CPPM-T-2017-06

UNIVERSITE D’AIX-MARSEILLE
SHANDONG UNIVERSITY
ECOLE DOCTORALE 352
School of Physics
Centre de Physique des Particules de Marseille

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Physique et Science de la Matière
Spécialité : Physique des Particules et Astroparticules

Chao WANG

Recherche de la production du boson de Higgs associé à une paire de quark top avec le détecteur ATLAS auprès du LHC

Search for the production of the Higgs boson associated with a pair of top quarks with the ATLAS detector at the LHC

Soutenue le 6/12/2017 devant le jury :

Djamel BOUMEDIENNE	LPC	Examinateur
Cristinel DIACONU	CPPM	Directeur de thèse
Cunfeng FENG	SDU	Examinateur
Eric KAJFASZ	CPPM	Examinateur invité
Yanwen LIU	USTC	Rapporteur
Lianliang MA	SDU	Directeur de thèse
Emmanuel MONNIER	CPPM	Directeur de thèse
Zhiqing ZHANG	LAL	Rapporteur
Xuai ZHUANG	IHEP	Examinateur

Numéro national de thèse/suffixe local : 2017AIXM0433/049ED352

CPPM-T-2017-06

SHANDONG UNIVERSITY

AIX MARSEILLE UNIVERSITY

DOCTORALE SCHOOL 352

School of Physics

Centre de Physique des Particules de Marseille

A dissertation to obtain the doctoral degree

Domain : Physics and Matter Science

Specialty : Particle Physics and Astroparticles

Chao WANG

Search for the production of the Higgs boson associated with a pair of top quarks with the ATLAS detector at the LHC

Recherche de la production du boson de Higgs associé à une paire de quark top avec le détecteur ATLAS auprès du LHC

Sustained on 6/12/2017 in front of a jury composed of:

Djamel BOUMEDIENNE	LPC	Examinateur
Cristinel DIACONU	CPPM	Directeur de thèse
Cunfeng FENG	SDU	Examinateur
Eric KAJFASZ	CPPM	Examinateur invité
Yanwen LIU	USTC	Rapporteur
Lianliang MA	SDU	Directeur de thèse
Emmanuel MONNIER	CPPM	Directeur de thèse
Zhiqing ZHANG	LAL	Rapporteur
Xuai ZHUANG	IHEP	Examinateur

Numéro national de thèse/suffixe local : 2017AIXM0433/049ED352

Abstract

The production of the Higgs boson associated with a pair of top quarks is still not observed and is one of the most important Higgs boson production modes. Hence, the discovery of this production mode is a very attracting search: not only will it be the first time we can observe this Higgs production mode but also we will be able to measure its Yukawa coupling to the top quark. The measured results can answer the basic question of the Standard Model (SM) and can also search for any hints of new physics by comparing them to the SM prediction.

An analysis of searching for the production of the Higgs boson associated with a pair of top quarks in three leptons final state is presented in this thesis. This analysis is performed with the collected data by the ATLAS detector in 2015 and 2016 during the so-called Run 2 campaign corresponding to an integrated luminosity of 36.1 fb^{-1} at a center of mass energy of $\sqrt{s} = 13$ TeV. It uses a boosted decision tree algorithm to discriminate between signal and background. The dominant background of fake leptons is estimated with the data-driven method (Matrix Method).

For a Standard Model Higgs boson of a 125 GeV mass, an excess of events over the expected background from other SM processes is found with an observed significance of 2.2 standard deviations, compared to an expectation of 1.5 standard deviations. The best fit for the $t\bar{t}H$ production cross section is $1.5^{+0.8}_{-0.7}$ times the SM expectation, consistent with the value of the Yukawa coupling to top quarks in the Standard Model.

Keywords : LHC, ATLAS, Higgs, $t\bar{t}H$, Matrix Method, Multivariate analysis

Résumé

Introduction

La production du boson de Higgs associée à une paire de quarks top est l'un des modes de production de boson de Higgs les plus importants bien que toujours pas encore observé. Par conséquent, découvrir ce mode de production est l'une des recherches les plus attrayantes après la découverte Higgs: non seulement cela sera la première fois que nous pourrons observer l'existence de ce mode de production du Higgs mais nous pourrons également en mesurer le couplage de Yukawa. Les résultats de ces mesures peuvent répondre aux questions fondamentales du Modèle Standard (SM) et peuvent également donner des indices de nouvelle physique en les comparants à la prédiction SM. Cela fait de l'étude de la recherche et de la mesure de la production l'une des analyses les plus importantes de l'ère post-Higgs.

Plusieurs canaux de désintégration du boson de Higgs ont été considérés pour mesurer le couplage de Yukawa dans la production de ttH avec des états finaux multileptons. Des lots de données de collision proton-proton produits par le grand collisionneur de hadrons, ou Large Hadron Collider (LHC), installé au CERN (Organisation européenne pour la recherche nucléaire) à Genève ont été utilisés. Ils correspondent à des luminosités intégrées par expérience de 5 fb^{-1} à une énergie dans le centre de masse de 7 TeV (2011) et de 20 fb^{-1} à 8 TeV (2012) et représentent le Run 1. Les collaborations ATLAS et CMS, opérant les détecteurs éponymes auprès du LHC, ont chacune produit des résultats de combinaisons basées sur ces données du Run 1. Les significances combinées de deux expériences pour l'observation du processus ttH sont de 4.4σ, alors que seulement 2.0σ est attendu, correspondant à un excès mesuré de 2.3σ par rapport à la prédiction SM. Le LHC produit depuis 2015 des collisions à $\sqrt{s} = 13$ TeV et la section efficace de production ttH augmente d'un facteur quatre par rapport à 8 TeV. La mise à niveau pendant deux ans du détecteur ATLAS a permis d'améliorer ses performances et la précision des analyses physiques, en particulier pour la mesure SM et la recherche de nouvelle physique.

Une analyse de la recherche de la production de boson de Higgs associée à une paire de quarks top dans des états finaux à trois leptons est présentée dans cette thèse. L'analyse est réalisée avec des données collectées par le détecteur ATLAS en 2015 et 2016 à 13 TeV pendant la campagne dite Run 2 et correspondant à une luminosité intégrée de 36.1 fb^{-1}. L'analyse multivariée, dans sa version en arbre de décision renforcé, est utilisée pour distinguer le signal du bruit de fond. Les faux-leptons et les processus $t\bar{t}V$ sont les bruits de fond principaux, en particulier le fond de faux leptons est estimé avec une méthode matricielle (MM) pilotée par les données (Méthode de la Matrice) dans le canal à trois leptons. Un ajustement simultané est utilisé pour estimer les résultats finaux et principalement la force du signal à valider la prédiction SM.

Modèle standard et le détecteur ATLAS

La théorie actuellement en cours, dominante et réussie, qui explique ces particules et champs fondamentaux, ainsi que leur dynamique, s'appelle le Modèle Standard (SM).

Le SM est devenu à une théorie complète décrivant les interactions de base entre les particules et les forces. La gravité, qui devrait probablement être médiée par une particule: le graviton, reste à observer.

Les particules fondamentales du SM sont des fermions et des bosons, avec respectivement des spins demi-entiers ou un entier (y compris zéro). Le spin est un nombre quantique signé intrinsèque à la particule avec une direction. Les fermions constituent la matière conventionnelle dans la nature. Les bosons sont des particules de spin entières et sont les médiateurs des forces fondamentales.

Quatre forces fondamentales sont connues aujourd'hui: les forces forte, faible, électromagnétique et gravitationnelle. Bien qu'il y ait des particules massives, la force gravitationnelle n'a aucune influence au niveau des particules et est généralement ignorée dans le cadre SM. Les forces électromagnétique et gravitationnelle ont une portée infinie et obéissent à une loi carrée inverse. La portée des forces forte et faible peut être estimée par le principe d'incertitude de Heisenberg et les masses des particules porteuses de force, qui sont les gluons de la force nucléaire forte et les bosons W et Z de la force faible.

Les théories modernes décrivent les forces physiques en termes de champs, par exemple le champ électromagnétique, le champ gravitationnel et les champs qui décrivent les forces entre les particules élémentaires. Dans les théories de champs, différentes configurations des champs non observables peuvent donner des quantités observables identiques. Une transformation d'une telle configuration de champ à une autre est appelée une transformation de jauge. L'absence de changement dans les grandeurs mesurables, malgré la transformation du champ, est une propriété appelée invariance jauge et parfois aussi appelée symétrie jauge. Dans SM, la symétrie joue un rôle très important et SM est une théorie de jauge qui est construite en utilisant la symétrie. Un groupe de transformations qui peuvent être effectuées sur des champs est la base de la symétrie et laisse l'invariant lagrangien. Deux types de symétries sont présentés: les symétries globales et les symétries locales. La signification de la symétrie est que les lois de conversation sont obtenues naturellement et les lois de conversation ont les symétries sous-jacentes.

Basé sur l'invariance de jauge locale, un lagrangien de symétrie peut être construit pour décrire l'électrodynamique quantique (QED) et la chromodynamique quantique (QCD). La symétrie de jauge peut garantir que la théorie est re-normalisée et que la théorie de perturbation peut être utilisée pour faire les prédictions. Mais elle interdit le terme de masse supplémentaire pour les bosons W et Z. Pour obtenir la masse, la symétrie de jauge doit être brisée tout en gardant le lagrangien symétrique. Cela peut se faire par le biais de la rupture de symétrie spontanée (SSB) en conduisant naturellement au boson de Higgs.

Les dernières mesures des propriétés du boson de Higgs, découvert avec succès par ATLAS et CMS dans les données du Run1 montrent une compatibilité remarquable avec les prédictions SM, et aucune des collaborations n'a observé d'écarts significatifs par rapport à cette théorie. Le SM est la théorie, qui peut décrire la relation entre les particules et les interactions fondamentales, de loin la plus réussie. L'accord couvre plusieurs ordres de grandeur en sections efficaces et sur une grande variété de processus de référence du SM.

Malgré l'énorme succès du SM, il reste de nombreuses questions non résolues en physique, comme la matière noire, l'énergie noire, les problèmes de hiérarchie, l'asymétrie matière-antimatière ou l'oscillation des neutrinos. De plus, le SM ne tient pas compte de la gravité et il n'y a pas encore de compréhension confirmée expérimentalement de cette force au niveau quantique. Beaucoup d'efforts ont été faits pour combiner la gravité avec le SM pour former une théorie de tout, mais jusqu'ici les tentatives n'ont qu'un succès limité. Des modèles théoriques au-delà du Modèle Standard (BSM) sont nécessaires pour traiter ces problèmes, comme le modèle SUper-SYmétrique dit SUSY.

Le Large Hadron Collider (LHC) est l'accélérateur et collisionneur d'hadrons le plus grand et le plus puissant du monde. Il s'agit de l'ajout le plus récent au complexe d'accélérateurs du CERN situé près de Genève à la frontière entre la France et la Suisse. Le LHC est situé dans un tunnel circulaire de 27 km à une profondeur variant entre 45 m et 170 m sous terre. Le LHC est conçu pour produire des collisions proton-proton jusqu'à une énergie dans le centre de masse de 14 TeV. La première collision proton-proton au LHC s'est produite en novembre 2009 et les premières collisions de 7 TeV ont débuté en 2010. Les collisions de 13 TeV ont débuté en 2015 après deux ans de fermeture pour améliorer le détecteur et le LHC. Quatre expériences principales ont été installées et développées au LHC. ATLAS et le CMS sont les plus grandes expériences et sont conçus pour comprendre le modèle standard avec précision. La physique du quark b est principalement étudiée par l'expérience LHCb et ALICE étudie le mécanisme de confinement des quarks.

ATLAS (A Toroidal LHC Apparatus) est une expérience à usage général conçue pour sonder un large éventail de données physiques allant de la QCD à basse énergie, des mesures de précision électrofaible et sur les quarks top, à la recherche du boson de Higgs et de la nouvelle physique à l'échelle du TeV. ATLAS mesure 44 m de long, 25 m de haut et pèse environ 7 000 tonnes. En commençant par son cœur et en se déplaçant vers l'extérieur, on trouve d'abord le détecteur interne (ID), structuré en trois couches et responsable de la mesure précise de l'impulsion des particules chargées dans un champ magnétique solénoïdal de 2 T. On a ensuite les calorimètres électromagnétiques et hadroniques, qui mesurent les dépôts d'énergie des électrons, des photons et des hadrons et déduisent également l'énergie transversale manquante des neutrinos et d'autres particules non détectées. Enfin on a un spectromètre à muons de précision (MS) qui permet le suivi et le déclenchement des muons dans un champ magnétique toroïdal. La collaboration ATLAS compte environ 3000 personnes provenant de 178 institutions dans 38 pays. ATLAS a enregistré ses premières collisions en 2009, après un arrêt d'un an dû à un accident sur un des aimants supraconducteurs du LHC, et a ensuite fonctionné avec succès jusqu'au début 2013 ou il a été arrêté pour une période de mise à niveau de deux ans. ATLAS a commencé à prendre des données à 13 TeV en 2015 et continuera à fonctionner jusqu'à la prochaine mise à niveau fin 2018.

Recherche de la production du boson de Higgs associé à une paire de quarks top

Le couplage du boson de Higgs à d'autres particules peut être mesuré en étudiant les différents processus de production et de modes de désintégration du boson de Higgs. La production associée d'un boson de Higgs et d'une paire de quarks top peut fournir une information directe du couplage de Yukawa Higgs-Top, probablement le couplage le plus crucial aux fermions. Ce processus implique deux gluons en collision, qui se décomposent en une paire de quark-antiquark. Un quark et un antiquark de chaque paire peuvent ensuite se combiner pour former une particule de Higgs. La mesure du couplage de Yukawa relie deux des particules les plus lourdes du modèle standard, le quark top (173 GeV) et le boson de Higgs (125 GeV). Le couplage Yukawa du boson de Higgs au quark top est un paramètre clé du SM. La comparaison de ces mesures aux prédictions du SM a le potentiel d'identifier sans ambiguïté de nouveaux effets physiques qui peuvent modifier la section efficace de la production de $t\bar{t}H$ par rapport à l'espérance SM.

