	$p_T^{max}(j\text{-pair})$	$p_T^{max}(j\text{-pair})$	H_T^{all}
2	$p_T((l,b)^{\Delta\eta^{min}})$	$m(b_1+b_2+l_1+l_2+E_T^{miss})$	$p_T((l,b)^{\Delta\eta^{min}})$
3	$\Delta p_T(j_1,j_3)$	$m^{min}(b\text{-pair})$	$\Delta p_T(j_1,j_3)$
4	$E(j_1)$	$E(j_1)$	$p_T(j_3+l_1)$
5	$\Delta m(j_2+l_2+E_T^{miss},j_1+j_3+l_1)$	$p_T((l,b)^{\Delta\eta^{min}})$	$m^{min}(b\text{-pair})$
6	$p_T((l,b)^{\Delta R^{min}})$	$\Delta p_T(l_2,u_1+b_2+E_T^{miss})$	$\Delta p_T(b_2,b_1+l_2)$
7	$m(j\text{-pair}^{\Delta\eta^{min}})$	$\Delta p_T(l_2,u_1+b_1+E_T^{miss})$	$\Delta p_T(j_2,j_3+l_1+E_T^{miss})$
8	$\Delta p_T(j_1,j_2+E_T^{miss})$	$\Delta p_T(l_2,l_1+E_T^{miss})$	$\Delta E(j_3,j_2+l_1+l_2+E_T^{miss})$
9	$p_T(j_1+j_2+j_3+l_1)$	$\Delta p_T(j_1,j_3+l_1+E_T^{miss})$	$\Delta m(j_2+l_2+E_T^{miss},j_1+j_3+l_1)$
10	$\Delta E(l_1+E_T^{miss},j_1+j_2)$	$\Delta E(l_1,j_2+E_T^{miss})$	–

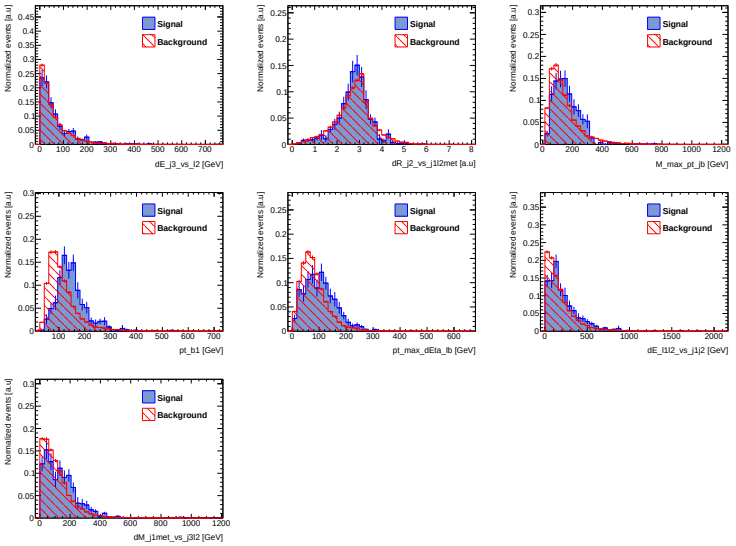
TABLE 4.5: List of variables used in the BDT, for the various SR and H^+ masses above 600 GeV, ranked by importance.

Variable	Definition	3j3b	≥4j3b	≥4j≥4b
$m((j,b)^{p_{\mathrm{T}}^{\mathrm{max}}})$	Invariant mass of the jet-b-jet pair with largest transverse momentum	✓		
$\Delta E(j_3, l_2)$	Energy difference between the third jet and the subleading lepton	✓		
$E(j_3)$	Energy of third jet	✓		
$\Delta m(j_2 + l_1, j_1 + j_3 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}})$	Invariant mass difference between $j_2 + l_1$ and $j_1 + j_3 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}$	✓		
$\Delta R(j_2, j_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}})$	Angular difference between subleading jet and $j_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}$	✓		
$p_{\mathrm{T}}(b_1)$	Transverse momentum of leading b-jet	✓		
$p_{\mathrm{T}}((l,b)^{\Delta\eta^{\mathrm{max}}})$	Transverse momentum of the lepton-b-jet pair with largest $\Delta\eta$	✓		
$m((l,b)^{\Delta\phi^{\mathrm{min}}})$	Invariant mass of the lepton-b-jet pair with smallest $\Delta\phi$		✓	
$\Delta E(b_1, l_1 + E_{\mathrm{T}}^{\mathrm{miss}})$	Energy difference between the leading b-jet and $l_1 + E_{\mathrm{T}}^{\mathrm{miss}}$		✓	
$\Delta m(j_2 + j_3, j_1 + l_1 + l_2)$	Invariant mass difference between $j_2 + j_3$ and $j_1 + l_1 + l_2$		✓	
$\Delta m(l_1 + j_3 + E_{\mathrm{T}}^{\mathrm{miss}}, j_1 + j_2 + l_2)$	Invariant mass difference between $l_1 + j_3 + E_{\mathrm{T}}^{\mathrm{miss}}$ and $j_1 + j_2 + l_2$		✓	
$\Delta p_{\mathrm{T}}(j_1, j_3)$	Transverse momentum difference between leading and third jet		✓	✓
$m^{\mathrm{min}}(b\text{-pair})$	Smallest invariant mass of any b-jet pair		✓	✓
$m^{\mathrm{min}}(l,b)$	Smallest invariant mass of any lepton-b-jet pair		✓	✓
$p_{\mathrm{T}}(b_2 + l_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}})$	Transverse momentum of $b_2 + l_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}$			✓
$\Delta R(l_2, j_2 + j_3 + l_1 + E_{\mathrm{T}}^{\mathrm{miss}})$	Angular difference between subleading lepton and $j_2 + j_3 + l_1 + E_{\mathrm{T}}^{\mathrm{miss}}$			✓
$H_{\mathrm{T}}^{\mathrm{all}}$	Transverse momentum sum of all jets and leptons from the primary vertex			✓

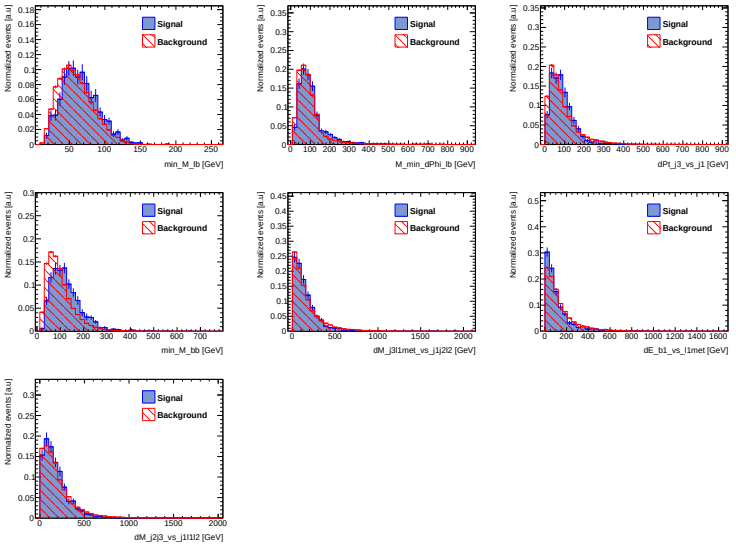
TABLE 4.6: Input variables to the BDT trained on $m_{H^+} \leq 600$ GeV signal samples. The indice of objects starts from 1.

Variable	Definition	3j3b	≥4j3b	≥4j≥4b
$p_{\mathrm{T}}^{\mathrm{max}}(j\text{-pair})$	Maximum transverse momentum of any jet pair	✓	✓	
$p_{\mathrm{T}}((l,b)^{\Delta\eta^{\mathrm{min}}})$	Transverse momentum of the lepton-b-jet pair with smallest $\Delta\eta$	✓		✓
$\Delta p_{\mathrm{T}}(j_1, j_3)$	Transverse momentum difference between leading and third jets	✓		✓
$E(j_1)$	Energy of the leading jet	✓	✓	
$\Delta m(j_2 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}, j_1 + j_3 + l_1)$	Invariant mass difference between $j_2 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}$ and $j_1 + j_3 + l_1$	✓		
$p_{\mathrm{T}}((l,b)^{\Delta R^{\mathrm{min}}})$	Transverse momentum of the lepton-b-jet pair with smallest ΔR	✓		
$m(j\text{-pair}^{\Delta\eta^{\mathrm{min}}})$	Invariant mass of the jet pair with smallest $\Delta\eta$	✓		
$\Delta p_{\mathrm{T}}(j_1, j_2 + E_{\mathrm{T}}^{\mathrm{miss}})$	Transverse momentum difference between leading jet and $j_2 + E_{\mathrm{T}}^{\mathrm{miss}}$	✓		
$p_{\mathrm{T}}(j_1 + j_2 + j_3 + l_1)$	Transverse momentum of $j_1 + j_2 + j_3 + l_1$	✓		
$\Delta E(l_1 + E_{\mathrm{T}}^{\mathrm{miss}}, j_1 + j_2)$	Energy difference between $l_1 + E_{\mathrm{T}}^{\mathrm{miss}}$ and $j_1 + j_2$	✓		
$m(b_1 + b_2 + l_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}})$	Invariant mass of $b_1 + b_2 + l_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}$		✓	
$m^{\mathrm{min}}(b\text{-pair})$	Smallest invariant mass of any b-jet pair		✓	✓
$p_{\mathrm{T}}((l,b)^{\Delta\eta^{\mathrm{min}}})$	Transverse momentum of the lepton-b-jet pair with smallest separation in η		✓	
$\Delta p_{\mathrm{T}}(l_2, u_1 + b_2 + E_{\mathrm{T}}^{\mathrm{miss}})$	Transverse momentum difference between subleading lepton and $u_1 + b_2 + E_{\mathrm{T}}^{\mathrm{miss}}$		✓	
$\Delta p_{\mathrm{T}}(l_2, u_1 + b_1 + E_{\mathrm{T}}^{\mathrm{miss}})$	Transverse momentum difference between subleading lepton and $u_1 + b_1 + E_{\mathrm{T}}^{\mathrm{miss}}$		✓	
$\Delta p_{\mathrm{T}}(l_2, l_1 + E_{\mathrm{T}}^{\mathrm{miss}})$	Transverse momentum difference between subleading lepton and $l_1 + E_{\mathrm{T}}^{\mathrm{miss}}$		✓	
$\Delta p_{\mathrm{T}}(j_1, j_3 + l_1 + E_{\mathrm{T}}^{\mathrm{miss}})$	Transverse momentum difference between leading jet and $j_3 + l_1 + E_{\mathrm{T}}^{\mathrm{miss}}$		✓	
$\Delta E(l_1, j_2 + E_{\mathrm{T}}^{\mathrm{miss}})$	Energy difference between leading lepton and $j_2 + E_{\mathrm{T}}^{\mathrm{miss}}$		✓	
$H_{\mathrm{T}}^{\mathrm{all}}$	Transverse momentum sum of all jets and leptons from the primary vertex			✓
$p_{\mathrm{T}}(j_3 + l_1)$	Transverse momentum of $j_3 + l_1$			✓
$\Delta p_{\mathrm{T}}(b_2, b_1 + l_2)$	Transverse momentum difference between subleading b-jet and $b_1 + l_2$			✓
$\Delta p_{\mathrm{T}}(j_2, j_3 + l_1 + E_{\mathrm{T}}^{\mathrm{miss}})$	Transverse momentum difference between subleading jet and $j_3 + l_1 + E_{\mathrm{T}}^{\mathrm{miss}}$			✓
$\Delta E(j_3, j_2 + l_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}})$	Energy difference between third jet and $j_2 + l_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}$			✓
$\Delta m(j_2 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}, j_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}})$	Invariant mass difference between $j_2 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}$ and $j_1 + l_2 + E_{\mathrm{T}}^{\mathrm{miss}}$			✓

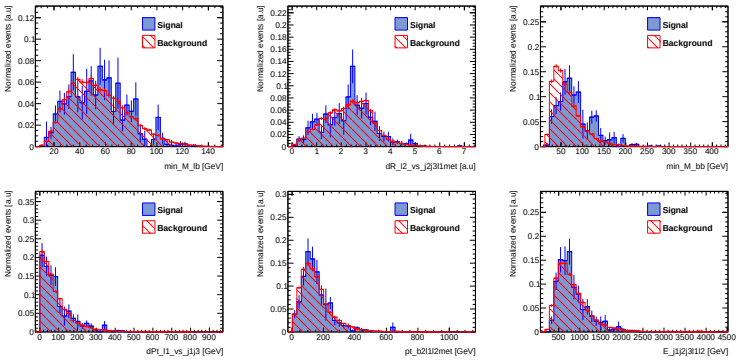
TABLE 4.7: Input variables to the BDT trained on $m_{H^+} > 600$ GeV signal samples. The indice of objects starts from 1.



(A) $3j3b$

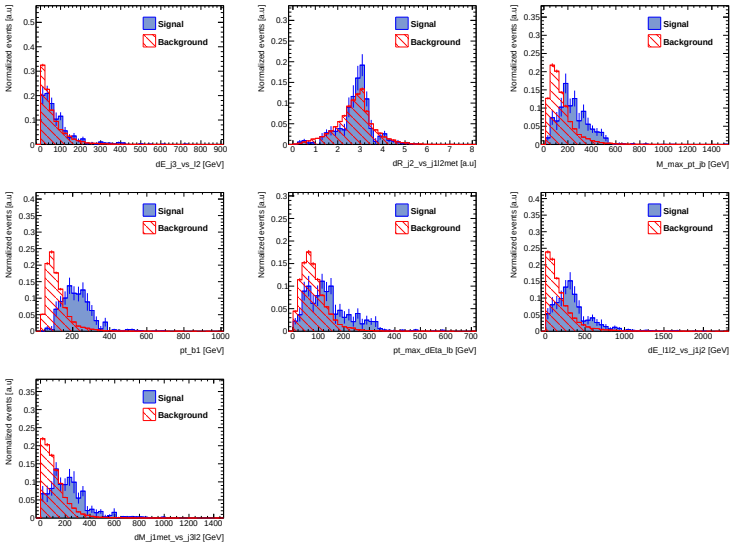


(B) $\geq 4j3b$

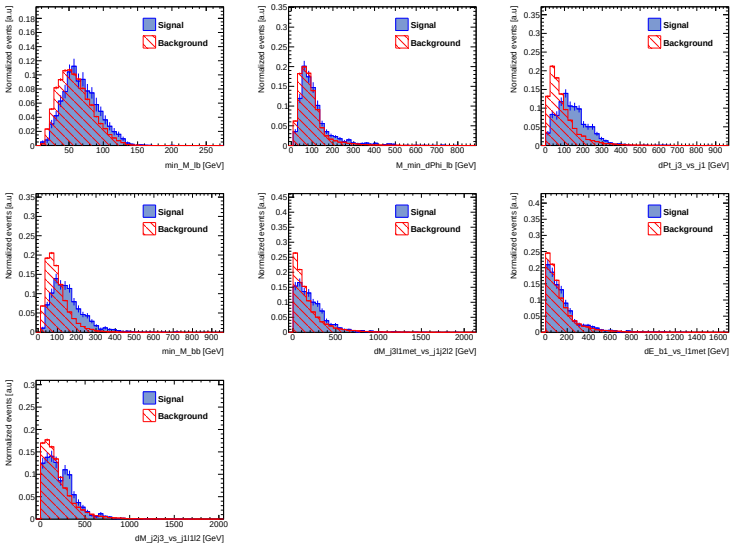


(C) $\geq 4j \geq 4b$

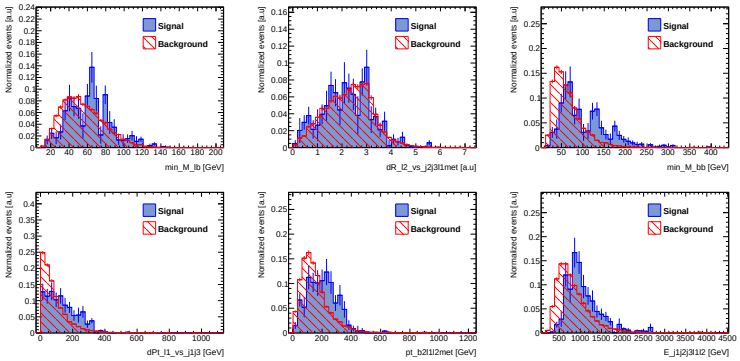
FIGURE 4.16: BDT input variable used in 400 GeV BDT, separated into different SRs. The signal is H^+ and background is inclusive $t\bar{t}$ sample. Both signal and backgrounds are normalized to unity.



(A) $3j3b$

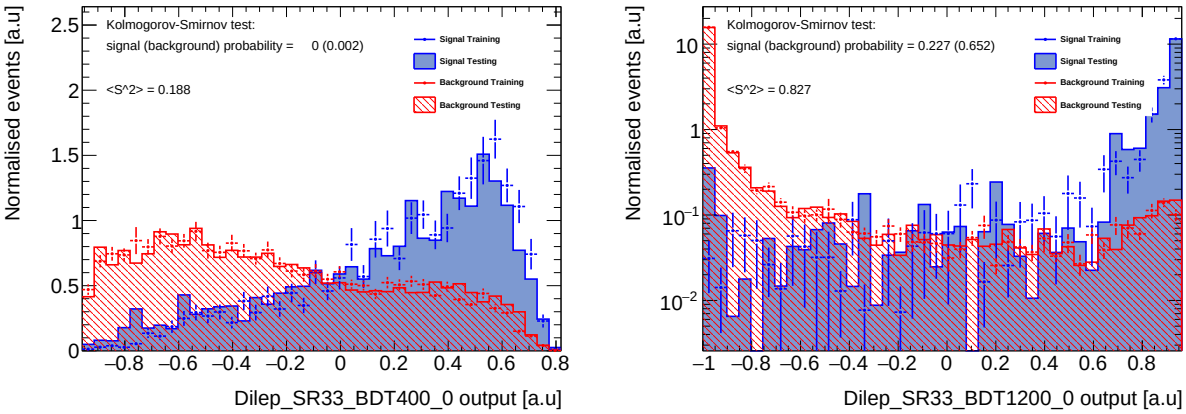


(B) $\geq 4j3b$



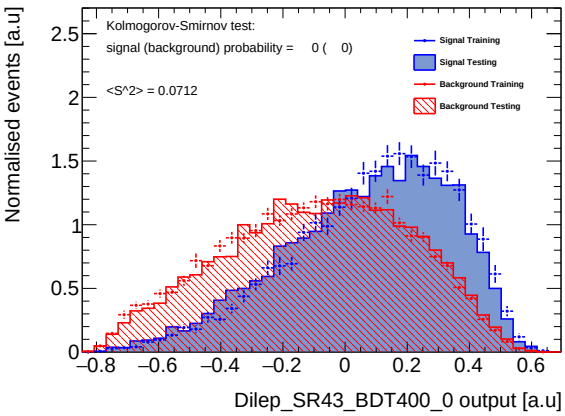
(C) $\geq 4j \geq 4b$

FIGURE 4.17: BDT input variable used in 600 GeV BDT, separated into different SRs. The signal is H^+ and background is inclusive $t\bar{t}$ sample. Both signal and backgrounds are normalized to unity.

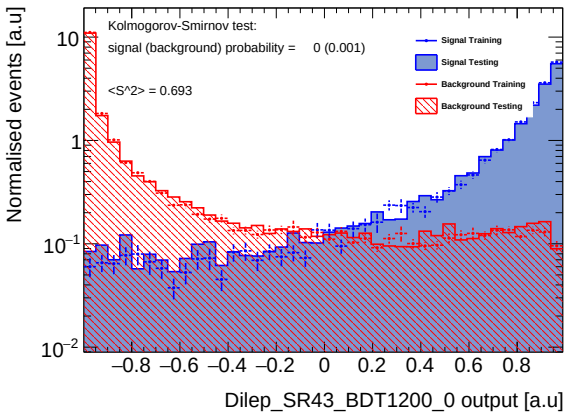


(A) 400 GeV3j3b

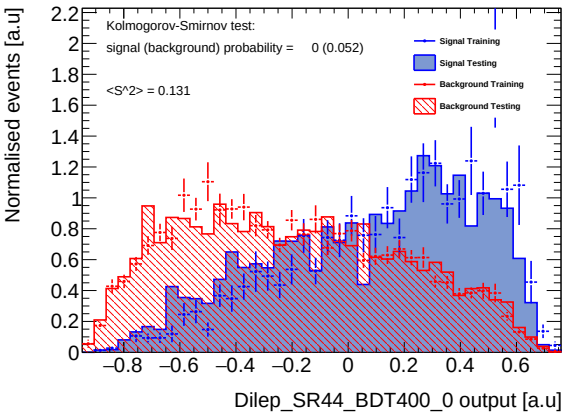
(B) 1200 GeV3j3b



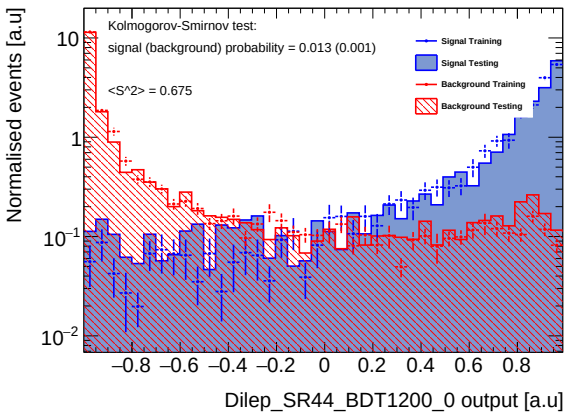
(C) 400 GeV$\geq 4j3b$



(D) 1200 GeV$\geq 4j3b$



(E) 400 GeV$\geq 4j \geq 4b$



(F) 1200 GeV$\geq 4j \geq 4b$

FIGURE 4.18: BDT outputs for 400 GeV(left) and 1200 GeV(right) H^+samples as examples. BDTs are trained for each SR separately.

4.4 Mass reconstrution in di-lepton final states

4.4.1 Introduction

As discussed at section 4.3.2 and 4.3.3, one of the biggest challenges of the analysis in the low mass range is to separate signal and $t\bar{t}b\bar{b}$ events since $t\bar{t}b\bar{b}$ is irreducible and can exactly mimic signal events. Increasing the separation of signal and $t\bar{t}b\bar{b}$ is a key point to improve the sensitivity and lower the correlation between the signal strength and the $t\bar{t}b\bar{b}$ normalization factor.

So far the BDT is used to discriminate signal events, but it is not powerful enough in the low mass range since there only kinematic variables are used as inputs which are roughly the same between signal and $t\bar{t}b\bar{b}$. This is due to the fact that a low mass H^+ is not energetic enough to provide large p_T decay products significantly different from the b-quarks in $t\bar{t}b\bar{b}$ events. Therefore a natural idea for searching new particles, which consists on reconstructing the particle's invariant mass, is proposed to increase the separation power.

Given the fact that in the dilepton final state there are two neutrinos originating from two W bosons, it is not possible to fully reconstruct the kinematic mass since all missing parts appear as one "missing transverse energy (MET)" in the detector. Thus it is not possible to reconstruct these two neutrinos in the first place. Beside this, the high jet multiplicity in the signal event topology also makes it hard to reconstruct the full event topology in a simple way.

Therefore, the multivariate method, BDT, is used to help the reconstruction, as well as taking the advantage of the so called "truth-matching" method. Jets are associated to quarks which are generated from the hard-scattering of pp collision, with $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.3$. Each jet can only be matched to one quark. This way, all objects (jets and leptons) are labelled by its truth information from truth matching. This matching result is called "Right" matching, which can also be considered as a full reconstruction in case all objects are matched. At the exception of this correct match, other "matchings" can be made by changing the objects to wrong labels by hand. All the possible permutations except the "Right" matching are called "Wrong" matching. The reconstruction BDT (RecoBDT) is then trained using "Right" matches as signal and "Wrong" matches as background.
In this analysis, the charges of two leptons are always opposite, they can always be identified by the charge and non-exchangeable. Therefore only jets are taken into account in the permutation. The definitions of object labels are listed below, as well as in figure 4.19:

- b_0: associated b quark
- b_1: b quark from H^+ top quark decay
- b_2: b quark from H^+ decay
- b_3: b quark from associated top quark decay

In addition to jets, MET also needs to be considered, since information of neutrinos is necessary to the reconstruction of W bosons. As two neutrinos cannot be reconstructed from MET, three different treatments of MET are considered here:

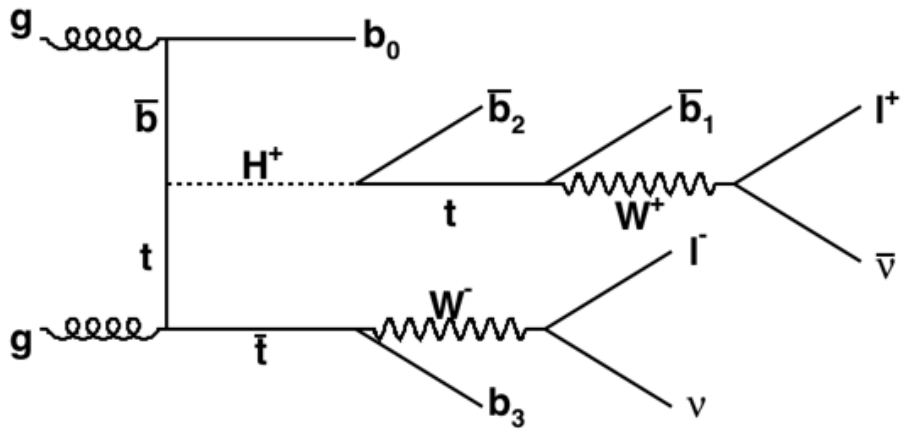


FIGURE 4.19: The Feynman diagram of H^+ di-lepton final states, with the labelling of objects indicated for reconstruction BDT training.

- Treatment 1: Assume neutrinos have the same direction as MET, and transverse momentum (p_T) of each neutrino is a half of MET.
- Treatment 2: Assume neutrinos have the same direction as the corresponding charged leptons, and then calculate the 4-momentum of each neutrino
- Treatment 3: Ignore neutrinos in the reconstruction of W bosons. In this case, W boson will be identity to the corresponding charged lepton.

Various W bosons and top quarks combinations are then reconstructed based on the above treatments separately and tested together in the BDT training. The best variables are selected in the training.

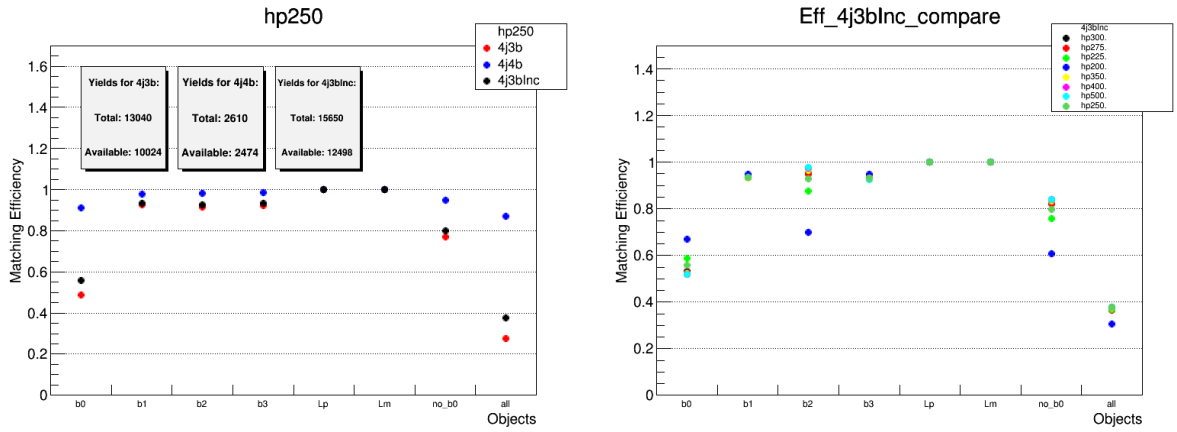
4.4.2 Reconstruction BDT Training

First, a check of matching efficiency on all jets and leptons is performed to determine the reconstruction strategy. The result is shown in figure 4.20. As expected, the matching efficiency of leptons are close to 100 % since they are required to pass tight selection. Since in addition they are also required to be isolated, they can be considered as exactly the truth leptons. But note that not all leptons (e, μ) come directly from W bosons decay in the signal events, they can also come from leptonic τ decay through the chain $W \to \tau\nu \to e(\mu)\nu\nu$. This will bring another neutrino which makes the topology even more complex. All events with a leptonic decay τ are simply removed from the training set.

In terms of jets, b_0 has significantly low efficiency. This is due to the fact that the associated b quark is used to be softest among all four b-quarks which makes it eliminated by jet p_T threshold more easily. Beside this softest b-quark, the other three b-quarks show good matching efficiency.

Based on the discussion above, there are three more cuts applied in the event selection in addition to the standard signal region selection.

- At least 3 jets, in which b_1, b_2, b_3 must be matched. But they are not required to be b-tagged to increase statistics.



(A) H^+ 250 GeV

(B) All mass points in $\geq 4j \geq 3b$ region

FIGURE 4.20: Matching efficiency of all objects, where "no_b0" means total efficiency without taking into account b0, "all" means total efficiency taking into account all objects. 4.20a shows the comparison of various regions of Hp 250 GeV sample. 4.20b shows the comparison of all different mass points in $\geq 4j \geq 3b$ region.

- Remove events with fake leptons (in this case leptons are not matched).
- Remove events with either of the W bosons decays to $\tau\nu$.

All possible kinematic properties of reconstructed intermediate particles ($W^{\pm}$, t, $\bar{t}$, H^+, etc) are tested in the BDT. Then the following 12 variables are selected as BDT training inputs:

- $\Delta_{B_m B_n}[m, n = 1, 2, 3]$: Angular distance between two b partons.
- m_t^{NoNu}: Invariant mass of top quark, reconstructed without neutrino information (treatment 3)
- $m_{\bar{t}}^{NoNu}$: Invariant mass of anti-top quark, reconstructed without neutrino information (treatment 3)
- $\Delta R_{t\bar{t}}^{NoNu}$: Angular distance between two reconstructed top quarks without neutrino information (treatment 3).
- $p_{B_n}^T[n = 1, 2, 3]$: Transverse momentum of a b parton
- $jetTagWeight_n[n = 1, 2, 3]$: Pseudo continuous b-tagging weight of a b parton. This weight is defined as binned mv2c10 weight used for b-tagging (i.e. first bin means this jet is not tagged, second bin means this jet passes 85 % tagger but not 77 % tagger).

The parameters of BDT training used here are as following (the meanings are described at 4.3.4.1):

- NTrees=1000
- nCuts=25
- MaxDepth=5

- Shrinkage=0.3
- BoostType=Grad
- GradBaggingFraction=0.7

These parameters are simply optimized by comparing separation power of BDT outputs trained with different set-ups.

The BDT training is done with 8 different H^+ masses samples (200 GeV, 225 GeV, 250 GeV, 275 GeV, 300 GeV, 350 GeV, 400 GeV, 500 GeV) in the $\geq 4j \geq 3b$ inclusive signal region to gain enough statistics. The plots showing the discriminating power of input variables are shown in figure 4.21 and 4.22 for 225 GeV and 400 GeV H^+ as examples. The training results are shown in figure 4.23, also taking 225 GeV and 400 GeV H^+ as examples. There is no obvious overtraining observed and the results show a relatively good separation between right and wrong matches.

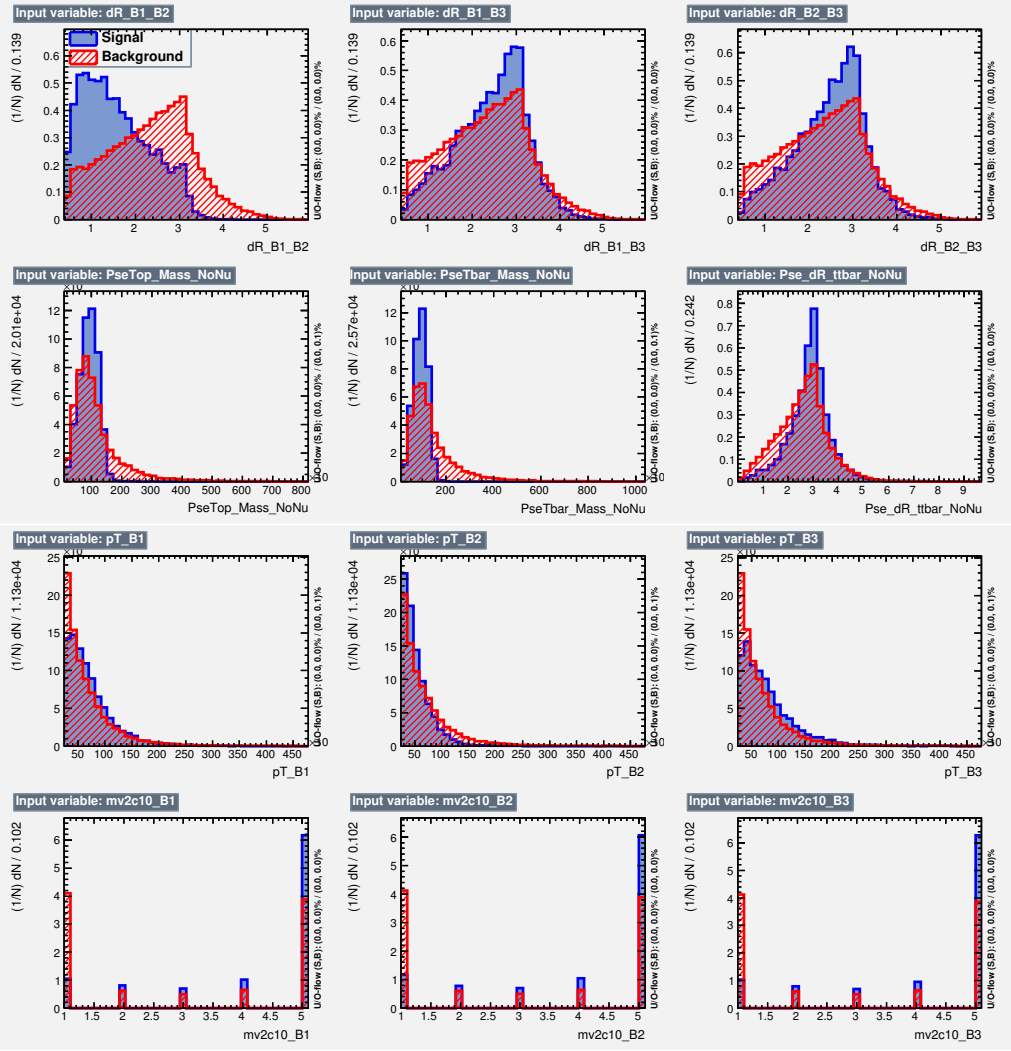


FIGURE 4.21: The plots showing the discriminating power of reconstruction BDT input variables for 225 GeVH^+. The signal is correct matching (blue) while the backgrounds are wrong matching (red).