La production de $t\bar{t}H$ peut être observée à travers différentes topologies selon les désintégrations de Higgs et de quarks supérieurs. L'analyse présentée dans cette thèse est conçue pour rechercher des désintégrations de Higgs en paires de bosons W, le quark top et le quark anti-top se désintégrant en Wb. Chaque boson W se désintègre, soit leptoniquement (ℓ = électron, muon, tau) avec la production d'énergie manquante, soit hadroniquement, dans de nombreuses topologies. Selon le nombre de leptons dans l'état final des événements de signal, les topologies pourraient être avec 2ℓSS (exactement deux leptons légers de mêmes signe de charges et avec un veto sur les taus hadroniques); avec 3ℓ (exactement 3 leptons légers); avec 4ℓ (exactement 4 leptons), et. L'orthogonalité des événements entrant dans les différents canaux est assurée dans la définition des canaux. La recherche de $t\bar{t}H$ dans l'état final avec exactement 3 leptons est présentée dans cette thèse, c'est mon travail principal pendant mon doctorat.

L'analyse présentée dans cette thèse utilise les 36.1 fb^{-1} de données, dites "Run 2", collectées à partir de collision proton-proton enregistrée par le détecteur ATLAS à $\sqrt{s}$ = 13 TeV entre 2015 et 2016. Les données sont ensuite préparées en utilisant un cadre dédié de réduction et compactification dédié, appelé dérivation. Cette dérivation fournit une réduction spécifique pour les événements ttH avec des multileptons dans les états finaux. La taille totale de l'ensemble de données a été réduite à 3,6% du total disponible pour les événements ttbars simulés et à 0.1% pour les ensembles de données de collision.

Tous les lots de données simulées par Monte Carlo sont traités par une simulation complète de la réponse du détecteur ATLAS basée sur Geant4. Tous les événements simulés ont ensuite été traités en utilisant les mêmes algorithmes de reconstruction et la même chaîne d'analyse que les données réelles de collisions. Les événements simulés sont corrigés afin que les efficacités de reconstruction et d'identification des objets, les échelles d'énergie et les résolutions d'énergie correspondent à celles déterminées à partir des données.

Après les sélections de base décrites ci-dessus, au niveau de l'événement, le vertex primaire avec l'exigence d'être celui ayant les traces associées de plus haut p_T est requis. Les événements avec un bruit important dans les calorimètres ou avec de données corrompues sont supprimés et une procédure appelée suppression de chevauchement est conçue pour éviter le double comptage d'objets et pour supprimer les leptons ayant une forte probabilité de provenir des désintégrations de hadrons. Deux stratégies de déclenchement différentes ont été adoptées en 2015 et 2016 pour tenir compte des différents schémas de collision pour chaque période.

Tous les critères de sélections des candidats leptons sont regroupés et appelées objets de niveau lâche. Les objets de niveau serré doivent, sous certaines conditions, satisfaire au point de fonctionnement de l'algorithme de maximum de vraisemblance (LH) utilisé pour les électrons afin de réduire les électrons faux ou non direct. Un algorithme dédié pour l'isolation, par arbre de décision renforcé et produisant la variable `PromptLeptonIso`, a été construit pour améliorer la réduction du bruit de fond de leptons non prompts produits dans les désintégrations de hadrons.

Les jets sont reconstruits à partir d'amas topologiques calibrés construits à partir des dépôts d'énergie dans les calorimètres, en utilisant l'algorithme anti-kt. Les jets ayant des contributions énergétiques susceptibles d'être générés par les effets de bruit ou de détecteur sont supprimés et seuls les jets satisfaisant une impulsion transverse, p_T, supérieure à 25 GeV sont conservés. Les jets contenant des b-hadrons sont identifiés par une méthode discriminante multivariée. Le point de fonctionnement utilisé dans cette recherche correspond à un rendement moyen de 70% pour b-tagged jets de p_T> 20 GeV.

Dans l'état final à trois leptons, la valeur absolue de la charge totale des trois leptons dans les événements doit être 1, aucun candidat tau n'est autorisé. Parmi ces trois leptons, celui de charge opposée à celles des deux autres est appelée "lepton 0". Le bruit de fond résultant de l'addition d'un lepton faux ou non prompt à un événement di-lepton de charges opposées aura un lepton supplémentaire comme "Lepton 1" ou "Lepton 2" parce que ce sera la même charge qu'un prompt lepton. En plus des exigences strictes sur "Lepton 1" et "Lepton 2", un $p_T > 15$ GeV pour les deux leptons est également requis. L'exigence supplémentaire de $p_T > 10$ GeV pour le "Lepton 0" est plus faible que "Lepton 1" et "Lepton 2", ainsi que l'isolation et l'identification du lepton faites avec un point de fonctionnement lâche. D'autre recommandations sont également appliquées pour éliminer les leptons de désintégration de quarks et les bruits de fond potentiels contenant des désintégrations de Z en leptons où un des leptons a une impulsion très faible et n'est pas reconstruit.

Les jets et les jets b-étiquetés sont contraint d'avoir au moins 2 jets et au moins 1 jets b-étiqueté.

Estimation des faux leptons

Plusieurs processus physiques peuvent produire la même signature que le signal. Une bonne estimation de ces bruits de fond est un des aspects les plus importants pour cette analyse. Les bruits de fond dans cette analyse sont classés en deux catégories:

Irréductible : Les événements qui peuvent conduire au même état final que le signal et l'état final contiennent trois leptons chargés. Les processus $t\bar{t}W$, $t\bar{t}Z$ et di-boson sont dans cette catégorie. L'estimation de ces fonds irréductibles s'appuie sur les simulations Monte Carlo.

Réductible : La plupart du temps, les événements pour lesquels un lepton non prompt ou un lepton faux est sélectionné comme lepton prompt sont appelés bruit de fond réductible. Les processus peuvent conduire à un état final compatible avec le signal avec l'objet mal reconstruit. Les principaux processus sont $t\bar{t}$, Z + jets.

La méthode basée sur les données (la méthode matricielle dans cette analyse) est utilisée pour obtenir l'estimation du fond réductible. A titre de référence, les résultats du Run 1 montrent que les bruits de fond dominant sont les processus non-prompt (faux), le processus $t\bar{t}V$ et dans une moindre mesure les contributions di-bosons VV.

En raison de la résolution limitée du détecteur, les objets ne sont pas reconstruits parfaitement. L'un des objets mal reconstruits, les faux leptons, est un problème commun dans les analyses. Notamment pour l'analyse qui implique des objets leptoniques comme l'étude présentée dans cette thèse. Les principales origines de ces faux leptons sont: la conversion des photons, et les jets de saveurs légères et lourdes.

Une étude spécifique de l'origine des faux leptons a été réalisée dans l'analyse 3ℓ pour les faux électrons et les muons en utilisant les informations de vérité de lot d'événements $t\bar{t}$ MC. En ce qui concerne la composition de l'origine des faux leptons dans la région de présélection de 3ℓ, les résultats montrent que les faux issus de saveurs lourdes (faux leptons qui proviennent des mésons b et c) dominent à la fois les faux électrons et les faux muons; une contribution importante des conversions de photons produisant des faux électrons, mais aucune contribution de ce type dans le cas des faux muons comme prévu.

La méthode matricielle (MM) est une technique basée sur les données pour avoir l'estimation de la fausse contamination dans l'analyse. L'estimation ne peut pas décrire les faux objets en utilisant les échantillons MC seulement car le MC n'est pas fiable dans tous les cas et il y a encore beaucoup de cas non prévus dans le jeu de données réelles. Cette technique basée sur les données peut utiliser les informations provenant d'un jeu de données de collision réel et élimine la dépendance au MC. La MM est une méthode basée sur les données largement utilisée dans les analyses d'ATLAS. Pour avoir une bonne estimation du bruit de fond de faux, la MM est donc également utilisé dans l'analyse à trois leptons dans l'état final. La MM y est utilisé pour estimer les faux électrons et les muons dans une région à contrainte serrée en utilisant l'information d'objet lâches dans une région à contrainte anti-serrée.

Les faux leptons sont estimés avec une matrice 2×2 dans l'analyse 2ℓSS. Une méthode matricielle similaire a donc été également développée pour l'analyse à 3ℓ mais avec une matrice 8×8. La méthode matricielle simplifiée dans 3ℓ est introduite avec une assomption: le "Lepton 0", qui est de signe opposé, a une très faible possibilité d'être le faux lepton. La simulation MC $t\bar{t}$ est utilisée pour vérifier cela avec une pré-sélection de 3ℓ. Il a été trouvé que la possibilité pour le "Lepton 0" d'être le faux dans la région de signal pré-MVA 3ℓ est seulement de 1%.

L'échantillon ttbar simulé est utilisé pour la mesure réelle et l'extraction des fausse pendant le test de cohérence. Un échantillon de $t\bar{t}$ dédié généré avec un filtre dileptonique est utilisé pour le cas 3ℓ pour améliorer les statistiques. En ce qui concerne le rayonnement de photon supplémentaire, un autre échantillon ciblant spécifiquement la production de $t\bar{t}$ avec un photon rapide supplémentaire rayonné par l'un des quarks top a été généré et utilisé.

Deux régions de contrôle, enrichies en vraies et faux leptons, ont été conçues pour mesurer l'efficacité des leptons réels et faux à passer les exigences de sélection serrées. Ces régions ont des statistiques suffisamment grandes et peuvent représenter la cinématique et la composition du bruit de fond dans la région de signal. Les critères de sélection des di-leptons assurent l'orthogonalité à la région de signal 3ℓ.

Un test de fermeture dédié est effectué dans l'analyse 3ℓ pour vérifier la performance de la MM avec la sélection pré-MVA. Les résultats montrent que l'accord est bon entre la prédiction MC et l'estimation MM eut égard aux statistiques faibles et les grandes incertitudes.

Méthode d'analyse multivariée

La technique d'analyse multi-variable a été utilisée dans l'expérience ATLAS pour de nombreuses analyses. En particulier pour l'analyse présentée dans cette thèse, les arbres de décision renforcés (BDT) sont optimisés pour rejeter davantage le faux fond réductible qui est principalement du ttbar et le fond irréductible qui sont des $t\bar{t}W$ et $t\bar{t}Z$. La sélection de base pré-MVA est requise et la contribution dominante provient du bruit de fond de faux et des $t\bar{t}V$. Pour les supprimer, une stratégie d'analyse BDT à deux dimensions a été est conçue pour augmenter le nombre d'événements de signal dans la région pré-MVA. Un ensemble de variables choisies est utilisé comme entrée du BDT. L'accord entre les données et les échantillons MC pour ces entrées a été mesuré comme bon et montrent le pouvoir discriminant prometteur entre les événements de signal et de fond.

Un algorithme d'arbre de décision renforcé par gradient (BDTG) est entrainé sur un lot de données simulées pour le fond irréductible et pour l'estimation basée sur les données avec un total de 10 variables. Le bruit de fond de faux estimé avec la méthode s'appuyant sur les données a été utilisé pendant l'entrainement du BDTG au lieu du MC. Les études ont été faites pour trouver une bonne combinaison de ces deux BDTG. Enfin le BDTG combiné discriminant a été ensuite divisé en six zones pour optimiser le signal et le fond en utilisant la fonction TransfoD appelée auto-binning. Généralement, le BDT combiné a un bon accord entre les données et le MC et aussi une bonne séparation pour les événements de signal.

Résultats

La distribution de sortie BDT combiné et segmentée, "binée" est utilisé comme entrée de l'ajustement. Toutes les incertitudes systématiques associées (systématique expérimentale, systématique théorique et incertitudes statistiques) sont considérées. Les bruits de fond de faux issu des données, $t\bar{t}V$ et dibosons sont contraints simultanément par l'ajustement.

Les résultats attendus sont d'abord réalisés avec l'ensemble de données Asimov. Puis, l'ensemble total des données est utilisé pour les résultats finaux. La valeur attendue de l'intensité du signal $t\bar{t}H$ qui a été obtenue pour le canal 3ℓ, en ajustant les données d'Asimov, est de: $1^{+0.77}_{-0.70}$, la signification correspondante est de 1.45 écart-type avec 1.21 d'incertitude. Les paramètres de nuisance (NP) correspondant aux incertitudes systématiques sont tous centrés sur zéro et les facteurs d'échelle de normalisation sont tous centrés autour de 1 comme prévu. Surtout en 3ℓ, les NPs de haut rang sont ceux qui sont liés à l'estimation de faux, comme la non-fermeture des faux en 3ℓ, la statistiques de faux muon, et aussi la différence entre 2ℓ et 3ℓ à cause des fausses conversions de photons. Généralement, il n'y a pas de problème de sur-contrainte dans le canal 3ℓ.

L'ajustement a été effectué pour les données dans une région de signal et les résultats sont: la force du signal observée est de $1.5^{+0.8}_{-0.7}$, la signification correspondante a atteint 2.18 écarts-types avec l'incertitude 1.48. L'amélioration est grande par rapport aux résultats du Run 1 en raison de la section efficace plus élevée et une stratégie d'analyse avancée et optimisée.

ATLAS a publié les résultats approuvés avec 36.1 fb^{-1} dans les canaux multileptons incluant celui à 3ℓ. Plus particulièrement dans le canal 3ℓ, le résultat publié utilise une stratégie d'analyse multi-classe, aussi basée sur une analyse multi-variée. La plus grande différence entre la méthode utilisée dans cette thèse, utilisée comme vérification dans le résultat publié, et cette version multi-classe retenue pour ce résultat est la procédure de catégorisation appliquée sur la sortie BDT finale pour avoir une région de signal et de contrôle pure. L'analyse multi-classes a une région de signal et quatre autres régions de contrôle qui couvrent les processus $t\bar{t}W$, $t\bar{t}Z$, fakes et diboson. Toutes ces régions sont basées sur la sortie de l'entrainement d'un BDT 5D.

La valeur la mieux ajustée de la force du signal observé (attendu), combinant tous les canaux multileptons, est de $1.54^{+0.49}_{-0.42}(1.00^{+0.43}_{-0.39})$. La significance observée (attendue) dans l'hypothèse de bruit de fond seul est de 4.1 écarts-types (2.7 écarts-types). Cet excès correspond à l'évidence de la production de ttH dans les états finaux à plusieurs leptons. Les résultats présentés par ATLAS sont compatibles avec les résultats présentés dans cette thèse avec même une légère amélioration obtenue avec l'analyse multivariée utilisée.

Conclusion

Le boson de Higgs a été découvert en 2012 par ATLAS et CMS. Cette dernière pièce du modèle standard trouvée, la structure du modèle standard est presque terminée. Mais encore beaucoup de questions reste à répondre, le couplage de Higgs-Top Yukawa en est l'une d'entre elle.