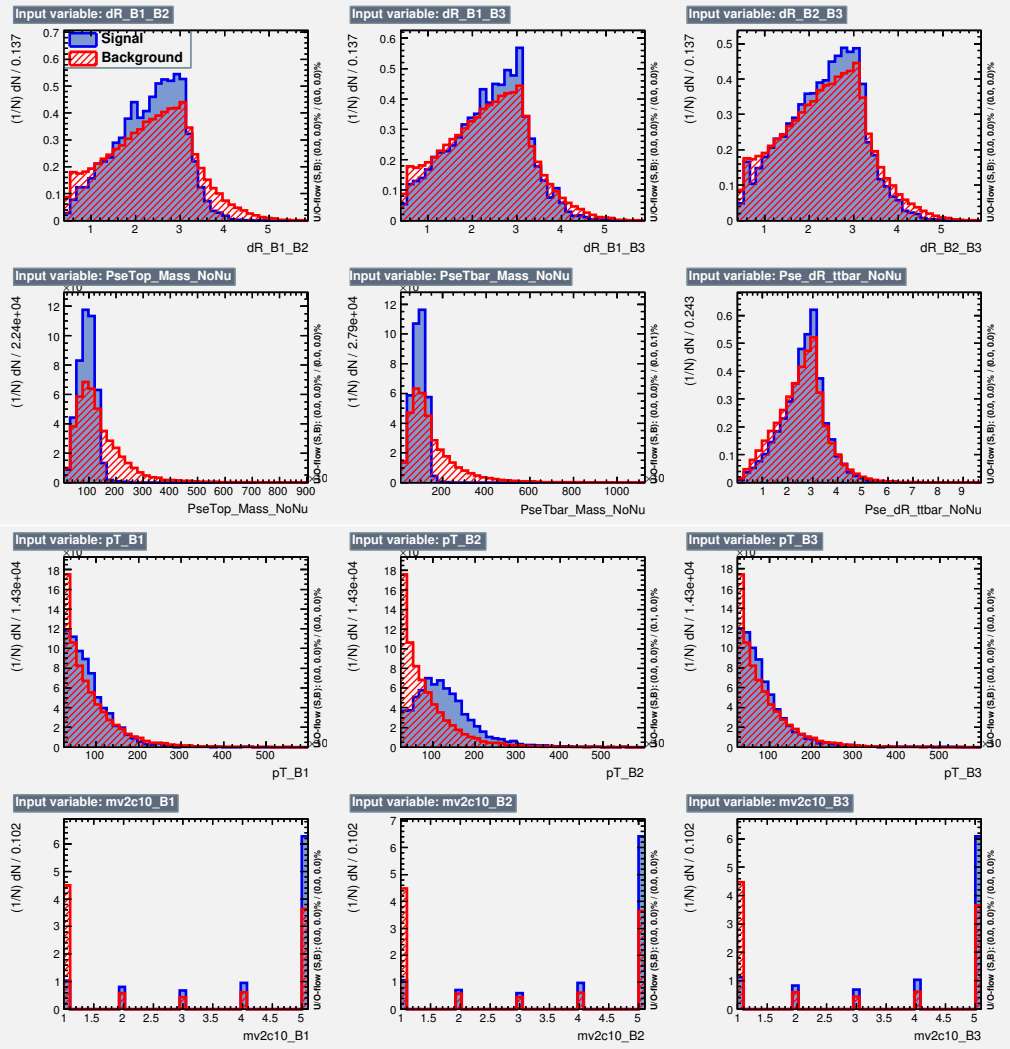
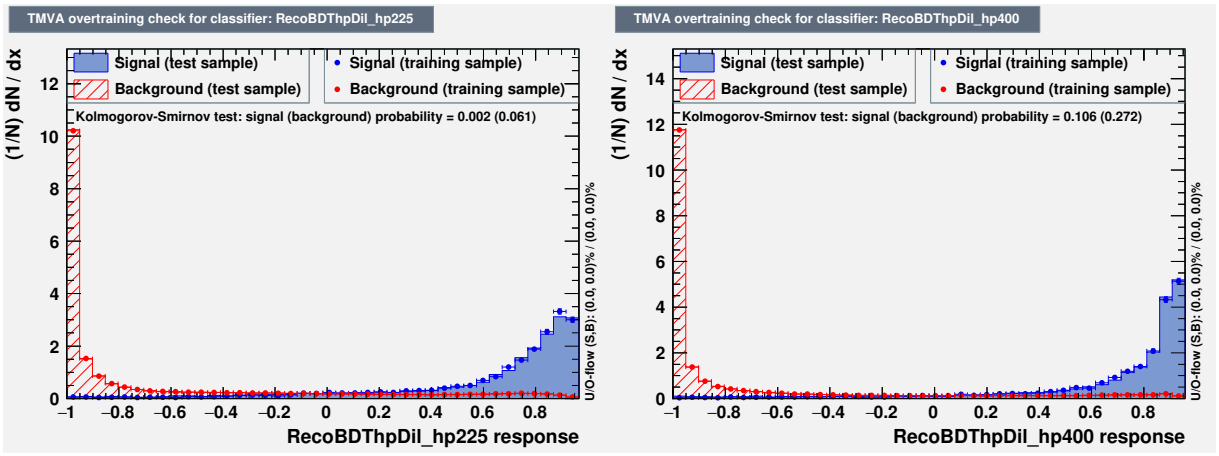


FIGURE 4.22: The plots showing the discriminating power of reconstruction BDT (RecoBDT) input variables for 400 GeVH^+. The signal is correct matching (blue) while the backgrounds are wrong matching (red).



(A) Overtraining check of RecoBDT trained on 225 GeV H^+

(B) Overtraining check of RecoBDT trained on 400 GeV H^+

FIGURE 4.23: Training results of 225 GeV (4.23a) and 400 GeV (4.23b) H^+ samples. No overtraining is seen in both cases since the testing sample agrees with training sample very well.

4.4.3 Application to the analysis

For a given event which passes the selection criteria described in section 4.4.2, every combination of these objects is assigned a BDT score evaluated from the RecoBDT trained before. The number of mappings depends on the jet multiplicity of the event since leptons are always considered to be matched, which is a permutation following the equation $N_{mapping} = P(nJets, 3)$. The mapping which gets the highest BDT score, or say which is closest to 1, is considered as the right match. And the value is taken as the output of this method.

Depending on this, there are two approaches of applying RecoBDT to the analysis as indicated below:

- Reconstruct the full signal event topology using RecoBDT.
- Put RecoBDT output into the classification BDT (the discriminant in the fit) as an additional input variable.

For the first approach, the reconstruction efficiency of considering the mapping with highest RecoBDT score as right match is estimated for signal samples, which is shown in figure 4.24. Despite the fact that the RecoBDT has a good separation, the reconstruction efficiency is not as good as expected. The reason is that there are likely several combinations that give very similar RecoBDT score in one event and the RecoBDT is not fine-tuned enough to distinguish them.

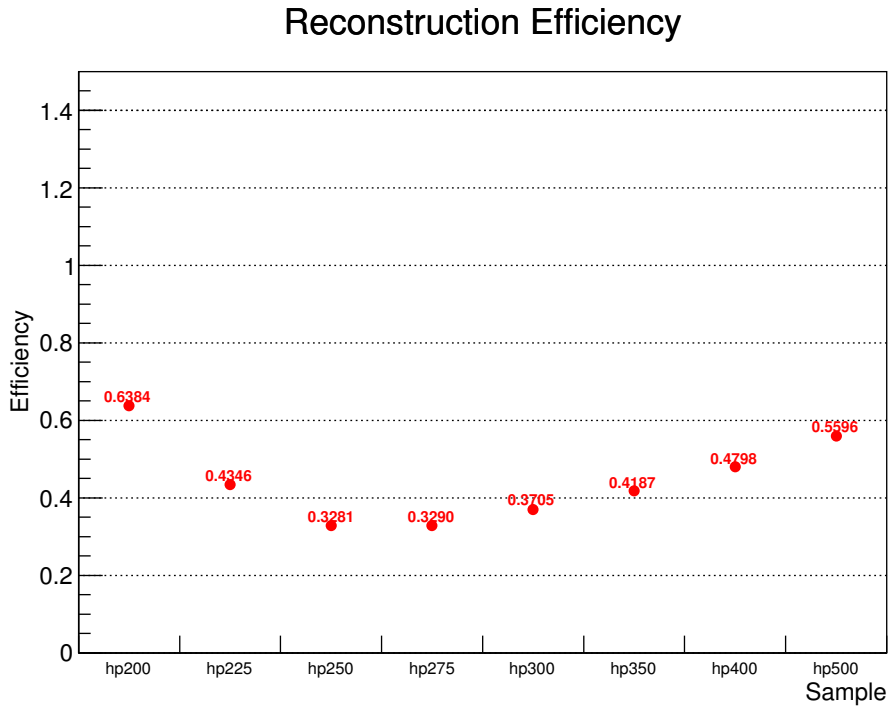
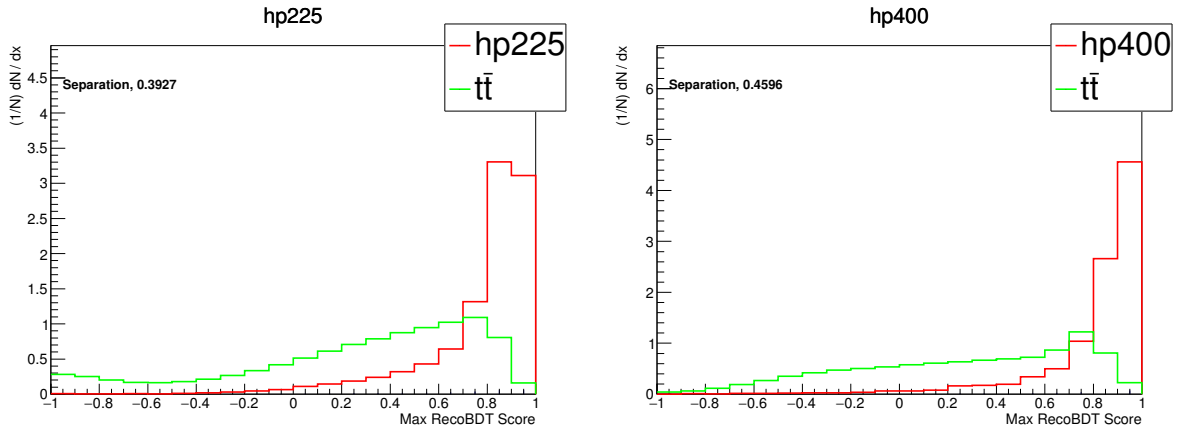


FIGURE 4.24: The reconstruction efficiency test for all mass points. The efficiency is derived from test samples.

For the second approach, first of all a test of separation power against inclusive $t\bar{t}$ background is done in figure 4.25. Note that in the inclusive $t\bar{t}$ sample, the dominant process is $t\bar{t} + ljets$, which usually falls into CRs but not SRs. Thus there is a reduction of the separation in SRs where the

dominant process is $t\bar{t} + b\bar{b}$, which is not quite sensitive to the most powerful variable in the inclusive $t\bar{t}$ sample, pseudo continuous b-tagging (PCBT). Then the RecoBDT output is plugged into the classification BDT training and compared to the original classification BDT without RecoBDT in figure 4.26 and 4.27 for 225 GeV and 400 GeV H^+respectively.

Furthermore, to save computing time, we also evaluate RecoBDT trained on a given mass to all signal samples (i.e. RecoBDT trained on 200 GeV signal is applied to all different H^+ samples from 200 GeV to 500 GeV and so for the RecoBDTs trained on the other H^+ samples). Therefore for a signal sample with given mass, there are in total nine classification BDT trained based on the different RecoBDT outputs, as shown in figure 4.28. The idea is to select one RecoBDT which has overall best performance among all signal samples and SRs. In particular the performance in 200 GeV to 300 GeV region is considered first, since classification BDTs for H^+ mass above 300 GeV already have a relatively good separation.

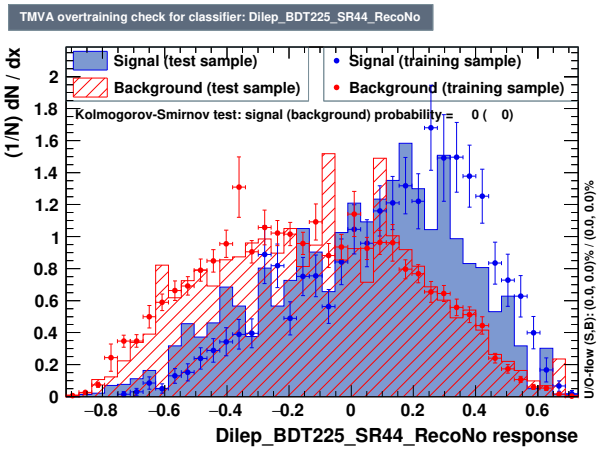


(A) Separation power against inclusive $t\bar{t}$ sample of 225 GeV H^+

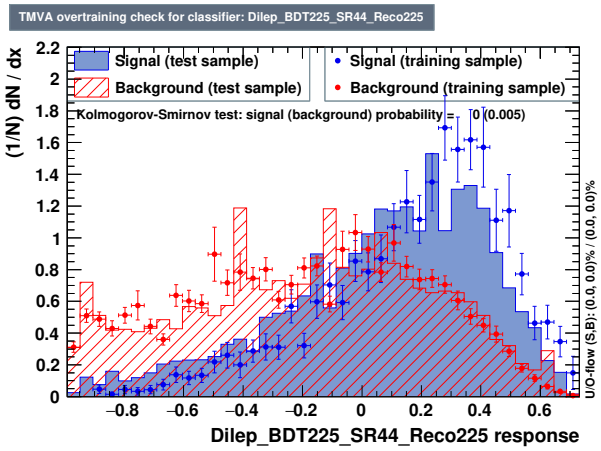
(B) Separation power against inclusive $t\bar{t}$ sample of 400 GeV H^+

FIGURE 4.25: Test of the separation power against inclusive $t\bar{t}$ sample for 225 GeV (4.25a) and 400 GeV (4.25b) H^+ respectively.

As a result, the RecoBDT trained on 225 GeV signal is selected to be used for all mass points. A test of fit comparing the fitted parameter-of-interest (POI which is signal strength of H^+ in this analysis) with and without RecoBDT is shown in 4.29. The result shows that RecoBDT does give improvements of 15 % . But due to the fact that PCBT variables are used in RecoBDT training, the corresponding systematics need to be taken into account in the fit, which will cover most of the improvements and reduct the improvements to around 1 % . As discussed in previous section, the calibration of PCBT is not available for jets with $p_T \leq 200$ GeV which makes the variable unusable in the training. This limited improvement also reflects that the current RecoBDT is not optimal and need to be further fine-tuned to get better performance. Given that this method is quite time consuming (around 0.1 seconds per event on average), it is not used to produce the final results.

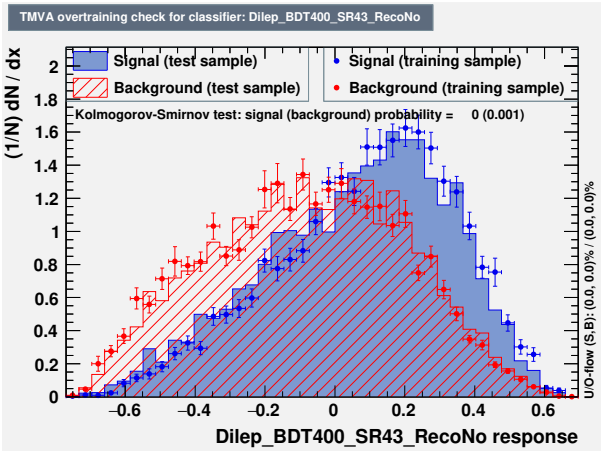


(A) classification BDT trained without Reco BDT input for 225 GeV H^+ sample in $\geq 4j \geq 4b$ SR

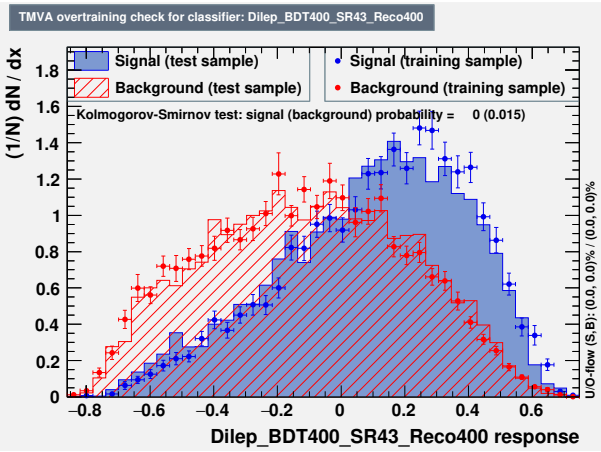


(B) classification BDT trained with 225 GeV Reco BDT input for 225 GeV H^+ sample in $\geq 4j \geq 4b$ SR

FIGURE 4.26: Comparison of classification BDT outputs trained without/with reconstruction BDT input of 225 GeV H^+ sample in $\geq 4j \geq 4b$ SR, shown in 4.26a and 4.26b respectively.

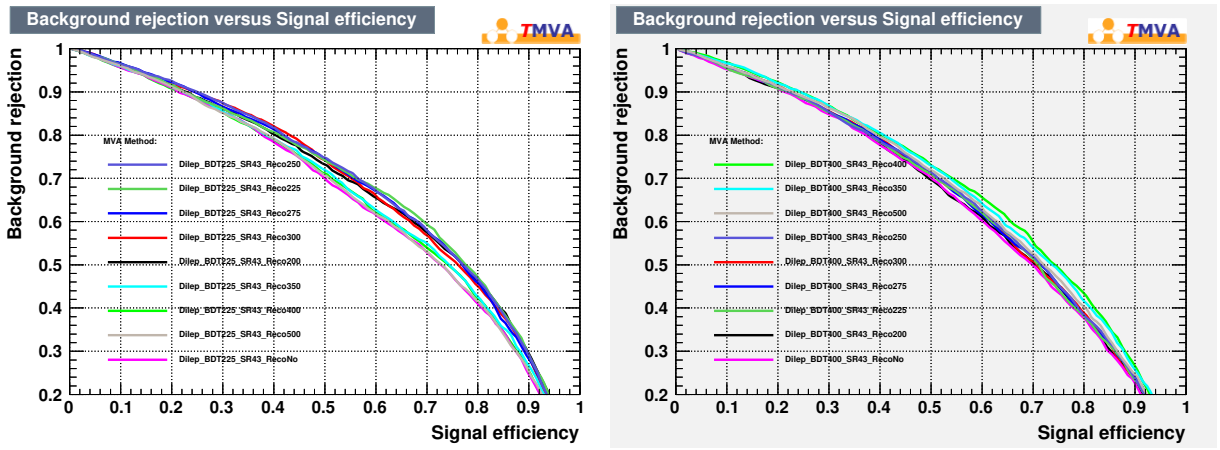


(A) classification BDT trained without Reco BDT input for 400 GeV H^+ sample in $\geq 4j3b$ SR



(B) classification BDT trained with 400 GeV Reco BDT input for 400 GeV H^+ sample in $\geq 4j3b$ SR

FIGURE 4.27: Comparison of classification BDT trained without/with reconstruction BDT input of 400 GeV H^+ sample in $\geq 4j3b$ SR, shown in 4.27a and 4.27b respectively.



(A) ROC curve of various classification BDT of 225 GeV H^+ trained on different RecoBDT

(B) ROC curve of various classification BDT of 400 GeV H^+ trained on different RecoBDT

FIGURE 4.28: Comparison of the performance of different mass RecoBDT applied to a given mass classification BDT, 4.28a shows for 225 GeVand 4.28b shows for 400 GeVboth in $\geq 4j3b$ SR. The meaning of the names in the legend, for instance "Dilep_BDT225_SR43_Reco200", is that a classification BDT of 225 GeV H^+in SR $\geq 4j3b$ is trained with the reconstrution BDT trained on 200 GeV H^+as an additional input. And so for the other legends.

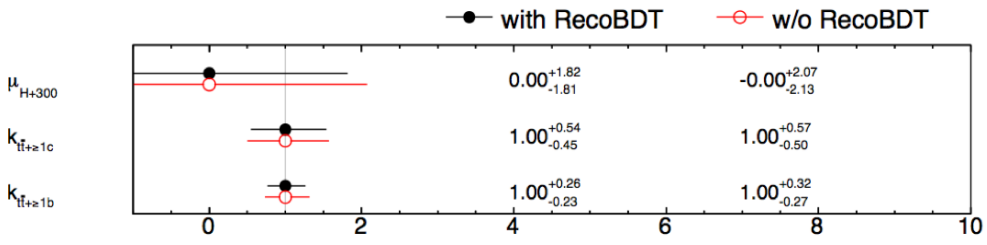


FIGURE 4.29: The comparison of asimov S+B fit using default classification BDT or RecoBDT + classification BDT as discriminant. The result shows some improvements of $\Delta\mu$.

4.5 Results

4.5.1 Systematic uncertainties

There are various systematic uncertainties from different sources affecting the results, such as the luminosity uncertainties, the calibration of objects and the modelling of simulated samples. Systematic uncertainties can modify the normalization of simulated samples, change the shape of distributions or both of them. The uncertainties are split into different components to get a precise estimation, as listed in table 4.8.

Systematic uncertainty	Type	Number of components
Luminosity	N	1
Pile-up	NS	1
Electron reconstruction	NS	6
Muon reconstruction	NS	13
Jet and E_T^{miss} reconstruction	NS	28
Flavour tagging, 70% efficiency calibration (*)	NS	27
Flavour tagging, step-wise efficiency calibration (*)	NS	126
Signal QCD scale and PDF	NS	31
Background modelling, $t\bar{t}$+ jets	NS	29
Background modelling, other top	NS	25
Background modelling, non-top (lepton-plus-jets final state)	N	13
Background modelling, non-top (di-lepton final state)	N	4

TABLE 4.8: List of systematic uncertainties considered. 'N' indicates that the uncertainty is taken as normalization-only, while 'NS' means that the uncertainty applies to both normalization and shape. Flavour-tagging uncertainties marked with '*' are different for the two sets of calibrations: the step-wise efficiency calibration for $m_{H^+} \leq 300$ GeV, and the 70% efficiency point calibration elsewhere.

4.5.1.1 Instrumental uncertainties

The combined uncertainty of the integrated luminosity of data collected in 2015 and 2016 is 2.1 %. It is applied as a normalization uncertainty for all simulated samples. The uncertainty is derived from a methodology that detailed in Ref [83]. An uncertainty is also included to account for the variations in pile-up, which covers the uncertainty on the inelastic cross-sections [84].

Uncertainties associated to charged leptons come from the trigger selections, the object reconstruction, the identification, the isolation criteria, and the lepton momentum scale and resolution. These uncertainties are estimated by comparing $Z \to \ell^+\ell^-$ ($\ell = e, \mu$) events in data and MC simulations [31, 85]. The charged lepton uncertainties have a small impact on the result.

Uncertainties associated to jets come from the jet reconstruction, identification efficiencies related to JES, jet energy resolution and jet vertex tagger efficiency [86]. The JES-related uncertainties are derived by combining information from test-beam data, LHC collision data (*in situ* techniques) and simulation [87]. Many source of JES uncertainties which are related to the *in situ* calibration using Z+jets, γ+jets and multi-jet data are reduced to eight uncorrelated components through an eigen-solve-decomposition. Other components are related to jet flavor, pile-up corrections, η-dependence and high p_T jets. Part of the uncertainties on E_T^{miss} are propagated from leptons and jets

since E_T^{miss} is calculated based on these objects. Other uncertainties on E_T^{miss} arise from soft objects which are not considered in the calculation of letpons and jets [88].

Another important source of uncertainties on jets is from flavor tagging. Differences between data and simulation in the b-tagging efficiency for b-jets, c-jets and light jets are taken into account using correction factors. The corrections for b-jets and mis-tag rate for c-jets are derived from $t\bar{t}$ events and the mis-tag rate of light jets is derived from multi-jet samples. Systematic uncertainties affecting the correction factors are derived in the p_T and η bins used for extracting the correction factors and then transformed to uncorrelated components using an eigenvector decomposition [89, 90, 91]. For $m_{H^+} > 300$ GeV, corrections of a fixed working point 70 % is used, as well as the corresponding systematics (6 for b-jets, 3 for c-jets and 16 for light jets). For $m_{H^+} \leq 300$ GeV, a step-wise correction is used to support the kinematic discriminant D in lepton-plus-jets channel and the number of eigen-variations is increased by a factor of 5 to account for the five b-tagging efficiency bins.

4.5.1.2 Theoretical uncertainties

The uncertainties due to the choice of different scales in the H^+ samples is estimated by varying the renormalization and factorisation scales up and down by a factor of two. The uncertainty ranges from 7 % at low mass to 15 % at high mass ($\geq$ 1300 GeV) in lepton-plus-jets final state and from 12 % to 16.5 % in di-lepton final state. The PDF uncertainty is esimated using the PDF4LHC15_30 PDF set [92].

The $t\bar{t}$ modelling uncertainty is one of the dominant uncertainties in the analysis and it is separated into several components. The uncertainty on the inclusive $t\bar{t}$ cross-section at NNLO+NNLL [54] is 6 %, including the effects of renormalization, factorisation scales and the PDF, the QCD coupling constant and the top quark mass. Due to the large difference between 4FS and 5FS predictions for the $t\bar{t}+3b$ process, an additional 50 % normalization uncertainty is assigned to it. The uncertainty due to the choice of generator is derived by comparing the nominal sample to a sample generated using SHERPA 2.2.1 and a POWHEG sample using HERWIG 7 [80, 93] for parton showering is used to derive the uncertainty due to the choice of parton shower and hadronisation model. Furthermore, the uncertainty due to the modelling of initial-state radiation and final-state radiation is derived with two different POWHEG+PYTHIA8 samples in which the radiation is increased or decreased [94].

For the $t\bar{t}+\geq 1b$ background, an additional uncertainty is assigned by comparing POWHEG+PYTHIA8 sample and SHERPA with 4FS sample. This includes the difference between a 5FS inclusive $t\bar{t}$ prediction at NLO and a 4FS NLO $t\bar{t}b\bar{b}$ prediction. For the $t\bar{t}+\geq 1c$ background, an additional uncertainty is assigned by comparing a MG5_AMC sample interfaced to Herwig++ [80] with the nominal sample. All the alternative samples are reweighted to the NLO prediction of $t\bar{t}b\bar{b}$ from SHERPA before the uncertainty is derived.

In addition, uncertainties due to the reweighting (4.3.1.3) to the SHERPA NLO prediction of $t\bar{t}b\bar{b}$ are considered. For these uncertainties, the $t\bar{t}+\geq 1b$ is reweighted to different SHERPA predictions with modified scale parameters, in particular where the renormalization scale is varied up and down by a factor of two. Two alternative PDF sets, MSTW2008NLO [95] and NNPDF2.3NLO [51], are used and uncertainties in the underlying event and parton shower are estimated from samples

with an alternative set of tuned parameters for the underlying event and an alternative shower recoil scheme. Because of the absence of b-jets from multi-parton interactions and final-state gluon radiation in the $t\bar{t}b\bar{b}$ prediction from SHERPA, a 50% uncertainty is assigned to the $t\bar{t}+b$ (MPI/FSR) category. An uncertainty due to the reweighting of the leading jet p_T is derived by comparing a reweighted event sample with an event sample without reweighting. Since the reweighting changes the normalization for jet with $p_T > 400$ GeV by 15 %, an additional normalization uncertainty of 15% is assigned.

For the single top production cross-section, an uncertainty of 5 % is applied, uncorrelated between Wt and t-channel. There is also an uncertainty due to initial-state and final-state radiation assigned by using samples with factorisation and renormalization scale variations and appropriate variations of the Perugia 2012 set of tuned parameters. The modelling uncertainties of parton showering and hadronisation are estimated by comparing with samples where the parton shower generator is Herwig++ instead of PYTHIA 6.428. The uncertainty in the interference between Wt and $t\bar{t}$ production at NLO is estimated by comparing the default 'diagram removal' scheme with an alternative 'diagram subtraction' scheme.

The uncertainty from $t\bar{t}V$ generation is estimated by comparison with samples generated with SHERPA. The uncertainty in the $t\bar{t}V$ production cross-section is about 15%, taken from the NLO predictions, uncorrelated between $t\bar{t}W$ and $t\bar{t}Z$ with PDF and QCD scale variations.

The $t\bar{t}H$ modelling uncertainty is estimated by an uncertainty in the cross-section, uncorrelated between QCD ($^{+5.8}_{-9.2}$%) and the PDFs ($\pm$3.6%), and the modelling of the parton shower and hadronisation by comparing PYTHIA8 with Herwig++. The minor $tH+X$ backgrounds, $tHjb$ and WtH are treated as one background and its cross-section uncertainty is 6 % due to PDF uncertainties and another 10 % due to factorisation and renormalization scale uncertainties.

The uncertainties from the data-driven estimation of non-prompt leptons are based on a comparison between data and the non-prompt lepton estimates in CRs. A 50 % uncertainty is assigned in the lepton-plus-jet final state. In the di-lepton final state, an uncertainty of 25% is assigned.

An uncertainty of 40 % is assumed for the W+jets cross-section, with an additional 30% for W+HF jets. These uncertainties are derived from variations of the renormalization and factorisation scales and matching parameters in SHERPA samples. An uncertainty in Z+jets of 35% is applied, including both the variation of the scales and matching parameters in SHERPA simulations and the data-driven correction factors applied to the Z+HF jets component.

4.5.2 Statistical study

4.5.2.1 Statistical model

To test for the presence of H^+ signal, a binned maximum-likelihood fit to the data is performed simultaneously in all regions. In control regions the number of events is taken as the input for the fit, while in the signal regions the binned BDT output is used. The parameter of interest is the signal strength μ, defined as the product of production cross-section $\sigma(pp \to tbH^+)$ and the

branching ratio $\mathcal{B}(H^+ \to tb)$. To model the normalization of the $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ backgrounds, two unconstrained fit parameters are employed. The procedures to quantify the agreement of background-only or background-plus-signal hypothesis and to determine the exclusion limits are based on the profile likelihood ratio test and the CL_s method [96, 97, 98].

A likelihood function, $\mathcal{L}(\mu, \theta)$, which is a product of Poisson probability terms, is constructed to estimate the signal strength μ.

$$\mathcal{L}(\mu) = \prod_k \prod_{l_k} \text{Poisson}(n_{l_k}|m_{l_k}) \prod_s \prod_{s_i} \text{Poisson}(n_{s_i}|m_{s_i}) \prod_j p(\tilde{\theta}_j|\theta_j), \tag{4.5}$$

Each Poisson term is corresponding to a CR (indice l_k) or a bin of BDT distribution in SR (indice s_i). 'n' is the observed number of events while 'm' is the expected number of events from simulations. It is also a function of a set of nuisance parameters θ. The nuisance parameters includes the effects from normalization of backgrounds, systematic uncertainties and statistical uncertainties of the MC samples. All nuisance parameters are constrained with Gaussian terms.

To extract the exclusion limit on μ, the following test statistic is used:

$$\tilde{t}_\mu = \begin{cases} -2\ln \dfrac{\mathcal{L}\left(\mu, \hat{\hat{\theta}}(\mu)\right)}{\mathcal{L}\left(0, \hat{\hat{\theta}}(0)\right)} & \hat{\mu} < 0, \\ -2\ln \dfrac{\mathcal{L}\left(\mu, \hat{\hat{\theta}}(\mu)\right)}{\mathcal{L}\left(\hat{\mu}, \hat{\theta}\right)} & \hat{\mu} \geq 0. \end{cases} \tag{4.6}$$

The values of the signal strength and nuisance parameters that maximise the likelihood function are represented by $\hat{\mu}$ and $\hat{\theta}$, respectively. For a given value of μ, the values of the nuisance parameters that maximise the likelihood function are represented by $\hat{\hat{\theta}}(\mu)$.

4.5.2.2 Fit results

The fit procedure is performed using the TRExFitter package which is developed based on RooStats. The post-fit yields under the background-plus-signal hypothesis is shown in table 4.9, and the post-fit distribution is shown in figure 4.30 for the di-lepton final state.

The total effect of uncertainties is summarized in table 4.10 for the combined fit of lepton-plus-jets and di-lepton final states. The influence of uncertainties varies as the H^+ mass. The dominant uncertainties are the modelling of $t\bar{t}+ \geq 1b$ background, the jet flavor-tagging uncertainties and the statistical uncertainty due to the limited size of simulated samples.

Process	CR $3j2b$	SR/CR $3j3b$	CR $\geq$ $4j2b$	SR $\geq$ $4j3b$	SR $\geq$ $4j$ $\geq$ $4b$
$t\bar{t}+\geq 1b$	2330 ± 330	940 ± 130	3300 ± 500	2050 ± 280	322 ± 35
$t\bar{t}+\geq 1c$	6100 ± 1300	520 ± 140	9900 ± 2000	1310 ± 290	30 ± 14
$t\bar{t}$ + light	$50\,700 \pm 2300$	260 ± 70	$32\,500 \pm 2100$	420 ± 120	4 ± 5
Non-prompt leptons	420 ± 110	6.7 ± 2.4	620 ± 160	48 ± 13	2.2 ± 0.8
$t\bar{t}W$	48 ± 7	1.48 ± 0.17	129 ± 7	9.8 ± 1.1	0.55 ± 0.21
$t\bar{t}Z$	43 ± 5	5.8 ± 1.1	174 ± 10	32.9 ± 2.0	7.0 ± 1.3
Single top Wt	1700 ± 500	40 ± 12	1110 ± 330	63 ± 26	3.9 ± 2.0
Other top	3.9 ± 0.5	0.12 ± 0.05	21.8 ± 3.5	5.8 ± 2.2	2.0 ± 0.9
Diboson	36 ± 4	1.2 ± 0.4	46 ± 6	3.1 ± 0.9	0.48 ± 0.28
Z + jets	1600 ± 500	42 ± 16	1300 ± 400	82 ± 29	5.3 ± 2.0
$t\bar{t}H$	26.2 ± 1.3	8.5 ± 0.5	116 ± 6	52.2 ± 3.5	16.0 ± 1.9
tH	1.95 ± 0.27	0.42 ± 0.10	5.7 ± 0.7	2.14 ± 0.32	0.48 ± 0.09
Total	$62\,800 \pm 2800$	1810 ± 110	$49\,300 \pm 2300$	4060 ± 200	390 ± 28
Data	62 399	1774	48 356	4047	376
H^+ (200 GeV)	92 ± 12	27 ± 4	72 ± 12	49 ± 8	9.0 ± 1.6
H^+ (800 GeV)	70 ± 12	32 ± 7	212 ± 33	157 ± 27	44 ± 9

TABLE 4.9: Event yields of the background processes and data in all regions of the di-lepton final state, after the fit to the data under the background-plus-signal hypothesis ($m_{H^+} = 200$ GeV). The expected event yields for the H^+ signal with masses of 200 GeV and 800 GeV are shown with pre-fit uncertainties and assuming a cross-section times branching ratio of 1 pb. The uncertainties include both the statistical and systematic uncertainties.

Uncertainty Source	$\Delta\mu(H^+_{200})$ [pb]	$\Delta\mu(H^+_{800})$ [pb]
Jet flavour tagging	0.70	0.050
$t\bar{t}+\geq 1b$ modelling	0.65	0.008
Jet energy scale and resolution	0.44	0.031
$t\bar{t}$+light modelling	0.44	0.019
MC statistics	0.37	0.044
$t\bar{t}+\geq 1c$ modelling	0.36	0.032
Other background modelling	0.36	0.039
Luminosity	0.24	0.010
Jet-vertex assoc., pile-up modelling	0.10	0.006
Lepton, E_T^{miss}, ID, isol., trigger	0.08	0.003
H^+ modelling	0.03	0.006
Total systematic uncertainty	1.4	0.11
$t\bar{t}+\geq 1b$ normalisation	0.61	0.022
$t\bar{t}+\geq 1c$ normalisation	0.28	0.012
Total statistical uncertainty	0.69	0.050
Total uncertainty	1.5	0.12

TABLE 4.10: The summary of the effects of the systematic uncertainties on the signal strength parameter for the combination of the lepton-plus-jets and di-lepton final states, shown for an H^+ signal with a mass of 200 and 800 GeV. The total systematic uncertainty can be different from the sum in quadrature of the individual sources due to the correlations between uncertainties. The normalisation factors for both $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ are included in the statistical uncertainties.

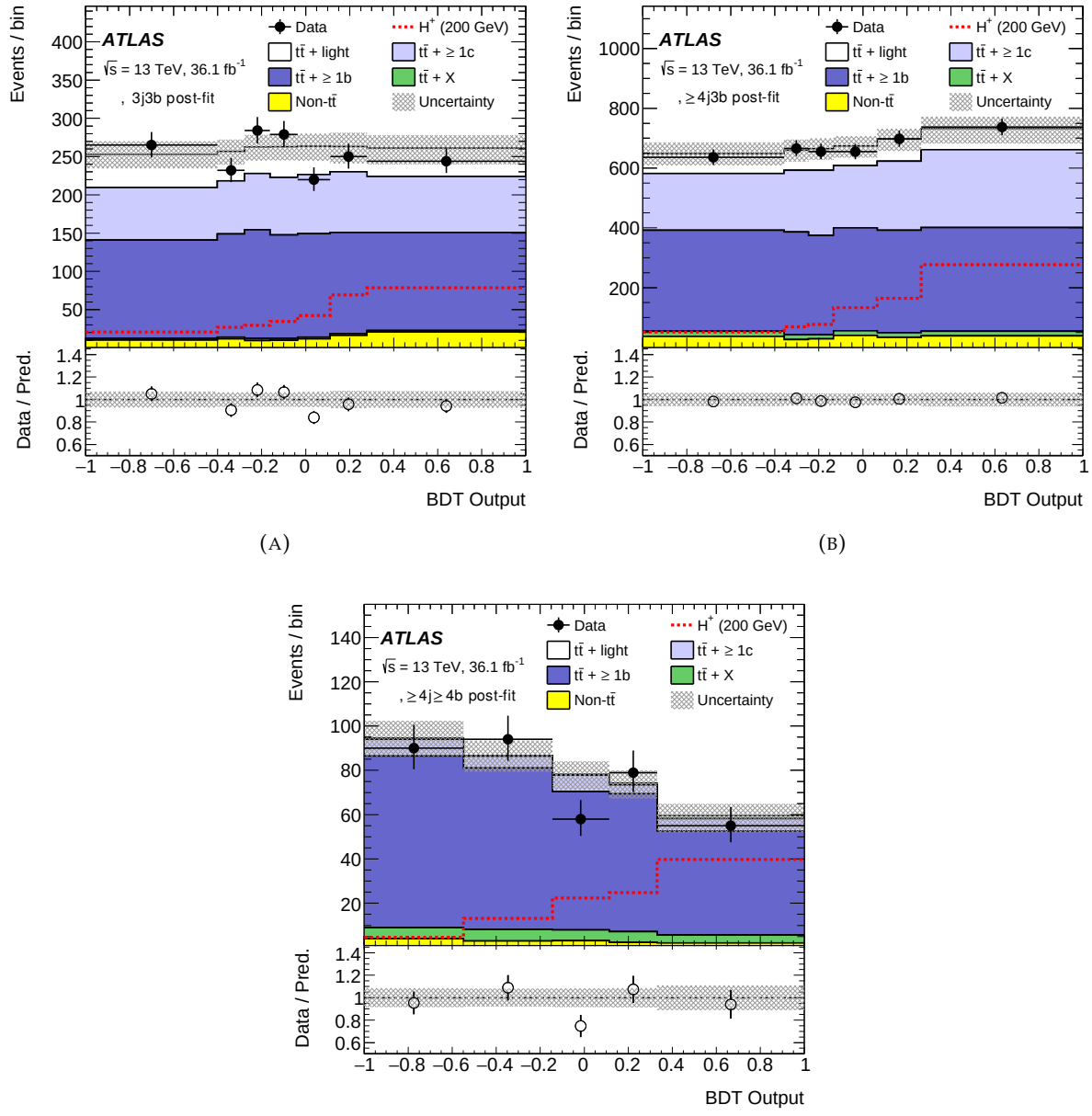


FIGURE 4.30: Distributions of the BDT output after the fit to data under the background-plus-signal hypothesis ($m_{H^+} = 200$ GeV) in the SRs of di-lepton final state: (4.30a) for $3j3b$, (4.30c) for $\geq 4j3b$ and (4.30c) for $\geq 4j \geq 4b$. The pre-fit signal distribution is shown superimposed as a dashed line with arbitrary normalisation. For this mass point (200 GeV) the signal strength is found to be -0.4 ± 1.5 pb.

4.5.2.3 interpretation

The 95% confidence level (CL) upper limits on $\sigma(pp \to tbH^+) \times \mathcal{B}(H^+ \to tb)$ are presented in Figure 4.31. The observed (expected) 95% CL upper limits on the $pp \to tbH^+$ production cross-section times the branching ratio $\mathcal{B}(H^+ \to tb)$ range from $\sigma \times \mathcal{B} = 2.9\,(3.0)$ pb at $m_{H^+} = 200$ GeV to $\sigma \times \mathcal{B} = 0.070\,(0.077)$ pb at $m_{H^+} = 2$ TeV. The largest deviation from the SM hypothesis obtained from the fit is observed at 300 GeV, corresponding to a local p_0 value of 1.13 %.

Figure 4.32 shows the 95% CL exclusion limits set on tan β for the $m_h^{\text{mod}-}$ scenario of the MSSM [99, 39, 8] and the hMSSM [100, 101, 102]. Beyond tree level, the Higgs sector is affected by the choice of parameters in addition to Higgs boson masses and tan β. For the $m_h^{\text{mod}-}$ benchmark scenario the top-squark mixing parameter is chosen such that the mass of the lightest CP-even Higgs boson, m_h, is close to the measured mass of the Higgs boson that has been discovered. In the hMSSM scenario, instead of adjusting the parameters of soft supersymmetry breaking, the value of m_h is used to predict the masses and couplings of the MSSM Higgs bosons.

For H^+ masses of 200–920 GeV (200–965 GeV), the observed exclusion of low values of tan β at 95% CL is in the range 0.5–1.91 (0.5–1.95) for the $m_h^{\text{mod}-}$ (hMSSM) scenario. The most stringent limits on tan β are set for H^+ masses around 250 GeV. High values of tan β between 36 and 60 are excluded in the H^+ mass range 200–520 GeV (220–540 GeV) for the $m_h^{\text{mod}-}$ (hMSSM) scenario. The most stringent exclusion, tan $\beta > 36$, is at 300 GeV for both the $m_h^{\text{mod}-}$ and hMSSM benchmark scenarios. In the $m_h^{\text{mod}-}$ scenario for tan $\beta = 0.5$, the observed (expected) exclusion of H^+ masses is $m_{H^+} < 920$ GeV ($m_{H^+} < 930$ GeV).

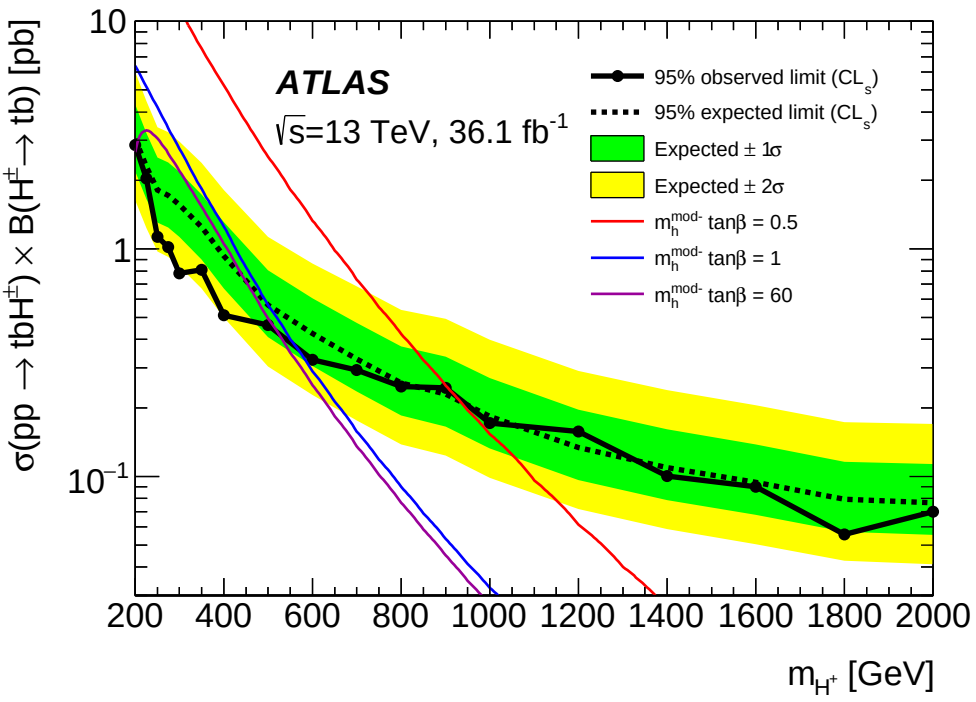
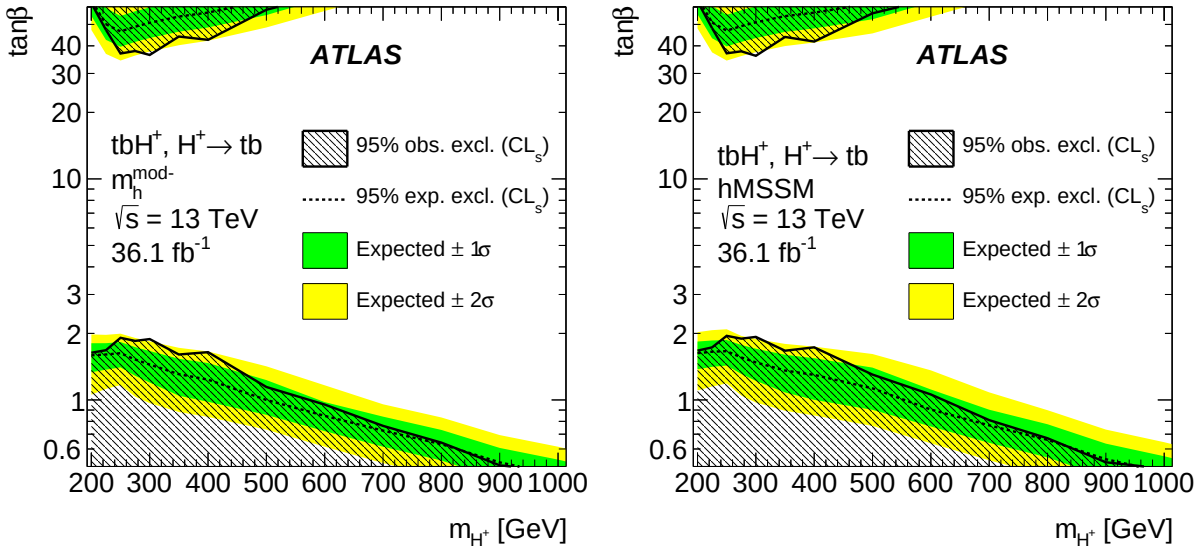


FIGURE 4.31: Expected and observed limits for the production of $pp \to tbH^+$. The bands surrounding the expected limit show the 68% (green) and 95% (yellow) confidence intervals. The limits are based on the combination of the lepton-plus-jets and di-lepton final states. Examples of theory predictions are shown for three representative values of tan β in the $m_h^{\text{mod}-}$ benchmark scenario.



(A) (B)

FIGURE 4.32: Expected and observed limits on tan β as a function of m_{H^+} in the $m_h^{\text{mod}-}$(4.32a) and the hMSSM (4.32b) scenarios of the MSSM. Limits are shown for tan β values in the range of 0.5–60, where predictions are available from both scenarios. The bands surrounding the expected limits show the 68% (green) and 95% (yellow) confidence intervals. The limits are based on the combination of the lepton-plus-jets and di-lepton final states.

4.5.2.4 CMS results

The other experiment at the LHC, CMS, also published their result on the same search for heavy H^+ boson at $\sqrt{s} = 8$ TeVwith a data sample corresponding to an integrated luminosity of 19.7 fb^{-1} [103]. The analysis is performed in $\mu\tau_h$ final state and dilepton (ee/eμ/$\mu\mu$) final state targeting at both $H^+ \to t\bar{b}$ and $H^+ \to \tau\nu$. Another single lepton (e/μ+jets) final state is dedicated to $H^+ \to t\bar{b}$. The mass of H^+varies from 180 GeV to 600 GeV.

The signal is extracted from transverse mass, b-tagged jet multiplicity and H_T distributions by performing a binned maximum likelihood fit. The upper limit on charged Higgs boson production with branching fraction is set by combining the three final states as shown in figure 4.33.

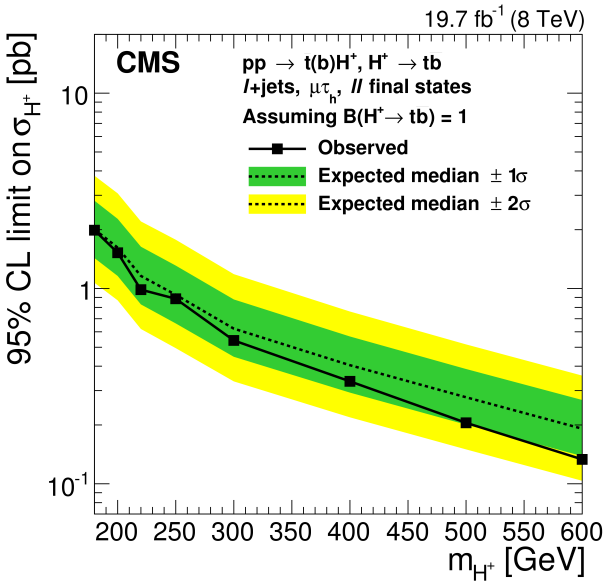


FIGURE 4.33: Expected and observed 95 % CL upper limits on $\sigma(\mathrm{pp}\rightarrow \bar{t}(b)H^{+})$ for the combination of the $\mu\tau_h$, l+jets and dilepton final states. The region above the solid line is excluded.

Chapter 5

Search for the production of Higgs boson associated with a pair of top quarks in multilepton final states

5.1 Introduction

There is an increasing interest in the measurements of the Higgs properties, such as its spin, parity and coupling to other particles, after the discovery in 2012 [1, 2, 3]. The couplings between the Higgs boson and other Standard Model (SM) particles are well predicted from the theory, and their measurements constitute a good test for new physics beyond SM. In the SM, the coupling strength of fermions to the Higgs boson is proportional to their mass. Hence the top quark has the strongest coupling, which is also a key parameter of the SM, among all the SM particles. The detailed discussion of theory is presented in section 1. Experimentally, the coupling can be determined from the cross section of $gg \to H$ production through a top quark loop, or from the cross section of $t\bar{t}H$ production which is a tree-level process (figure 5.1) at the lowest order in perturbation theory. The coupling is expected to be around the unity and sensitive to potential new physics effects which can modify the value predicted by the SM.

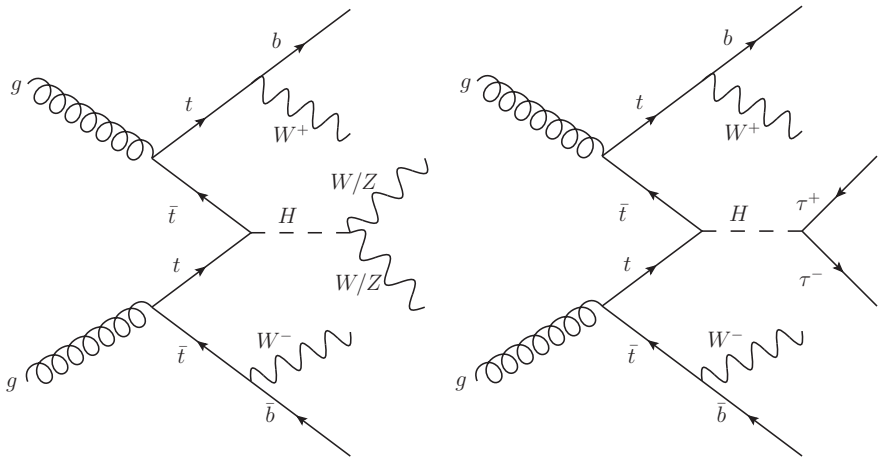


FIGURE 5.1: Examples of tree-level Feynman diagrams for the production of the Higgs boson in association with a pair of top quarks. Higgs boson decays to WW/ZZ (left) or $\tau\tau$ (right) are shown.

The $t\bar{t}H$ production can be observed through different event topologies according to the decay products of top quarks and the Higgs boson. So far, several different analyses concerning various Higgs decay modes have been performed, such as $H \to b\bar{b}$, $H \to \gamma\gamma$ and so on. In terms of multilepton final states, it mainly targets Higgs decays to WW^*, $\tau\tau$ and ZZ^* while top quarks decay either leptonically or hadronically. It is relatively statistically limited but can still get good sensitivity since multiple leptons can help to suppress backgrounds from other SM processes. According to the number of reconstructed leptons (e, μ) and hadronic τs, there are different sub-channels in this analysis targeting different Higgs decay modes, as shown in figure 5.2.

In this chapter, first an overview of previous $t\bar{t}H$ multilepton results is presented, followed by the analysis strategy. It is concluded by a detailed discussion of the sub-channel I have been involved in, 2ℓSS+$1\tau_{\text{had}}$ (the final state including two same-sign light leptons (electrons or muons) and one hadronic tau). Fianlly the results of combined fit and comparison to the CMS results will be presented. Only the three most powerful final state 2ℓSS, 3ℓ, 2ℓSS+$1\tau_{\text{had}}$ will be described in detail.

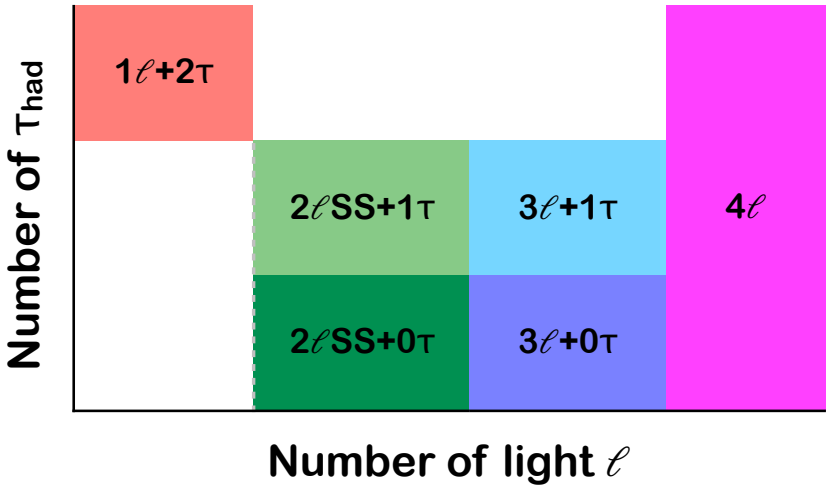


FIGURE 5.2: $t\bar{t}H$ multilepton channels in terms of light lepton (electrons or muons) multiplicity and hadronic tau multiplicity. SS means same-sign.

5.2 Personal contributions

I contributed to define the 2ℓSS+$1\tau_{\text{had}}$ optimal signal region by optimizing the joint sensitivity of 2ℓSS and 2ℓSS+$1\tau_{\text{had}}$. The signal region definitions of the ttH analysis are discussed in section 5.3.2. I produced input data in which I computed reconstructed variables for another group to optimize their MVA on. These reconstructed high level variables and the MVA analysis using it are described in section 5.4.4.1. With the same reconstructed variables I implemented a new robust categorisation, described in section 5.4.2. I compared three different methods for the reducible background

Process	Generator (alternative)	Parton Shower (alternative)	PDF	Tune
$t\bar{t}H$	POWHEG-BOX [104]	PYTHIA 8	NNPDF 3.0 NLO [51]/ NNPDF 2.3 LO [46]	A14
	(-)	(HERWIG++)		
$tHqb$	MG5_AMC	PYTHIA 8	CT10 [105]	A14
tHW	MG5_AMC	HERWIG++	CT10 /CTEQ6L1 [81, 107]	UE-EE-5 [106]
$t\bar{t}W$	SHERPA 2.2.1 [108] (MG5_AMC)	SHERPA 2.2.1 (PYTHIA 8)	NNPDF 3.0 NNLO	SHERPA default
$t\bar{t}(Z/\gamma^*)$	MG5_AMC	PYTHIA 8	NNPDF 3.0 NLO /2.3 LO	A14
	(SHERPA)	(SHERPA)		
$t(Z/\gamma^*)$	MG5_AMC	PYTHIA 8	CTEQ6L1	Perugia2012 [76]
$tW(Z/\gamma^*)$	MG5_AMC	PYTHIA 8	NNPDF 2.3 LO	A14
$t\bar{t}t$, $t\bar{t}t\bar{t}$	MG5_AMC	PYTHIA 8	NNPDF 2.3 LO	A14
$t\bar{t}W^+W^-$	MG5_AMC	PYTHIA 8	NNPDF 2.3 LO	A14
$t\bar{t}$	POWHEG-BOX [104]	PYTHIA 8	CT10/CTEQ6L1	Perugia2012
$t\bar{t}\gamma$	MG5_AMC	PYTHIA 8	NNPDF 2.3 LO	A14
s-, t-channel, Wt single top	POWHEG-BOX [109, 110]	PYTHIA 6	CT10 /CTEQ6L1	Perugia2012
VV, $qqVV$, VVV	SHERPA 2.2.2 [108]	SHERPA	NNPDF 3.0 NNLO	SHERPA default
$Z \to \ell^+\ell^-$	SHERPA 2.2	SHERPA	NNPDF 3.0 NLO	SHERPA default

TABLE 5.1: The configurations used for event generation of signal and background processes. If only one parton distribution function (PDF) is shown, the same one is used for both the matrix element (ME) and parton shower generators; if two are shown, the first is used for the matrix element calculation and the second for the parton shower. "V" refers to production of an electroweak boson (W or Z/γ^*). "Tune" refers to the underlying-event tune of the parton shower generator. "MG5_AMC" refers to MADGRAPH5_AMC@NLO 2.2.1 [44]; "PYTHIA 6" refers to version 6.427 [111]; "PYTHIA 8" refers to version 8.2 [112]; "HERWIG++" refers to version 2.7 [80]. Samples using PYTHIA 6 or PYTHIA 8 have heavy flavour hadron decays modelled by EVTGEN 1.2.0 [60]. All samples include leading-logarithm photon emission, either modelled by the parton shower generator or by PHOTOS [113].

estimates : the template fit, Matrix method and fake factor,described in sections 5.4.3. I performed the full statistical analysis for the channel in order to chose the most sensitive and robust to systematics anlysis, which is described in section 5.4.4. In order to be thorough for the reader I also describe other channels in section 5.3.2.

5.3 Analysis strategy of $t\bar{t}H$ multilepton with 80 fb^{-1} Data

5.3.1 Data and Monte Carlo samples

This analysis uses 80 fb^{-1} of data collected from proton-proton collision recorded by the ATLAS detector at $\sqrt{s} = 13$ TeV during 2015-17. The simulated samples are listed in table 5.1, the details are basically the same as already explained in H^+ chapter 4.3.1.

5.3.2 Signal region definition and signal extraction

The sub-channels in $t\bar{t}H$ multilepton are determined in terms of the loosely identified light lepton (e/μ) multiplicity and the τ_{had} multiplicity. Therefore all these sub-channels have no overlap event between each other, which allows for an easy combination to gain sensitivy. Signal region of each sub-channel is optimized separately to get maximum significance. For sub-channels with high statistics, multivariate analysis method is employed to extract signal from backgrounds while for low statistic sub-channels, 2ℓSS+$1\tau_{\text{had}}$, 3ℓ+$1\tau_{\text{had}}$, 4ℓ where expected yields are lower than 30, only cut-and-count analysis is performed. The basic information is summarized in table 5.2.

	Non-tau channels			Tau channels		
	2ℓSS	3ℓ	4ℓ	1ℓ+$2\tau_{\text{had}}$	2ℓSS+$1\tau_{\text{had}}$	3ℓ+$1\tau_{\text{had}}$
Light lepton	2T	1L*, 2T	4L*	1L*	2T	1L*, 2T
τ_{had}	0M	0M	-	2T	1M	1M
N_{jets}, $N_{b-\text{jets}}$	$\geq 4, \geq 1$	$\geq 2, \geq 1$	$\geq 2, \geq 1$	$\geq 3, \geq 1$	$\geq 4, \geq 1$	$\geq 2, \geq 1$
Non-prompt lepton strategy	semi-DD (TF)	semi-DD (TF)	semi-DD (SF)	MC	semi-DD (TF)	MC
Fake tau strategy	–	–	–	DD (SS data)	semi-DD (SF)	semi-DD (SF)
BDT trained against	Fakes and $t\bar{t}V$	$t\bar{t}$, $t\bar{t}W$, $t\bar{t}Z$, VV	$t\bar{t}Z$ / -	$t\bar{t}$	-	-
Discriminant	2D BDT	5D BDT	Event count	BDT	Event count	Event count
Number of bins in SR	4	5	1 / 1	3	1	1
Control regions	6	4	-	-	-	-

TABLE 5.2: Summary of basic characteristics and strategies of the six analysis channels. In lepton selection, *T* stands for Tight lepton definition, *L* stands for Loose lepton definition, and *L** stands for Loose lepton definition with an additional requirement to pass "FixedCutLoose" isolation. For the fake lepton and tau background estimates, *DD* means data-driven, from which *TF* is the template fit method and *SF* refers to the fake scale factor method.

5.3.3 MVA analysis

To distinguish signal from background processes, MVA methods (see section 4.3.4.1) have been developed in 2ℓSS, 3ℓ and 1ℓ+$2\tau_{\text{had}}$ channels. For the 4ℓ channel, a BDT is trained to help defining the signal region (SR). The BDT of 1ℓ+$2\tau_{\text{had}}$ is trained based on jets and tau related variables, to separate signal (mainly $H \to \tau\tau$) from $t\bar{t}$ backgrounds (with one or two fake τ_{had}).

5.3.3.1 2ℓSS 2D analysis strategy

Two BDTs are trained to further reject events with fake/non-prompt leptons and charge flip leptons (BDT_{ttbar}), as well as $t\bar{t}W$ (BDT_{ttV}). Nine input variables are used:

- Number of jets with $p_T > 25$ GeV, N_{jets};
- Number of b-jets with $p_T > 25$ GeV, $N_{b\text{-jets}}$, tagged with 70% b-tag efficiency;
- Leptonic flavour, ee, $e\mu$ μe and $\mu\mu$ where the first lepton corresponds to the one with the highest p_T;
- Distance between leading lepton and its closest jet, $\Delta R(\ell_0, \text{jet})$;
- Distance between sub-leading lepton and its closest jet, $\Delta R(\ell_1, \text{jet})$;

- Maximum eta between lepton $|\eta_{\ell_0}|$ and $|\eta_{\ell_1}|$, Max($|\eta_{\ell}|$);
- Sub-leading lepton p_T, $p_T(\ell_1)$;
- Missing transverse energy, E_T^{miss}
- Distance between the two same-sign leptons, $\Delta R(\ell_0, \ell_1)$;

In order to maximise the discriminant power against the $t\bar{t}W$ and $t\bar{t}$ backgrounds, a categorization procedure is developed in the two dimensional BDT plane. The categories are defined as following:

- SR: BDT$_{t\bar{t}} > 0$ and BDT$_{t\bar{t}V} > 0$
- ttW CR: BDT$_{t\bar{t}} > 0$ and BDT$_{t\bar{t}V} < 0$
- ttbar CR: BDT$_{t\bar{t}} < 0$, split by lepton flavor $ee + \mu e$ and $\mu\mu + e\mu$

In the SR, the remaining BDT$_{t\bar{t}V}$ shape, binned into 3 bins, is used for the final fit.

5.3.3.2 3ℓ multiclass BDT

Events passing the pre-selection are trained using a five-dimensional BDT (signal, $t\bar{t}W$, fakes, $t\bar{t}Z$, dibosons) with the xgboost package [114]. There are 26 variables used as input for the training, based on topological aspects of the events. As a result, a five dimensional multiclass discriminant is formed and categorized into five categories in which each of the trained processes has highest purity. Therefore the signal enriched category is the SR and the other four categories are regarded as CRs of corresponding process.

5.3.4 Background estimate

Backgrounds events can be categorized as irreducible backgrounds and reducible backgrounds. Irreducible backgrounds are events with the same number of prompt leptons as the $t\bar{t}H$ signal, such as $t\bar{t}W$, $t\bar{t}Z$, diboson. The estimate of irreducible backgrounds basically relies on simulated samples. Reducible backgrounds contain at least one charge-flip electron or one fake lepton or one fake hadronic tau depending on the channel. These backgrounds mainly arise from $t\bar{t}$ events. The light lepton fake estimate procedure is unified between the three channels where this background is relevant and important - 2ℓSS, 2ℓSS+$1\tau_{\text{had}}$ and 3ℓ. The fake hadronic tau backgrounds in 2ℓSS+$1\tau_{\text{had}}$ and 3ℓ+$1\tau_{\text{had}}$ are estimated from a dedicated 2ℓOS+$1\tau_{\text{had}}$ fake tau control region. The charge flip (QMisID) background in 2ℓSS and 2ℓSS+$1\tau_{\text{had}}$ is estimated by a data-driven method from $Z \to e^+e^-$ events. Fake tau and charge flip background will be discussed in detail later in section 5.4.3.

5.3.4.1 Template fit

The template fit method is employed to estimate fake light leptons in all 0 tau and 2ℓSS+$1\tau_{\text{had}}$ analyses. The fake lepton background is a mixture of leptons from semi-leptonic heavy flavor decay and photon conversions. The normalization of the different fake contribution templates, as given

by the MC of all processes contributing to non-prompt lepton background, are left free-floating in a fit to data, and these normalization factors are used to correct the fakes from the MC estimates. The template fit method is a semi-data-driven method, i.e. it relies on the truth information from $t\bar{t}$ and $t\bar{t}\gamma$ MC samples to define different types of fake leptons, and on the general description of fake kinematics by MC, but gets each type of fake leptons normalization from data.

The main contribution to non-prompt lepton background comes from $t\bar{t}$, followed with a much smaller contribution by $V + jets$ and single top. Based on the truth classification of events containing a fake lepton, the following main contributions are distinguished, and a free-floating normalization factor (NF) is assigned to each of them:

- $NF_e^{externalCO}$: normalization factor applied to events with one fake electron from external/material photon conversion.
- NF_e^{HF}: normalization factor applied to events with one non-prompt electron from B decay, C decay or light hadron (dominated by B decay).
- NF_μ^{HF}: normalization factor applied to events with one non-prompt muon from B decay, C decay or light hadron (dominated by B decay).

A non-negligible contribution from internal conversions ($\gamma* \rightarrow \ell\ell$) is also predicted by MC in the control and signal regions. Since the normalization of this background might not be correct from MC, an additional normalization factor, $NF^{internalCO}$, is introduced in the template fit.

The classification of $t\bar{t}$ and $t\bar{t}\gamma$ Monte Carlo samples in the aforementioned categories is based on their truth origin as follows:

- Prompt leptons: leptons from Top, Bremsstrahlung radiation or rare Top decay.
- Conversion: Conversion photon fakes to electron
  - Internal Conversion: electron with decay radius below 20 mm
  - External Conversion: electron with decay radius larger than 20 mm
- B decay: non-prompt leptons from B decay
- C decay: non-prompt leptons from C decay
- Other decay: leptons from light quarks or other processes.

In addition, $t\bar{t}$ and $t\bar{t}\gamma$ Monte Carlo events containing a charge-flip electron are vetoed, since they are estimated with dedicated data-driven methods. Due to the large contamination in all control and signal regions from the $t\bar{t}W$ background, and since its cross section has been generally measured higher than the expectation, its normalisation is also left free-floating in the fit with the normalisation factor $NF_{t\bar{t}W}$.

There are 17 regions used in the template fit, as described in section 5.3.2. For the 2ℓSS channel, two more CRs are defined in addition to the 4 CRs in the pre-MVA region, requiring only 2 or 3 jets while keeping other selections. In order to improve the constraint on the NFs of the internal and

external conversions and decrease the correlation between these two, dedicated internal conversion and external conversion CRs are defined based on the electron properties: the conversion radius, the invariant mass of the track associated to the electron and its closest track (originating from the conversion) calculated at conversion vertex ($m_{trk-trk,CV}$), and the same invariant mass calculated at primary vertex ($m_{trk-trk,PV}$). Figure 5.3 shows the final region setup in the template fit.