Des mesures précises sur ses propriétés confirmeront la nature du boson de Higgs, et tout écart par rapport à la prédiction du modèle standard représentera un signe clair d'une nouvelle physique. La production associée du boson de Higgs avec une paire de quarks top en permet une mesure directe et est le sujet principal de cette dissertation.

Le travail présenté dans cette thèse porte sur la recherche de la production de bosons de Higgs associée à une paire de quarks top avec un état final à trois leptons.

Une technique multidimensionnelle (BDT) est employée pour avoir un bon pouvoir de discrimination entre le signal et le bruit de fond. L'un des principaux bruits de fond du aux faux leptons est estimé avec une méthode matricielle s'appuyant sur les données. Une bonne modélisation de ce fond et une meilleure séparation avec la sortie BDT sont obtenues. Les résultats finaux sont obtenus avec un ajustement simultané sur un BDT 2D combiné.

Pour une masse du boson de Higgs de 125 GeV, un excès d'événements par rapport au bruit de fond attendu provenant d'autres processus du modèle standard est observé avec une significance observée de 2.18 écarts-types, contre une espérance de 1.45 écart-type. Le meilleur ajustement pour la section efficace de production de $t\bar{t}H$, en supposant une masse de boson de 125 GeV de Higgs, est $1.5^{+0.8}_{-0.7}$ fois la valeur SM, et est compatible avec la valeur du couplage Yukawa au quark top dans le Modèle standard. Les résultats sont compatibles avec les résultats publiés d'ATLAS en utilisant le même ensemble de données mais avec une stratégie d'analyse différente pour les états finaux à trois leptons et des améliorations peuvent être introduites dans l'analyse présentée dans cette thèse.

Comparé aux résultats de Run 1, une grande amélioration a été obtenue à la fois sur la significance et sur l'incertitude sur la mesure de l'intensité du signal. Cela est dû entre autre à l'énergie élevée dans le centre de masse qui apporte une plus grande section efficace, aux améliorations apportées au détecteur ATLAS, ainsi qu'à la stratégie d'analyse optimisée. Des incertitudes toujours importantes sur les mesures finales sont tout de même observées et nécessitent une méthode ré-optimisée pour réduire l'impact de ces sources d'incertitude. Le résultat prometteur montre l'évidence de la production de $t\bar{t}H$ mais plus de données seront nécessaires pour une mesure plus précise.

Après le redémarrage du LHC en 2015, le LHC a fourni davantage de données et plus de possibilités pour valider les prédictions du SM et rechercher de nouvelles physiques. Des mises à niveau du détecteur ATLAS et de ses sous-systèmes ont été apportées de façon limitées. En 2018, le LHC sera arrêté pour un nouveau cycle d'amélioration et d'ATLAS fera de même en apportant entre autre des améliorations au système de déclenchement afin de faire face à des taux de déclenchement plus élevés du à l'accroissement de la luminosité instantanée qu'aura le LHC. Après un redémarrage en 2019, le LHC fonctionnera pendant trois ans à un débit de 14 TeV et recueillera 300 fb^{-1} de données.

L'arrêt définitif du LHC se fera en 2023, période au cours de laquelle ATLAS installera un nouveau système de trajectographie intérieur plus résistant aux rayonnements et plus granulaire. Un nouveau schéma de déclenchement, ainsi que des améliorations de l'électronique du spectromètre à muon et du calorimètre, seront réalisés. Le LHC à haute luminosité amélioré (HL-LHC) débutera en 2025, permettant au collisionneur de produire 3000 fb^{-1} de données. Les nouveaux jeux de données permettront alors des mesures des propriétés du boson de Higgs beaucoup plus précises afin de valider les prédictions par SM.

Acknowledgements

The thesis writing is almost finished when I am wiring this acknowledgements and many previous memories come to eyes. I enjoyed the three years PhD studying life in Marseille which is a beautiful coastal city with lots of friends. I like the sea and enjoy the sunshine here. The three years of PhD studying is the most unforgettable memory in my whole life. I spent almost the whole time of PhD studying in CPPM as a co-PhD student between Aix-Marseille University and Shandong University.

Firstly, I should say thank to my supervisor from Shandong University, Prof. Lianliang Ma and the supervisor from CPPM, Dr. Emmanuel Monnier. They provide me the opptunity of studying the projects I am interested in and give me the chance to do this research in CPPM which has lots of experience on particle physics. As the leader of the ATLAS analysis team in Shandong University, Prof. Lianliang Ma taught me many things on data analysis. Dr. Emmanuel Monnier is as a coordinator of the LAr detector team and helped me a lot understand the knowledge on hardware. Besides, Dr. Emmanuel is a so kind person who offers a lot helps in both studying and ordinary life.

I should say thanks to Dr. Cristineal Diaconu who is my co-supervisor in CPPM. He guided me into the exact data analysis work during the master stage and we worked together when I did my qualification task. I learnt a lot from him and am benefit a lot from his working style.

Sincerely thanks to Prof. Cunfeng Feng who leads me into the experimental high energy physics and guides me the whole master stage. He let me know this beautiful field and provided me the chance to access the LHC as well as the ATLAS experiment.

Many thanks to the Dr. Kun Liu who is the convenor of the analysis team and colleague at the same time. His charming ideas and helpful advice helped me when I am confused.

Thanks to Jian Liu and Ruiqi Zhang who are my friends in Marseille. I got lots of experience on hardware and analysis from them. Also we helped each other a lot in ordinary life.

Special thanks to Bo Liu, Xin Xia, Yongke Zhao, Mengting Yang who are my friends and colleagues in Shandong University and they are so kind friends who I met.

Thanks to Venugopal Ellajosyula and Robert Wolff who are my colleagues in CPPM and cooperators of the analysis.

Finally, I should thank to my parents who support me all the time.

Many people and friends helped me a lot and encouraged me during those three years and I could not express my thanks any more to them who I do not mention here.

Contents

List of Figures

List of Tables

Introduction

The production of the Higgs boson associated with a pair of top quarks is one of the most important Higgs boson production modes and it is still not observed so far. Therefore, to discover this production mode is one of the most charming searches after the discovery of the Higgs boson: not only for the first time we can observe the existence of this Higgs production mode but also we can measure the Yukawa coupling of the Higgs boson to the top quark. The measured results can answer the basic question of the Standard Model (SM) and can also search for any hints of the new physics by comparing to the SM prediction. This makes the study of production searching and measurement to be one of the most important analyses in Run 2 operation.

In the ATLAS and CMS experiments several Higgs boson decay channels were considered in Run 1 to measure the top Yukawa coupling within $t\bar{t}H$ production: $H \to \gamma\gamma$, $H \to bb$, and $H \to$ multileptons final states. Collision datasets corresponding to integrated luminosities per experiment of 5 fb^{-1} at $\sqrt{s} = 7$ TeV (recorded in 2011) and 20 fb^{-1} at $\sqrt{s} = 8$ TeV (recorded in 2012) were employed. The ATLAS and CMS both have the individual results based on these Run 1 data. The combined significance of two experiments for the observation of the $t\bar{t}H$ process is 4.4σ, whereas only 2.0σ is expected, corresponding to a measured excess of 2.3σ with respect to the SM prediction.

The Large Hardon Collider (LHC) has entered the $\sqrt{s} = 13$ TeV energy scale since 2015 and the cross section of $t\bar{t}H$ increases by a factor of 4 compared to 8 TeV. The two years upgrade of the ATLAS detector brings better improvements to the physics analysis, especially for the SM measurement and new physics searching.

The analysis of searching for the production of Higgs boson associated with a pair of top quarks in the three leptons final state is presented in this thesis. The collected data by the ATLAS detector corresponding to an integrated luminosity of 36.1 fb^{-1} at a centre of mass energy of $\sqrt{s} = 13$ TeV is employed. The multi-variate analysis (Boosted Decision Tree) is employed to discriminate between signal and background. Fake leptons and $t\bar{t}V$ processes are main backgrounds, especially the fake leptons background is estimated with the data-driven method (Matrix Method). A simultaneous fit is used to estimate the final result which is mainly the signal strength ($\mu_{t\bar{t}H}$) to validate the SM prediction.

The overview of the particle physics and theoretical framework of the Standard

Model are presented in Chapter 1. Besides, the measured results of $t\bar{t}H$ production from the ATLAS and CMS with full Run 1 data are also discussed in this chapter.

The basic knowledge of the LHC and ATLAS experiment is introduced in Chapter 2. Since the importance of the physical objects which appear in the collision are reconstructed by the ATLAS detector, the definition of physical objects and the corresponding reconstruction algorithms are detailed in Chapter 3.

The analysis of searching for the Higgs boson associated with a pair of top quarks in the three leptons final state is presented in Chapter 4, where the detailed physical objects, the definition of the signal region, the estimation of fakes as well as the multi-variate optimization study are introduced.

Chapter 5 shows the results and the interpretations.

The Chapter Conclusion summaries the analysis results and gives the prospects of this analysis and the particle physics.

1 The Standard Model

Human beings are around with different kinds of matter, and keep thinking about the nature of those matter all the time. The idea that all matter is composed of elementary particles can date back to at least the 6th century BC. But those early ideas are simple and most of them were founded with only imaginations. Many different theories were proposed and tested over the centuries. Electron was discovered between 1879 and 1897 and became the first discovered elementary, truly fundamental particle. Since then particle physics as one of the basic part of the physics came to the public and affects us from time to time. The currently dominant and successful theory explaining these fundamental particles and fields, along with their dynamics, is called the Standard Model (SM). This model was built with many generation physicists' effort. Modern particle physics generally investigates the Standard Model and its various possible extensions, from the newest "known" particle, the Higgs boson, even to the oldest known force field, gravity. The following sections are the introduction to the Standard Model.

1.1 The Standard Model particles

Since the first particle of the SM, electron, is discovered, more and more elementary particles are found during the recent 100 years. The SM is come to a complete theory of describing the basic interactions between particles and forces. The SM describes the fundamental particles which are listed in Figure 1.1.

The fundamental particles of the SM are fermions and bosons, with half-integer and integer (including zero) spin, respectively. Spin is an intrinsic quantum number of a particle, and it has a sign (a directional quantity). Fermions make up conventional matter in the nature. Bosons are particles with integer spin and they are the mediators of the fundamental forces. The fermions are naturally organized into three families/generations of increasing mass. All known stable matter is actually made up of only first generation fermions, electrons, up-quarks and down-quarks. The second and third generation fermions are unstable and decay into lighter particles.

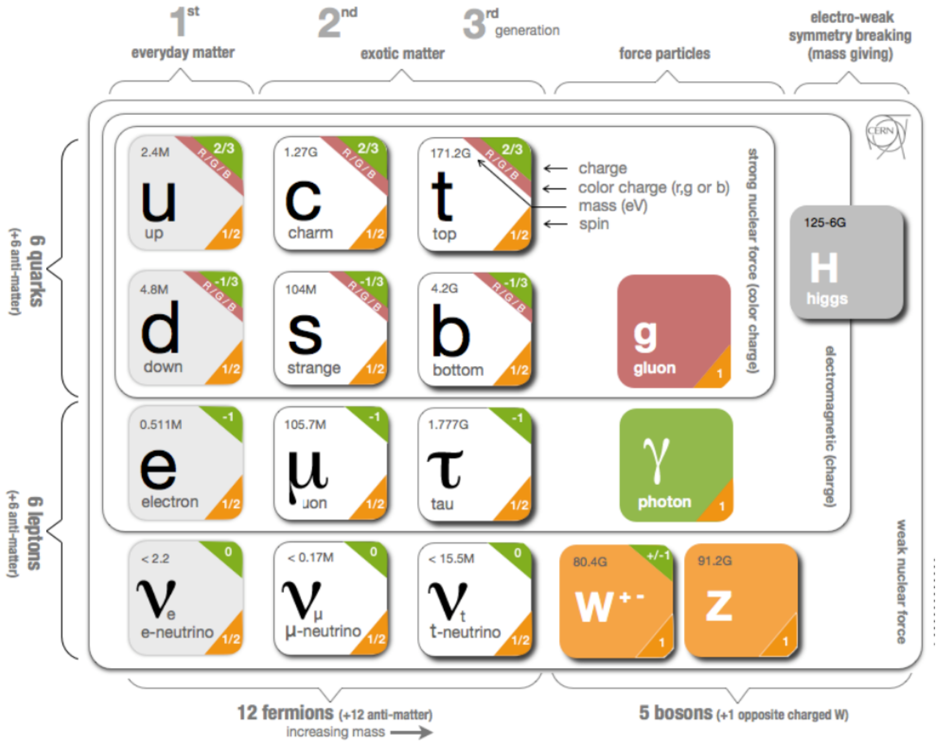


Figure 1.1: Standard Model of elementary particles.

Leptons are the elementary, half spin particle which do not undergo the strong interactions. Two sets of leptons exist: charged leptons (electron, muon, tau(τ)) and neutral leptons (neutrinos). There are six types of leptons divided by the flavours, forming three generations. The first generation is the electronic leptons, including the electron and electron neutrino; the second is the muonic leptons, comprising the muon and muon neutrino; and the third is the tauonic leptons, as the τ and the τ neutrino. Leptons have various intrinsic properties, including electric charge, spin, and mass. Leptons are not subject to the strong interaction, but they are subject to the other three fundamental interactions: gravitation, electromagnetism (excluding neutrinos, which are electrically neutral) and the weak interaction. Leptons are an important part of the Standard Model, e.g., electrons are one of the components of atoms, alongside protons and neutrons.

Quarks are spin-$\frac{1}{2}$ particles, which are fermions, and the constituent of the proton and neutron. They carry fractional charges and there exist six different types of quarks (up (u), down (d), charm (c), strange (s), top (t), bottom (b)) which are commonly referred to as flavours. Quarks have an intrinsic property named color, that can be viewed as a charge that is conserved in the SM, just like the electric charge. Each flavour of quark comes in three different colors: red (R), green (G) and blue (B), tripling the actual number of quarks in the SM. Quarks can experience all four fundamental interactions. Due to a phenomenon known as color confinement, quarks are never directly observed or found in isolation; they can be found only within hadrons, such as baryons and mesons.