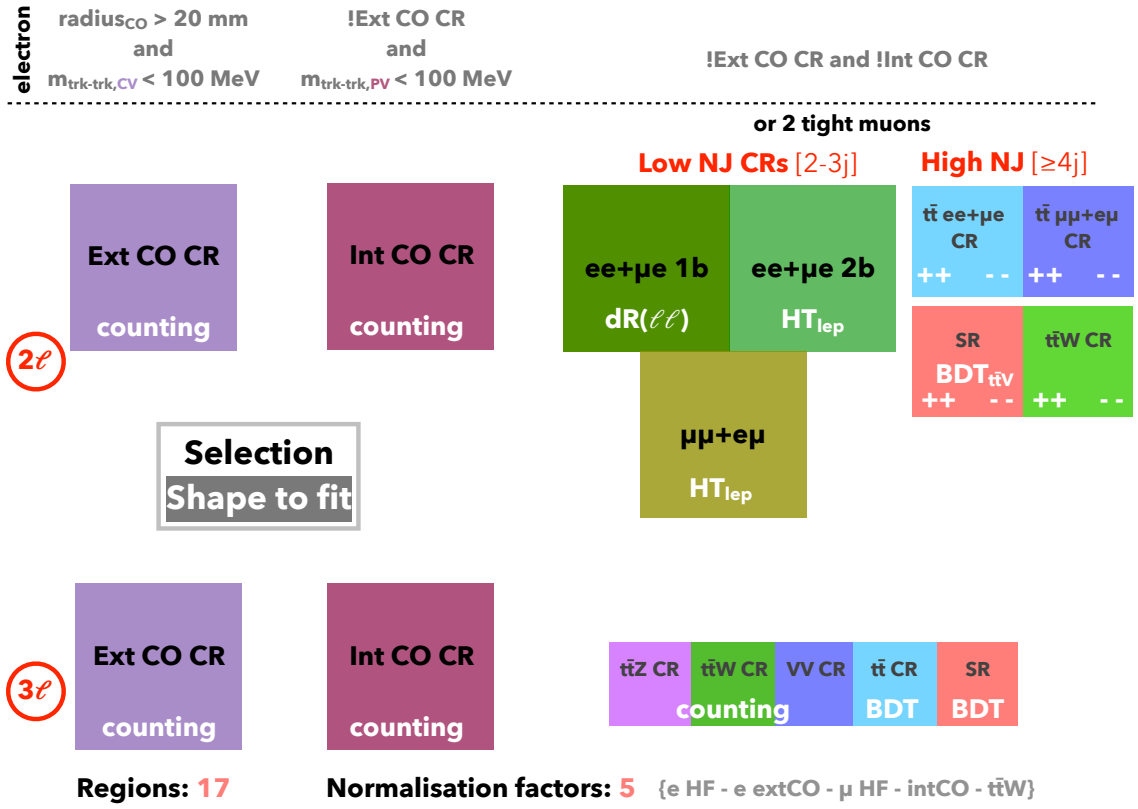


FIGURE 5.3: Final fit setup tested in the implementation of the template fit fakes method.

Before unblinding, the template fit is only performed in CRs to extract the NFs, and then the NFs are inserted into SRs to estimate the expected sensitivity. While after unblinding the data, the template fit and sensitivity estimate are performed in one simultaneous fit. The fitted NFs are summarized in table 5.3.

Parameter	$NF_{t\bar{t}W}$	NF_e^{ExtCO}	NF_e^{IntCO}	NF_e^{HF}	NF_μ^{HF}
Stat-Only	1.51 ± 0.12	1.67 ± 0.45	0.75 ± 0.25	1.11 ± 0.27	1.31 ± 0.14
Syst.+ Stat.	1.44 ± 0.18	1.70 ± 0.51	0.75 ± 0.26	1.09 ± 0.32	1.28 ± 0.17

TABLE 5.3: Fitted values of the 5 NFs from the simultaneous fit including all available CRs in 2ℓSS and 3ℓ channels: first row with statistic-only errors, second row with statistic errors and all other systematic error except ttW instrumental systematics.

Since the template fit relies on MC distributions, there are mainly three types of systematics associated:

- $t\bar{t}$ modelling

- $t\bar{t}W$ modelling
- shape systematics on the fake template

The $t\bar{t}$ modelling systematics follow the common Top group procedure, estimated by using samples with various amount of radiation and scale choice. They have very small impacts on the result. The $t\bar{t}W$ modelling systematics are obtained with the scale variation of the nominal SHERPA2.2.1 multileg NLO 0,1j@NLO+2j@LO sample, plus the comparison to AMC@NLO+PYTHIA8(0j@NLO). The shape systematics on the fake template are derived by loosening requirements on the lepton isolation to enrich photon conversion and heavy flavors separately.

5.4 Study of 2ℓSS+1$\tau_{\mathbf{had}}$ final state

The general strategy and result have been presented in previous sections. The following contents will focus on my contributions in 2ℓSS+1τ_{had} final state.

5.4.1 Event selection

The basic definition of tight signal region for 2ℓSS+1τ_{had} is as follows:

- Passing di-lepton triggers
- Trigger objects matching with offline leptons
- Two same sign, tight identified leptons with $p_T \geq 20\ GeV$
- Exactly one medium identified τ_{had} candidate originating from primary vertex with $p_T \geq 25\ GeV$, and charge opposite to light leptons
- At least 4 jets with $p_T \geq 25\ GeV$, at least one of them is b-tagged at 70 % working point

To suppress fake electron arising from photon conversions, additional requirements on leptons, unified to 2ℓSS channel, are applied:

- Pseudorapidity of electrons must be smaller than 2, $|\eta_{electron}| < 2.0$
- Invariant mass of two light leptons must be larger than 12 GeV, $M_{l,l} > 12GeV$
- Angular distance of two light leptons must be larger than 0.5 for events with at least one electron, $\Delta R_{l,l} > 0.5$

The final fit is performed in the region with all above selections taken into account. A prefit plots of the SR, fake leptons estimated from $t\bar{t}$ simulation samples, can be found in figure 5.4.

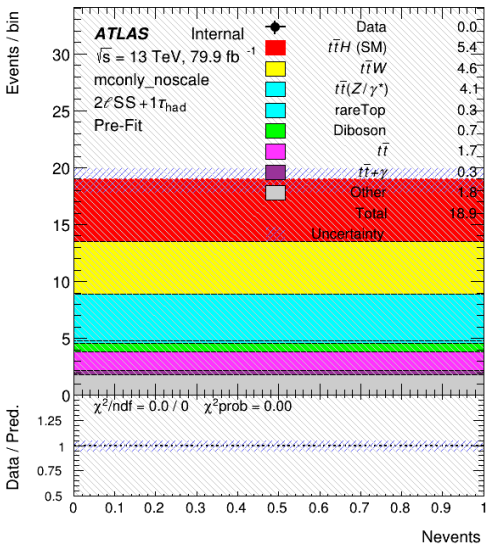


FIGURE 5.4: Prefit distributions with MC fakes in the signal region.

There is an additional control region to validate the analysis, namely "low nJets CR", defined as follows:

- Passing di-lepton triggers
- Trigger objects matching with offline leptons
- Two same sign, tight identified leptons with $p_T \geq 20\ GeV$
- Exactly one medium identified τ_{had} candidate originating from primary vertex with $p_T \geq 25\ GeV$, and charge opposite to light leptons
- Contains 2 or 3 jets with $p_T \geq 25\ GeV$, at least one of them is b-tagged at 70 % working point
- $|\eta_{electron}| < 2.0$
- $M_{l,l} > 12GeV$
- $\Delta R_{l,l} > 0.5$ for events with at least one electron

The selection is basically the same as the tight signal region, but loosen the jet multiplicity to 2-3.

5.4.2 Categories analysis optimization

In 2ℓSS+$1\tau_{\text{had}}$ analysis, the default approach to extract the signal in previous results was the use of BDT (see section **??**). Given the fact that this channel is statistically limited, a simpler approach will be a good validation as well as a backup option. This motivates a cut-and-count analysis, namely categorization.

In the categorization analysis, the whole signal region (introduced in section 5.4.1) is categorized into several smaller regions (categories) by applying cuts on discriminating variables (some of them are also BDT input variables), and then a simultaneous fit on the events counts across all the categories is performed to extract the signal strength. The motivation of this study is to get a better sensitivity with respect to the inclusive cut-and-count analysis while being more robust than the multivariate analysis. All the following optimizations are done taking into account only inclusive theoretical systematics (listed in table 5.4).

The optimization is split into several steps:

1. Choose variables and rank them
2. Attempt to combine two, three or four variables to build categories
3. Choose the best (in terms of both sensitivy and robustness) combination
4. Merge low statistic categories to avoid large fluctuations
5. Check overtraining

First of all, a simple scanning of 20 variables, listed in table 5.5, with 10 cut points for each is done separately. These variables are built based on the features of signal topology, so that they can distinguish signal from backgrounds. As an example, figure 5.5 shows the distributions of three variables in the SR. The significance which is then computed from Asimov fits against $t\bar{t}V$ only backgrounds is employed to rank the variables and the corresponding cut points. An example is shown in figure 5.6. This gives the sensitivity of each variable and meanwhile ensures the smoothness of cuts. Ten variables are selected for next step of optimization, listed in table 5.6.

Next, every combination of 2, 3 or 4 variables is used to build categories and again all of them are tested by Asimov fits against $t\bar{t}V$ only backgrounds to get the best combination. For each combination of variables, 3 cut values are tested for each variable, as shown in figure 5.7 for example. Using more cut points for each variable for optimization consumes too much time. The cut value optimization allows to take into account, though in a limited discrete way, possible correlations between the 3 chosen variables. In the optimization, combinations of 2 variables are found to be suboptimal, while combinations of 4 variables are found to add limited sensitivity for an additional complication. The variables chosen as a result to the optimization are:

- Mll: Invariant mass of the two same sign leptons
- TM: Transverse mass of the leading lepton and the missing transverse energy
- DR: ΔR between the leading lepton and the closest jet.

Systematic	Value
ttbar cross-section	50%
ttgamma cross-section	50%
3top cross-section	50%
4top QCD scale	30.8%/-25.6%
4top PDFunc	5.5%/-5.9%
VV cross-section	50%
ttWW QCD scale	10.9%/-11.8%
ttWW PDFunc	2.10%
ttZ QCD scale	9.6%/-11.3%
ttZ PDFunc	4%
ttW QCD scale	12.9%/-11.5%
ttW PDFunc	3.40%
tZ cross-section	50%
WtZ cross-section	50%
rareTop cross-section	50%
tHjb QCD scale	6.5%/-14.7%
tHjb PDFunc	3.7%
WtH QCD scale	4.9%/-6.7%
WtH PDFunc	6.3%
ttH QCD scale	5.8%/-9.2%
ttH PDFunc	3.60%

TABLE 5.4: Systematics included in the optimisation

Events from the SR are then split into whether they are above or below the optimum cut value which leads to $2^3 = 8$ categories. Given the fact that 2ℓSS+$1\tau_{had}$ is still a rather low statistic channel, and also for simplicity reasons, those 8 categories are further merged into only 3 categories, depending on whether they show a high purity ($>$ 55 %), medium purity (20 % to 55 %) or a low purity (rest category). Then the 3 categories are defined as such:

- High purity category: Mll $\leq$ 135 GeV&&TM $\leq$ 115 GeV&&DR $\leq$ 1.25||Mll $\leq$ 135 GeV&&TM $>$ 115 GeV&&DR $\leq$ 1.25
- Medium purity category: Mll $\leq$ 135 GeV&&TM $\leq$ 115 GeV&&DR $>$ 1.25||Mll $>$ 135 GeV&&TM $\leq$ 115 GeV&&DR $\leq$ 1.25
- Rest category: All the rest.

The signal purity (S/B) of these three categories are presented in figure 5.8.

Variable Name	Meaning
Mll01	Invariant mass of 2 leading leptons
HT_lep	Sum of Pt of all leptons
TransverseMassLeadLepMET	Transverse mass of leading lepton and missing transverse energy
DRll01	Angular distance of 2 leading leptons
MvisnonH	Visible mass between τ_{had} and furthest lepton
lep_Pt_0	leading lepton pT
MvisH	Visible mass between τ_{had} and closest lepton
DeltaRLeadLepClosestJet	Angular distance between leading lepton and closest jet
lep_Pt_1	Transverse momentum of subleading lepton
max_eta	maximum eta of light leptons
DeltaRSubLepClosestJet	Angular distance between subleading lepton and closest jet
Mjj_closeW	Invariant mass of two light jets which are close to the W boson
nJets_OR_T	Number of jets
LD_HTmiss_ETmiss	Sum of missing transverse energy and Sum of Pt of all jets: 0.6*ETmiss + 0.4*HTmiss
MtopW	Invariant mass of the two leading non b-tag jets which invariant mass is closest to mass of W boson and the leading b-tag
tempDP2l	$\Delta\phi$ between leading lepton and subleading lepton
lep_Eta_1	Pseudorapidity of subleading lepton
HT	Sum of Pt of all objects
jet2_eta	Pseudorapidity of 3rd jet
jet3_eta	Pseudorapidity of 4th jet

TABLE 5.5: Input variables for the categorization

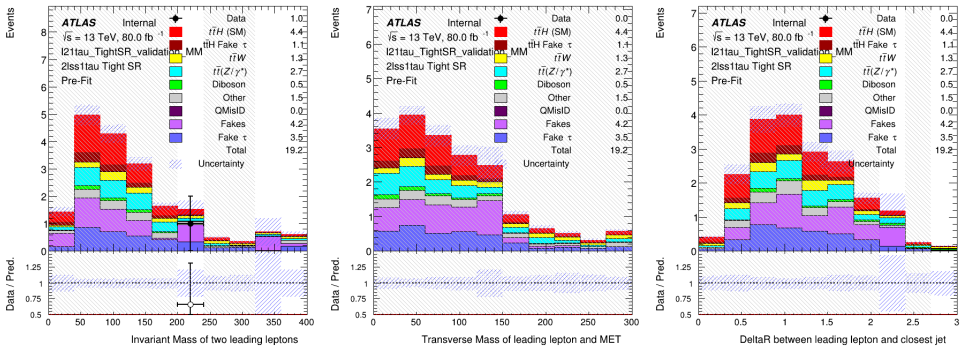


FIGURE 5.5: Distributions of the three variables which are chosen to build the final categorisation in the SR. Invariant mass of 2 leading leptons(left), Transverse mass of leading lepton and missing transverse energy (middle), DeltaR between leading lepton and closest jet (right). Bins where S/B > 15 % are blinded.

The raw yields and weighted yields of events in each category are summarized in table 5.7 and table 5.8, respectively. No obvious fluctuation is observed in the categorization setup. The

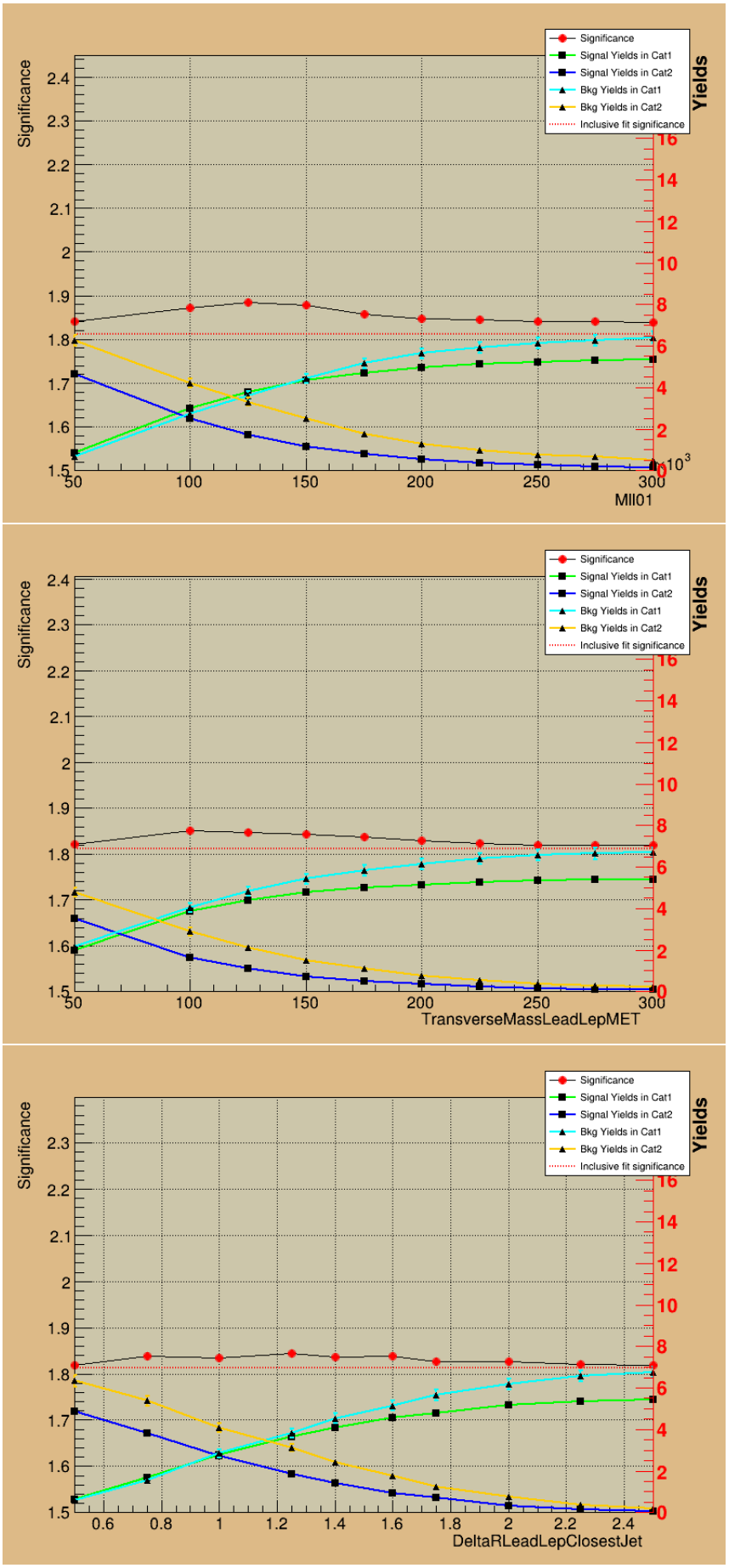


FIGURE 5.6: Results of 1-D scanning shown for 3 variables which are chosen to build the final categorisation. Invariant mass of 2 leading leptons(first), Transverse mass of leading lepton and missing transverse energy (second), DeltaR between leading lepton and closest jet (third).

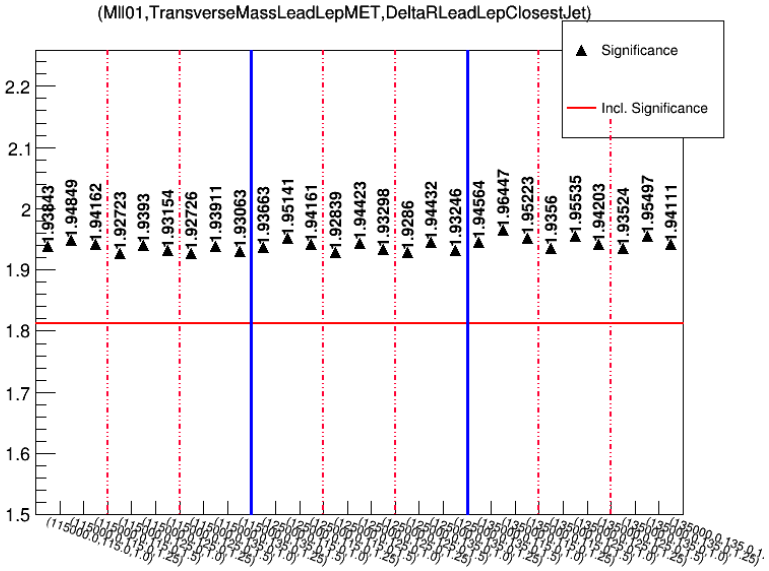


FIGURE 5.7: Results of the 3-D scanning of the variables which are chosen to build the final categorisation. The blue line splits the whole range according to the value of Mll01(left to right, 115 GeV, 125 GeV and 135 GeV). Then The red dash line split each smaller range according to the value of TransverseMassLeadLepMET (left to right, 115 GeV, 125 GeV and 135 GeV). Finally the three bins in each region defined by red dash lines represent the value of DeltaRLeadLepClosestJet (from left to right, 1, 1.25, 1.5).

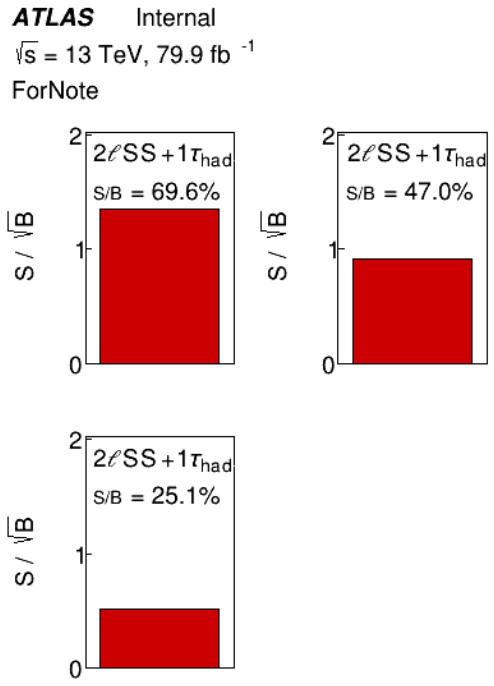


FIGURE 5.8: Signal purity (S/B) in High purity category (top left), Medium purity category (top right) and Rest category (bottom left).

ID	Name	Cut 1	Cut 2	Cut 3
1	Mll01	115 GeV	125 GeV	135 GeV
2	TransverseMassLeadLepMET	115 GeV	125 GeV	135 GeV
3	MvisH	175 GeV	200 GeV	225 GeV
4	HT_lep	100 GeV	125 GeV	150 GeV
5	DRll01	1.25	1.5	1.75
6	lep_Pt_0	85 GEV	100 GeV	115 GeV
7	MvisnonH	175 GeV	200 GeV	225 GeV
8	max_eta	1.25	1.4	1.65
9	lep_Pt_1	40 GeV	50 GeV	60 GeV
10	DeltaRLeadLepClosestJet	1	1.25	1.5

TABLE 5.6: Top 10 ranking variables after the first procedure, 1-D scanning. To save computing time, only three cut points of each variable is used for further optimization instead of ten cut points in the 1-D scanning. The explanation of variables can be found in table 5.5.

comparisons of categorization fit to inclusive signal region fit and to BDT fit in several conditions are presented in table 5.9. The non-prompt lepton backgrounds are estimated from MC simulation ($t\bar{t}$ and $t\bar{t}\gamma$).

	$2\ell SS+1\tau_{had}HighPurity$	$2\ell SS+1\tau_{had}MediumPurity$	$2\ell SS+1\tau_{had}Rest$
$t\bar{t}W$	284 ± 40.1	269 ± 38.2	405 ± 55.7
$t\bar{t}Z/\gamma$	1170 ± 138	1090 ± 128	1000 ± 118
$t\bar{t}lllowmass$	25 ± 5.75	10 ± 3.36	20 ± 5.02
rareTop	24 ± 13.0	13 ± 7.44	2 ± 1.73
VV	80 ± 41.0	132 ± 67.0	83 ± 42.5
Three top	88 ± 45.0	77 ± 39.5	161 ± 81.6
Four top	119 ± 36.0	87 ± 26.8	146 ± 43.8
$t\bar{t}WW$	16 ± 4.42	14 ± 4.09	22 ± 5.35
tZ	6 ± 3.87	8 ± 4.90	6 ± 3.87
WtZ	26 ± 14.0	35 ± 18.5	36 ± 19.0
$tHjb$	1 ± 1.01	8 ± 2.97	2 ± 1.43
WtH	35 ± 6.67	37 ± 6.90	29 ± 5.96
ttbar	3 ± 2.29	2 ± 1.73	4 ± 2.83
ttgamma	1 ± 1.12	1 ± 1.12	1 ± 1.12
$t\bar{t}H$ (SM)	2770 ± 0	1700 ± 0	1210 ± 0
Total	4650 ± 43.4	3480 ± 42.2	3130 ± 43.8

TABLE 5.7: Raw yields in each category. "Raw yields" simply counts the number of events (one event counts 1). The optimization was only made against ttV but applied to all processes.

	2ℓSS+$1\tau_{\text{had}}$ *HighPurity*	2ℓSS+$1\tau_{\text{had}}$ *MediumPurity*	2ℓSS+$1\tau_{\text{had}}$ *Rest*
$t\bar{t}W$	0.680 ± 0.143	0.868 ± 0.146	1.30 ± 0.177
$t\bar{t}Z/\gamma$	1.27 ± 0.174	1.46 ± 0.163	1.24 ± 0.151
$t\bar{t}lllowmass$	0.0395 ± 0.0288	0.0113 ± 0.0171	0.0528 ± 0.0243
rareTop	0.169 ± 0.0504	0.0785 ± 0.0326	0.0138 ± 0.0138
VV	0.183 ± 0.0393	0.350 ± 0.0579	0.199 ± 0.0522
Three top	$0.0260 \pm 0.004\,09$	$0.0236 \pm 0.003\,97$	$0.0492 \pm 0.005\,72$
Four top	0.197 ± 0.0269	0.142 ± 0.0226	0.252 ± 0.0305
$t\bar{t}WW$	0.105 ± 0.0390	0.102 ± 0.0398	0.158 ± 0.0495
tZ	0.0461 ± 0.0274	0.0726 ± 0.0369	0.0542 ± 0.0336
WtZ	0.273 ± 0.0778	0.163 ± 0.0805	0.170 ± 0.0810
$tHjb$	$0.001\,75 \pm 0.002\,47$	$0.0142 \pm 0.007\,14$	$0.004\,09 \pm 0.004\,11$
WtH	0.0397 ± 0.0155	0.0552 ± 0.0166	0.0423 ± 0.0140
ttbar	0.675 ± 0.567	0.447 ± 0.447	0.563 ± 0.501
ttgamma	0.0325 ± 0.0459	0.0280 ± 0.0396	0.0291 ± 0.0412
$t\bar{t}H$ (SM)	2.60 ± 0	1.79 ± 0	1.04 ± 0
Total	6.34 ± 0.623	5.60 ± 0.514	5.16 ± 0.567

TABLE 5.8: Weighted yields in each category. "Weighted yields" takes into account the event weights those are introduced to correct the simulated distribution. The optimization was only made against ttV but applied to all processes.

		Expected Sig (σ)	Improvement (%)	BDT w.r.t Cat (%)
Inclusive SR	ttV stat only	1.855	-	-
	ttV syst	1.80	-	-
	full bkg stat only	1.45	-	-
	full bkg syst	1.365	-	-
Categorisation	ttV stat only	1.99	7.28	-
	ttV syst	1.94	7.78	-
	full bkg stat only	1.53	5.52	-
	full bkg syst	1.45	6.23	-
6-variable BDT	ttV stat only	2.03	9.43	2.01
	ttV syst	1.98	10	2.06
	full bkg stat only	1.53	5.52	0
	full bkg syst	1.44	5.49	-0.69

TABLE 5.9: Improvement in terms of sensitivity with respect to inclusive signal region fit and comparing with BDT fit. "ttV" means using only $t\bar{t}V$ events as backgrounds in the fit while "full bkg" means using all available simulated samples as backgrounds. "syst" means taking into account the inclusive theoretical systematic (listed in table 5.4) while "stat only" means that no systematic uncertainty is taken into account in the fit.

Ranking	(Var1, Var2, Var3)	(Cut1, Cut2, Cut3)	Best sensitivity (σ)
1	(TransverseMassLeadLepMET,DRll01,DeltaRLeadLepClosestJet)	(115.0,1.5,1.25)	1.68707
2	(Mll01,TransverseMassLeadLepMET,DeltaRLeadLepClosestJet)	(135000.0,115.0,1.25)	1.68415
3	(TransverseMassLeadLepMET,DRll01,MvisnonH)	(115.0,1.5,200000.0)	1.67641
4	(TransverseMassLeadLepMET,DRll01,max_eta)	(115.0,1.5,1.25)	1.67343
5	(DRll01,MvisnonH,DeltaRLeadLepClosestJet)	(1.5,200000.0,1.25)	1.67136
6	(TransverseMassLeadLepMET,MvisnonH,DeltaRLeadLepClosestJet)	(115.0,200000.0,1.25)	1.66908
7	(HT_lep,DRll01,DeltaRLeadLepClosestJet)	(125000.0,1.5,1.25)	1.66882
8	(Mll01,DRll01,DeltaRLeadLepClosestJet)	(135000.0,1.25,1.25)	1.66856
9	(DRll01,lep_Pt_1,DeltaRLeadLepClosestJet)	(1.25,40000.0,1.0)	1.66684
10	(MvisH,DRll01,DeltaRLeadLepClosestJet)	(175000.0,1.5,1.25)	1.66393

TABLE 5.10: The optimization use only events with even event number. The chosen combination ranks 2nd here.

Ranking	(Var1, Var2, Var3)	(Cut1, Cut2, Cut3)	Best sensitivity (σ)
1	(HT_lep,DRll01,DeltaRLeadLepClosestJet)	(125000.0,1.5,1.0)	1.64269
2	(Mll01,lep_Pt_0,lep_Pt_1)	(115000.0,85000.0,60000.0)	1.64201
3	(HT_lep,DRll01,lep_Pt_1)	(150000.0,1.25,60000.0)	1.63989
4	(Mll01,HT_lep,lep_Pt_0)	(125000.0,125000.0,85000.0)	1.63735
5	(Mll01,TransverseMassLeadLepMET,DeltaRLeadLepClosestJet)	(135000.0,115.0,1.0)	1.63314
6	(Mll01,DRll01,MvisnonH)	(115000.0,1.25,175000.0)	1.62989
7	(TransverseMassLeadLepMET,HT_lep,DRll01)	(115.0,125000.0,1.25)	1.62843
8	(Mll01,HT_lep,lep_Pt_1)	(115000.0,150000.0,60000.0)	1.62765
9	(TransverseMassLeadLepMET,DRll01,DeltaRLeadLepClosestJet)	(115.0,1.5,1.0)	1.62728
10	(Mll01,HT_lep,DeltaRLeadLepClosestJet)	(115000.0,100000.0,1.5)	1.62619

TABLE 5.11: The optimization use only events with odd event number. The chosen combination ranks 5th here.

5.4.2.1 Validation of categorization optimization

Just as for normal MVA methods, i.e. BDT, there is also a risk for the categorization to be over-trained, since the optimization uses the whole SR dataset and in the end apply to the same dataset. In order to ensure the robustness of the optimization procedure, a cross-check is performed by splitting the whole dataset into subsets. The split mimics the usual method widely used in the multivariate analysis, splitting by even or odd event number. The optimization procedure is repeated on two separate datasets, "odd-events" and "even-events". The results of the cross-check are summarized in table 5.10 and table 5.11. Considering only $t\bar{t}W$ and $t\bar{t}Z$ as backgrounds, the ranking of the selected combination shifts a bit, but still among the best combinations with very small difference in terms of sensitivity.

5.4.3 Non-prompt backgrounds

The $2\ell SS+1\tau_{had}$ final state suffers from both irreducible and reducible backgrounds, as shown in figure 5.9. The irreducible backgrounds basically follow the strategy of $t\bar{t}H$ multilepton analysis. In this part, fake lepton, including fake tau, charge flip and fake lepton estimates in the $2\ell SS+1\tau_{had}$ channel will be discussed in detail.

Due to the imperfections of the detectors, the physical objects are not always reconstructed correctly. As a result, leptons can be mis-reconstructed, namely non-prompt backgrounds. Non-prompt backgrounds consist of non-prompt light leptons (e, *mu*) (fake leptons), non-prompt tau leptons (fake tau) and electron charge flip (QmisID). Electrons, muons and taus all present in the $2\ell SS+1\tau_{had}$ final states, thus all types of non-prompt backgrounds are important in the analysis.

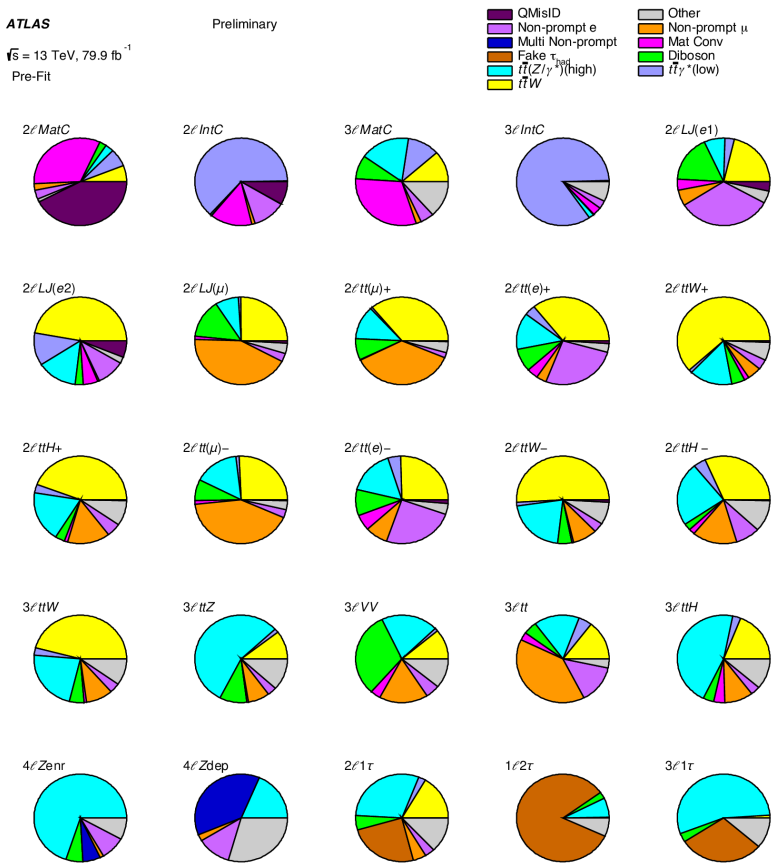


FIGURE 5.9: Fractional contributions of the various backgrounds to the total predicted background in each of the six analysis regions, including also the control regions. The piechart for 2ℓSS+$1\tau_{\text{had}}$ is at the third column of the fifth row. $t\bar{t}W$ is shown in yellow, $t\bar{t}Z$ in cyan, diboson in green.

Most of the fakes are originated from $t\bar{t}$ events, either with both fake lepton and fake tau or only fake lepton. Those non-prompt backgrounds are estimated separately, which will be discussed in the following sections.

5.4.3.1 Electron charge flip (QMisID)

There are two main processes contributing to QMisID:

- Hard Bremsstrahlung ($e^{\pm} \rightarrow e^{\pm}\gamma \rightarrow e^{\pm}e^{+}e^{-}$). In this case, EM cluster is coupled to the track of the opposite-sign electron in the trident. The QMisID rate is expected to be dependent on $|\eta|$ since the probability of this process depends on the transverse detector material. It is widely preponderant.
- Mismeasurement of the electron track-curvature. This effect is more important in the hight p_T range where the curvature is samller. Therefore the QMisID rate is expected to be dependent on p_T.

The QMisID rate is derived from data, based on the fraction of $Z \rightarrow ee$ events that are reconstructed as a same-sign electron pair.