Gauge bosons are force carriers. All known gauge bosons have a spin of 1. Therefore, all known gauge bosons are vector bosons. Different interactions are with different gauge bosons as the force medium. Such as photons carry the electromagnetic interaction, W and Z bosons carry the weak interaction and gluons carry the strong interaction. Isolated gluons do not appear because they are color charged and subject to color confinement. Gluons are bicolored, carrying one unit of color and one unit of a different anticolor. A total of eight gluons exists with different colors. All the particles except photon (neutrino does have mass as the phenomenon of neutrino oscillations) interact with Higgs filed and get the mass because of the Brout-Englert-Higgs mechanism. The Higgs boson is one of the quantum excitations of the Higgs field and discovered by ATLAS and CMS in 2012. More details are in Section 1.3.

1.1.1 Fundamental forces

Forces affect our daily life, like gravitational and electromagnetic forces. They affect the nature and life since the Universe was born. Four fundamental forces are known so far: strong, weak, electromagnetic and the gravitational force. These forces together with their approximate range of interaction are listed in Table 1.1. The electromagnetic and gravitational forces have an infinite range and obey an inverse square law. The range of the strong and weak forces can be estimated through Heisenberg's uncertainty principle and the masses of the force-carrying particles, which are the gluons for the strong nuclear force and the W and Z bosons for the weak force.

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

Table 1.1: The four fundamental interactions known in the nature.

The strong interaction is "felt" by quarks, and is mediated through massless particles called gluon. Quarks can bind with each other because they have color charge. A combination of all three colors creates a colorless particle, such as a baryon (proton for instance), which is a quark triplet state. A quark doublet is meson, which contains a quark and anti-quark pair and is color neutral. Gluons exist in eight independent color states. The model which describes the interactions of colored particles through the exchange of gluons is called Quantum Chromo Dynamics (QCD). C, P and CP-symmetry are conserved in strong interactions.

The Electromagnetic (EM) interaction is mediated by photons. Charged particles interact through the exchange of photon. Photons are massless and with no charge. The model which describes this interact is known as Quantum Electrodynamics (QED). C, P and CP-symmetry are conserved in electromagnetic interactions.

The weak interaction is mediated through the W^-, W^+ and the Z gauge bosons. Flavored particles interact weakly through the exchange of one of these weak bosons. However, at high energy, the weak and the EM forces are indistinguishable, and a combined theory to describe both, known as Electroweak Theory (EWT) is formed. Weak interactions allow quarks to change flavor, by unit charge e. Only in the weak interactions, C, P and CP-symmetry are violated.

The interactions of the strong, weak and electromagnetic can be described by $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ gauge symmetries. The subscripts C, L and Y represent the color, weak and hypercharge symmetries respectively. $SU(3)_C$ transformation describes the strong interaction, and $SU(2)_L \otimes U(1)_Y$ transformations describe the electroweak interaction.

1.2 Gauge symmetry

Modern theories describe physical forces in terms of fields, e.g., the electromagnetic field, the gravitational field, and fields that describe forces between the elementary particles. These fundamental fields cannot be directly measured, however, some associated quantities can be measured, like charges, energies. In field theories, different configurations of the unobservable fields can result in identical observable quantities. A transformation from such a field configuration to another is called a gauge transformation; the lack of change in the measurable quantities, despite the field being transformed, is a property called gauge invariance. Since any kind of invariance under a field transformation is considered a symmetry, gauge invariance is sometimes called gauge symmetry. Generally, any theory that has the property of gauge invariance is considered a gauge theory. In the SM, symmetry plays a very important role and the SM is a gauge theory which is built using symmetry. A group of transformations which can be performed on fields are the base of the symmetry, and leaving the Lagrangian invariant. Two kinds of symmetries are presented: global symmetries which allow one to change a field in the same way all over space-time; local symmetries where the field can be changed differently in each space-time point. The meaning of the symmetry is that conversation laws are obtained naturally and conversation laws have the underlying symmetries.

1.2.1 Quantum electrodynamics

In the quantum filed theory, particles can be described as a local field $\phi(x)$. The properties and interactions of the field are within a Lagrangian density, a function of the field and its space-time derivatives:

$$\mathcal{L}(x) = \mathcal{L}(\phi, \partial_\mu \phi). \tag{1.1}$$

Then applying the principle of the least action, the Euler-Lagrange equation can be achieved:

$$\partial_\mu\left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)}\right) - \frac{\partial \mathcal{L}}{\partial \phi} = 0. \tag{1.2}$$

Considering a Dirac Lagrangian which can describe the free fermion of mass m:

$$\mathcal{L}_{Dirac} = \bar{\psi}(x)(i\gamma^\mu \partial_\mu - m)\psi(x), \tag{1.3}$$

where $\psi(x)$ is the spinor field representing the fermions, $\bar{\psi} = \psi(x)^\dagger \gamma^0$ is its adjoint, and the γ^μ are the Dirac matrices.

Dirac motion equation for a fermion is obtained with the Euler-Lagrange Equation 1.2:

$$(i\gamma^\mu \partial_\mu - m)\psi(x) = 0. \tag{1.4}$$

With the transformation of U(1) group the $\mathcal{L}_{Dirac}$ can be invariant:

$$\psi(x) \rightarrow U\psi(x) = e^{i\theta}\psi(x), \tag{1.5}$$

where the θ is an arbitrary real constant.

If the local gauge transformation, θ is a function of x^μ, the Lagrangian is not invariant any more. Thus, to replace the derivative ∂_μ, the covariant derivative D_μ and gauge vector field A_μ are introduced as:

$$D_\mu \psi(x) \rightarrow e^{i\theta(x)} D_\mu \psi(x), \tag{1.6}$$

$$D_\mu \psi(x) \equiv \partial_\mu + iqA_\mu(x). \tag{1.7}$$

Then the local invariance of the Lagrangian is obtained by replacing ∂_μ by D_μ, then:

$$\begin{aligned} \mathcal{L} &= \bar{\psi}(x)(i\gamma^\mu \partial_\mu - m)\psi(x) \\ &= \mathcal{L}_{Dirac} - q\bar{\psi}(x)\gamma^\mu A_\mu \psi(x). \end{aligned} \tag{1.8}$$

The second item represents the QED interaction between a fermion and charge q and an external electromagnetic potential $A_\mu(x)$. $A_\mu(x)$ can be tuned into a true propagating field with the Proca Lagrangian:

$$\mathcal{L}_{Kinematics} = -\frac{1}{4} F_{\mu\nu}(x) F^{\mu\nu}(x), \tag{1.9}$$

where $F_{\mu\nu}(x) = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the usual electromagnetic field strength tensor.

This describes the free photon field and mass term is missing as this would violate the gauge invariance. Thus the full QED Lagrangian is:

$$\mathcal{L}_{QED} = \bar{\psi}(x)(i\gamma^\mu\partial_\mu - m)\psi(x) - q\bar{\psi}(x)\gamma^\mu A_\mu\psi(x) - \frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x) \tag{1.10}$$

with terms represent the kinetic energy and the mass of the fermion, the kinetic energy of the photon, and the interaction item.

1.2.2 Quantum chromodynamics

In colored quark model, each quark carries one of the three color states: red, green, blue. Quarks are triplets of the color under SU(3) and can be written as:

$$q_f = \begin{pmatrix} q_f^R \\ q_f^G \\ q_f^B \end{pmatrix} \quad \text{and} \quad \bar{q}_f = (\bar{q}_f^{\bar{R}}, \bar{q}_f^{\bar{G}}, \bar{q}_f^{\bar{B}}), \tag{1.11}$$

where f is the flavour. The Lagrangian of free quarks is presented as:

$$\mathcal{L} = \sum_f \bar{q}_f(i\gamma^\mu\partial_\mu - m)q_f. \tag{1.12}$$

In color space the Lagrangian is invariant under SU(3) transformations:

$$q(x) \to Uq(x) = e^{i\frac{\lambda_a}{2}\theta_a}q(x), \tag{1.13}$$

where U are 3×3 matrices, with $UU^\dagger = U^\dagger U = 1$ and $\det U = 1$.

The generator $\frac{\lambda_a}{2}$ (a = 1,...,8) is the Gell-Mann matrices. θ_a are a set of arbitrary parameters. Local gauge invariance can be achieved by replacing θ_a by $\theta_a(x)$. The gluons $G_\mu^a(x)$ are introduced to define the covariant derivative:

$$D_\mu\psi(x) \equiv \partial_\mu + ig_s\frac{\lambda_a}{2}G_\mu^a(x), \tag{1.14}$$

where g_s is the strong coupling constant. Thus Lagrangian for free quarks can be rewritten with covariant derivative, as well as an introduced gauge-invariant kinetic term for gluons:

$$G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a - g_s f_{abc}G_\mu^b G_\mu^c \tag{1.15}$$

and the final Lagrangian is:

$$\mathcal{L}_{QCD} = \sum_f \bar{q}_f(i\gamma^\mu\partial_\mu - m)q_f - g_s\sum_f(\bar{q}_f\gamma^\mu\frac{\lambda_a}{2}q_f)G_\mu^a - \frac{1}{4}G_{\mu\nu}^a G_a^{\mu\nu}. \tag{1.16}$$

The second term is the interaction between color currents of the quarks and the gluons fields G^a_μ. The $G^a_{\mu\nu}G^{\mu\nu}_a$ represents the interactions between gluons. The SM coupling constants show a dependence on energy, for EM interaction, the coupling becomes stronger with higher energy, but the QCD coupling is opposite. Here, it should be noted that, for photons, gauge invariance requires the gluons to be massless.

1.2.3 Weak interaction

Starting from the fermions, especially one generation of up and down quarks, with spin fields denotes by *u(x)* and *d(x)*,

$$\psi_1(x) = \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \quad \psi_2(x) = u_R,\, \psi_3(x) = d_R. \tag{1.17}$$

For the lepton section:

$$\psi_1(x) = \begin{pmatrix} \nu_{lL} \\ l_L \end{pmatrix}, \quad \psi_2(x) = \nu_{lR},\, \psi_3(x) = l_R, \tag{1.18}$$

here L and R represent the left-handed and right-handed fermion. Then the Lagrangian is obtained as:

$$\begin{aligned} \mathcal{L} &= i\bar{u}(x)\gamma^\mu \partial_\mu u(x) + i\bar{d}(x)\gamma^\mu \partial_\mu d(x) \\ &= \sum_{j=1}^{3} i\bar{\psi}_j(x)\gamma^\mu \partial_\mu \psi_j(x). \end{aligned} \tag{1.19}$$

Under the global gauge transformations of the $SU(2)_L \bigotimes U(1)_Y$ group, the Lagrangian is invariant. To fulfil the requirement of local invariant, the covariant derivatives is defined as:

$$D_\mu \equiv \partial_\mu + ig'\frac{Y}{2}B_\mu(x) + ig\frac{\sigma_a}{2}W^a_\mu(x), \tag{1.20}$$

where gauge $B_\mu(s)$ is associated to $U(1)_Y$, $W^a_\mu(x)$ is the three gauge fields, g and g' are the coupling constant of the $SU(2)_L$ and $U(1)_Y$.

After introducing the strength tensor:

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu, \quad W^a_{\mu\nu} = \partial_\mu W^a_\mu - g\varepsilon_{abc}W^b_\mu W^c_\mu, \tag{1.21}$$

the final EW interaction is:

$$\mathcal{L}_{EW} = \sum_{j=1}^{3} i\bar{\psi}(x)\gamma^\mu D_\mu \psi(x) - \frac{1}{4}B_{\mu\nu}B_{\nu\mu} - \frac{1}{4}W^a_{\mu\nu}W^{\mu\nu}_a. \tag{1.22}$$

The local gauge invariance assures the Lagrangian to describe the EM and weak interaction in a unified way, the physical interactions between fermions and photons, W and Z bosons. However, the gauge invariance requires massless terms for the gauge bosons or the fermions. Even we know that fermions and bosons have the mass from any experiment results.

1.3 Brout-Englert-Higgs mechanism

Based on local gauge invariance, a symmetry Lagrangian can be built to describe the QED and QCD. Gauge symmetry can guarantee that theory is renormalized and perturbation theory can be used to make predictions. But it forbids the additional mass term for the W and Z bosons. To get the mass the gauge symmetry must be broken while keeping the Lagrangian symmetric. This could be done through the spontaneous symmetry breaking (SSB).

Considering a complex scalar field $\phi(x)$:

$$\mathcal{L} = \partial_\mu \phi^* \partial^\mu \phi - V(\phi), \tag{1.23}$$

where

$$V(\phi) = \mu^2 \phi^* \phi + \lambda(\phi^* \phi)^2. \tag{1.24}$$

Here $\lambda < 0$ is to ensure the potential to have the bounded ground state. The case of $\mu^2 > 0$ and minimum $V(\phi)$ at ground state represents the spin zero particle, mass of μ and quartic coupling λ. If $\mu^2 < 0$, the potential has a minimum value at ground state:

$$|\phi_0| = \sqrt{\frac{-\mu^2}{2\lambda}} = \frac{\nu}{\sqrt{2}}, \tag{1.25}$$

the quantity ν ($\nu^2 = -\mu^2/\lambda$) is called vacuum expectation value. Because of the $U(1)$ phase invariance, the ϕ_0 is written as:

$$\phi_0 = \frac{\nu}{\sqrt{2}} e^{i\theta}. \tag{1.26}$$

The ground state solution, $\theta = 0$, causes the symmetry to be spontaneously broken. Now try to let W and Z bosons have the mass meanwhile photon is massless. Considering an $SU(2)$ doublet of complex fields:

$$\Phi(x) = \begin{pmatrix} \phi^\dagger(x) \\ \phi^0(x) \end{pmatrix}, \tag{1.27}$$

the Lagrangian of Higgs fields can be written as:

$$\mathcal{L}_{Higgs} = (D_\mu \Phi)^\dagger D^\mu \Phi - \mu^2 \Phi^\dagger \Phi - \lambda(\Phi^\dagger \Phi)^2, \tag{1.28}$$

where D_μ is the covariant derivative which is defined in Equation 1.20 to describe the filed and it is invariant under the $SU(2)_L \bigotimes U(1)_Y$ group transformations. Φ

can be parametrized as a perturbation of the ground state. $H(x)$ and $\theta_a(x)$ ($a = 1, 2, 3$) and local $SU(2)$ invariance are introduced to allow the elimination of the dependence on $\theta_a(x)$. The field $\theta_a(x)$ is the massless Nambu-Goldstone bosons associated with the SSB. By using the unitary gauge, these fields will be eaten by the gauge bosons and get the mass. So the Lagrangian will not have the terms related to $\theta_a(x)$. Rewriting the covariant derivative as:

$$D_\mu = \begin{pmatrix} \partial_\mu - \frac{i}{2}(gW^3_\mu + g'B_\mu) & -\frac{ig}{2}(W^1_\mu - iW^2_\mu) \\ -\frac{ig}{2}(W^1_\mu + iW^2_\mu) & \partial_\mu + \frac{i}{2}(gW^3_\mu - g'B_\mu) \end{pmatrix}, \tag{1.29}$$

then get the kinetic part:

$$(D_\mu\Phi)^\dagger D^\mu\Phi = \frac{1}{2}\partial_\mu H\partial^\mu H + \frac{1}{8}g^2(\nu + H)^2|W^1_\mu + iW^2_\mu|^2 + \frac{1}{8}(\nu + H)^2|gW^3_\mu - g'B_\mu|^2, \tag{1.30}$$

the final Higgs Lagrangian would be:

$$\mathcal{L}_{Higgs} = \frac{1}{2}\partial_\mu H\partial^\mu H + (\nu + H)^2(\frac{g^2}{4}W^\dagger_\mu W^\mu + \frac{g^2}{8\cos^2\theta_W}Z_\mu Z^\mu) - \lambda\nu^2H^2 - \lambda\nu H^3 - \frac{\lambda}{4}H^4, \tag{1.31}$$

where $W_\mu = \frac{W^1_\mu + iW^2_\mu}{\sqrt{2}}$, $Z_\mu = \frac{g'W^3_\mu - gB_\mu}{\sqrt{g^2+g'^2}}$ and $\cos\theta_W = \frac{g}{\sqrt{g^2+g'^2}}$.

Taking a look at the quadratic term of Equation 1.30 which is related to gauge bosons, the W and Z bosons mass is derived:

$$m_W = \frac{1}{2}\nu g, \quad m_Z = \frac{1}{2}\nu\sqrt{g^2 + g'^2} \tag{1.32}$$

and the relation between them is: $m_W = m_Z\cos\theta_W$. The Higgs boson (H) has been added as the generation of masses for the gauge bosons. And the SM Higgs is spin-0 scalar boson with no electric charge, and the mass is:

$$m_H = \sqrt{-2\mu^2} = \nu\sqrt{2\lambda}. \tag{1.33}$$

Although the SM gives the mass of the Higgs, the quartic coupling λ is a free parameter. The mass of the SM Higgs is still unknown until the measurements provided by the experiments. Through the spontaneous breaking of EW symmetry, the gauge bosons get the mass, however, fermions are still massless. Fermions can obtain the mass by adding interaction term between fermions and the scalar field, namely Higgs field. This term is called the Yukawa term and is given by:

$$\mathcal{L}_{Yukawa} = -(\Gamma^l_{i,j}\bar{\Psi}^l_L\Phi\psi^l_R + \Gamma^D_{i,j}\bar{\Psi}^Q_L\Phi\psi^D_R + \Gamma^U_{i,j}\bar{\Psi}^Q_L\Phi^\dagger\psi^U_R) + h.c, \tag{1.34}$$

the field doublets and singlets are split into their lepton and quark components.

The lepton doublet and singlet are represented by Ψ^l_L and Ψ^l_R, the left-handed up- and down-type quark doublets are represented by Ψ^Q_L, and the right-handed up- and down-type quark singlets are represented by ϕ_R^U and ϕ_R^D respectively. The $\Gamma_{i,j}$ terms are Yukawa coupling matrices for leptons and up and down quarks. These matrices contain terms of the Yukawa couplings $y_{i,j}$ for each fermion flavour i, j. The Yukawa couplings are not predicted by theory and must be obtained by experiment. After symmetry breaking in the unitary gauge the Yukawa term takes the form:

$$\begin{aligned}
\mathcal{L}'_{Yukawa} &= -\Gamma^l_{i,j}\bar{\Psi}'^l_L\frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ \nu + H \end{pmatrix}\psi'^l_R - \Gamma^D_{i,j}\bar{\Psi}'^Q_L\frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ \nu + H \end{pmatrix}\psi'^U_R \\
&\quad - \Gamma^U_{i,j}\bar{\Psi}'^Q_L\frac{1}{\sqrt{2}}\begin{pmatrix} \nu + H \\ 0 \end{pmatrix}\psi^D_R \\
&= \frac{\nu}{\sqrt{2}}\Gamma^l_{i,j}\bar{\Psi}'^l_L\psi'^l_R - \frac{1}{\sqrt{2}}\Gamma^l_{i,j}\bar{\Psi}'^l_L\psi'^l_R H - \frac{\nu}{\sqrt{2}}\Gamma^D_{i,j}\bar{\Psi}'^Q_L\psi'^D_R - \frac{1}{\sqrt{2}}\Gamma^D_{i,j}\bar{\Psi}'^Q_L\psi'^D_R H \\
&\quad - \frac{\nu}{\sqrt{2}}\Gamma^U_{i,j}\bar{\Psi}'^Q_L\psi'^U_R - \frac{1}{\sqrt{2}}\Gamma^U_{i,j}\bar{\Psi}'^Q_L\psi'^U_R H + h.c\,.
\end{aligned} \tag{1.35}$$

The fermion mass terms can be interpreted as $m = \frac{y_{ij}\nu}{\sqrt{2}}$. The primed fields represent the fermion fields in the unitary gauge. The terms containing H represent the fermion couplings to the Higgs boson.

Now the final SM Lagrangian can be written as the sum of the Lagrangian of QCD (Equation 1.16), EW (Equation 1.22), Higgs (Equation 1.31), Yukawa (Equation 1.35):

$$\mathcal{L}_{SM} = \mathcal{L}_{QCD} + \mathcal{L}_{EW} + \mathcal{L}_{Higgs} + \mathcal{L}_{Yukawa}. \tag{1.36}$$

1.4 Current status and remaining issues

The Large Hadron Collider (described in Section 2.1) is a proton-proton collider which provides several production modes of the Higgs boson. The dominant production mechanisms for the SM Higgs boson are through gluon-gluon fusion (ggF), vector boson fusion (VBF), associated production involving a W or Z boson (VH), and associated production with a $t\bar{t}$ pair ($t\bar{t}H$). Figure 1.2 shows a set of Feynman diagrams for these processes. The ggF and VBF production make the two greatest contributions to the total production cross section, and account for around 87% and 7% of the total for a Higgs boson of mass 125 GeV.

The Higgs boson is an unstable particle with a predicted mean lifetime of about 1.6×10^{-22} s for a mass of 125 GeV, and decays primarily into pairs of fermions or gauge bosons. The SM Higgs boson decay branching ratios are shown in Figure 1.3.

As shown in Figure 1.3 $H \to b\bar{b}$ decay has the largest branching fraction, about 57% for a Higgs boson of mass 125 GeV, but the ggF process is dominated by a background from the SM di-jet production that contaminates the signal and makes

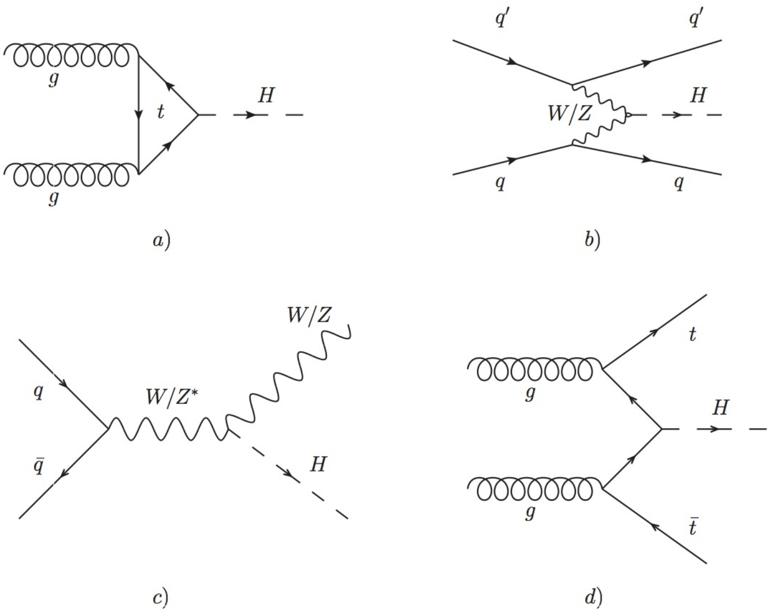


Figure 1.2: Summary of Higgs boson production mechanisms at the LHC (a) ggF, (b) VBF, (c) VH, (d) $t\bar{t}H$.

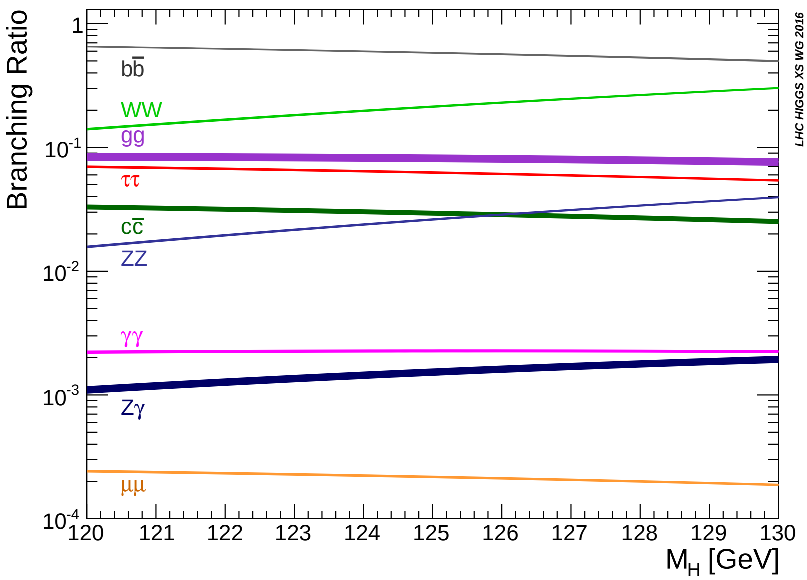


Figure 1.3: The SM Higgs boson decay branching ratios.

analysis challenging. Preferred channels for study are those with good signal to background ratios, $H \to \gamma\gamma$, $H \to ZZ^*$ and $H \to WW^*$ channels are chosen to search for and study Higgs currently although some of them is with small branching ratio. The $H \to WW^*$ channel is the main studied decay mode especially in this thesis.

On the 4th of July 2012, using only a fraction of the full dataset provided by the LHC, the ATLAS and CMS collaborations at CERN announced the discovery of a new particle, when searching for the SM Higgs boson. The excess of events observed was compatible with the production and decay of the SM Higgs boson, with a mass of approximately 125 GeV [1, 2].

Combining the ATLAS and CMS data for the $H \to \gamma\gamma$ and $H \to ZZ \to 4\ell$ channels, the mass of the Higgs boson is determined to be [3]

$$m_H = 125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{syst.})\ \text{GeV}. \tag{1.37}$$

Measurements of the Higgs boson cross section in the Higgs boson diphoton decay channel are performed using pp collision data recorded by the ATLAS experiment taken at a centre of mass energy of $\sqrt{s} = 13$ TeV and correspond to an integrated luminosity of 13.3 fb^{-1}. The corresponding signal strengths measured for the different production processes, and globally (i.e. assuming one common signal strength parameter for all production processes), are summarised in Figure 1.4, which also shows the global signal strength measured in Run 1. The presented measurements are dominated by the statistical uncertainties. The measurements agree with the SM expectations within 1 to 2σ, and no significant deviation from the Standard Model expectations is observed [4].

The latest measurements of the Higgs boson show remarkable compatibility with the SM expectations, and neither of the collaborations has observed any significant deviations from this theory. The SM is the most successful theory, which can describe the relationship between particles and fundamental interactions so far. The latest results with ATLAS experiment are shown in Figure 1.5 [5]. The agreement spans several orders of magnitude in cross-section and a variety of the SM benchmark processes.

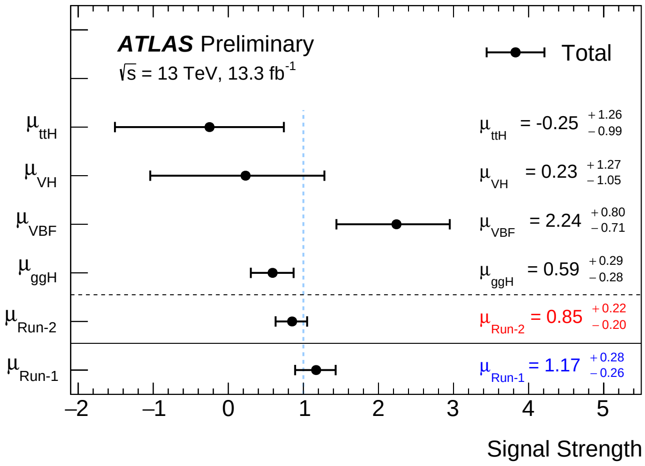


Figure 1.4: The signal strength measured for the different production processes (ggH, VBF, VH and $t\bar{t}H$) and globally (μ_{Run2}), compared to the global signal strength measured at 7 and 8 TeV (μ_{Run1}). The error bar shows the total uncertainty. The gluon-gluon fusion production cross section is larger by approximately 10%.

Despite the tremendous success of the SM, there are still many unsolved issues in physics, from dark matter, dark energy, hierarchy problem, matter-antimatter asymmetry to the neutrino oscillation. The absence of gravity and implicitly of its mediator in the SM, and there is no experimentally confirmed understanding of this force at the quantum level yet. Much effort has been done to combine gravity with the SM to form a theory of everything, but so far attempts have only limited progresses. The Beyond Standard Model (BSM) is needed to deal with those issues, like SUperSYmmetry (SUSY). Generally the BSM physics is expected at the TeV scale. For example, SUSY introduces a new symmetry relating fermions and bosons at the TeV scale, is presented as an elegant solution to the hierarchy problem, as it cancels out the quadratic divergences. Moreover, SUSY has the advantage of introducing a candidate for dark matter. The all 7, 8 and 13 TeV data recorded by ATLAS are employed to search for the SUSY particles. No obvious evidence of the superpartner (also sparticle) are found so far.