5.4.3.2 Fake tau

Fake tau could arise from mis-reconstructed jets or electrons. Fake tau contribution is estimated from specified 2lOS+1τ_{had} control region, which is dominated by fake tau background. Despite two lepton signal regions which require same sign lepton pair, this fake tau control region requires opposite sign lepton pair, therefore $t\bar{t}$ events with one jet mis-reconstructed as fake τ_{had} contribute a lot to this CR, as shown in figure 5.10. A data-driven (DD) scale factor (SF) is derived by comparing the normalization of fake tau events in Monte Carlo (MC) to data in 2lOS+1τ_{had} region. This SF is applied to each event with a jet faking tau individually to correct the normalization of MC yield in 3ℓ+1τ_{had} and 2ℓSS+1τ_{had} SR. For lepton faking tau events, they are already scaled by a dedicated SF from the tau combined performance group. The fake tau SF is calculated as such:

$$\theta_{fake\tau} = \frac{DD_{fakes}}{MC_{fakes}} = \frac{Data - MC_{RealTau} - MC_{FakeTauFromLeptons}}{MC_{FakeTauFromJets}} \tag{5.1}$$

It is found that the composition of the fake tau background significantly depends on the number of prongs, as shown in figure 5.11. Therefore the fake tau SF is derived and applied separately for 1- and 3-prong tau. Finally the difference of composition in different regions is also checked, as shown in figure 5.12, to guarantee that the SF can be applied to 2ℓSS+1τ_{had} region, and a systematics is derived to account for differences in fake tau composition across regions.

The fake tau proportion in each MC sample is shown in figure **??**. There is no significant fluctuation observed by comparing the fractions in CR to that in SR.

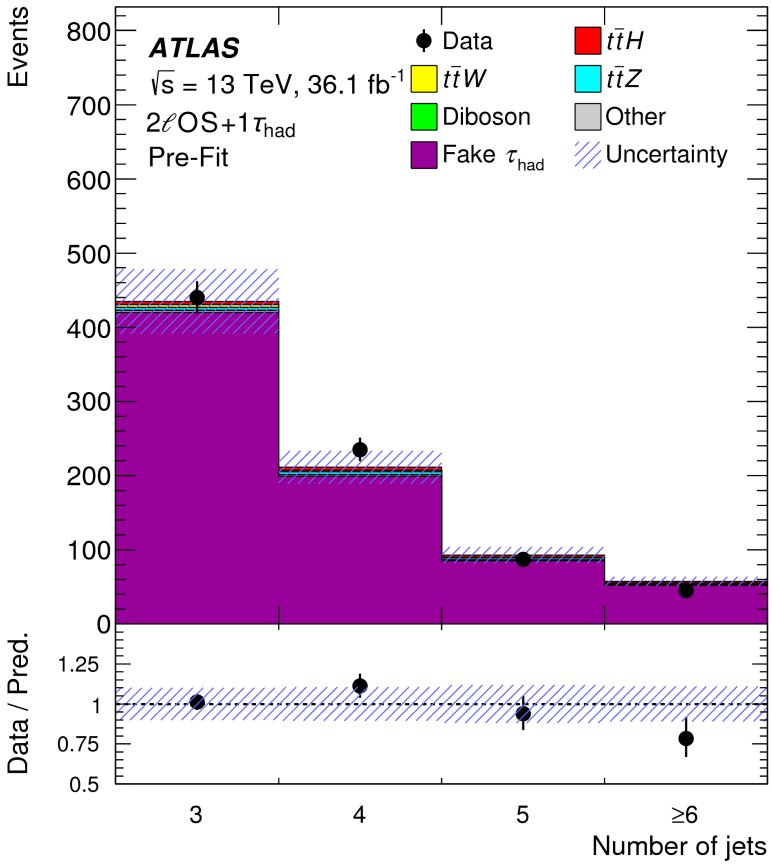


FIGURE 5.10: The post-fit BDT distribution in 2lOS+1τ_{had} from last round paper [4]. It was used to be one of the SRs. This region is dominated by fake τ_{had} contribution.

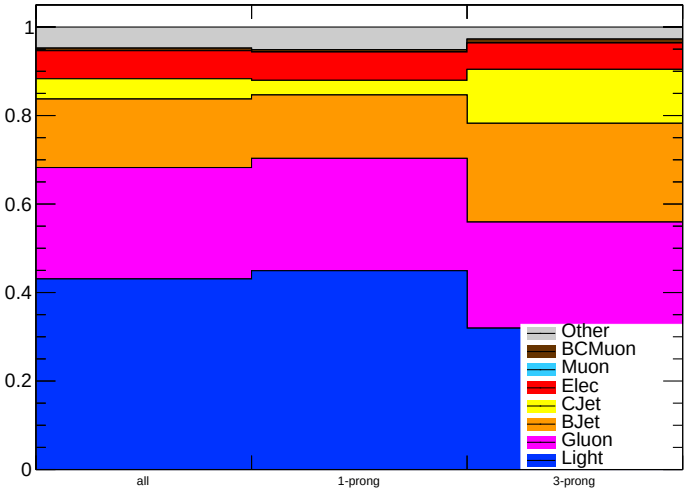


FIGURE 5.11: Composition of the fake tau background, shown for all taus, 1-prong (have only one charged track) and 3-prong (have three charged tracks) taus.

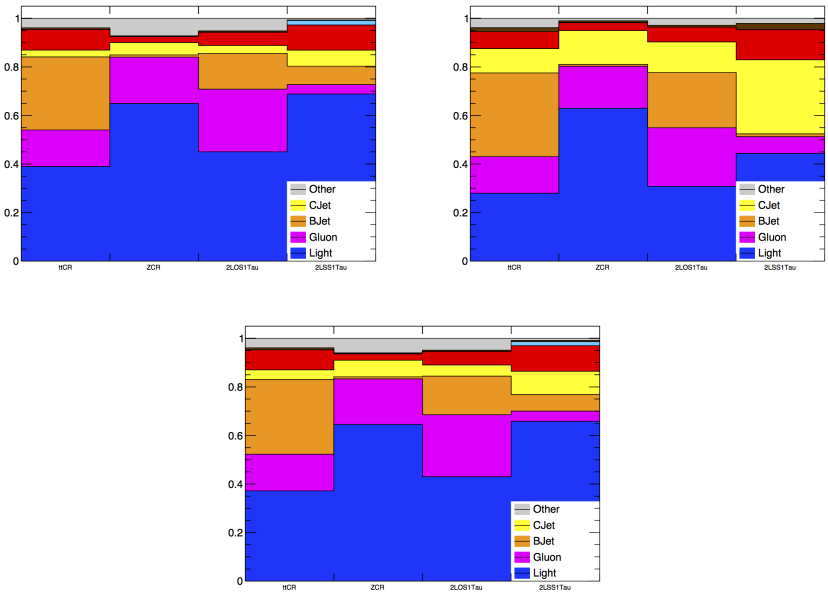


FIGURE 5.12: Composition plots of fake τ origin for 1-prong (left), 3-prong (middle) and all-prong (right).

5.4.3.3 Fake light leptons

Fake lepton is another important background in the analysis. It usually originates from $t\bar{t}$ events with a fake lepton from semi-leptonic b hadron decay, or from a converted photon. The default method, template fit, has already been discussed in section 5.3.4.1. There are two other data-driven methods, fake factor (FF) and matrix method (MM). MM was the default method for 2ℓSS and FF was the default for 2ℓSS+1τ_{had} in the last round analysis [4]. They are explained in the following sections since they were plugged into the statistical framework and I had to compare them to choose the best description. In the current analysis, FF estimate gives an additional uncertainty on the fake lepton in 2ℓSS+1τ_{had} channel only, while MM is aborted since template fit gives relatively better description of fake leptons and the composition of fake leptons is different in CR and SR, especially coming from fake electrons from conversions.

5.4.3.3.1 Fake factor The fake factor (FF) method is a data-driven technique to estimate fake lepton contribution in the signal region. This procedure is as follows (ABCD method): a control region is constructed by reversing the identification or isolation variables of the lepton (anti-tight). This control region, also known as "denominator", is enriched in the fake leptons and is orthogonal to the signal region (tight). Then the fake factor is calculated in a CR constructed by lowering the number of jets. Finally the fake factor is applied to a CR which is identical to the SR but with at least one anti-tight lepton. The regions used in 2ℓSS+1τ_{had} fake factor estimation is illustrated in figure 5.13.

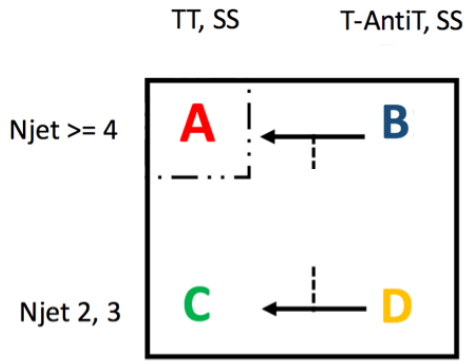


FIGURE 5.13: Illustration of the ABCD regions in 2ℓSS+1τ_{had}. "T" denotes tight lepton selections and "anti-T" inverse the tight selections. The fake factor (θ) is derived from region C and D. The estimate in the SR (region A) is obtained by applying the fake factor to region B.

The fake factor is derived separately for electrons and muons, as following:

$$FF_{muon} = \frac{CRC(TT)Data - PromptBackground - ttH}{CRD(T\bar{T})Data - PromptBackground - ttH} \tag{5.2}$$

$$FF_{electron} = \frac{CRC(TT)Data - PromptBackground - DD\ QmisID - ttH}{CRD(T\bar{T})Data - PromptBackground - DD\ QmisID - ttH} \tag{5.3}$$

where "T" refers to tight lepton and "$\bar{T}$" refers to anti-tight lepton. The contribution from signal and background events with prompt leptons are subtracted from data. The fake factor is parameterized as a function of p_T, as shown in figure 5.14. The final fake estimation is defined as:

$$N_{fake} = N_B[Data - PromptBackground - ttH] \times (FF_{electron} or FF_{muon}) \tag{5.4}$$

where N_B is the event yields in CRB.

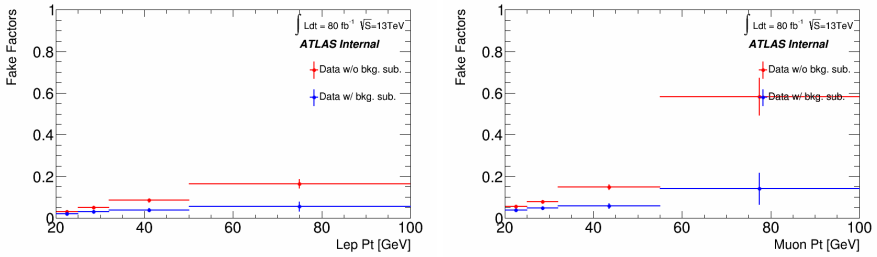


FIGURE 5.14: The fake factors as a funtion of p_T for electrons (left) and muons (right).

5.4.3.3.2 Matrix method The Matrix Method (MM) is another data-driven technique to estimate the fake lepton contribution. It has a similar idea to the FF method, estimating the fake contribution in tight region (denoted as T) from the loose object information in an anti-tight region (denoted as $\bar{T}$). To explain the basic strategy of the MM, a simplified case where only one lepton is considered would be helpful. The number of events with a tight lepton (referred to N^T) and with an anit-tight lepton (referred to $N^{\bar{T}}$) can be expressed in terms of efficiencies and inefficiencies for the basline loose real (referred to N^r) or fake (referred to N^f) leptons to pass the tight selection, as following:

$$N^T = \varepsilon_r N^r + \varepsilon_f N^f \tag{5.5}$$

$$N^{\bar{T}} = \bar{\varepsilon}_r N^r + \bar{\varepsilon}_f N^f \tag{5.6}$$

where ε_r (ε_f) represents the efficiency for a real (fake) lepton to pass the tight selection, and $\bar{\varepsilon}_r N^r \equiv (1-\varepsilon_r)$ ($\bar{\varepsilon}_f \equiv (1-\varepsilon_f)$) represents the probability for a real (fake) lepton to fail the tight selection but still pass the baseline selection. The above two equations can be formed into a matrix:

$$\begin{pmatrix} N^T \\ N^{\bar{T}} \end{pmatrix} = \begin{pmatrix} \varepsilon_r & \varepsilon_f \\ \bar{\varepsilon}_r & \bar{\varepsilon}_f \end{pmatrix} \begin{pmatrix} N^r \\ N^f \end{pmatrix} \tag{5.7}$$

By inverting this equation, the unkown number of real and fake leptons can be related to a number of observable quantities, i.e., the number of tight and anti-tight leptons and the efficiencies to pass the tight selection. Both real and fake lepton efficiencies can be measured directly in dedicated control regions using data.

In the multilepton analysis, there are two lepton candidates satisfying the baseline selection which could be fakes. Depending on whether or not each lepton passes the tight selection, each event can be categorised into any of four orthogonal (sidebands) regions:

- TT: event with both leptons passing tight selection (Total events: N^{TT})

- $T\not{T}$: event with leading lepton passing tight selection and subleading lepton failing tight selection (Total events: $N^{T\not{T}}$)
- $\not{T}T$: event with leading lepton failing tight selection and subleading lepton passing tight selection (Total events: $N^{\not{T}T}$)
- $\not{T}\not{T}$: event with both leptons failing tight selection (Total events: $N^{\not{T}\not{T}}$)

A 4×4 efficiency matrix can be defined to map the total number of such events into the total number of events in four dileptonic regions characterised by different real and fake lepton composition, as following:

- rr : event with both leptons being real (Total events: N^{rr}).
- rf : event with leading lepton being real and subleading lepton being fake (Total events: N^{rf}).
- fr : event with leading lepton being fake and subleading lepton being real (Total events: N^{fr}).
- ff : event with both leptons being fake (Total events: N^{ff}).

Then the 4×4 matrix equation can be written as:

$$\begin{pmatrix} N^{TT} \\ N^{T\not{T}} \\ N^{\not{T}T} \\ N^{\not{T}\not{T}} \end{pmatrix} = \begin{pmatrix} \varepsilon_{r,1}\varepsilon_{r,2} & \varepsilon_{r,1}\varepsilon_{f,2} & \varepsilon_{f,1}\varepsilon_{r,2} & \varepsilon_{f,1}\varepsilon_{f,2} \\ \varepsilon_{r,1}\not{\varepsilon}_{r,2} & \varepsilon_{r,1}\not{\varepsilon}_{f,2} & \varepsilon_{f,1}\not{\varepsilon}_{r,2} & \varepsilon_{f,1}\not{\varepsilon}_{f,2} \\ \not{\varepsilon}_{r,1}\varepsilon_{r,2} & \not{\varepsilon}_{r,1}\varepsilon_{f,2} & \not{\varepsilon}_{f,1}\varepsilon_{r,2} & \not{\varepsilon}_{f,1}\varepsilon_{f,2} \\ \not{\varepsilon}_{r,1}\not{\varepsilon}_{r,2} & \not{\varepsilon}_{r,1}\not{\varepsilon}_{f,2} & \not{\varepsilon}_{f,1}\not{\varepsilon}_{r,2} & \not{\varepsilon}_{f,1}\not{\varepsilon}_{f,2} \end{pmatrix} \begin{pmatrix} N^{rr} \\ N^{rf} \\ N^{fr} \\ N^{ff} \end{pmatrix} \tag{5.8}$$

Again, the number of fakes in signal region as a function of observables can be obtained by inverting the 4×4 matrix equation:

$$\begin{pmatrix} N^{rr} \\ N^{rf} \\ N^{fr} \\ N^{ff} \end{pmatrix} = \begin{pmatrix} \varepsilon_{r,1}\varepsilon_{r,2} & \varepsilon_{r,1}\varepsilon_{f,2} & \varepsilon_{f,1}\varepsilon_{r,2} & \varepsilon_{f,1}\varepsilon_{f,2} \\ \varepsilon_{r,1}\not{\varepsilon}_{r,2} & \varepsilon_{r,1}\not{\varepsilon}_{f,2} & \varepsilon_{f,1}\not{\varepsilon}_{r,2} & \varepsilon_{f,1}\not{\varepsilon}_{f,2} \\ \not{\varepsilon}_{r,1}\varepsilon_{r,2} & \not{\varepsilon}_{r,1}\varepsilon_{f,2} & \not{\varepsilon}_{f,1}\varepsilon_{r,2} & \not{\varepsilon}_{f,1}\varepsilon_{f,2} \\ \not{\varepsilon}_{r,1}\not{\varepsilon}_{r,2} & \not{\varepsilon}_{r,1}\not{\varepsilon}_{f,2} & \not{\varepsilon}_{f,1}\not{\varepsilon}_{r,2} & \not{\varepsilon}_{f,1}\not{\varepsilon}_{f,2} \end{pmatrix}^{-1} \begin{pmatrix} N^{TT} \\ N^{T\not{T}} \\ N^{\not{T}T} \\ N^{\not{T}\not{T}} \end{pmatrix} \tag{5.9}$$

The final number of fakes, N^{f}_{TT}, the total number of TT events with at least one fake lepton, can be obtained from the definition:

$$N^{f}_{TT} = N^{rf}_{TT} + N^{fr}_{TT} + N^{ff}_{TT} = \varepsilon_{r,1}\varepsilon_{f,2}N^{rf} + \varepsilon_{r,2}\varepsilon_{f,1}N^{fr} + \varepsilon_{f,1}\varepsilon_{f,2}N^{ff}. \tag{5.10}$$

It is important to mention that the real and fake lepton efficiencies are different for various sources of fake leptons, i.e., the most two important sources, from semi-leptonic b-hadron decay and from photon convention. Given the factor that 2ℓSS+$1\tau_{\text{had}}$ is very statistical limited, it is not possible to measure specific efficiencies. Therefore the difference of fake composition between 2ℓSS+$1\tau_{\text{had}}$ and 2ℓSS is checked, as shown in figure 5.15, to make sure that the configurations in 2ℓSS regions can be applied to 2ℓSS+$1\tau_{\text{had}}$ regions.

Then a closure test, by applying the MM weights to $t\bar{t}$ and $t\bar{t}\gamma$ MC events instead of data to see if you find what the predicted MC, is done both in the signal region and a low nJet region ($2 \leq$

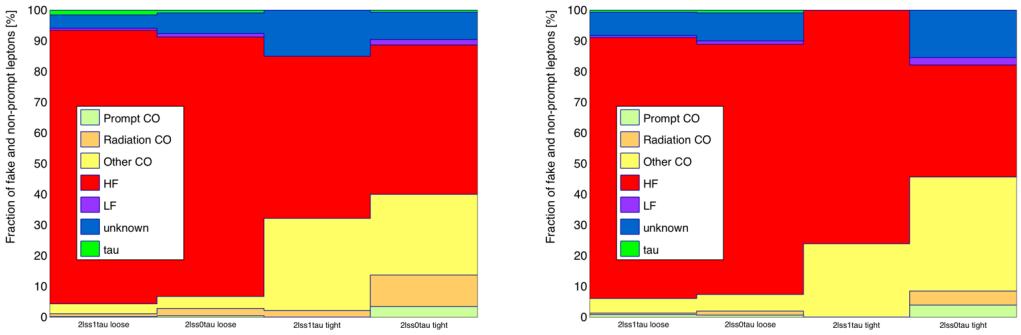


FIGURE 5.15: Fake compositions of loose and tight leptons in low nJets control region (left) and signal region (right).

$nJets \leq 3$), as shown in figure 5.16. The non-closure in signal region is summarized in table 5.12. Discriminating variables distributions of background from non prompt leptons estimated with the matrix method can be found in the right plot of figure 5.16.

	MC $t\bar{t}$+ $t\bar{t}\gamma$	MM applied on $t\bar{t}$+ $t\bar{t}\gamma$	MC/MM
SR ($nJets \geq 4$)	1.66 ± 0.58	1.41 ± 0.11	1.18 ± 0.417
LowNj SR ($2 \leq nJets \leq 3$)	7.49 ± 1.28	6.1 ± 0.22	1.23 ± 0.22

TABLE 5.12: Results of closure test on $t\bar{t}$+$t\bar{t}\gamma$ sample for MM in 2ℓSS+$1\tau_{\text{had}}$.

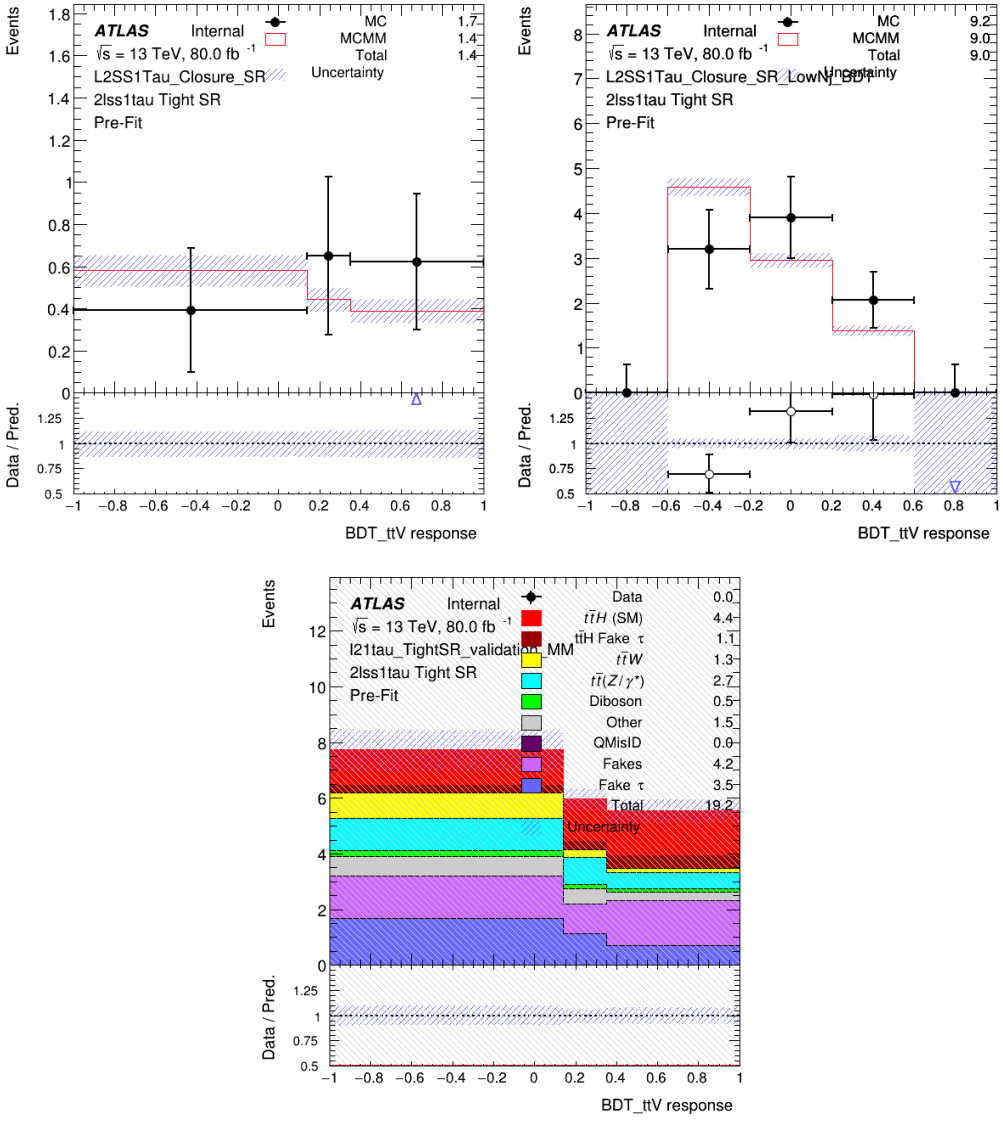


FIGURE 5.16: Closure test of MVA distribution (described later in section 5.4.4.1) in the SR (left) and CR (middle). Validation plot of MVA distribution using MM estimations is shown in the right.

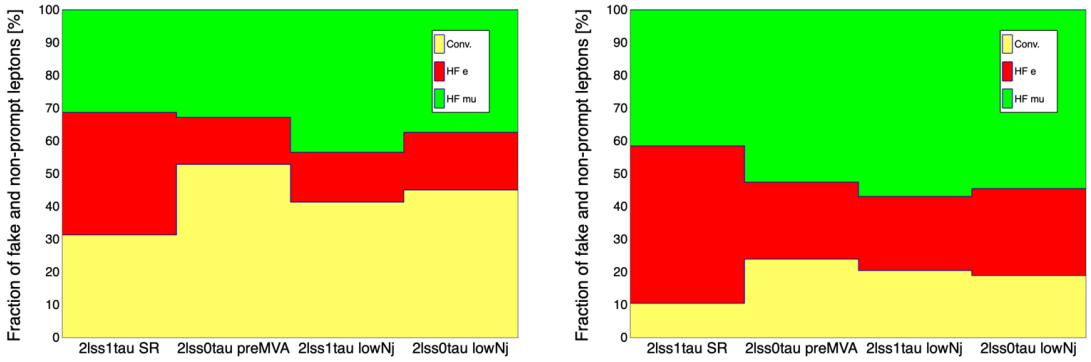
5.4.3.3.3 Template fit The template fit is performed across all multilepton channels, where $2\ell SS$ and 3ℓ channels are the dominant source of statistics. The template fit estimate in $2\ell SS+1\tau_{had}$ channel is driven by those two high statistic channels.

In the following studies, the NFs derived from $2\ell SS$ and 3ℓ CRs by a fit to data, listed in table 5.13, are directly propagated to $2\ell SS+1\tau_{had}$ channel.

	normalization factor	uncertainty
ttW	1.4447	0.21
$t\bar{t}$ γ ExtCnv	1.6985	1.04
$t\bar{t}$ γ IntCnv	0.746632	0.58
$t\bar{t}$ e HF	1.0937	0.32
$t\bar{t}$ mu HF	1.28327	0.16

TABLE 5.13: Normalization factors of template fit used in $2\ell SS+1\tau_{had}$ study. These numbers are different from those shown in table 5.3 due to the fact that these studies in $2\ell SS+1\tau_{had}$ were done before the final NFs were derived.

A comparison of the fake origin composition between $2\ell SS$ and $2\ell SS+1\tau_{had}$ is shown in Figure 5.17, table 5.14 and 5.15, both in the signal region and low nJets VR (same as SR but contains only 2 or 3 jets). The template fit SFs are not applied.



(A) raw event yields

(B) weighted event yields

FIGURE 5.17: Comparison of fake origin composition between $2\ell SS$ and $2\ell SS+1\tau_{had}$ in SR and low nJets VR.

	$2\ell SS$ preMVA	Fraction (%)	$2\ell ss+0\tau_{had}$ LowNj	Fraction (%)	$2\ell SS+1\tau_{had}$	Fraction (%)	$2\ell SS+1\tau_{had}$ VR	Fraction (%)
$t\bar{t}$ Conv.	92 ± 9.59	7.63 ± 0.83	112 ± 10.6	6.80 ± 0.66	0 ± 0	0 ± 0	4 ± 2	8.70 ± 4.53
$t\bar{t}+\gamma$ Conv.	544 ± 23.3	45.15 ± 2.33	628 ± 25.1	38.11 ± 1.79	5 ± 2.24	31.25 ± 16.01	15 ± 3.87	32.61 ± 9.70
$t\bar{t}$ HF e	163 ± 12.8	13.53 ± 1.13	281 ± 16.8	17.05 ± 1.10	6 ± 2.45	37.50 ± 17.95	7 ± 2.65	15.22 ± 6.17
$t\bar{t}+\gamma$ HF e	11 ± 3.32	0.91 ± 0.28	10 ± 3.16	0.61 ± 0.19	0 ± 0	0 ± 0	0 ± 0	0 ± 0
$t\bar{t}$ HF μ	376 ± 19.4	31.20 ± 1.84	580 ± 24.1	35.19 ± 2.24	5 ± 2.24	31.25 ± 16.78	19 ± 4.36	41.30 ± 11.26
$t\bar{t}+\gamma$ HF μ	19 ± 4.36	1.58 ± 0.36	37 ± 6.08	2.25 ± 0.37	0 ± 0	0 ± 0	1 ± 1	2.17 ± 2.20
Total	1210 ± 34.7	-	1650 ± 40.6	-	16 ± 4	-	46 ± 6.78	-

TABLE 5.14: Raw yields comparison of fake composition between $2\ell SS$ and $2\ell SS+1\tau_{had}$.

	2ℓSS preMVA	Fraction (%)	2ℓss + 0τ_{had} LowNj	Fraction (%)	2ℓSS+1τ_{had}	Fraction (%)	2ℓSS+1τ_{had} VR	Fraction (%)
$t\bar{t}$ Conv.	23.1 ± 2.53	13.52 ± 1.56	28.1 ± 2.83	10.52 ± 1.11	0 ± 0	0 ± 0	0.798 ± 0.425	10.52 ± 5.91
$t\bar{t} + \gamma$ Conv.	17.9 ± 0.839	10.44 ± 0.63	22.4 ± 1.03	8.37 ± 0.46	0.228 ± 0.133	10.34 ± 6.79	0.608 ± 0.195	8.01 ± 2.93
$t\bar{t}$ HF e	39.7 ± 3.24	23.21 ± 2.08	70.7 ± 4.33	26.44 ± 1.81	1.06 ± 0.495	48.06 ± 26.85	1.76 ± 0.694	23.23 ± 10.03
$t\bar{t} + \gamma$ HF e	0.461 ± 0.168	0.27 ± 0.10	0.284 ± 0.100	0.11 ± 0.04	0 ± 0	0 ± 0	0 ± 0	0 ± 0
$t\bar{t}$ HF μ	89.4 ± 4.81	52.19 ± 3.42	144 ± 6.22	53.93 ± 2.85	0.918 ± 0.444	41.60 ± 23.83	4.38 ± 1.04	57.69 ± 17.13
$t\bar{t} + \gamma$ HF μ	0.637 ± 0.159	0.37 ± 0.09	1.69 ± 0.339	0.62 ± 0.13	0 ± 0	0 ± 0	0.0413 ± 0.0413	0.54 ± 0.55
Total	171 ± 6.39	-	268 ± 8.16	-	2.21 ± 0.678	-	7.59 ± 1.34	-

TABLE 5.15: Weighted yields comparison of fake composition between 2ℓSS and 2ℓSS+1τ_{had}.

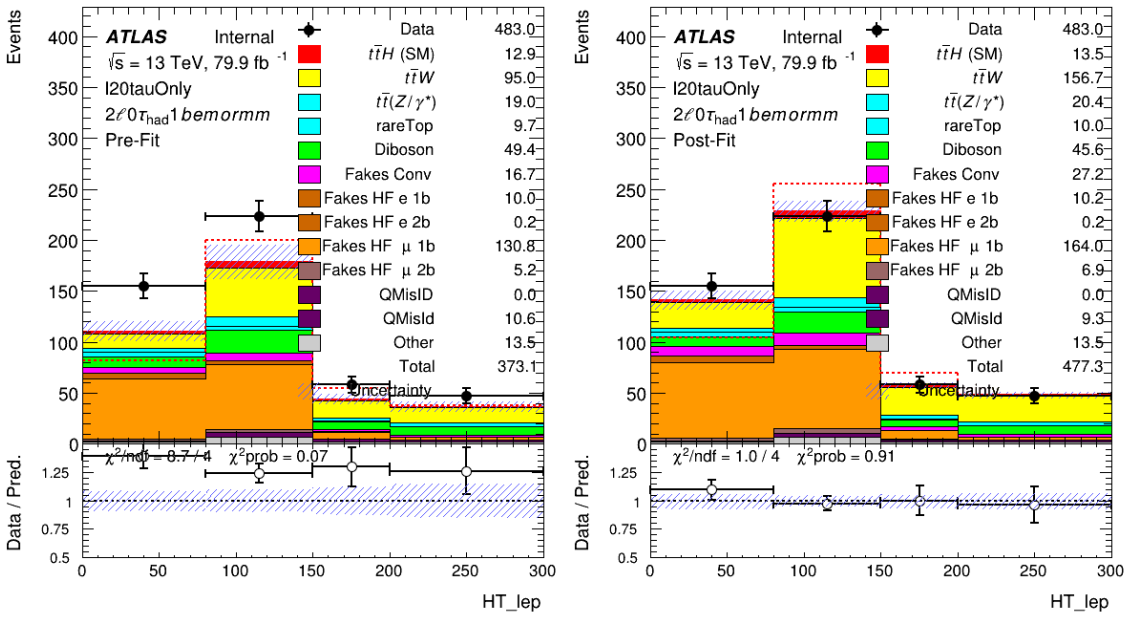
As stated above, 2ℓSS+1τ_{had} will only use the NFs derived from 2ℓSS channel fit. To check the effect without taking into account 2ℓSS+1τ_{had} regions, we compare 2 scenarios to derive the TF NFs, as follows:

- Scenario 1 : use 2ℓSS regions only to get the normalization factor out of the template fit method.
- Scenario 2 : use both 2ℓSS and 2ℓSS+1τ_{had} regions.

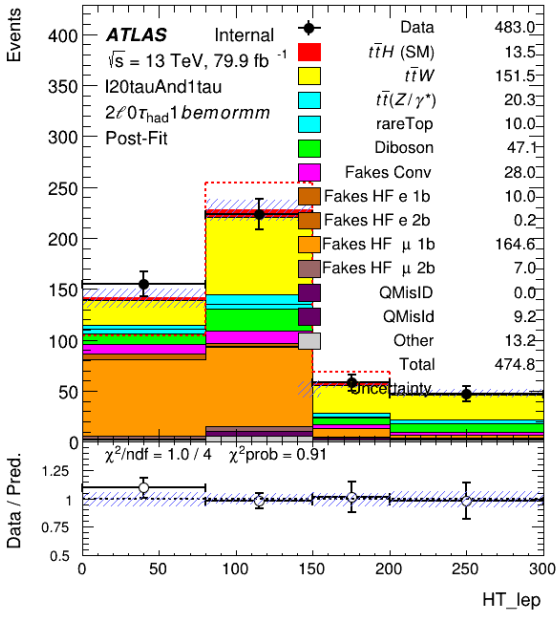
The fitted regions are as follows:

- preMVA CRs only defined in 2ℓSS channel
- Low nJets regions for both 2ℓSS and 2ℓSS+1τ_{had} channels, the only difference is tau requirement. There are 3 low nJets CRs:
  - CR1: $\mu^{\pm}\mu^{\pm}$ or $e^{\pm}\mu^{\pm}$ pair, fitting HT_{lep}
  - CR2: $e^{\pm}e^{\pm}$ or $\mu^{\pm}e^{\pm}$ pair, only 1 btagged jet, fitting $DRll01$
  - CR3: $e^{\pm}e^{\pm}$ or $\mu^{\pm}e^{\pm}$ pair, 2 or more btagged jets, fitting HT_{lep}

The prefit and postfit plots of CR1 in 2ℓSS and 2ℓSS+1τ_{had} as an example are shown in figure 5.18 and figure 5.19 respectively. The fitted normalization factors are shown in figure 5.20. As expected, all these plots show that 2ℓSS+1τ_{had} regions have very limited influence to the normalization factors and uncertainties. The template fit is driven by 2ℓSS regions which have much higher statistics.