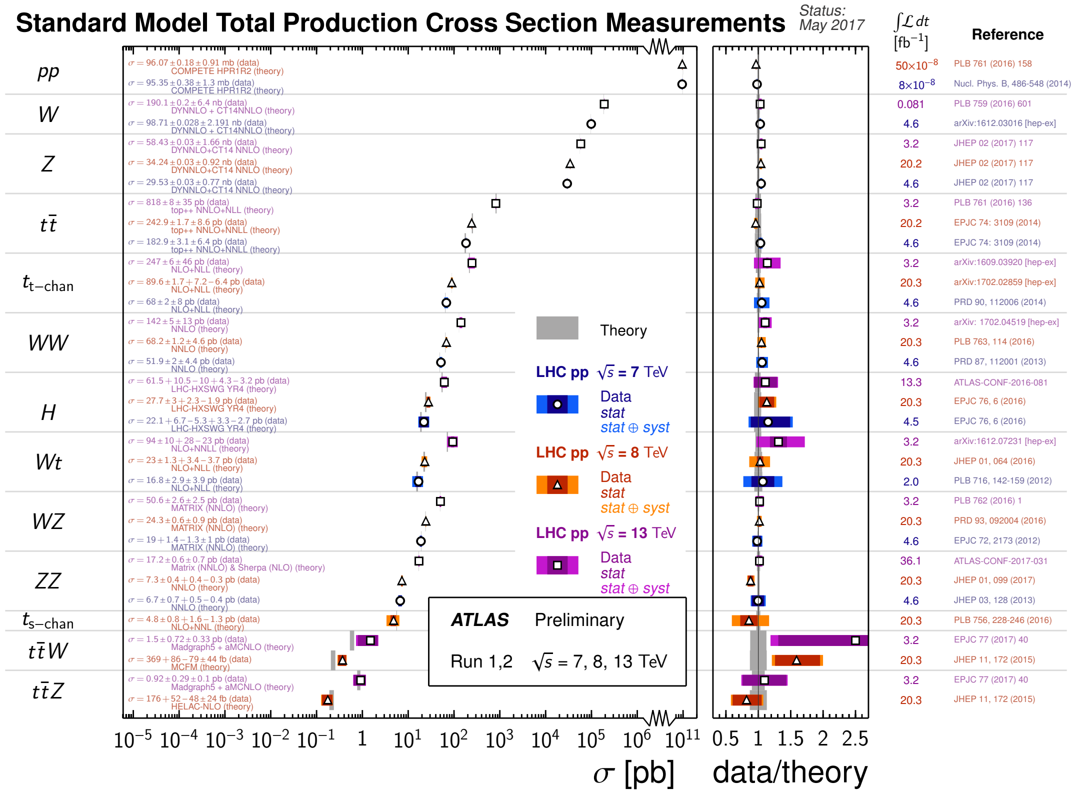


Figure 1.5: Summary of several Standard Model total production cross section measurements, corrected for leptonic branching fractions, compared to the corresponding theoretical expectations. All theoretical expectations were calculated at NLO or higher. The dark-color error bar represents the statistical uncertainty. The lighter-color error bar represents the full uncertainty, including systematics and luminosity uncertainties. The data/theory ratio, luminosity used and reference for each measurement are also shown. Uncertainties for the theoretical predictions are quoted from the original ATLAS papers. They were not always evaluated using the same prescriptions for PDFs and scales. Not all measurements are statistically significant yet.

1.5 The production of the Higgs boson associated with a pair of top quarks ($t\bar{t}H$)

After the discovery of the Higgs, one of the most important studies is to measure Higgs couplings to other particles to figure out whether the couplings are consist with the Standard Model Higgs boson, or those of a Higgs boson from an extended Higgs sector in Run 2. This section provides the basic information and meaning of the $t\bar{t}H$ production. The following Section 1.6 is with combined results of the ATLAS and CMS by using the 7 and 8 TeV data which are collected during the Run 1. The coupling of the Higgs boson to other particles can be measured by studying the various production processes and decay modes of the Higgs boson. The associated production of a Higgs boson with a pair of top quarks ($t\bar{t}H$, Figure 4.1) can provide direct information of the Yukawa coupling of the Higgs boson to the top quark, probably the most crucial coupling to fermions. This process involves two colliding gluons. Those gluons through the top-Higgs coupling form a Higgs particle and a top quark pair. Top quark Yukawa coupling measurement connects two of the heaviest Standard Model particles, the top quark (173 GeV) and the Higgs boson (125 GeV) : $\lambda_t = \frac{\sqrt{2}m_t}{\nu}$. Its value is predicted to be equal to unity and any deviation might give a hint on the possible existence of the new physics. The absence of a deviation can constrain and even exclude the BSM theories.

The SM Higgs cross section of different production processes versus energy are shown in Figure 1.6, especially the $t\bar{t}H$ process increases the most significantly compared to other productions around 10 TeV, which is the designed energy range for the Run 2 and the further LHC running. The Yukawa coupling of the Higgs boson to the top quark is a key parameter of the SM. It can be determined from the ratio of the top quark mass and Higgs field vacuum expectation value, from the cross section of $gg \to H$ production through a top quark loop, or from the cross section of the process $gg \to t\bar{t}H$, which is a tree-level process at lowest order in perturbation theory. Comparison of these measurements has the potential to identify and disambiguate new physics effects that can modify the $t\bar{t}H$ production cross section relative to the SM expectation.

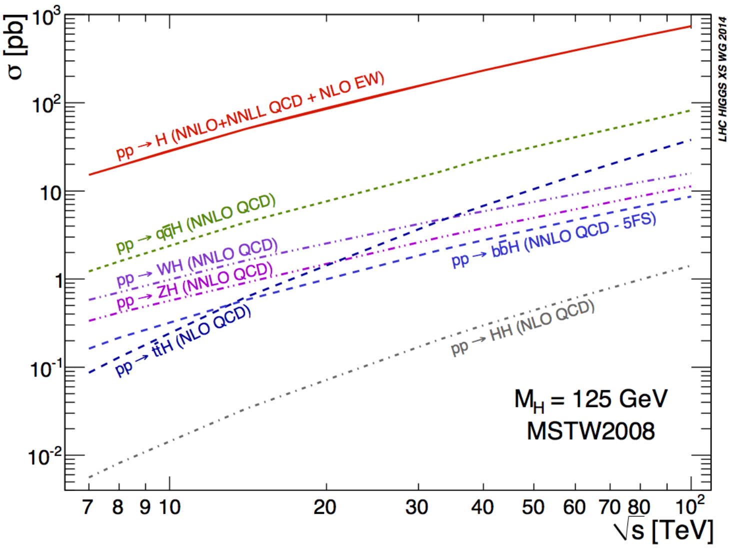


Figure 1.6: SM Higgs boson cross sections of different productions versus the centre mass of energy.

1.6 $t\bar{t}H$ searching at the LHC in Run 1

At the ATLAS and CMS several Higgs boson decay channels were considered in Run 1 to measure the top Yukawa coupling using $t\bar{t}H$ production: $H \rightarrow \gamma\gamma$, $H \rightarrow bb$ and $H \rightarrow$ multileptons final states. The combined results shown below are based on the complete Run 1 collision data collected by the ATLAS and CMS experiments. These data correspond to integrated luminosities per experiment of approximately 5 fb^{-1} at $\sqrt{s} = 7$ TeV (recorded in 2011) and 20 fb^{-1} at $\sqrt{s} = 8$ TeV (recorded in 2012). The results of the ATLAS and CMS individual combinations based on the Run 1 data are reported in Refs [6, 7]. Especially the signal strength of $t\bar{t}H$ production in multilepton final state by using 20.3 fb^{-1} of proton-proton collision data recorded by the ATLAS experiment at $\sqrt{s}$ = 8 TeV are shown in Figure 1.7, and the p-value of the likelihood ratio test of the background-only hypothesis corresponds to 1.8σ; the expectation in the presence of a Standard Model signal is 0.9σ [6].

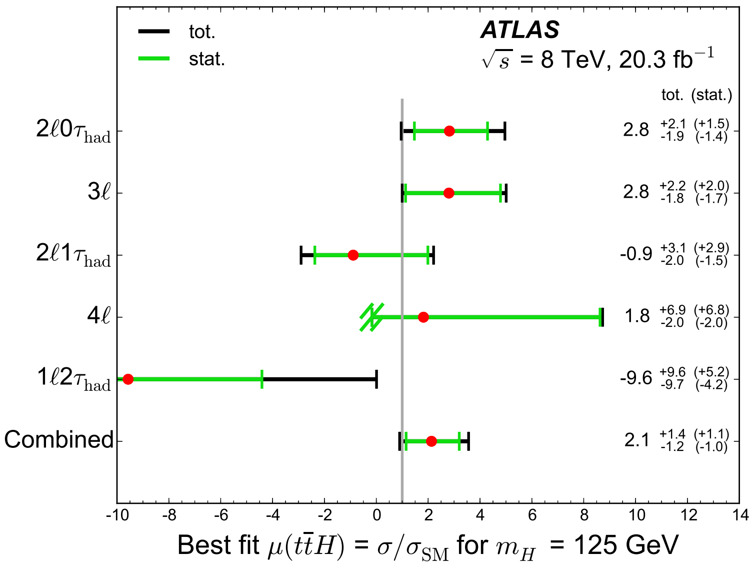


Figure 1.7: Best-fit values of the signal strength parameter $\mu = \sigma_{t\bar{t}H,\text{obs}}/\sigma_{t\bar{t}H,\text{SM}}$. For the 4ℓ Z-depleted category, μ < -0.17 results in a negative expected total yield and so the lower uncertainty is truncated at this point.

The global signal strength is the most precisely measured Higgs boson coupling-related observable, but this simple parametrisation is very model dependent. All Higgs boson production and decay measurements are combined assuming that all their ratios are the same as in the SM. The compatibility of the measurements with the SM can be tested in a less model-dependent way by relaxing these assumptions separately for the production cross sections and the decay branching fractions. Assuming the SM values for the Higgs boson branching fractions, the combined results of the ATLAS and CMS data of $t\bar{t}H$ signal strengths are listed in Table 1.2 [8].

Production process	ATLAS + CMS	ATLAS	CMS
$\mu_{t\bar{t}H}$	$2.3^{+0.7}_{-0.6}$	$1.9^{+0.8}_{-0.7}$	$2.9^{+1.0}_{-0.9}$

Table 1.2: Measured signal strengths μ and their total uncertainties for $t\bar{t}H$ production process. The results are shown for the combination of the ATLAS and CMS, and separately for each experiment, for the combined $\sqrt{s}$ = 7 and 8 TeV data. These results are obtained assuming that the Higgs boson branching fractions are the same as in the SM.

The combined likelihood is used to evaluate the significance for the observation of the $t\bar{t}H$. The combination of the data from the two experiments corresponds to summing their recorded integrated luminosities and consequently increases the sensitivity by approximately a factor of $\sqrt{2}$, since the theoretical uncertainties in the Higgs boson signal are only weakly relevant for this evaluation and all the other significant uncertainties are uncorrelated between the two experiments. The combined significances for the observation of the $t\bar{t}H$ process is 4.4σ, whereas only 2.0σ is expected, corresponding to a measured excess of 2.3σ with respect to the SM prediction [8].

2 Experimental apparatus

The Large Hadron Collider (LHC) is the largest and highest energy hadron collider in the world so far. It is the most recent addition to the CERN's accelerator and lies in a tunnel buried around 50 to 175 m beneath the France–Switzerland border near Geneva, Switzerland. According to the particle physics, when measuring the properties of nature at the smallest distance scales ever recorded with a terrestrial apparatus the highest energy particles accelerator ever built is required. The size of a structure is related to the probe energy via the de Broglie relation $\lambda = 1/\rho$. The same principle governs optical microscopes, limiting their resolution to hundreds of nano meters. However, the LHC can produce protons with energies up to $\sqrt{s} = 13$ TeV and A Toroidal LHC ApparatuS (ATLAS) is able to capture the products of the proton-proton collisions to probe distance scales as small as 10^{-20} m scale. Besides the ATLAS, CMS experiment has the similar physics goal as the ATLAS, while the other two experiments, the ALICE and LHCb are designed to have different physics goals and locate at the different parts of the LHC. In this chapter, the design and performance of the LHC are presented in Section 2.1 and the ATLAS detector is described in details in Section 2.2.

2.1 The Large Hadron Collider

The LHC inherited the tunnel of the Large Electron–Positron Collider (LEP) which took place from 1984 to 1989. The ring has a circumference of 26.7 km and a gradient of 1.4%. It was decided to build the LHC in this existing tunnel to considerably lower the construction costs, even though a hadron collider generally benefits from a larger radius and does not suffer as much from synchrotron radiation as a circular lepton collider. The LHC is designed to produce proton-proton collisions up to a center of mass energy of $\sqrt{s} = 14$ TeV. To accelerate the particles to desired energy, the beam is injected through a succession of machines which accelerate the particles to gradually higher energies. This accelerator chain is shown in Figure 2.1 [9].