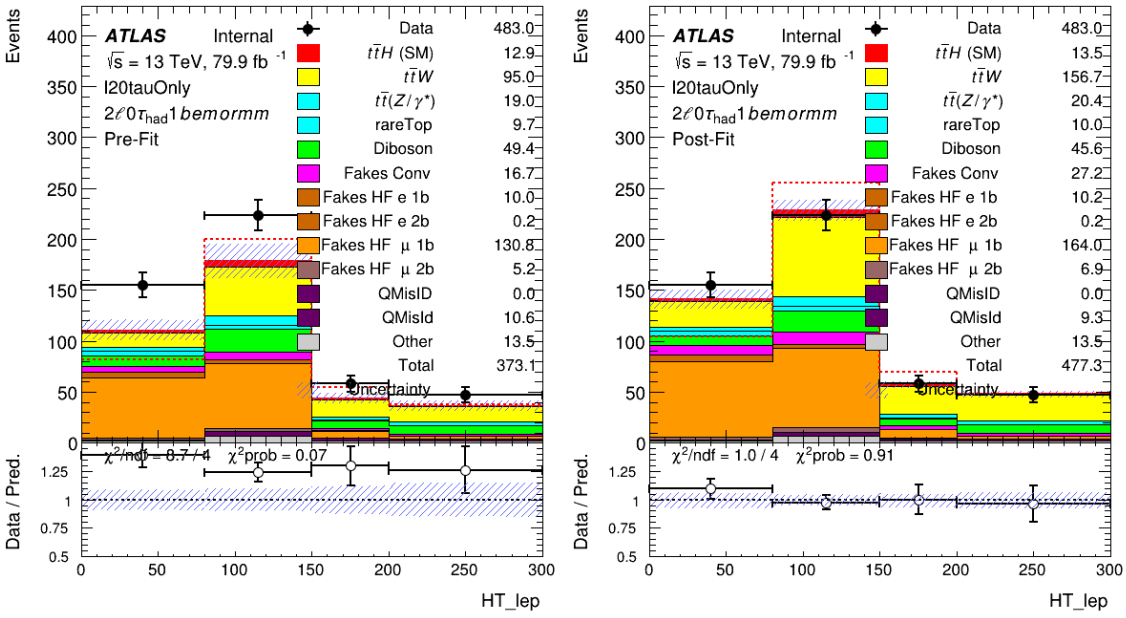


(A) (B)

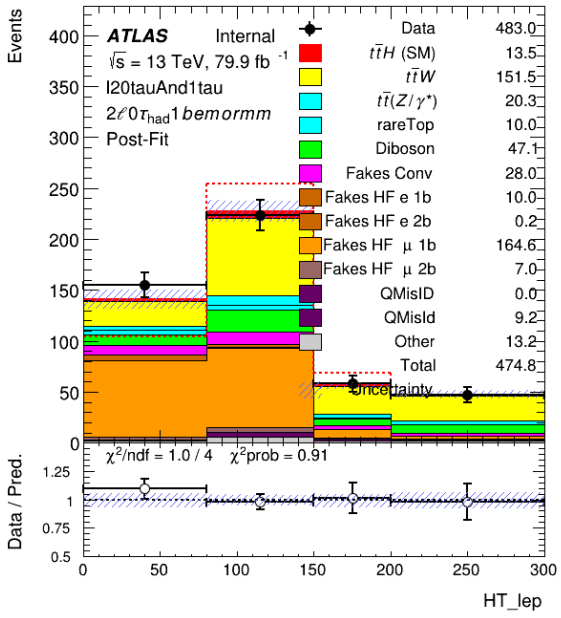


(C)

FIGURE 5.18: Prefit plot of CR1 in 2ℓSS channel used in the template fit validation (A), Postfit plots after 2ℓSS regions only fit (scenario 1) (B) and after 2ℓSS plus 2ℓSS+$1\tau_{had}$ regions fit (scenario 2) (C).



(A) (B)



(C)

FIGURE 5.19: Prefit plot of CR1 in 2ℓSS+1τ_{had} channel used in the template fit validation (A), Postfit plots after 2ℓSS regions only fit (scenario 1) (B) and after 2ℓSS plus 2ℓSS+1τ_{had} regions fit (scenario 2) (C).

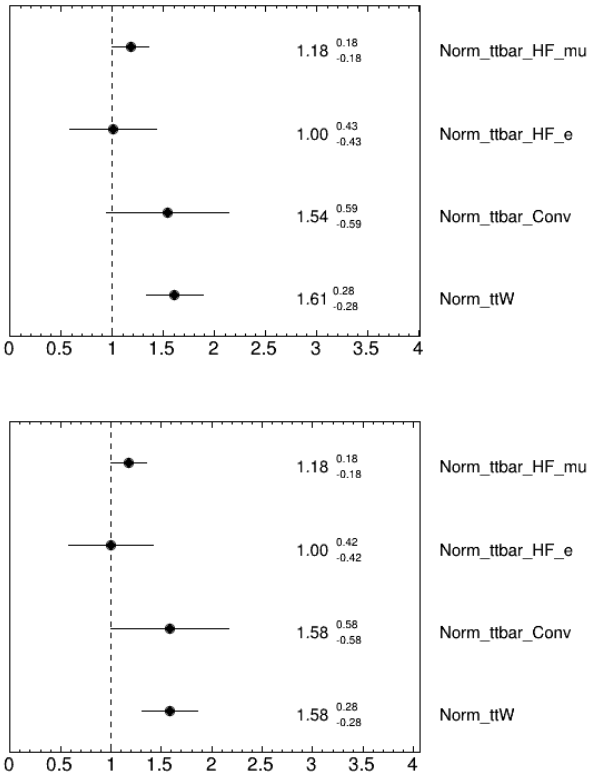


FIGURE 5.20: Normalization factors from template fit without $2\ell SS+1\tau_{had}$ (scenario 1) (top) and with $2\ell SS+1\tau_{had}$ (scenario 2) (bottom).

5.4.4 Statistical interpretation

5.4.4.1 Choice of analysis configurations

In previous results a BDT was trained to extract signal in 2ℓSS+$1\tau_{\text{had}}$ since it had the optimal sensitivity. The BDT is trained against a sum of all backgrounds, using six input variables as listed in table 5.16. The separation is defined as:

$$\langle S^2 \rangle = \frac{1}{2} \sum_{i=1}^{N_{bins}} \frac{(S_i - B_i)^2}{S_i + B_i} \Delta x_i \tag{5.11}$$

Variable Name	Rank	Seperation	Description
DeltaRLeadLepClosestJet	1	8.150×10^{-2}	ΔR between leading lepton and closest jet
DRll01	2	4.279×10^{-2}	ΔR between the two leptons
Mll01	3	2.811×10^{-2}	Invariant mass of the two leptons
TransverseMassLeadLepMET	4	2.234×10^{-2}	iTransvwerse mass of leading lepton and missing transverse energy
MvisnonH	5	1.705×10^{-2}	Visible mass between τ_{had} and furthest lepton
lep_Pt_0	6	1.651×10^{-2}	Transverse momentum of the leading lepton

TABLE 5.16: Ranking and separation of the variables used in the training.

Therefore there are several different options available for this analysis: BDT, categorization or event counting.

To archieve a optimal sensitivity and also robustness, various configurations have been tested by performing fits to Asimov data and checking the agreement with data in CR. The fits are performed in the SR (5.4.1), using `TRexFitter` package. The parameter of interest is the signal strength of $t\bar{t}H$ ($\mu_{t\bar{t}H}$). Systematic uncertainties have been discussed in previous section.

The nominal option that was chosen for this analysis is:

- Non-prompt leptons estimate: Template fit
- Discriminant variable: Event counting

The choice of non-prompt lepton estimate was decided for reasons of harmonisation with the 2ℓSS channel. To make decision on discriminant variable, the modelling of MC samples are checked in the CR (same as SR but with only 2 or 3 jets), as shown in figure 5.21, 5.22 and 5.23. As statistics increased from 36 fb^{-1} to 80 fb^{-1}, data-MC agreement was deteriorated in some discriminating variables distribution. Therefore it was chosen not to make such an extensive use of shape as an MVA analysis required. Further work is required to achieve a better understanding of discrepancies. Data and MC agreement does not seem to be too problematic apart from the invariant mass of the two leading leptons and the leading lepton pT where there seems to be a mismodelling around 100 GeV. This mis-modelling could be due to a statistical fluctuation in data. The large number of scanned distributions also increases the probability to have at least a couple of bins presenting a 2-sigma discrepancy (from the statistical Look Elsewhere effect). However to be conservative it was decided not to rely on these shapes neither with a BDT nor a categorisation and stick to a counting analysis until more statistics is analysed in order to rule out any significant mis-modelling.

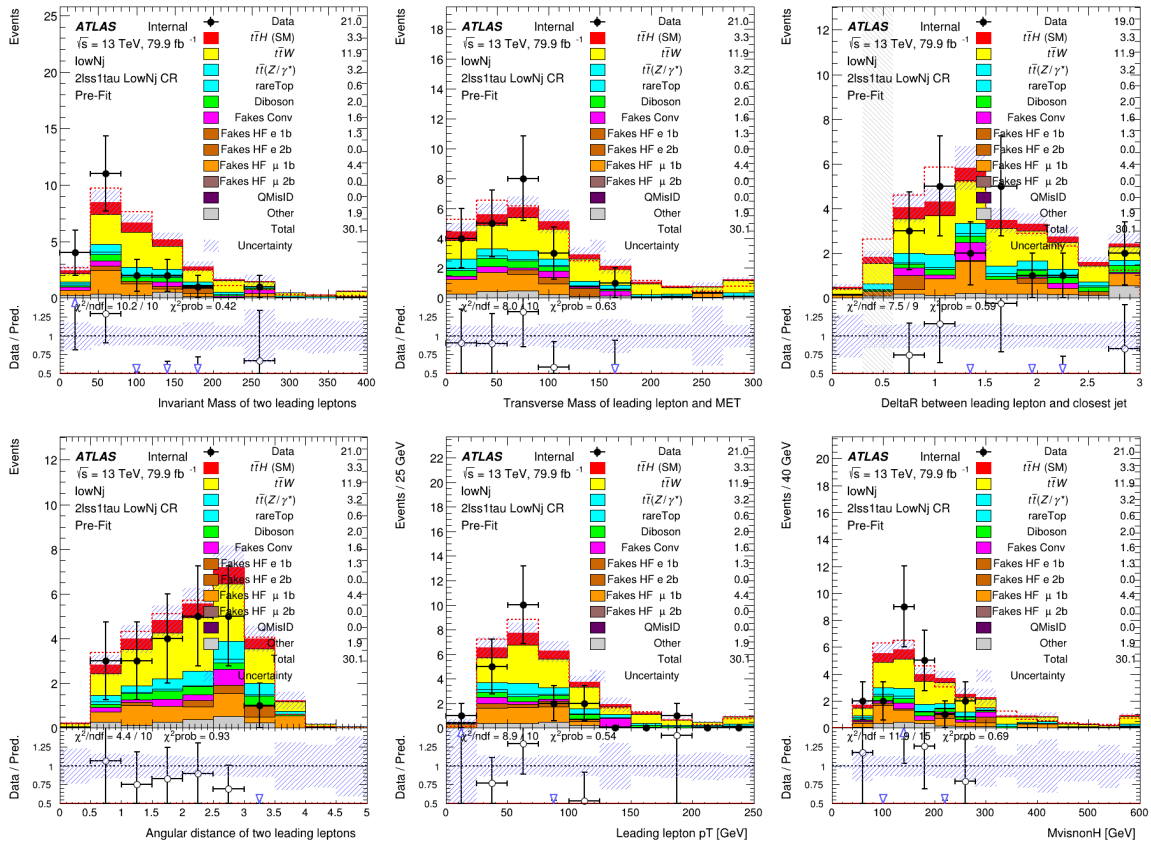


FIGURE 5.21: Validation plots of BDT input variables in the lowNj validation region.

In the following sections, the old results (BDT, categorization, fake factor and matrix method) will be presented in section 5.4.4.2 as references. The nominal strategy result will be presented in section 5.4.4.3. All these results are based on Asimov fit, real data is not used since it was blinded when these studies were performed. Looking into real data in SRs is forbidden before getting a robust analysis model, in order to avoid fine-tuning the analysis strategy that could introduce bias in the results.

5.4.4.2 Previous results

From the past experience, 2ℓSS+1τ_{had} channel is dominated by the non-closure of fakes since the non-closure value is derived from MC closure test using $t\bar{t}$ and $t\bar{t}\gamma$ which are very statistical limited in SR. Given the fact that now there are three methods, a new systematic called "method difference" which takes the difference of two non prompt lepton estimates (Fake Factor and Matrix Method) is being considered to replace the overall non-closure systematic. Thus not only the overall non-closure but also the shape impact is taken into account. Here the "method difference" is the difference between fake factor and matrix method.

The comparison plots between matrix method and fake factor estimations are shown in Figure 5.24. Only the non-closure uncertainties are included, the yields of fake factor and matrix method are 2.95 ± 1.2 and 4.64 ± 1.94 respectively. Furthermore, the "method difference" systematic is split into shape effect and normalization effect. In this way, the normalization impact can be

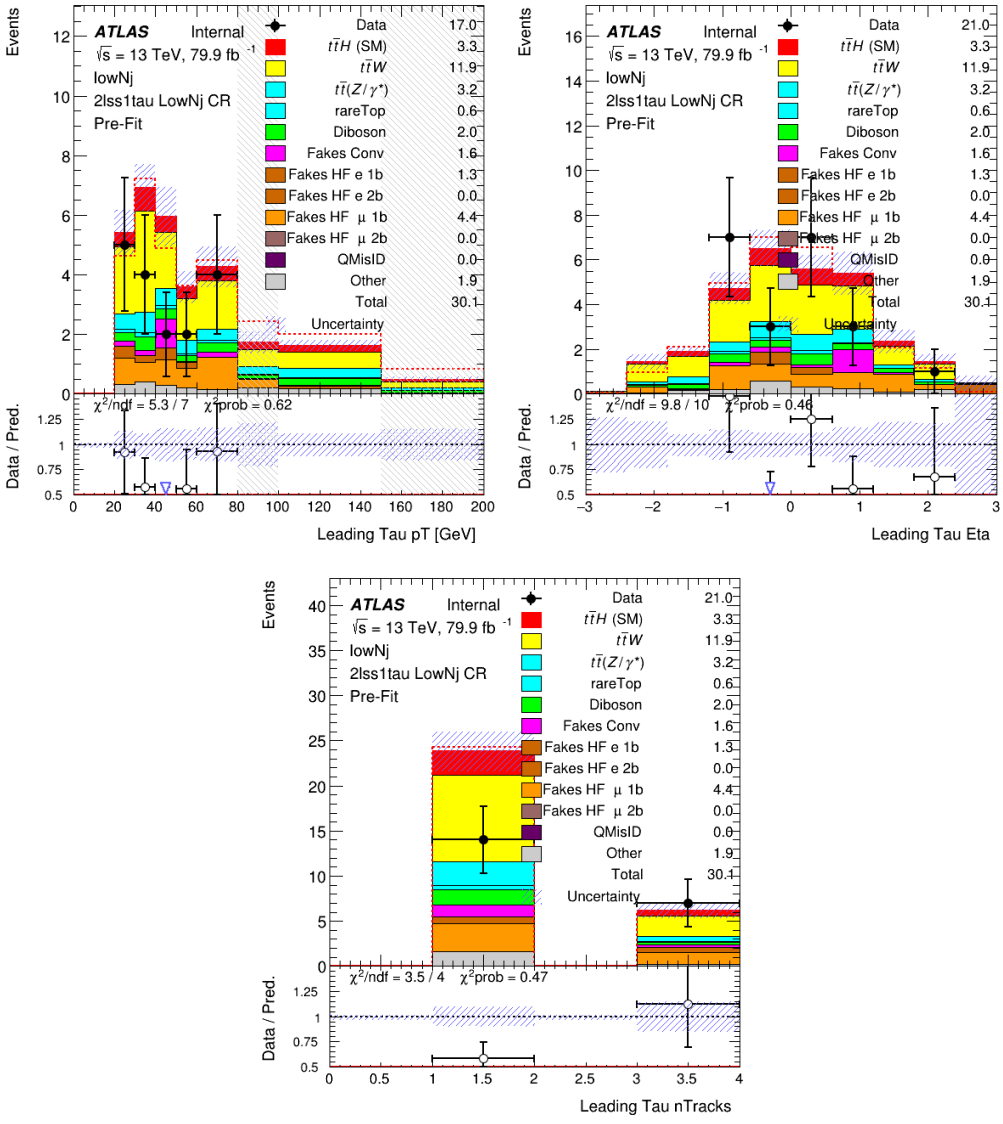


FIGURE 5.22: Validation plots of τ properties in the lowNj validation region.

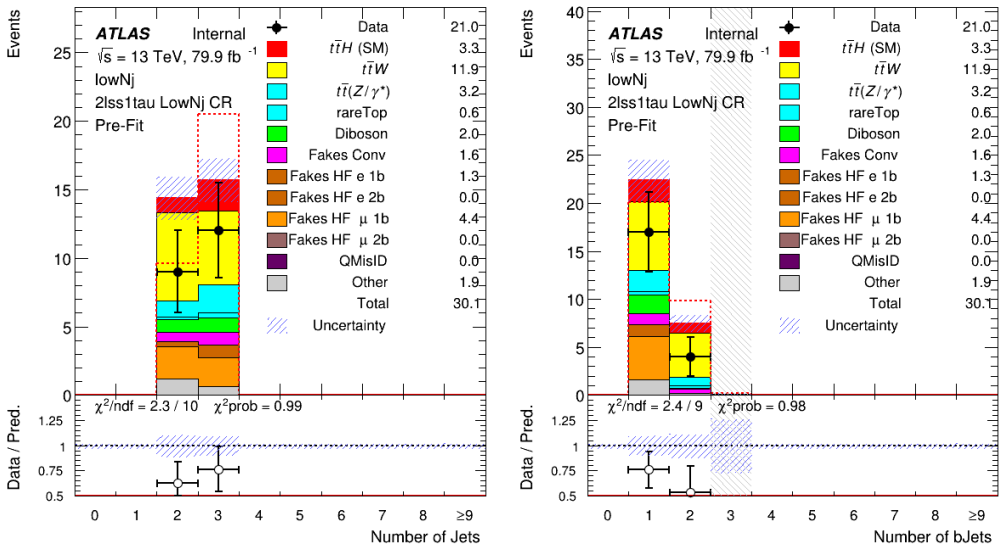


FIGURE 5.23: Validation plots of jet multiplicities in the lowNj validation region.

compared to non-closure and the shape impact can be seen more clearly.

Figure 5.25 shows a comparison between two fits using the overall non-closure or the "method difference" systematics. The BDT shape and FF estimate are used here. The impact of "method difference" is smaller comparing to that of the overall non-closure. The systematic control plot are shown in figure 5.26 and figure 5.27, where we can see a smaller non-closure value than the ones from MC closure test. And additionally the shape impact is also taken into account in the fit though it is small compared to the normalization impact. Figure 5.28 shows the "method difference" systematics in categorisation. It also gives a smaller non-closure in high and medium purity categories. The non-closure of "method difference" is found to be larger in the "Rest" category. But since the "method difference" systematics in each category are correlated and the "Rest" category does not play an important role in the fit due to a low signal purity, this large non-closure does not have a large impact on the result.

The BDT distribution is re-binned to 3 bins, $(-1, 0.14), (0.14, 0.35), (0.35, 1.0)$ which gives the optimal expected significance in stat only and expected fit on Asimov dataset. For the categorization fit, the whole signal region phase space is divided in 3 categories. Detailed description can be found in section 5.4.2. The yields are summarized in table 5.17 for BDT and table 5.18 for categorization, respectively. The distributions of BDT are shown in figure 5.29 and categorisation are shown in figure 5.30.

A summary of expected significance is shown in table 5.19. In both cases, the non-closure or normalization of method difference is still one of the top ranking systematics. Another dominant systematic is fake tau systematic which can be seen in the ranking plots 5.25.

The fit results among all cases show consistent result, the expected significances are around 1.3σ, slightly different depending on the choice of fake estimation and analysis method. Fits using FF estimate give a higher significance due to a lower central value.

	2ℓSS+$1\tau_{\text{had}}$
$t\bar{t}W$	1.33 ± 0.581
$t\bar{t}Z/\gamma$	2.71 ± 0.662
$t\bar{t}lllowmass$	0.0538 ± 0.0276
rareTop	0.0947 ± 0.0558
VV	0.479 ± 0.720
QMisID	$0.005\,83 \pm 0.001\,43$
Faketau	3.61 ± 1.56
Three top	0.0712 ± 0.0715
Four top	0.451 ± 0.136
$t\bar{t}WW$	0.237 ± 0.0725
tZ	0.125 ± 0.133
WtZ	0.470 ± 0.479
WtH	$3 \times 10^{-5} \pm 4.36 \times 10^{-6}$
$t\bar{t}H$ (SM)	4.38 ± 0.516
$t\bar{t}H$ Faketau	1.03 ± 0.0474
$Fakes\_FF$	2.95 ± 1.98
$Fakes\_MM$	4.64 ± 2.76
$Total\_FF$	18.0 ± 2.52
$Total\_MM$	19.7 ± 2.83

TABLE 5.17: Yields of signal region combining fake factor and matrix method estimation. Uncertainties include statistics and systematics in which method difference is used.

	2ℓSS+$1\tau_{\text{had}}HighPurity$	2ℓSS+$1\tau_{\text{had}}MediumPurity$	2ℓSS+$1\tau_{\text{had}}Rest$
$t\bar{t}W$	0.382 ± 0.250	0.321 ± 0.191	0.628 ± 0.294
$t\bar{t}Z/\gamma$	0.880 ± 0.297	0.996 ± 0.294	0.832 ± 0.245
$t\bar{t}lllowmass$	$0.008\,15 \pm 0.0223$	$0.005\,31 \pm 0.0105$	0.0404 ± 0.0225
rareTop	0.0583 ± 0.0444	0.0364 ± 0.0286	$1 \times 10^{-5} \pm 5.10 \times 10^{-6}$
VV	0.128 ± 0.195	0.212 ± 0.320	0.139 ± 0.214
QMisID	$0.001\,37 \pm 0.000\,384$	$0.002\,01 \pm 0.000\,596$	$0.002\,45 \pm 0.000\,725$
Faketau	0.968 ± 0.438	1.30 ± 0.580	1.34 ± 0.598
Three top	0.0179 ± 0.0182	0.0159 ± 0.0162	0.0375 ± 0.0379
Four top	0.146 ± 0.0492	0.103 ± 0.0364	0.202 ± 0.0660
$t\bar{t}WW$	0.0646 ± 0.0346	0.0650 ± 0.0395	0.107 ± 0.0501
tZ	0.0386 ± 0.0477	0.0577 ± 0.0689	0.0290 ± 0.0364
WtZ	0.231 ± 0.242	0.129 ± 0.148	0.110 ± 0.131
WtH	$1 \times 10^{-5} \pm 1.67 \times 10^{-6}$	$1 \times 10^{-5} \pm 1.67 \times 10^{-6}$	$1 \times 10^{-5} \pm 1.67 \times 10^{-6}$
$t\bar{t}H$ (SM)	2.05 ± 0.238	1.44 ± 0.206	0.884 ± 0.262
$t\bar{t}H$ Faketau	0.545 ± 0.0461	0.332 ± 0.0346	0.151 ± 0.0236
$Fakes\_FF$	1.31 ± 2.10	1.14 ± 0.407	0.502 ± 2.26
$Fakes\_MM$	1.99 ± 2.29	1.16 ± 0.766	1.50 ± 3.24
$Total\_FF$	6.83 ± 0.957	6.16 ± 0.722	5.01 ± 1.06
$Total\_MM$	7.51 ± 1.24	6.18 ± 0.918	6.00 ± 1.66

TABLE 5.18: Yields of each category combining fake factor and matrix method estimation. Uncertainties include statistics and systematics in which method difference is used.

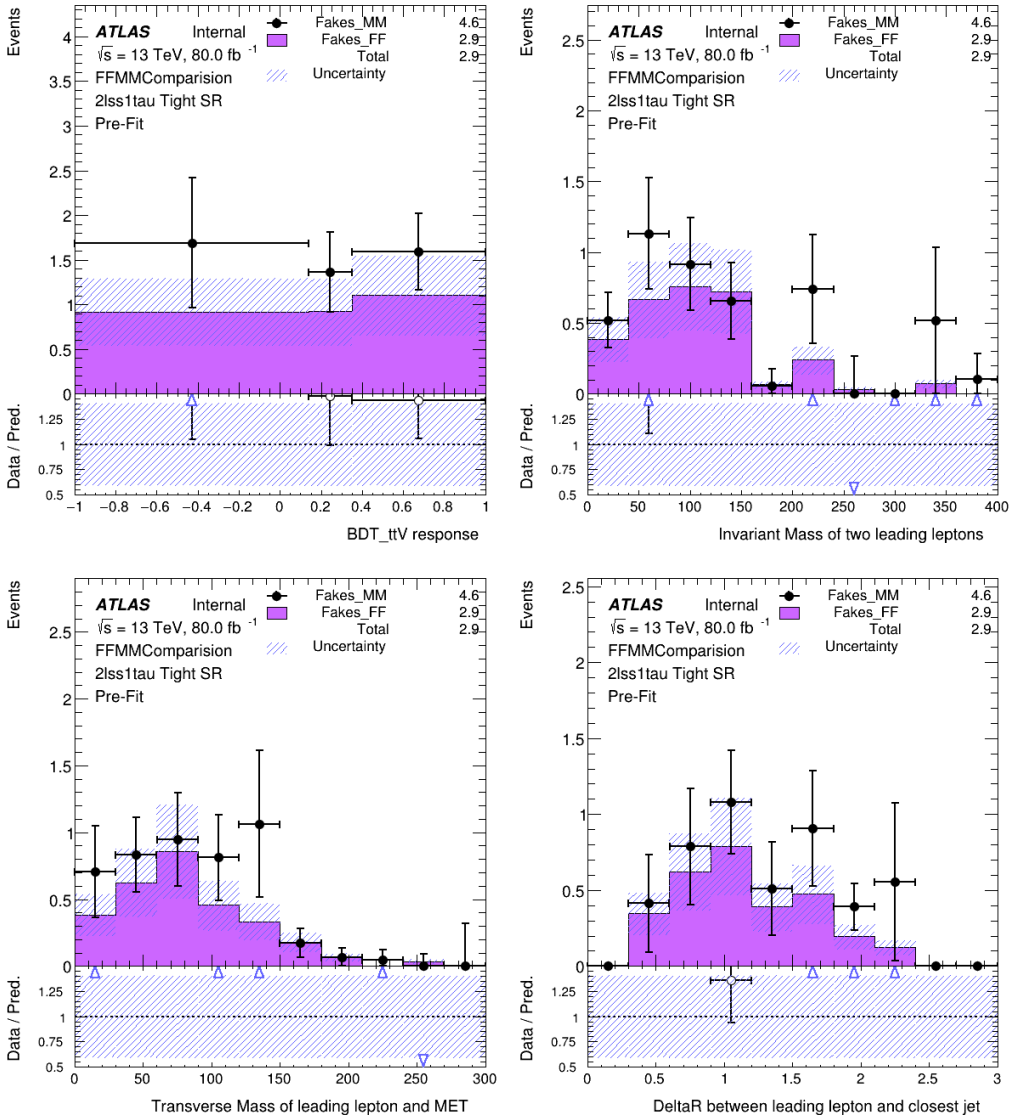


FIGURE 5.24: Comparison between matrix method and fake factor shown for MVA distribution and categorisation variables.

Expected Sig	Matrix Method		Fake Factor		Stat & Asimov
	Method Difference	Overall non-closure	Method Difference	Overall non-closure	
BDT	1.26σ	1.19σ	1.33σ	1.31σ	1.51σ
Categorisation	1.29σ	1.20σ	1.35σ	1.30σ	1.53σ

TABLE 5.19: The summary of expected significance calculated from the fit.

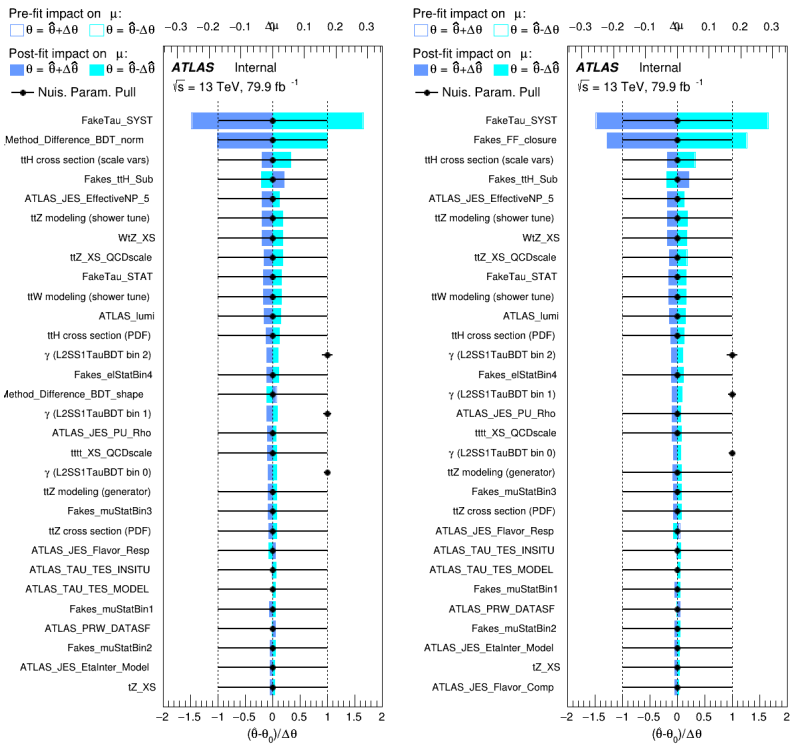
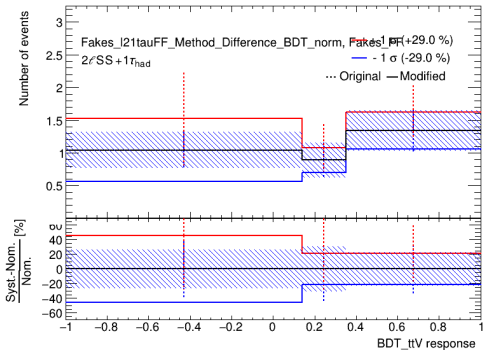
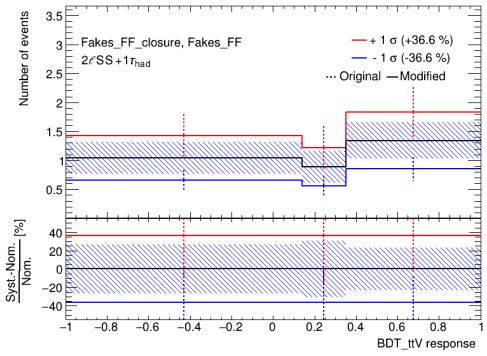


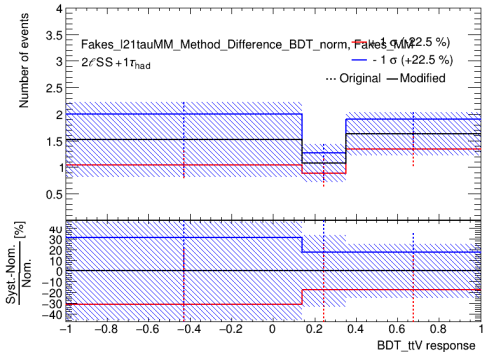
FIGURE 5.25: The ranking plots when using fake factor estimations and fitting on BDT. Left hand plot uses "method difference" systematic, right hand plot uses overall non-closure systematic. The impact of "method difference" systematic is smaller.



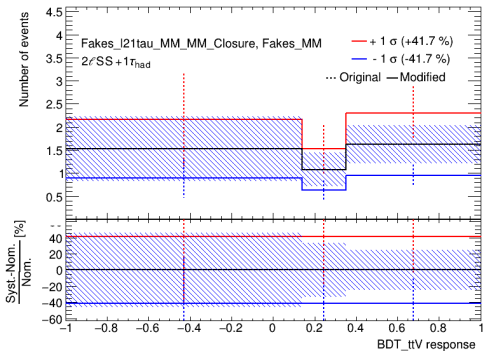
(A) Use FF fake and method difference



(B) Use FF fake and overall non-closure of 36.65 %

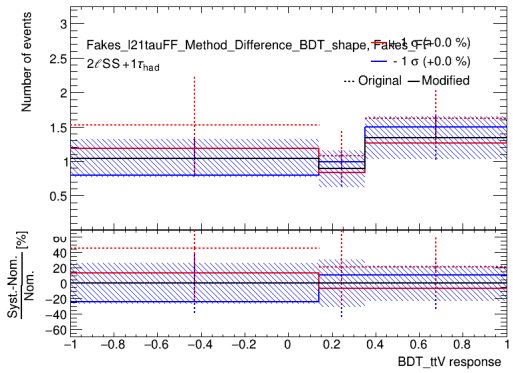


(C) Use MM fake and method difference



(D) Use MM fake and overall non-closure of 41.7 %

FIGURE 5.26: Systematic control plots of BDT distribution comparing the overall non-closure and normalization of method difference.



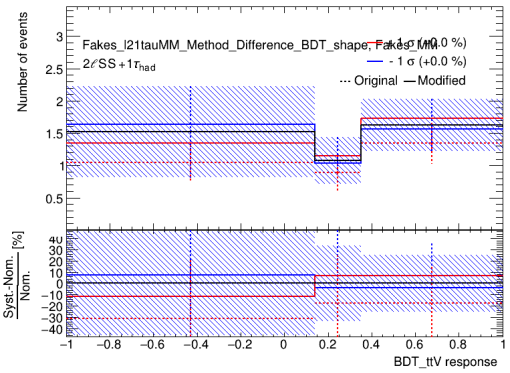


FIGURE 5.27: Shape impact of method difference systematics. Plot on the left hand uses fake factor estimation, on the right hand uses matrix method estimation.

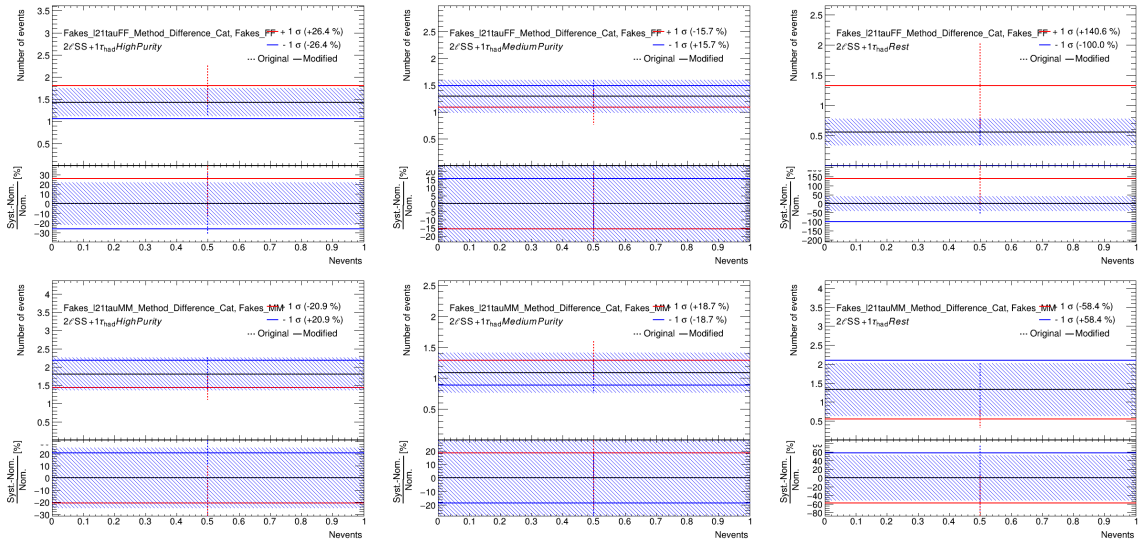


FIGURE 5.28: Systematic control plots of method difference systematics in each categories. Plots on the top use FF estimation, plots on the bottom use MM estimation.