The protons, produced by a duoplasmatron source through stripping hydrogen atoms of their orbiting electrons, are injected in the linear accelerator in the chain, the Linac 2, where they get accelerated to an energy of 50 MeV. The beam is then firstly fed into the Proton Synchrotron (PS) Booster, which has a radius of 25 m

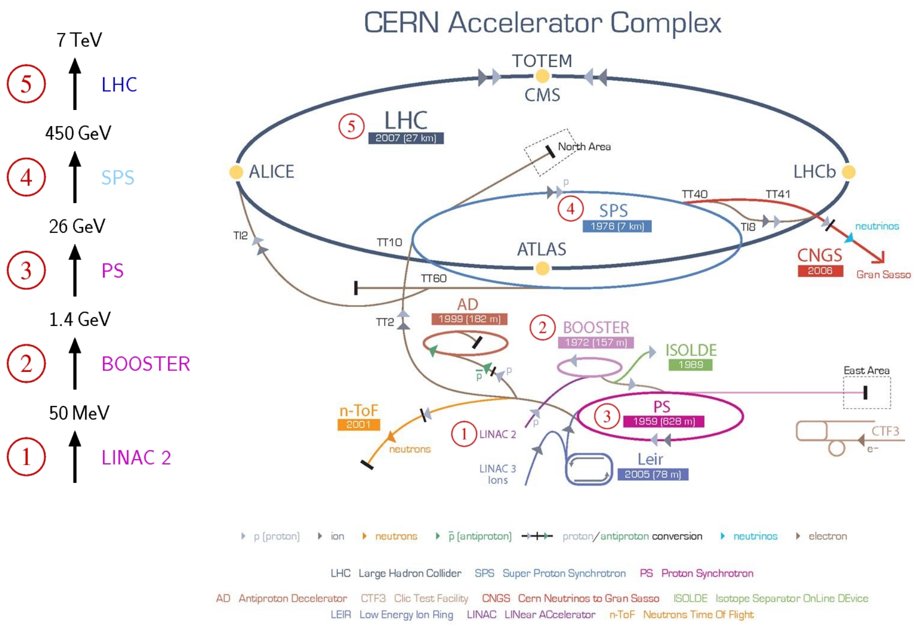


Figure 2.1: The accelerator complex at CERN.

and accelerates the particles to 1.4 GeV, secondly into the PS where an energy of 26 GeV is obtained, and finally into the Super PS (SPS), which has a circumference of 6.9 km, where the particles are accelerated to 450 GeV. The particles have now reached the LHC injection energy and are fed into the LHC ring via two transfer lines that circulate in opposite directions. It takes approximately 4 minutes to fill each of the two LHC rings. The particles are then accelerated for 20 minutes by electric fields in superconducting radio-frequency (RF) cavities operating at 400 MHz with a 5 MV/m gradient to reach the design energy of 14 TeV. Once the particles are circulating around the ring at the nominal beam energy, they will lose about 7 KeV of synchrotron radiation per turn. The RF cavities provide energy corrections to the beams to account for this loss. The accelerator chain can also accelerate lead ions which are passed through a Low Energy Ion Ring and are from there transferred through the PS and the SPS. In the LHC ring, they get to be accelerated to half of the maximally collided energy per nucleon. The LHC has not reached its design energy yet, it has been running at 7, 8 and 13 TeV, in 2011, 2012 and 2015-2016, respectively.

The instantaneous luminosity and the center of mass energy are important parameters for the LHC. The number of events per second, N_{event}, for a particular physics process is given by:

$$N_{event} = L \times \sigma_{event}, \tag{2.1}$$

where σ_{event} is the cross section of the physics process and L is instantaneous luminosity. The instantaneous luminosity is defined as:

$$\mathcal{L}_{inst} = \frac{\mu n_b f_r}{\sigma_{inel}}, \tag{2.2}$$

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. The instantaneous luminosity depends on the collider parameters, namely the density of protons in the colliding bunches, the collision transverse area (defined by the accelerator optics in the collision zone) as well as on the frequency of collisions and the number of bunches. The design luminosities for pp collisions at the four main LHC experiments are 10^{34} $\text{cm}^{-2}\text{s}^{-1}$ for the ATLAS and CMS, 10^{32} $\text{cm}^{-2}\text{s}^{-1}$ for the LHCb and 10^{30} $\text{cm}^{-2}\text{s}^{-1}$ for the ALICE.

An increasing challenge for the LHC data is in-time pile-up and out-of-time pile-up with higher center of mass energy. In-time pile-up refers to multiple protons from each bunch interacting in a given bunch crossing whereas out-of-time pile-up refers to the presence of collisions from surrounding bunch crossings in the read-out window of the considered bunch crossing. Pile-up increases the track and vertex multiplicity as well as the overall energy in an event. This renders the reconstruction of physics objects like tracks, vertices and jets very challenging. To avoid large systematic uncertainties, a precise pile-up modelling is therefore crucial for successful physics analyses. Generally pile-up events are modelled by using the minimum bias events which are based on Monte Carlo simulation.

Holding together the protons in a circular beam pipe, a powerful magnetic field is needed. To reach a high intensities, superconducting magnets are used. The magnet coils are based on niobiumtitan superconductor cables, which are able to conduct the electricity without resistance or energy loss. With the superconductivity facilities, the LHC magnetic system is able to be operated at required intensities of up to 8.4 Tesla, given an electric current of 11850 A. Highest constraints on the cryostat system design are from the LEP tunnel and its facilities to construct the LHC collider. A total of eight refrigeration plants, filled with liquid Helium, are supplying the required cooling superfluid.

The proton beam bending is ensured by 1232 dipole magnets, each piece weighting 27.5 tons, a length of 16.5 m and a diameter of 570 mm. Each magnet has two apertures, for the two proton beams tubes (two-in-one magnet design, Figure 2.2 [10]) and provides a magnetic field of 8.33 T. The required temperature is 1.9 K most of time, but it can also go up to 4.5 K. As the protons are charged particles, it is important to focus the beam and retain it in the ultra high vacuum chamber. Other thousands of multipoles magnets are used to improve the beam focusing and also to reduce the unwanted interactions between the bunch constituents. Finally, inner triplet magnets are used to squeeze the protons in the collision points.

Besides the mentioned magnets, many more normal magnets are used in order to reach the target beam parameters, leading to more than 9000 magnets surrounding the LHC ring. They represent around 90% (40000 tons) of the cold mass. The magnet powering is performed in eight independent and symmetric sectors to ensure a limitation of the stored energy in a magnet. Besides the powerful magnetic system, the LHC is characterized by the biggest vacuum system in the world. The vacuum system can reach a vacuum pressure of 10^{-6} mbar at cryogenic temperatures (5-20 K).

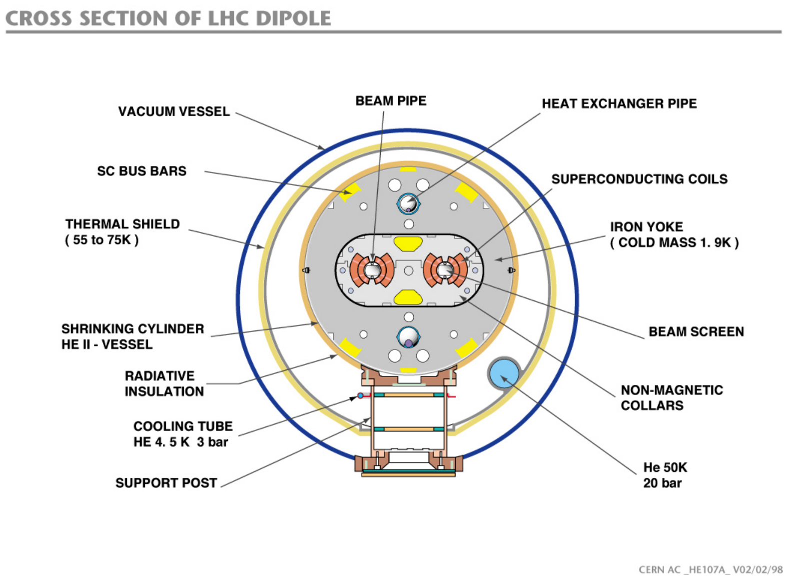


Figure 2.2: Cross section of the LHC dipole.

The first proton-proton collision at LHC occurred in November 2009, at energy of 900 GeV in the centre of mass system and increased to 2.37 TeV within few days. The first 7 TeV collisions started at 2010 with validation tests after test period. Delivered luminosity increased rapidly in following years in Figure 2.3 (left), especially in the year 2012 and year 2016. The situation in 2016 is much different because of the two years update from 2013 to 2014 (Section 2.2.7). The last Run 1 data taking started in May 2012, with an increase in energy in the center mass to 8 TeV. In Figure 2.3 (right) a total integrated luminosity and data quality are presented. With the good performance of the detector, data recorded efficiently by the ATLAS. Comparing to 7 TeV, the simultaneous interactions increased considerably (Figure 2.4) in 8 TeV. Many Standard Model measurements are published with better precise results by using the 7 TeV and 8 TeV data recorded in 2011 and 2012. Especially in 2012, Higgs boson is announced to be discovered by the ATLAS and CMS experiments separately.

In the LHC, four main experiments were installed and developed. The ATLAS and CMS are the biggest experiments and are designed to understand the Standard

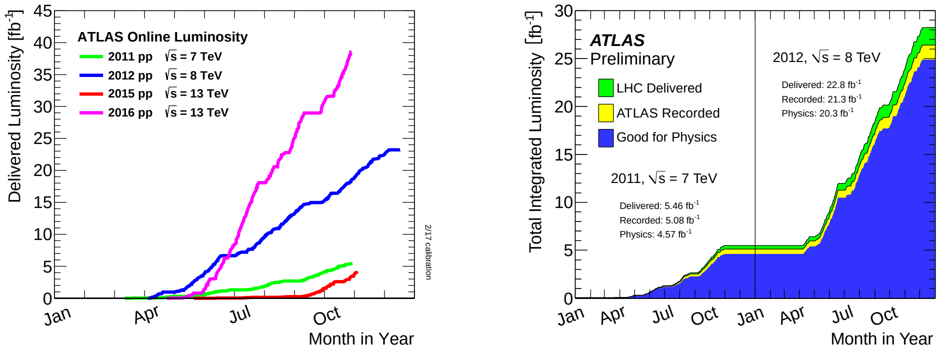


Figure 2.3: Delivered Luminosity versus time for 2011-2016 (pp data only) (left) and total integrated luminosity and data quality in 2011 and 2012 [11].

Model precisely, especially for the Higgs physics and also to discover the physics implying unknown particles (like supersymmetric particles predicted by SUSY) around the TeV scale. The b physics is primarily studied at the LHCb experiment, which covers the main subjects of heavy flavour and electroweak physics. It allows the study of $B_s \to \mu^+\mu^-$ the rare SM processes (i.e. flavour changing neutral current) and the CP violation in different processes involving B_s, D_0 and Kaon meson states. It also led to the discovery of new baryon resonance, Ξ_b^-. A Large Ion Collider Experiment (ALICE) is studying the quark confining mechanism, by creating a quark-gluon plasma. The considered collisions are lead-lead or lead-proton.

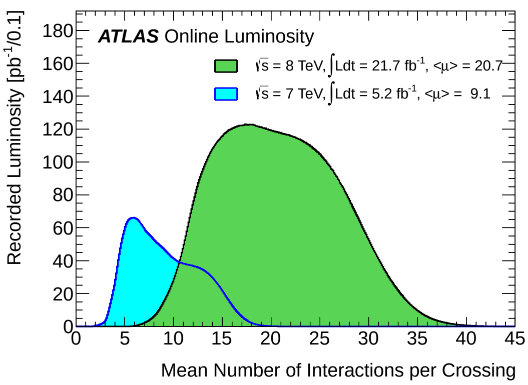


Figure 2.4: Number of interactions per crossing in 2011 and 2012 [11].

2.2 The ATLAS detector

The ATLAS is a general purpose experiment that is designed to probe a large range of physics from soft QCD, precision electroweak and top quark measurements to searches for the Higgs boson and new physics at TeV scale. To this point, the ATLAS is built in a layered structure with cylindrical geometry and almost complete hermetic coverage. A schematic of the detector can be seen in Figure 2.5 [12].

The ATLAS detector is 44 m long, 25 m tall and weighs about 7000 tons. Starting from its core and moving outwards, first is the inner detector (ID) that is structured in three layers and is responsible for precisely measuring the momentum of charged particles moving in a 2 T solenoidal magnetic field, followed by the electromagnetic and hadronic calorimeters which measure the energy deposition of electrons, photons and hadrons and also infer the missing transverse energy from neutrinos and other undetected particles, and finally a precise muon spectrometer (MS) which provides tracking and triggering of muons in a toroidal magnetic field. The detector took fifteen years to design, build and install.

The ATLAS collaboration counts about 3000 people from 178 institutions in 38 countries. The ATLAS recorded its first collisions in 2009, after a one year shut down due to an LHC magnet quenching accident, and has been running successfully until the beginning of 2013 when it shut down for a two years upgrade period. The ATLAS begun taking 13 TeV data in 2015 and will continue full running till the end of 2018 for the next upgrade.

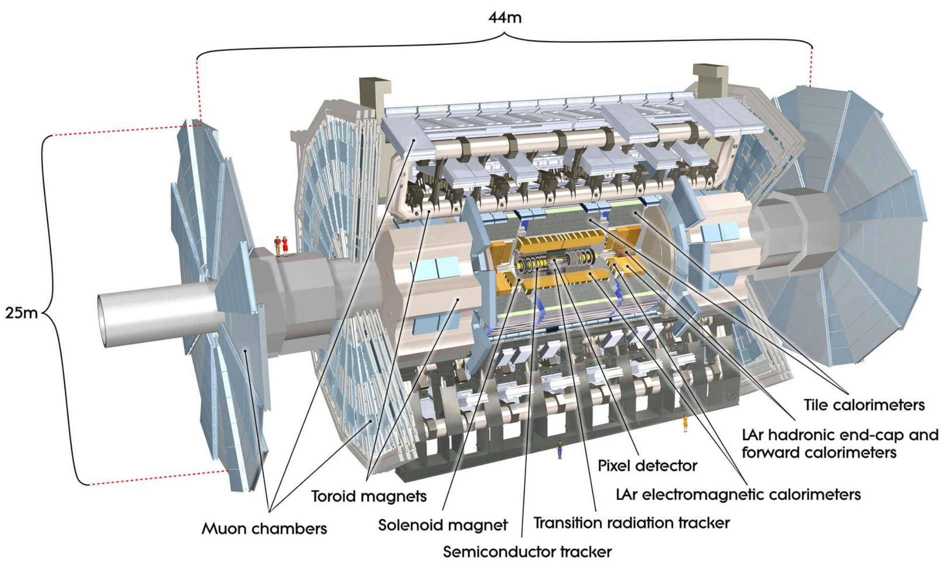


Figure 2.5: The overview of the ATLAS detector.

The detector is designed to test the Standard Model predictions, including the existence of the SM Higgs boson, and to search for new physics phenomena beyond the SM. These searches require accurate particle identification and precise measurements close to the interaction point. The physics goals and the high energy and luminosity of the LHC can be expressed in the following general requirements:

1. The high energy and flux radiation environment require the use of fast and radiation-hard electronics and sensors. Besides, a high sub-detector granularity is crucial to handle the high particle multiplicity and to cope with the pile-up arising from high luminosity.
2. An almost full coverage in the polar θ, and azimuthal ϕ, angles is required to include the entire collision event to enable the calculation of the missing transverse energy (MET) of the undetected particles.
3. Very good charged particle momentum resolution and high track reconstruction efficiency in the inner detector are essential for the precise measurements which are close to the interaction point.
4. Very good electromagnetic calorimeter to identify and measure the energy of both electrons and photons is needed.
5. Good hadronic calorimeter with full coverage is needed to measure the jet energy and calculate the MET with good resolution.
6. Muon identification and the ability to measure momentum with good resolution over a wide range of momenta are necessary.
7. Stable magnetic field with strong bending power is fundamental for momentum measurements of charged particles in the inner detector and muon spectrometer.
8. High efficiency triggering is needed to achieve a high and stable data taking rate with sufficient background rejection.