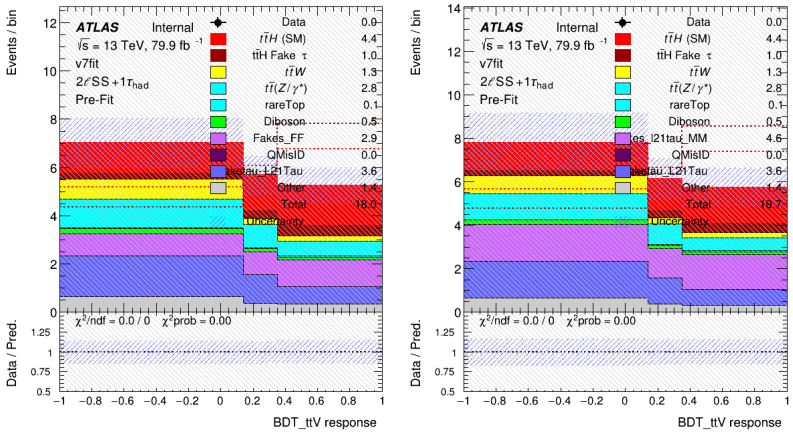


FIGURE 5.29: The prefit plots of BDT distribution using fake factor estimations (left) or matrix method estimations (right). The "method difference" systematic is used.

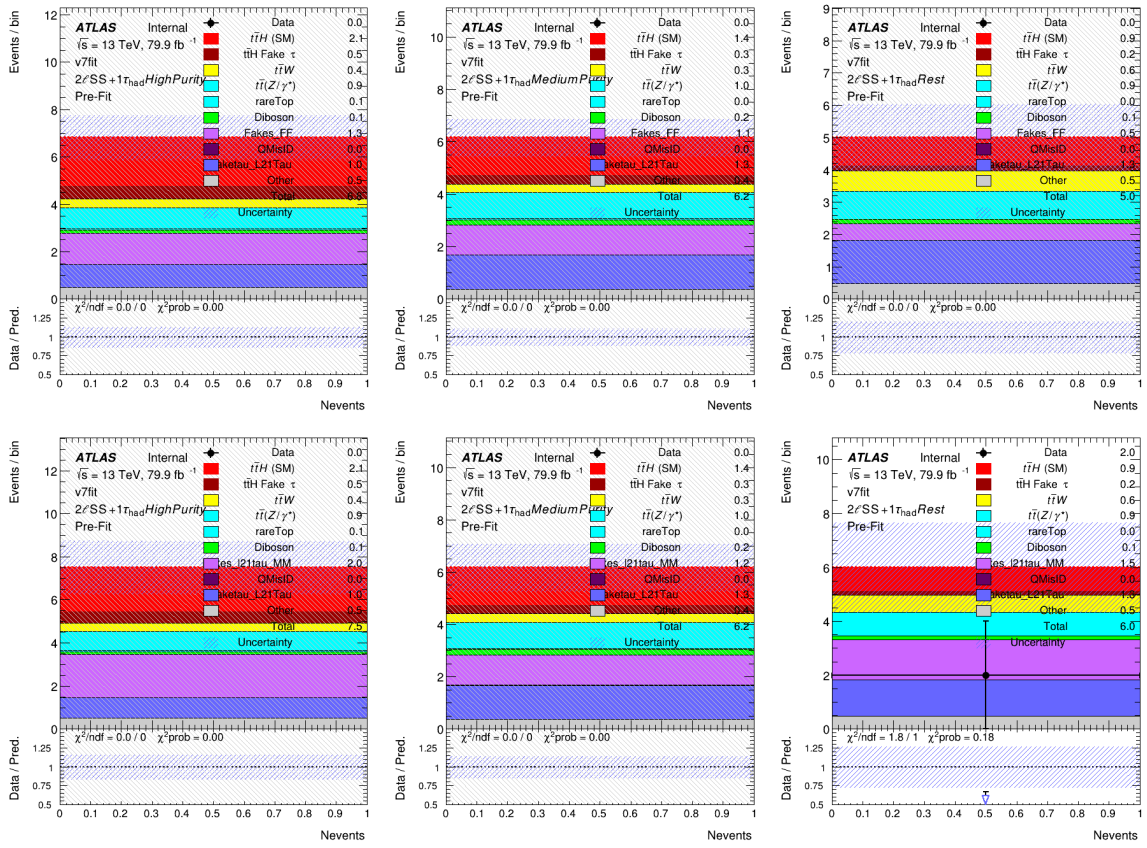


FIGURE 5.30: The prefit plots of each category using fake factor estimations (top) or matrix method estimations (bottom). The "method difference" systematic is used.

5.4.4.3 Current results

In the default strategy, template fit is chosen to estimate the central value of non-prompt lepton Background and fake factor estimate is used to derive the "method difference" systematic. It accounts for the possible discrepancies between a DD estimate and the pseudo-DD estimate based on TF, as shown in figure 5.31 and table 5.20. The $t\bar{t}W$ contribution in these plots is scaled by a factor of 1.6, according to the template fit results in table 5.13.

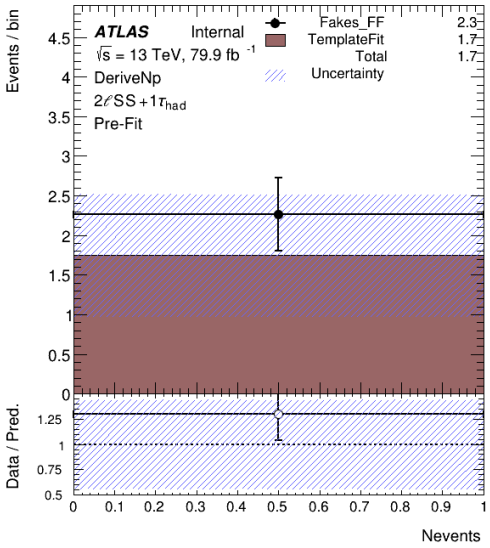


FIGURE 5.31: Comparison of template fit and fake factor estimation in SR.

	Fact factor yields	Template fit yields	Ratio (FF/TF)
SR	2.27 ± 0.32	1.75 ± 0.55	1.30 ± 0.45

TABLE 5.20: Yields and ratio of fake facor and template fit shown in SR.

Yields of each sample are shown in table 5.21. Now fake τ contribution is included in each monte carlo sample individually, any event which contains a fake τ is scaled by the fake τ scale factor. Expected significances and fitted μ are summarized in table 5.22. Prefit plots are shown in figure 5.32b. NP ranking, NP constrains and correlations are shown in figure 5.33b, figure 5.34b and figure 5.35b respectively.

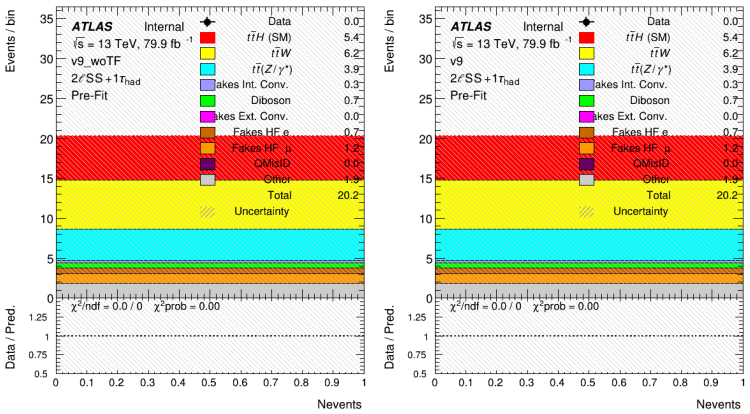
To test the impact of the "method difference" systematic on the significance, another fit without this systematic is performed. The only systematics remained on lepton fakes would be those uncertainties on template fit NFs, listed in table 5.13. Expected significances and fitted μ are summarized in table 5.23 to compare with normal fit. Prefit plots are shown in figure 5.32a. NP ranking, NP constrains and correlations are shown in figure 5.33a, figure 5.34a and figure 5.35a respectively.

	$2\ell SS+1\tau_{had}$
$t\bar{t}W$	6.16 ± 2.56
$t\bar{t}Z/\gamma$	3.92 ± 0.629
$t\bar{t}lllowmass$	0.0840 ± 0.0496
rareTop	0.183 ± 0.110
VV	0.712 ± 0.367
$t\bar{t}$ internal conversion	$7.47 \times 10^{-6} \pm 4.32 \times 10^{-6}$
$t\bar{t}+\gamma$ internal conversion	0 ± 0
$t\bar{t}$ Conv.	$1.70 \times 10^{-5} \pm 8.77 \times 10^{-7}$
$t\bar{t}+\gamma$ Conv.	$1.70 \times 10^{-5} \pm 8.77 \times 10^{-7}$
$t\bar{t}$ HF e	0.707 ± 0.546
$t\bar{t}+\gamma$ HF e	$1.09 \times 10^{-5} \pm 3.52 \times 10^{-6}$
$t\bar{t}$ HF μ	1.17 ± 0.902
$t\bar{t}+\gamma$ HF μ	$1.28 \times 10^{-5} \pm 2.14 \times 10^{-6}$
QMisID	$0.002\,42 \pm 0.000\,984$
Three top	0.0995 ± 0.0501
Four top	0.575 ± 0.170
$t\bar{t}WW$	0.387 ± 0.0596
tZ	0.166 ± 0.0881
WtZ	0.602 ± 0.311
VVV	0.0240 ± 0.0125
VH	$1 \times 10^{-5} \pm 5.16 \times 10^{-7}$
$t\bar{t}H$ (SM)	5.45 ± 5.67
Total	20.2 ± 4.55

TABLE 5.21: Yields of $2\ell SS+1\tau_{had}$ signal region.

Analysis Strategy	Expected significance	Fitted μ
One-bin fit	0.99	$1.00^{+1.11}_{-1.01}$

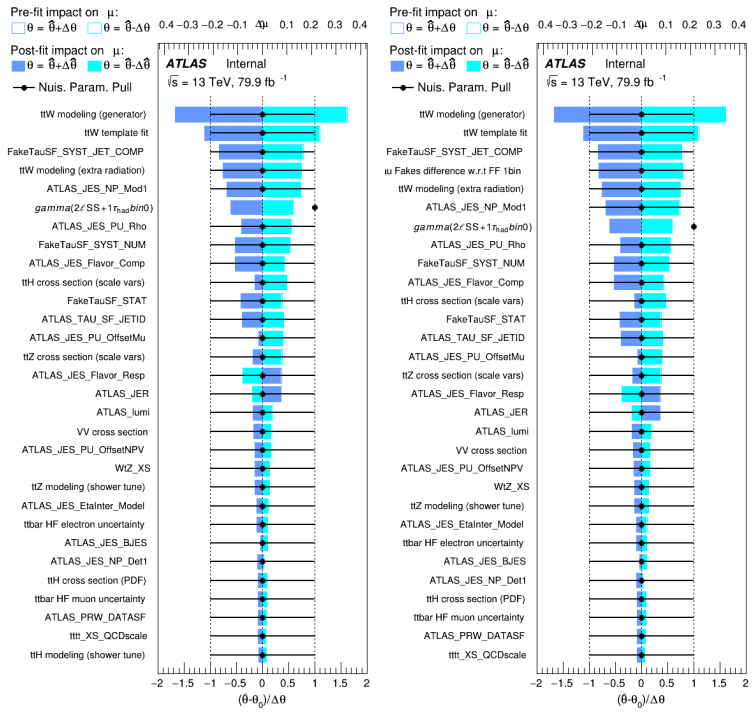
TABLE 5.22: Summary of fit results.



(A) w/o TF-FF method difference

(B) w/i TF-FF method difference

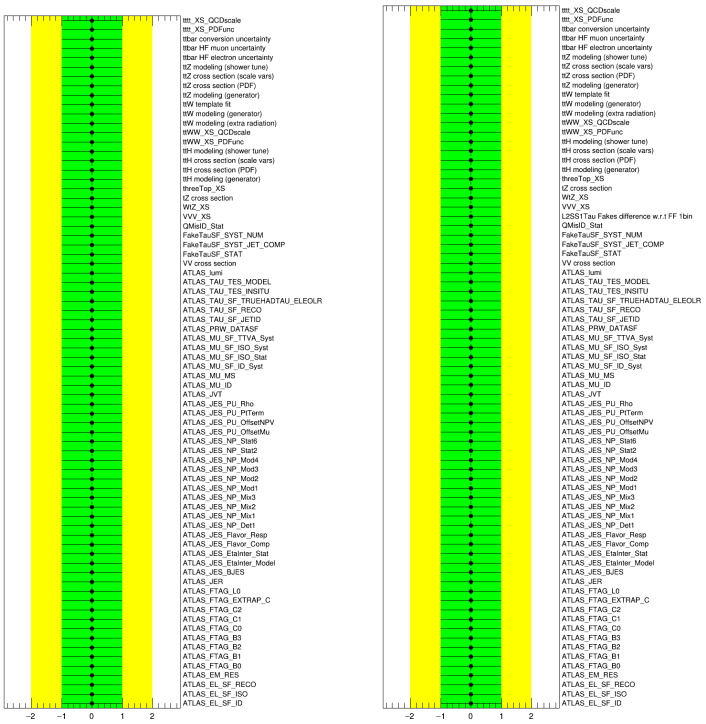
FIGURE 5.32: Prefit plots of fit without (left) and with (right) taking the Template Fit and Fake Factor difference as data-driven estiamte systematics.



(A) w/o TF-FF method difference

(B) w/i TF-FF method difference

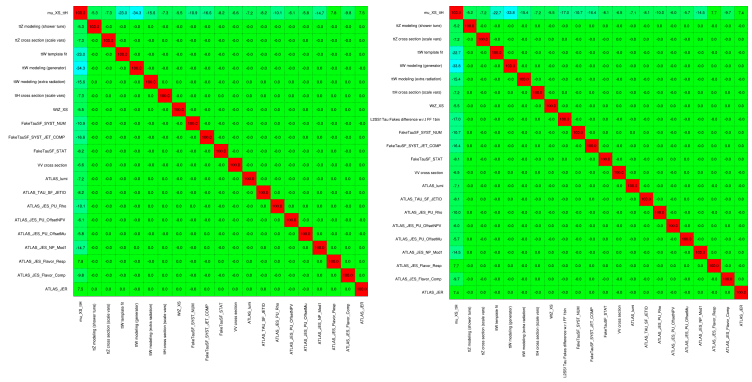
FIGURE 5.33: Ranking plots of fit without (left) and with (right) taking the Template Fit and Fake Factor difference as data-driven estiamte systematics.



(A) w/o TF-FF method difference

(B) w/i TF-FF method difference

FIGURE 5.34: NP pull of fit without (left) and with (right) taking the Template Fit and Fake Factor difference as data-driven estiamte systematics.



(A) w/o TF-FF method difference (B) w/i TF-FF method difference

FIGURE 5.35: Correlation plots of fit without (left) and with (right) taking the Template Fit and Fake Factor difference as data-driven estiamte systematics

Analysis Strategy	Expected significance	Fitted μ	Difference w.r.t one-bin
One-bin fit	0.99	$1.00^{+1.11}_{-1.01}$	-
One-bin w/o methodDiff	1.00	$1.00^{+1.10}_{-0.99}$	1.02 %

TABLE 5.23: Fit results without method difference comparing to normal one-bin fit.

5.4.4.4 Alternative approach

In this section, the same fit but using BDT or categorization as discriminant variable is presented as a reference to check the sensitivy loss. The "method difference" systematic is also applied to both BDT and categorization, as shown in figure 5.36 and table 5.24. Expected significances and fitted μ are summarized in table 5.25. Prefit plots are shown in figure 5.37.

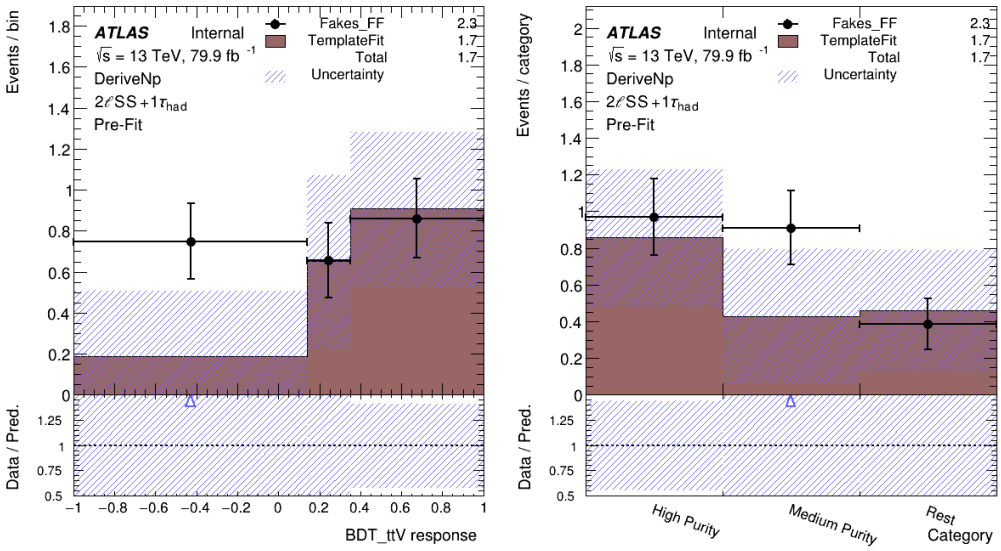


FIGURE 5.36: Comparison of template fit and fake factor estimation in SR, as a function of BDT (left) or category (right).

		Fake factor yields	Template fit yields	Ratio (FF/TF)
BDT	bin 1	0.75 ± 0.18	0.19 ± 0.32	3.98 ± 6.90
	bin 2	0.66 ± 0.18	0.65 ± 0.42	1.01 ± 0.71
	bin 3	0.86 ± 0.19	0.91 ± 0.38	0.95 ± 0.45
Categorisation	High purity	0.97 ± 0.21	0.86 ± 0.38	1.13 ± 0.55
	Medium purity	0.91 ± 0.20	0.43 ± 0.37	2.13 ± 1.89
	Rest	0.39 ± 0.14	0.46 ± 0.37	0.84 ± 0.68

TABLE 5.24: Yields and ratio of fake facor and template fit shown in each bin.

Analysis Strategy	Expected significance	Fitted μ	Difference w.r.t one-bin
One-bin fit	1.13	$1.00^{+1.01}_{-0.89}$	-
BDT	1.19	$1.00^{+0.97}_{-0.85}$	5.38 %
Categorisation	1.23	$1.00^{+0.96}_{-0.82}$	9.18 %

TABLE 5.25: Summary of fit results in alternative approaches comparing to one-bin fit.

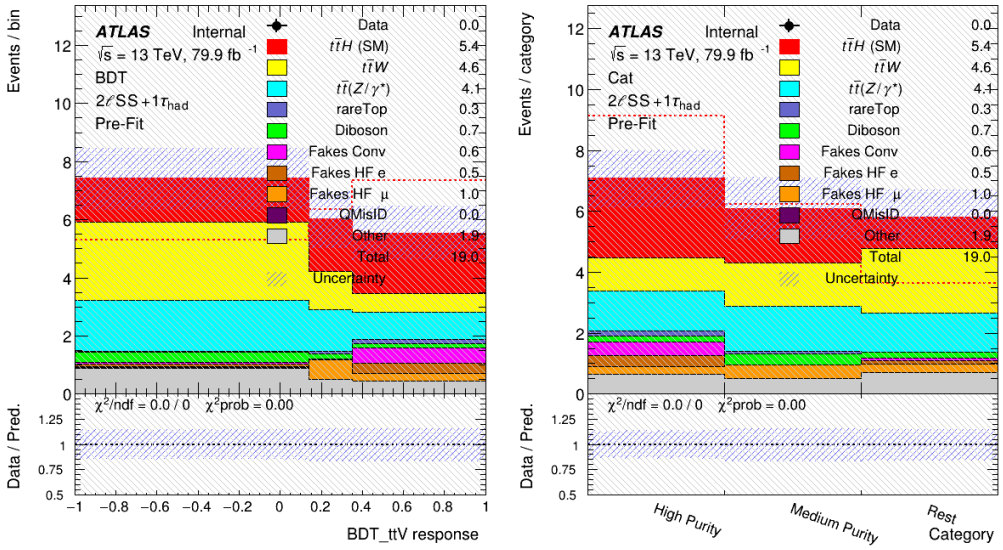


FIGURE 5.37: Prefit plots of BDT (left) and categorisation (right).

Systematic uncertainty	Components
Luminosity (N)	1
Pileup modelling	1
Physics objects	
Electron	8
Muon	11
Tau	7
Jet energy scale and resolution	28
Jet vertex fraction	1
Jet flavour tagging	17
E_T^{miss}	3
Total (Experimental)	77
Data-driven background estimates	
Non-prompt light-lepton estimates (3ℓ, 3ℓ+$1\tau_{had}$)	1
Fake τ_{had} estimates	6
Electron charge misassignment	2
Total (Data-driven reducible background)	9
Template fit uncertainties	
Material conversions	1
Internal conversions	1
HF non-prompt leptons	18
LF non-prompt leptons	2
Total (Template fit)	22

Systematic uncertainty	Components
$t\bar{t}H$**modelling**	
Renormalisation and factorisation scales	3
Parton shower and hadronisation model	1
Higgs boson branching ratio	4
Shower tune	1
PDF	32
$t\bar{t}W$**modelling**	
Radiation	1
Generator	1
PDF	32
Extrapolation	4
$t\bar{t}(Z/\gamma^*)$ **(high mass) modelling**	
Cross section (N)	2
Generator	1
Renormalisation and factorisation scales	3
Shower tune	1
$t\bar{t}$**modelling**	
Radiation	1
WZ **modelling**	
HF composition (N)	3
Shower tune	1
Other background modelling	
Cross section (N)	22
Total (Signal and background modelling)	120
Total (Overall)	218

TABLE 5.26: Sources of systematic uncertainty considered in the analysis. "N" means that the uncertainty is taken as normalisation only for all processes and channels affected. Some of the systematic uncertainties are split into several components, as indicated by the number in the rightmost column. The PDF uncertainties are correlated between the $t\bar{t}H$ signal and the $t\bar{t}W$ background.

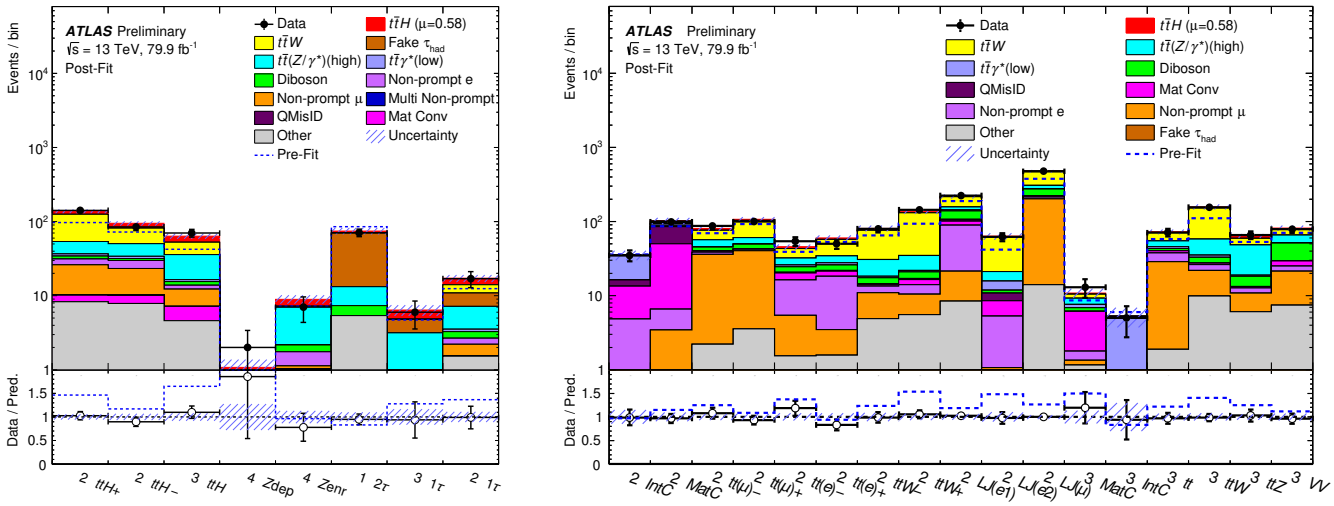
5.4.4.5 Summary

In conclusion, several approaches proposed for 2ℓSS+$1\tau_{had}$ have been tested. Using event counting, instead of advanced technologies such as BDT and categorization, only loses about 10 % sensitivity. The loss is acceptable, especially considering that no shape systematic is needed. The whole unblinded fit (using real data) will be presented in the following section 5.5.

5.5 Results of $t\bar{t}H$ multilepton analysis

In this section, the result of the $t\bar{t}H$ multilepton analysis is presented, following by a comparison to the result from CMS collaboration. The statistic model is the same maximum likelihood fit as introduced in section 4.5.2.1. In total 25 event categories as described in section 5.3.2 are used to determine the $t\bar{t}H$ cross section and the normalization of $t\bar{t}W$ and other backgrounds. Seventeen categories of 2ℓSS and 3ℓ are used as CRs for the template fit to determine and constrain the fake leptons and $t\bar{t}W$, the remaining categories are used as SRs to measure the $t\bar{t}H$ cross section. The systematics taken into account are summarized in table 5.26, including instrumental uncertainties, theoretical uncertainties and template fit associated uncertainties.

The fitted signal strength, $\hat{\mu}$, is calculated by maximising the likelihood function with respect to all parameters and the total uncertainty, while the uncertainty, $\Delta\hat{\mu}$, is obtained from the variation of q_0 $(-2\ln(\mathcal{L}(0,\hat{\hat{\lambda}}_0,\hat{\hat{\theta}}_0)/\mathcal{L}(\hat{\mu},\hat{\hat{\lambda}}_{\hat{\mu}},\hat{\hat{\theta}}_{\hat{\mu}})))$ by one unit from its minimum value. The estimate of the



(A) (B)

FIGURE 5.38: Comparison between data and prediction for the event yields in (a) the eight signal regions and (b) the 17 control regions. The background contributions after the fit ("Post-fit") are shown as filled histograms. The total background before the fit ("Pre-Fit") is shown as a dashed blue histogram. The $t\bar{t}H$ signal, which is scaled according to the results of the fit, is shown as a filled red histogram added to the post-fit background.

impact of the systematic uncertainties is obtained by subtracting in quadrature the statistical uncertainty from $\Delta\hat{\mu}$, which is determined by redoing the fit to data after fixing all NPs to their best-fit values. The contribution from the background NFs is included in the statistical uncertainty. The expected results are obtained in the same way as the observed results but replacing the data in each input bin by the prediction from simulation and the data-driven fake and non-prompt estimates with all NPs set to their best-fit values obtained from the fit to data. The significance is obtained from the test statistic using the asymptotic formulae given in Refs. [115].

Figure 5.38 shows the comparison of data to the predictions after the fit. In all the categories, the predictions agree with the observed data within uncertainties. The observed (expected) significance is 1.8 (3.1) σ. The best-fit value of μ is:

$$\hat{\mu} = 0.58^{+0.26}_{-0.25}\ (\text{stat.})^{+0.19}_{-0.15}\ (\text{exp.})^{+0.13}_{-0.11}\ (\text{bkg. th.})^{+0.08}_{-0.07}\ (\text{sig. th.}) = 0.58^{+0.36}_{-0.33}. \tag{5.12}$$

The best-fit value of μ for each individual channel and the combination of all channels are shown in Figure 5.39. The individual channel results are extracted from the full fit but with a separate parameter of interest for each channel. The probability that the six fitted signal strengths are compatible with a single value is 98%.

Meanwhile, the normalization factors of $t\bar{t}W$ background in 2ℓSS and 3ℓ regions are found to be: $\hat{\lambda}^{2\ell\text{LJ}}_{t\bar{t}W} = 1.56^{+0.30}_{-0.28}$, $\hat{\lambda}^{2\ell\text{HJ}}_{t\bar{t}W} = 1.26^{+0.19}_{-0.18}$, and $\hat{\lambda}^{3\ell}_{t\bar{t}W} = 1.68^{+0.30}_{-0.28}$. The result shows that they are compatible with each other and that observed data prefers a higher $t\bar{t}W$ cross section than the prediction from current theory calculation.

The impact of systematic uncertainties are summarized in table 5.27. The leading systematic uncertainty is jet energy scale and resolution, followed by theoretical modelling uncertainties. The

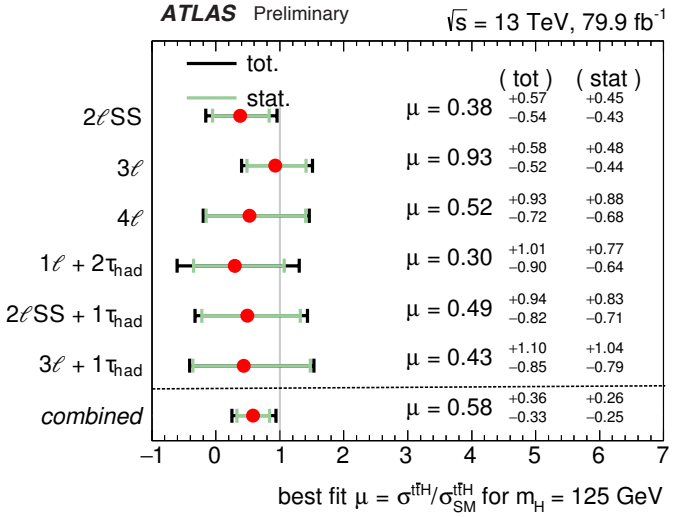


FIGURE 5.39: The observed best-fit values of the $t\bar{t}H$ signal strength μ and their uncertainties by analysis channel and combined. The individual μ values for the channels are obtained from a simultaneous fit with the signal-strength parameter for each channel floating independently. The SM prediction corresponds to μ=1.

measured $t\bar{t}H$ cross section is found to be:

$$\hat{\sigma}(t\bar{t}H) = 294^{+132}_{-127}\ (\text{stat.})^{+94}_{-74}\ (\text{exp.})^{+73}_{-56}\ (\text{bkg. th.})^{+41}_{-39}\ (\text{sig. th.})\ \text{fb} = 294^{+182}_{-162}\ \text{fb}, \tag{5.13}$$

which agrees with the predicted SM cross section $\sigma(t\bar{t}H) = 507^{+35}_{-50}$ fb.

5.5.1 Comparison with CMS results

The CMS collaboration presented the same search for multilepton final states, but requiring at least one hadronically decayed tau on a dataset corresponding to an integrated luminosity of 35.9 fb^{-1}at $\sqrt{s} = 13$ TeV [116]. In the analysis, three channels, $1\ell{+}2\tau_h$, $3\ell{+}1\tau_h$ and $2\ell\text{SS}{+}1\tau_h$, are included which are basically the same as those tau channels in ATLAS $t\bar{t}H$ multilepton analysis. As a result, the combined observed (expected) signal rate is $\mu = 0.72^{+0.62}_{-0.53}$ ($1.00^{+0.67}_{-0.57}$) times the SM $t\bar{t}H$ production rate, with an observed (expected) significance of 1.4 σ (1.8 σ), as shown in figure 5.40.

Another search without requiring the presence of hadronic tau, targeting $H \to WW^*$ and $H \to ZZ^*$, was presented on the same dataset [117]. Three channels, two same-sign leptons, three leptons and four leptons, are included. The best fit $t\bar{t}H$ yield is 1.5 ± 0.5 times the standard model prediction and the observed (expected) significance is 3.3 σ (2.5 σ) as shown in figure 5.40.

A search using data collected during 2017, corresponding to a integrated luminosity of 41.5 fb^{-1}was also presented [118]. The six channels mentioned above and a new final states, $2\ell\text{SS}{+}2\tau_h$, are included in this result. The observed (expected) $t\bar{t}H$ production rate is $0.75^{+0.46}_{-0.43}$ ($1.00^{0.39}_{-0.35}$) times the SM $t\bar{t}H$ expectation, corresponding to a significance of 1.7 σ (2.9 σ), as shown in figure 5.41. This result is also combined with the two previous results. The combined fit obtained an observed (expected) singal rate of $0.96^{0.34}_{-0.31}$ ($1.00^{0.30}_{-0.27}$) times the SM $t\bar{t}H$ production rate, corresponding to a significance of 3.2 σ (4.0 σ).

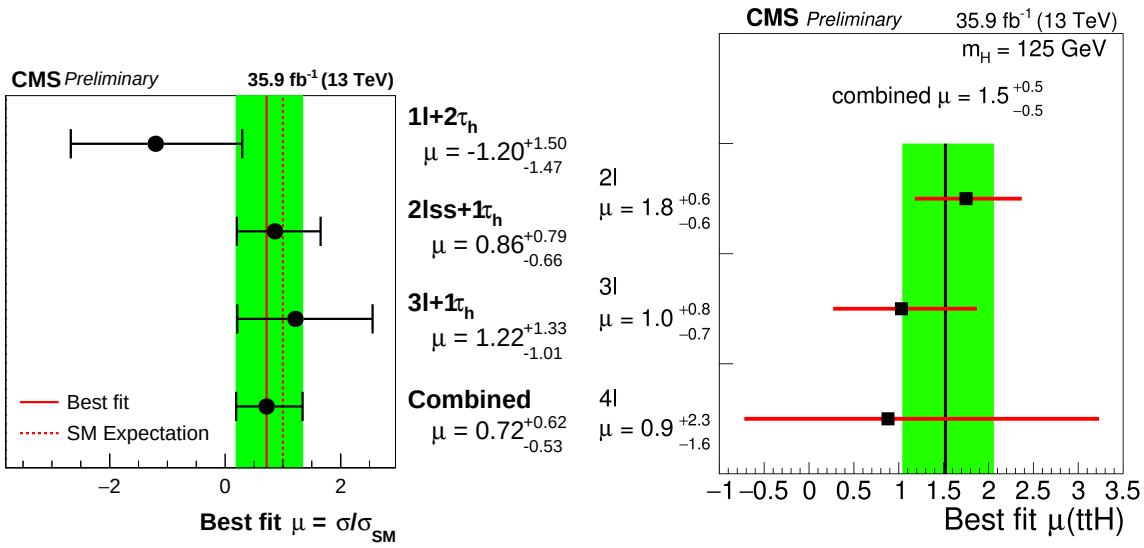


FIGURE 5.40: Signal rates μ, in units of the SM $t\bar{t}H$ production rate, measured in each category individually and for the combination of all three categories. Final states with hadronic taus in left and without hadronic taus in right.

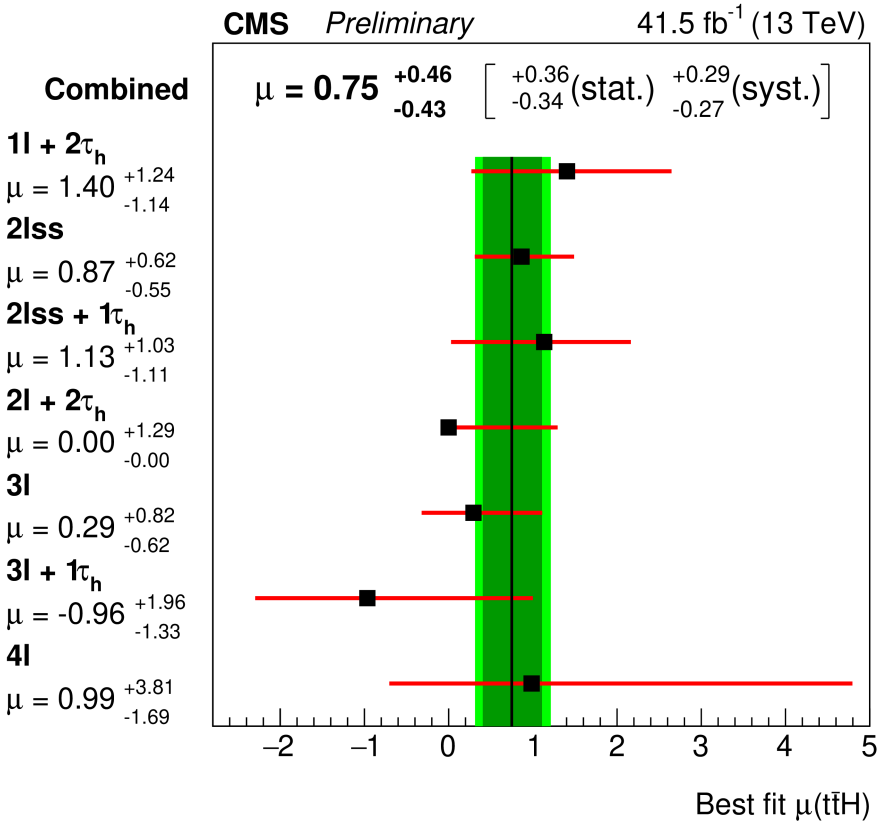


FIGURE 5.41: Signal rates μ, in units of the SM $t\bar{t}H$ production rate, measured using data collected in 2017.

Uncertainty source	$\Delta\hat{\mu}$	
Jet energy scale and resolution	$+0.13$	-0.13
$t\bar{t}(Z/\gamma^*)$ (high mass) modelling	$+0.09$	-0.09
$t\bar{t}W$modelling (radiation, generator, PDF)	$+0.08$	-0.08
Fake τ_{had} background estimate	$+0.07$	-0.07
$t\bar{t}W$modelling (extrapolation)	$+0.05$	-0.05
$t\bar{t}H$cross section	$+0.05$	-0.05
Simulation sample size	$+0.05$	-0.05
$t\bar{t}H$modelling	$+0.04$	-0.04
Other background modelling	$+0.04$	-0.04
Jet flavour tagging and τ_{had} identification	$+0.04$	-0.04
Other experimental uncertainties	$+0.03$	-0.03
Luminosity	$+0.03$	-0.03
Diboson modelling	$+0.01$	-0.01
$t\bar{t}\gamma^*$ (low mass) modelling	$+0.01$	-0.01
Charge misassignment	$+0.01$	-0.01
Template fit (non-prompt leptons)	$+0.01$	-0.01
Total systematic uncertainty	$+0.25$	-0.22
Intrinsic statistical uncertainty	$+0.23$	-0.22
$t\bar{t}W$normalisation factors	$+0.10$	-0.10
Non-prompt leptons normalisation factors (HF, material conversions)	$+0.05$	-0.05
Total statistical uncertainty	$+0.26$	-0.25
Total uncertainty	$+0.36$	-0.33

TABLE 5.27: Breakdown of the contributions to the uncertainties in $\hat{\mu}$. The total statistical uncertainty is evaluated by fixing all the nuisance parameters in the fit except for the free-floating background normalisation factors. The contribution from the uncertainty in the normalization factors is included in the total statistical uncertainty. The statistical uncertainty evaluated after also fixing the background normalisation factors is then indicated as “intrinsic statistical uncertainty". Statistical uncertainties from data-driven background estimates are included in the experimental uncertainties. The other quoted numbers are obtained by repeating the fit after having fixed a certain set of nuisance parameters corresponding to a group of systematic uncertainty sources, and subtracting in quadrature the resulting total uncertainty of μ from the uncertainty from the full fit. The same procedure is followed for quoting the individual effects of background normalisation factors.

Chapter 6

Conclusion

The observation of the Higgs boson brought the last piece of the SM. The search and measurements for different Higgs production modes and the search for new particles are now popular topics. In this thesis, the studies targeting $t\bar{t}H$ to leptons and heavy charged Higgs boson have been presented.

The search for heavy charged Higgs boson is presented in chapter 4. The events are categorized into signal regions and control regions based on the jet and b-jet multiplicities to get optimal sensitivity. Multivariate methods are employed to distinguish H^+ from backgrounds. The dominant background, $t\bar{t}+b\bar{b}$, is irreducible and not well-modelled by simulation. Therefore, $t\bar{t}$ events are classified into $t\bar{t}+ljets$, $t\bar{t}+c\bar{c}$ and $t\bar{t}+b\bar{b}$ based on the flavor of additional jets. $t\bar{t}+c\bar{c}$ and $t\bar{t}+b\bar{b}$ are left free-floating in the fit. An additional re-weighting to SHERPA 2.1.1 is applied on $t\bar{t}+b\bar{b}$ events to get better modelling. A mass reconstruction method based on BDT is developed for low mass H^+. This method shows improvement on the sensitivity, but was not used in the analysis due to its impact on the computing time. The search is performed on 18 mass points, from 200 GeV to 2000 GeV. No significant excess is observed and upper limits on the cross section are set. The parameter space is also constrained.

The search for $t\bar{t}H$ production in multilepton final states is presented in chapter 5. Several sub-channels were considered according to the number of light leptons and hadronic taus. The template fit method is developed to estimate the fake leptons and $t\bar{t}W$ contribution. Multivariate methods are used to extract the $t\bar{t}H$ signal. My contribution focuses on the 2ℓSS+$1\tau_{\text{had}}$ final state. A cut-and-count method is developed and optimized as an alternative approach. The sensitivity reaches basically the same level as BDT. The matrix method is also adapted to 2ℓSS+$1\tau_{\text{had}}$ as an alternative fake lepton estimate. A new systematic study that takes into account the difference of two fake lepton estimates is used, to reduce the non-closure systematic and to take into account the shape effect. Finally a simultaneous fit across all sub-channels is performed to extract the signal strength. The result is found to be:

$$\hat{\mu} = 0.58^{+0.26}_{-0.25}\ (\text{stat.})^{+0.19}_{-0.15}\ (\text{exp.})^{+0.13}_{-0.11}\ (\text{bkg. th.})^{+0.08}_{-0.07}\ (\text{sig. th.}) = 0.58^{+0.36}_{-0.33}. \tag{6.1}$$

And the measured $t\bar{t}H$ cross section agrees with the SM prediction within uncertainties.

LHC has finished run2 operation at the end of 2018. The total pp data collected during run2 corresponds to an integrated luminosity of 140 fb^{-1}. Next data taking operation (run3) will start in 2021 and then LHC will upgrade to HL-LHC in 2026. I was also involved in the development of the new Inner Tracker (ITk) which is designed to replace the current Inner Detector (ID). My task was

to develop a debug tool to check the compability of the Geant4 detector simulation. HL-HLC will collect 3000 fb^{-1} pp collision data. This will allow us to measure the Higgs properties precisely and search for rare decays ($\mu\mu$, $Z\gamma$, $c\bar{c}$). Any deviation of Higgs properties from the SM predictions will be a hint for new physics. On the other hand, there is still room for direct search for new particles. Almost every BSM theory includes additional Higgs bosons, especially charged Higgs bosons. The observation of these new scalar particles would be direct evidence of new physics. We can also improve our understanding of simulation and systematics to get better results. Developing new analysis methods, such as machine learning techniques, is also a way to increase the sensitivity in the future. All these efforts will help us to understand our world better.

List of Figures

List of Tables

Bibliography

[1] Georges Aad et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Lett.* B716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[2] Serguei Chatrchyan et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Phys. Lett.* B716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[3] Georges Aad et al. "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments". In: *Phys. Rev. Lett.* 114 (2015), p. 191803. DOI: 10.1103/PhysRevLett.114.191803. arXiv: 1503.07589 [hep-ex].

[4] Morad Aaboud et al. "Evidence for the associated production of the Higgs boson and a top quark pair with the ATLAS detector". In: *Phys. Rev.* D97.7 (2018), p. 072003. DOI: 10.1103/PhysRevD.97.072003. arXiv: 1712.08891 [hep-ex].

[5] Albert M Sirunyan et al. "Observation of $t\bar{t}H$ production". In: *Phys. Rev. Lett.* 120.23 (2018), p. 231801. DOI: 10.1103/PhysRevLett.120.231801. arXiv: 1804.02610 [hep-ex].

[6] M. Tanabashi et al. "Review of Particle Physics". In: *Phys. Rev. D* 98 (3 Aug. 2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[7] Antonio Pich. "The Standard model of electroweak interactions". In: *The Standard model of electroweak interactions*. [,1(2007)]. 2008, pp. 1–49. arXiv: 0705.4264 [hep-ph]. URL: http://doc.cern.ch/yellowrep/2007/2007-005/cern-2007-005.pdf.

[8] D. de Florian et al. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". In: (2016). DOI: 10.2172/1345634, 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[9] C. Grojean. "Higgs Physics". In: *Proceedings, 8th CERN–Latin-American School of High-Energy Physics (CLASHEP2015): Ibarra, Ecuador, March 05-17, 2015*. 2016, pp. 143–158. DOI: 10.5170/CERN-2016-005.143. arXiv: 1708.00794 [hep-ph].

[10] Georges Aad et al. "Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment". In: (2019). arXiv: 1909.02845 [hep-ex].

[11] Morad Aaboud et al. "Measurement of the Higgs boson mass in the $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ channels with $\sqrt{s} = 13$ TeV pp collisions using the ATLAS detector". In: *Phys. Lett.* B784 (2018), pp. 345–366. DOI: 10.1016/j.physletb.2018.07.050. arXiv: 1806.00242 [hep-ex].

[12] G. C. Branco et al. "Theory and phenomenology of two-Higgs-doublet models". In: *Phys. Rept.* 516 (2012), pp. 1–102. DOI: 10.1016/j.physrep.2012.02.002. arXiv: 1106.0034 [hep-ph].

[13] A. G. Akeroyd et al. "Prospects for charged Higgs searches at the LHC". In: *Eur. Phys. J.* C77.5 (2017), p. 276. DOI: 10.1140/epjc/s10052-017-4829-2. arXiv: 1607.01320 [hep-ph].

[14] Esma Mobs. "The CERN accelerator complex. Complexe des accélérateurs du CERN". In: (July 2016). General Photo. URL: https://cds.cern.ch/record/2197559.

[15] LHC. *LHC luminosity*. 2018. URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2.

[16] Joao Pequenao. "Computer generated image of the whole ATLAS detector". Mar. 2008. URL: https://cds.cern.ch/record/1095924.

[17] Joao Pequenao. "Computer generated image of the ATLAS inner detector". Mar. 2008. URL: http://cds.cern.ch/record/1095926.

[18] Joao Pequenao. "Computer Generated image of the ATLAS calorimeter". Mar. 2008. URL: http://cds.cern.ch/record/1095927.

[19] G. Aad et al. "Commissioning of the ATLAS Muon Spectrometer with Cosmic Rays". In: *Eur. Phys. J.* C70 (2010), pp. 875–916. DOI: 10.1140/epjc/s10052-010-1415-2. arXiv: 1006.4384 [physics.ins-det].

[20] Mark S. Neubauer. "A Fast Hardware Tracker for the ATLAS Trigger System". In: *Particles and fields. Proceedings, Meeting of the Division of the American Physical Society, DPF 2011, Providence, USA, August 9-13, 2011*. 2011. arXiv: 1110.1910 [hep-ex].

[21] A. Ruiz Martínez. "The Run-2 ATLAS Trigger System". In: *J. Phys. Conf. Ser.* 762.1 (2016), p. 012003. DOI: 10.1088/1742-6596/762/1/012003.

[22] G. Aad et al. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *JINST* 3 (2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003.

[23] G. Duckeck et al. "ATLAS computing: Technical design report". In: (2005).

[24] The ATLAS Collaboration et al. "The ATLAS Simulation Infrastructure". In: *The European Physical Journal C* 70.3 (Dec. 2010), pp. 823–874. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-010-1429-9. URL: https://doi.org/10.1140/epjc/s10052-010-1429-9.

[25] S. Agostinelli et al. "GEANT4: A Simulation toolkit". In: *Nucl. Instrum. Meth.* A506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[26] W. Lukas. "Fast Simulation for ATLAS: Atlfast-II and ISF". In: 2012. URL: http://cdsweb.cern.ch/record/1458503/files/ATL-SOFT-PROC-2012-065.pdf.

[27] *The High-Luminosity LHC Project. 298th Meeting of Scientific Policy Committee*. Tech. rep. CERN/SPC/1068. CERN/FC/6014. CERN/3255. June 2016. URL: https://cds.cern.ch/record/2199189.

[28] "Technical Design Report for the ATLAS Inner Tracker Strip Detector". In: (2017).

[29] LHC. *MassDebugger*. 2018. URL: https://twiki.cern.ch/twiki/bin/viewauth/Atlas/G4MassDebugger.

[30] Morad Aaboud et al. "Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s}$ = 13 TeV". In: *Submitted to: Eur. Phys. J.* (2019). arXiv: 1902.04655 [physics.ins-det].

[31] Georges Aad et al. "Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$ =13 TeV". In: *Eur. Phys. J.* C76.5 (2016), p. 292. DOI: 10.1140/epjc/s10052-016-4120-y. arXiv: 1603.05598 [hep-ex].

[32] The ATLAS collaboration. "Measurement of the tau lepton reconstruction and identification performance in the ATLAS experiment using pp collisions at $\sqrt{s} = 13$ TeV". In: (2017).

[33] Morad Aaboud et al. "Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run 2 data collected in 2015 and 2016". In: *Eur. Phys. J.* C79.3 (2019), p. 205. DOI: 10.1140/epjc/s10052-019-6650-6. arXiv: 1810.05087 [hep-ex].

[34] M. Aaboud et al. "Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev.* D96.7 (2017), p. 072002. DOI: 10.1103/PhysRevD.96.072002. arXiv: 1703.09665 [hep-ex].

[35] Morad Aaboud et al. "Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s} = 13$ TeV". In: *JHEP* 08 (2018), p. 089. DOI: 10.1007/JHEP08(2018)089. arXiv: 1805.01845 [hep-ex].

[36] *Performance of missing transverse momentum reconstruction for the ATLAS detector in the first proton-proton collisions at at $\sqrt{s}$= 13 TeV*. Tech. rep. ATL-PHYS-PUB-2015-027. Geneva: CERN, July 2015. URL: http://cds.cern.ch/record/2037904.

[37] T. D. Lee. "A Theory of Spontaneous T Violation". In: *Phys. Rev. D* 8 (4 Aug. 1973), pp. 1226–1239. DOI: 10.1103/PhysRevD.8.1226. URL: https://link.aps.org/doi/10.1103/PhysRevD.8.1226.

[38] John F. Gunion and Howard E. Haber. "The CP conserving two Higgs doublet model: The Approach to the decoupling limit". In: *Phys. Rev.* D67 (2003), p. 075019. DOI: 10.1103/PhysRevD.67.075019. arXiv: hep-ph/0207010 [hep-ph].

[39] J R Andersen et al. "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties". In: (2013). Ed. by S Heinemeyer et al. DOI: 10.5170/CERN-2013-004. arXiv: 1307.1347 [hep-ph].

[40] Marcela Carena et al. "Suggestions for benchmark scenarios for MSSM Higgs boson searches at hadron colliders". In: *Eur. Phys. J.* C26 (2003), pp. 601–607. DOI: 10.1140/epjc/s2002-01084-3. arXiv: hep-ph/0202167 [hep-ph].

[41] N Andari et al. *Higgs Production Cross Sections and Decay Branching Ratios*. Tech. rep. ATL-PHYS-INT-2010-030. Geneva: CERN, Mar. 2010. URL: https://cds.cern.ch/record/1255653.

[42] Georges Aad et al. "Search for charged Higgs bosons in the $H^{\pm} \to tb$ decay channel in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector". In: *JHEP* 03 (2016), p. 127. DOI: 10.1007/JHEP03(2016)127. arXiv: 1512.03704 [hep-ex].

[43] Elin Bergeaas Kuutmann et al. *Search for charged Higgs bosons in the $H^{\pm} \to tb$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector*. Tech. rep. ATLAS-COM-CONF-2016-047. Geneva: CERN, July 2016. URL: https://cds.cern.ch/record/2196988.

[44] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[45] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. "A Brief Introduction to PYTHIA 8.1". In: *Comput. Phys. Commun.* 178 (2008), pp. 852–867. DOI: 10.1016/j.cpc.2008.01.036. arXiv: 0710.3820 [hep-ph].

[46] Richard D. Ball et al. "Parton distributions with LHC data". In: *Nucl. Phys.* B867 (2013), pp. 244–289. DOI: 10.1016/j.nuclphysb.2012.10.003. arXiv: 1207.1303 [hep-ph].

[47] Paolo Nason. "A New method for combining NLO QCD with shower Monte Carlo algorithms". In: *JHEP* 11 (2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040. arXiv: hep-ph/0409146 [hep-ph].

[48] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method". In: *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph].

[49] Simone Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *JHEP* 06 (2010), p. 043. DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581 [hep-ph].

[50] John M. Campbell et al. "Top-Pair Production and Decay at NLO Matched with Parton Showers". In: *JHEP* 04 (2015), p. 114. DOI: 10.1007/JHEP04(2015)114. arXiv: 1412.1828 [hep-ph].

[51] Richard D. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[52] *Studies on top-quark Monte Carlo modelling for Top2016*. Tech. rep. ATL-PHYS-PUB-2016-020. Geneva: CERN, Sept. 2016. URL: https://cds.cern.ch/record/2216168.

[53] Torbjörn Sjöstrand et al. "An Introduction to PYTHIA 8.2". In: *Comput. Phys. Commun.* 191 (2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph].

[54] Michal Czakon and Alexander Mitov. "Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders". In: *Comput. Phys. Commun.* 185 (2014), p. 2930. DOI: 10.1016/j.cpc.2014.06.021. arXiv: 1112.5675 [hep-ph].

[55] Matteo Cacciari et al. "Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation". In: *Phys. Lett. B* 710 (2012), pp. 612–622. DOI: 10.1016/j.physletb.2012.03.013. arXiv: 1111.5869 [hep-ph].

[56] Peter Bärnreuther, Michal Czakon, and Alexander Mitov. "Percent Level Precision Physics at the Tevatron: First Genuine NNLO QCD Corrections to $q\bar{q} \to t\bar{t} + X$". In: *Phys. Rev. Lett.* 109 (2012), p. 132001. DOI: 10.1103/PhysRevLett.109.132001. arXiv: 1204.5201 [hep-ph].

[57] M. Czakon and A. Mitov. "NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels". In: *JHEP* 12 (2012), p. 054. DOI: 10.1007/JHEP12(2012)054. arXiv: 1207.0236 [hep-ph].

[58] M. Czakon and A. Mitov. "NNLO corrections to top-pair production at hadron colliders: the quark-gluon reaction". In: *JHEP* 01 (2013), p. 080. DOI: 10.1007/JHEP01(2013)080. arXiv: 1210.6832 [hep-ph].

[59] Michał Czakon, Paul Fiedler, and Alexander Mitov. "Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $\mathcal{O}(\alpha_S^4)$". In: *Phys. Rev. Lett.* 110 (2013), p. 252004. DOI: 10.1103/PhysRevLett.110.252004. arXiv: 1303.6254 [hep-ph].

[60] D. J. Lange. "The EvtGen particle decay simulation package". In: *Nucl. Instrum. Meth. A* 462 (2001), pp. 152–155. DOI: 10.1016/S0168-9002(01)00089-4.

[61] Fabio Cascioli, Philipp Maierhofer, and Stefano Pozzorini. "Scattering Amplitudes with Open Loops". In: *Phys. Rev. Lett.* 108 (2012), p. 111601. DOI: 10.1103/PhysRevLett.108.111601. arXiv: 1111.5206 [hep-ph].

[62] T. Gleisberg et al. "Event generation with SHERPA 1.1". In: *JHEP* 02 (2009), p. 007. DOI: 10.1088/1126-6708/2009/02/007. arXiv: 0811.4622 [hep-ph].

[63] Fabio Cascioli et al. "NLO matching for $t\bar{t}b\bar{b}$ production with massive b-quarks". In: *Phys. Lett. B* 734 (2014), p. 210. DOI: 10.1016/j.physletb.2014.05.040. arXiv: 1309.5912 [hep-ph].

[64] Morad Aaboud et al. "Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector". In: *Phys. Rev.* D97.7 (2018), p. 072016. DOI: 10.1103/PhysRevD.97.072016. arXiv: 1712.08895 [hep-ex].

[65] Tanju Gleisberg and Stefan Höche. "Comix, a new matrix element generator". In: *JHEP* 12 (2008), p. 039. DOI: 10.1088/1126-6708/2008/12/039. arXiv: 0808.3674 [hep-ph].

[66] Steffen Schumann and Frank Krauss. "A Parton shower algorithm based on Catani-Seymour dipole factorisation". In: *JHEP* 03 (2008), p. 038. DOI: 10.1088/1126-6708/2008/03/038. arXiv: 0709.1027 [hep-ph].

[67] Stefan Höche et al. "QCD matrix elements + parton showers: The NLO case". In: *JHEP* 04 (2013), p. 027. DOI: 10.1007/JHEP04(2013)027. arXiv: 1207.5030 [hep-ph].

[68] Kirill Melnikov and Frank Petriello. "Electroweak gauge boson production at hadron colliders through $\mathcal{O}(\alpha_S^4)$". In: *Phys. Rev. D* 74 (2006), p. 114017. DOI: 10.1103/PhysRevD.74.114017. arXiv: hep-ph/0609070.

[69] Ryan Gavin et al. "FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order". In: *Comput. Phys. Commun.* 182 (2011), pp. 2388–2403. DOI: 10.1016/j.cpc.2011.06.008. arXiv: 1011.3540 [hep-ph].

[70] Ye Li and Frank Petriello. "Combining QCD and electroweak corrections to dilepton production in FEWZ". In: *Phys. Rev. D* 86 (2012), p. 094034. DOI: 10.1103/PhysRevD.86.094034. arXiv: 1208.5967 [hep-ph].

[71] D. Bardin et al. "SANC integrator in the progress: QCD and EW contributions". In: *JETP Lett.* 96 (2012), pp. 285–289. DOI: 10.1134/S002136401217002X. arXiv: 1207.4400 [hep-ph].

[72] A. B. Arbuzov, R. R. Sadykov, and Z. Was. "QED bremsstrahlung in decays of electroweak bosons". In: *Eur. Phys. J. C* 73.11 (2013), p. 2625. DOI: 10.1140/epjc/s10052-013-2625-1. arXiv: 1212.6783 [hep-ph].

[73] Stefano Frixione et al. "Single-top hadroproduction in association with a W boson". In: *JHEP* 07 (2008), p. 029. DOI: 10.1088/1126-6708/2008/07/029. arXiv: 0805.3067 [hep-ph].

[74] Pierre Artoisenet et al. "Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations". In: *JHEP* 03 (2013), p. 015. DOI: 10.1007/JHEP03(2013)015. arXiv: 1212.3460 [hep-ph].

[75] Torbjorn Sjöstrand, Stephen Mrenna, and Peter Z. Skands. "PYTHIA 6.4 physics and manual". In: *JHEP* 05 (2006), p. 026. DOI: 10.1088/1126-6708/2006/05/026. arXiv: hep-ph/0603175.

[76] Peter Zeiler Skands. "Tuning Monte Carlo generators: The Perugia tunes". In: *Phys. Rev. D* 82 (2010), p. 074018. DOI: 10.1103/PhysRevD.82.074018. arXiv: 1005.3457 [hep-ph].

[77] Nikolaos Kidonakis. "Two-loop soft anomalous dimensions for single top quark associated production with a W^- or H^-". In: *Phys. Rev. D* 82 (2010), p. 054018. DOI: 10.1103/PhysRevD.82.054018. arXiv: 1005.4451 [hep-ph].

[78] Nikolaos Kidonakis. "Next-to-next-to-leading logarithm resummation for s-channel single top quark production". In: *Phys. Rev. D* 81 (2010), p. 054028. DOI: 10.1103/PhysRevD.81.054028. arXiv: 1001.5034 [hep-ph].

[79] Nikolaos Kidonakis. "Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production". In: *Phys. Rev. D* 83 (2011), p. 091503. DOI: 10.1103/PhysRevD.83.091503. arXiv: 1103.2792 [hep-ph].

[80] M. Bähr et al. "Herwig++ physics and manual". In: *Eur. Phys. J. C* 58 (2008), p. 639. DOI: 10.1140/epjc/s10052-008-0798-9. arXiv: 0803.0883 [hep-ph].

[81] J. Pumplin et al. "New generation of parton distributions with uncertainties from global QCD analysis". In: *JHEP* 07 (2002), p. 012. DOI: 10.1088/1126-6708/2002/07/012. arXiv: hep-ph/0201195.

[82] Andreas Höcker et al. "TMVA - Toolkit for Multivariate Data Analysis". In: *PoS* ACAT (2007), p. 040. DOI: 10.22323/1.050.0040. arXiv: physics/0703039.

[83] ATLAS Collaboration. "Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC". In: *Eur. Phys. J. C* 76 (2016), p. 653. DOI: 10.1140/epjc/s10052-016-4466-1. arXiv: 1608.03953 [hep-ex].

[84] ATLAS Collaboration. "Measurement of the Inelastic Proton–Proton Cross Section at $\sqrt{s} =$ 13 TeV with the ATLAS Detector at the LHC". In: *Phys. Rev. Lett.* 117 (2016), p. 182002. DOI: 10.1103/PhysRevLett.117.182002. arXiv: 1606.02625 [hep-ex].

[85] ATLAS Collaboration. *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton–proton collision data*. ATLAS-CONF-2016-024. 2016. URL: https://cds.cern.ch/record/2157687.

[86] ATLAS Collaboration. "Jet energy resolution in proton–proton collisions at $\sqrt{s} = 7$ TeV recorded in 2010 with the ATLAS detector". In: *Eur. Phys. J. C* 73 (2013), p. 2306. DOI: 10.1140/epjc/s10052-013-2306-0. arXiv: 1210.6210 [hep-ex].

[87] ATLAS Collaboration. "Jet energy scale measurements and their systematic uncertainties in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev. D* 96 (2017), p. 072002. DOI: 10.1103/PhysRevD.96.072002. arXiv: 1703.09665 [hep-ex].

[88] ATLAS Collaboration. "Performance of missing transverse momentum reconstruction with the ATLAS detector using proton–proton collisions at $\sqrt{s} = 13$ TeV". In: (2018). arXiv: 1802.08168 [hep-ex].

[89] ATLAS Collaboration. "Performance of *b*-jet identification in the ATLAS experiment". In: *JINST* 11 (2016), P04008. DOI: 10.1088/1748-0221/11/04/P04008. arXiv: 1512.01094 [hep-ex].

[90] ATLAS Collaboration. *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*. ATL-PHYS-PUB-2016-012. 2016. URL: https://cds.cern.ch/record/2160731.

[91] ATLAS Collaboration. *Measurement of b-tagging efficiency of c-jets in $t\bar{t}$ events using a likelihood approach wth the ATLAS detector*. ATLAS-CONF-2018-001. 2018. URL: https://cds.cern.ch/record/2306649.

[92] Jon Butterworth et al. "PDF4LHC recommendations for LHC Run II". In: *J. Phys. G* 43 (2016), p. 023001. DOI: 10.1088/0954-3899/43/2/023001. arXiv: 1510.03865 [hep-ph].

[93] Johannes Bellm et al. "Herwig 7.0/Herwig++ 3.0 release note". In: *Eur. Phys. J. C* 76.4 (2016), p. 196. DOI: 10.1140/epjc/s10052-016-4018-8. arXiv: 1512.01178 [hep-ph].

[94] ATLAS Collaboration. *Simulation of top-quark production for the ATLAS experiment at $\sqrt{s} = 13$ TeV*. ATL-PHYS-PUB-2016-004. 2016. URL: https://cds.cern.ch/record/2120417.

[95] A. D. Martin et al. "Parton distributions for the LHC". In: *Eur. Phys. J. C* 63 (2009), pp. 189–285. DOI: 10.1140/epjc/s10052-009-1072-5. arXiv: 0901.0002 [hep-ph].

[96] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *Eur. Phys. J. C* 71 (2011), p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727 [physics.data-an]. Erratum: Eur. Phys. J. C **73** (2013) 2501.

[97] Alexander L. Read. "Presentation of search results: the CL_S technique". In: *J. Phys. G* 28 (2002), p. 2693. DOI: 10.1088/0954-3899/28/10/313.

[98] Thomas Junk. "Confidence level computation for combining searches with small statistics". In: *Nucl. Instrum. Meth. A* 434 (1999), pp. 435–443. DOI: 10.1016/S0168-9002(99)00498-2. arXiv: hep-ex/9902006.

[99] M. Carena et al. "MSSM Higgs boson searches at the LHC: benchmark scenarios after the discovery of a Higgs-like particle". In: *Eur. Phys. J. C* 73.9 (2013), p. 2552. DOI: 10.1140/epjc/s10052-013-2552-1. arXiv: 1302.7033 [hep-ph].

[100] A. Djouadi et al. "The post-Higgs MSSM scenario: habemus MSSM?" In: *Eur. Phys. J.* C73 (2013), p. 2650. DOI: 10.1140/epjc/s10052-013-2650-0. arXiv: 1307.5205 [hep-ph].

[101] A. Djouadi et al. "Fully covering the MSSM Higgs sector at the LHC". In: *JHEP* 06 (2015), p. 168. DOI: 10.1007/JHEP06(2015)168. arXiv: 1502.05653 [hep-ph].

[102] Emanuele Bagnaschi et al. *Benchmark scenarios for low* tan β *in the MSSM*. LHCHXSWG-2015-002. 2015. URL: https://cds.cern.ch/record/2039911.

[103] Vardan Khachatryan et al. "Search for a charged Higgs boson in pp collisions at $\sqrt{s} = 8$ TeV". In: *JHEP* 11 (2015), p. 018. DOI: 10.1007/JHEP11(2015)018. arXiv: 1508.07774 [hep-ex].

[104] S. Frixione, G. Ridolfi, and P. Nason. "A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction". In: *JHEP* 09 (2007), p. 126. DOI: 10.1088/1126-6708/2007/09/126. arXiv: 0707.3088 [hep-ph].

[105] H.-L. Lai et al. "New parton distributions for collider physics". In: *Phys. Rev. D* 82 (2010), p. 074024. DOI: 10.1103/PhysRevD.82.074024. arXiv: 1007.2241 [hep-ph].

[106] Michael H. Seymour and Andrzej Siodmok. "Constraining MPI models using σ_{eff} and recent Tevatron and LHC Underlying Event data". In: *JHEP* 10 (2013), p. 113. DOI: 10.1007/JHEP10(2013)113. arXiv: 1307.5015 [hep-ph].

[107] P. M. Nadolsky et al. "Implications of CTEQ global analysis for collider observables". In: *Phys. Rev. D* 78 (2008), p. 013004. DOI: 10.1103/PhysRevD.78.013004. arXiv: 0802.0007 [hep-ph].

[108] T. Gleisberg et al. "Event generation with SHERPA 1.1". In: *JHEP* 02 (2009), p. 007. DOI: 10.1088/1126-6708/2009/02/007. arXiv: 0811.4622 [hep-ph].

[109] E. Re. "Single-top Wt-channel production matched with parton showers using the POWHEG method". In: *Eur. Phys. J. C* 71 (2011), p. 1547. DOI: 10.1140/epjc/s10052-011-1547-z. arXiv: 1009.2450 [hep-ph].

[110] S. Alioli et al. "NLO single-top production matched with shower in POWHEG: $s-$ and $t-$channel contributions". In: *JHEP* 09 (2009), p. 111. DOI: 10.1088/1126-6708/2009/09/111. arXiv: 0907.4076 [hep-ph].

[111] T. Sjöstrand et al. "High-energy-physics event generation with Pythia 6.1". In: *Comput. Phys. Commun.* 135 (2001), p. 238. DOI: 10.1016/S0010-4655(00)00236-8. arXiv: hep-ph/0010017 [hep-ph].

[112] T. Sjöstrand, S. Mrenna, and P. Skands. "A brief introduction to PYTHIA 8.1". In: *Comput. Phys. Commun.* 178 (2008), p. 852. DOI: 10.1016/j.cpc.2008.01.036. arXiv: 0710.3820 [hep-ph].

[113] Piotr Golonka and Zbigniew Was. "PHOTOS Monte Carlo: A Precision tool for QED corrections in Z and W decays". In: *Eur. Phys. J.* C45 (2006), pp. 97–107. DOI: 10.1140/epjc/s2005-02396-4. arXiv: hep-ph/0506026 [hep-ph].

[114] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *CoRR* abs/1603.02754 (2016). URL: http://arxiv.org/abs/1603.02754.

[115] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *Eur. Phys. J. C* 71 (2011). Erratum: Eur. Phys. J. C **73** (2013) 2501, p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727 [physics.data-an].

[116] CMS Collaboration. "Search for the associated production of a Higgs boson with a top quark pair in final states with a τ lepton at $\sqrt{s} = 13$ TeV". In: (2017).

[117] CMS Collaboration. "Search for Higgs boson production in association with top quarks in multilepton final states at $\sqrt{s} = 13$ TeV". In: (2017).

[118] "Measurement of the associated production of a Higgs boson with a top quark pair in final states with electrons, muons and hadronically decaying τ leptons in data recorded in 2017 at $\sqrt{s} = 13$ TeV". In: (Nov. 2018).