The ATLAS satisfies these requirements to a large extent. Its extended length provides a large acceptance in pseudorapidity and the high granularity of the inner detector allows an efficient track reconstruction. Both the electromagnetic and hadronic calorimeters provide good energy resolution. The ATLAS muon spectrometer yields a good muon identification and the trigger system has a very high efficiency. The ATLAS coordinate system, sub-detectors and trigger system will be described as follows.

2.2.1 The ATLAS coordinate system

The coordinate system defined here is used to describe tracks of particles in the ATLAS detector. The origin of this coordinate system is defined to be at the interaction point (IP), where the collision occurs. The positive x-axis points from the IP towards the center of the LHC ring, while the y-axis points upwards from the IP. The beam direction, which is transverse to the x-y plane, defines the z-axis with the positive direction pointing towards LHCb. Since the ATLAS is cylindrical shaped, it is also convenient to define a cylindrical coordinate system. In this system the polar angle θ is measured between the z-axis and the x-y plane and the azimuthal angle ϕ is measured from the x-axis around the beam in the x-y plane.

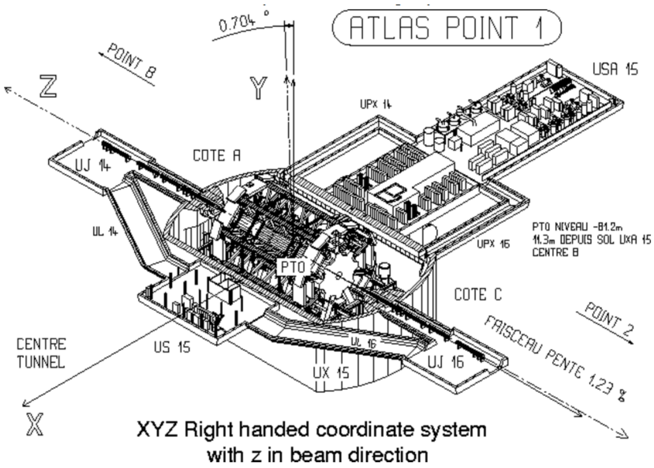


Figure 2.6: The coordinate system in the ATLAS detector. The general tilt of the LEP/LHC tunnel causes the y-axis to be slightly different from vertical [13].

The pseudorapidity variable η defined by Equation 2.3 is used to describe the object's trajectory in the detector,

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

It depends only on the angle θ, but it can also be defined in terms of the particle's momentum as Equation 2.4:

$$\eta = \frac{1}{2}\ln(\frac{|\mathbf{p}| + p_z}{|\mathbf{p}| - p_z}). \tag{2.4}$$

It should be noted that the rapidity of a particle is defined as Equation 2.5:

$$\eta = \frac{1}{2}\ln(\frac{E + p_z}{E - p_z}). \tag{2.5}$$

The angular distance ΔR between two objects in the detector is usually defined in the η - ϕ space as Equation 2.6:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}). \tag{2.6}$$

The transverse momentum, p_T, is defined in the x-y plane as Equation 2.7:

$$p_T = |\mathbf{p}| \sin\theta. \tag{2.7}$$

2.2.2 Inner tracking detector

The Inner Detector (ID), shown in Figure 2.7 [14], is contained within a cylinder of 7 m length and 1.15 m radius and is fully immersed within the 2 T magnetic field provided by the solenoid. The large track density at the LHC needs a very good momentum resolution and precise vertex reconstruction, both for primary vertices from the hard collision and for secondary vertices from long-lived particles such as kaons, τ leptons or jets produced by heavy flavour quarks. The ID detects particles by measuring the interaction of a particle with the surrounding material at discrete space points up to $|\eta| < 2.5$. It can only detect electromagnetically charged particles, neutral particles are not detected in the ID. The direction of the curvature gives the charge of the particle, the degree of curvature gives the momentum.

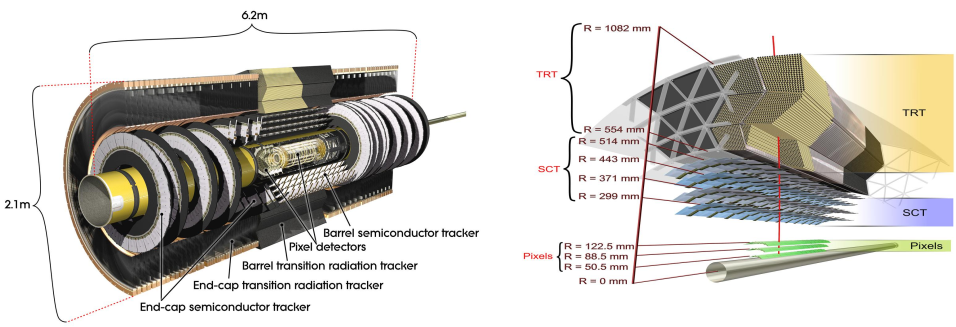


Figure 2.7: View of the ATLAS inner detector (left) and zoomed view on the layered structure of the pixel, SCT and TRT subsystems.

The ID is composed of three subsystems, the pixel detector, the semiconductor tracker (SCT) and the transition radiation tracker (TRT). It combines high resolution silicon pixel layers and silicon microstrip detectors in the inner part and continuous straw-tube detectors in the outer part. Out of the three ID subsystems, the pixel system is the one located closest to the beam line. Its innermost layer, often called the B-layer, is positioned at 5 cm from the IP. The pixel detector has the highest granularity and gives very precise measurements of track impact parameters which help with the identification of displaced vertices from b-hadron or τ lepton decays. The TRT offers substantial discrimination between electrons and charged hadrons over a wide energy range through the detection of transition radiation photons. During Run 1, several leaks in the TRT exhaust system led to a large loss of highly expensive Xenon gas. To reduce the operation costs, a few parts of the TRT (internal layer of the barrel TRT and one end-cap wheel on each side of the detector) were running with Xenon-Argon gas mixture since 2015.

2.2.3 The ATLAS calorimeter

The ATLAS calorimeter (Figure 2.8 [15]) is separated in one barrel ($|\eta| < 1.475$), and two end-cap regions. The former is divided in the LAr electromagnetic barrel and the Tile barrel sub-systems. Each end-cap calorimeter includes the LAr electromagnetic end-cap (EMEC), the LAr hadronic end-cap (HEC) and the LAr forward (FCal) subsystems. The HEC calorimeter is situated behind the EMEC, while the FCal is placed near the inner detector. Generally, different particles have the different behaviours in each detector (Figure 2.9 [16]). Electrons can leave a full track in the inner detector, then they deposit its energy in the LAr electromagnetic calorimeter by interacting with the materials. But muons can go through the whole detector and a clear full track can be found.

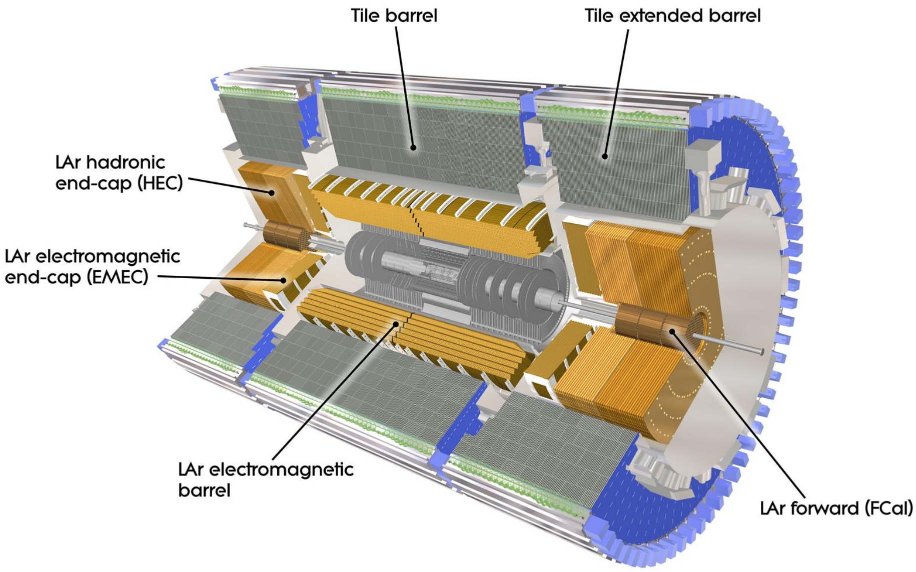


Figure 2.8: Overview of the ATLAS calorimeter. Near the beam-pipe the tracker is visible, surrounding it is the EM calorimeter and beyond the hadronic calorimeter. Both barrel and end-caps elements are displayed.

2.2.3.1 Electromagnetic calorimeter

The Electromagnetic (EM) calorimeter is divided into a barrel component up to $|\eta| < 1.475$ and two end-cap sections covering $1.375 < |\eta| < 3.2$. All three of

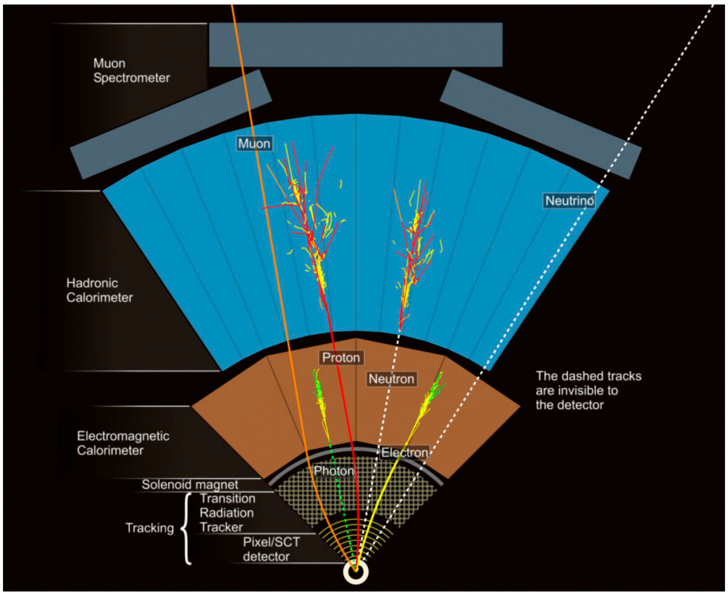


Figure 2.9: Overview of particles passage through the ATLAS tracker, electromagnetic calorimeter, hadronic calorimeter and eventually through the first layers of the muon spectrometer.

these sections have their own cryostat. The barrel part comprises two half-barrels divided by a 4 mm gap at $z = 0$. Each half-barrel is 3.2 m long, weighs 57 tons and the inner (outer) radius measures 2.8 m (4 m). Each of the two end-cap parts is made of two coaxial wheels where the inner and outer wheels cover the ranges $1.375 < |\eta| < 2.5$ and $2.5 < |\eta| < 3.2$, respectively. A wheel has a width of 63 cm and weighs 27 tons.

The EM calorimeter is a Liquid Argon (LAr) detector with accordion-shaped electrodes made of Kapton, a polyimide, and lead absorber plates whose thickness is optimized for energy resolution. The accordion-shaped arrangement allows for a full coverage in without any cracks and ensures that each particle travelling through the detector will cross approximately the same amount of material. As charged particles hit the absorber plates, they produce EM showers of electrons and photons, the latter maybe turn into pair-produce electrons. These showers ionize the LAr and the ionized electrons drift to the read-out electrodes. The latter are installed between the absorbers and are surrounded by copper plates that are held at a potential of 2000 V, leading to a drift time of 450 ns. The signal size on the electrode, which is roughly proportional to the number of electrons reaching the electrode, determines the energy measurement and so directly in influence the resolution. It is therefore important to prevent large resolution from leakage fluctuations. The additional electrons from the containment of the full shower in the longitudinal direction also improve the sampling resolution. A presampler detector is installed within $|\eta| < 1.8$ to correct for inhomogeneous energy losses of electrons and photons travelling through the ID and support structures. It is made of a thin, active LAr layer of 1.1 cm (0.5 cm) width in the barrel (end-cap).

2.2.3.2 Hadronic calorimeter

The purpose of the hadronic calorimeter is to measure the position and energy of hadronic jets. Hadronic jets are narrow cones of protons, neutrons, mesons and other particles produced by the fragmentation and hadronization of quarks and gluons. These jets pass through the ID and ECAL without significant loss of energy. The HCAL is located outside of the ECAL and comprises a barrel and two end-cap regions. In the barrel region the HCAL uses steel layers as an absorber and plastic scintillator sampling sheets, called tiles, as an active material. The barrel itself is divided into a central section which covers $|\eta| < 1$ and two extended barrels covering $0.8 < |\eta| < 1.7$. The tile calorimeter extends radially from 2.28 m to an outer radius of 4.25 m making a total thickness of 9.7 interaction lengths (λ) at $\eta = 0$. Hadrons interact with absorber nuclei producing a shower of particles. Light is emitted when the shower particles pass through the scintillator. Fibers at the end of each tile collect this light and carry it to photomultiplier tubes which convert it into an electric signal. This signal is used to determine the energy of the incident particles. The Hadronic End-cap Calorimeter (HEC) uses the same LAr technology as the ECAL, but with copper plates as an absorber. Each end-cap region consists of two wheels fixed behind the end-cap ECAL. With a small overlap with the tile calorimeter, the HEC covers $1.5 < |\eta| < 3.2$.

2.2.4 Muon spectrum system

The Muon Spectrometer (MS) as shown in Figure 2.10 [17] is the outer-most layer of the detector and has the capability of accurately measuring the muon momentum independent of the inner detector tracking system and provide an independent muon trigger. The spectrometer consists of two precision detectors and two triggering detectors, which are embedded in a toroidal magnetic field. The two precision detectors are the Monitored Drift Tubes (MDT) and the Cathode Strip Chambers (CSC) and the two triggering detectors: the Thin Gap Chambers (TGC) and the Resistive Plate Chambers (RPC). With the use of all these sub-detectors the muon momentum resolution is designed to be 3% for 10-200 GeV and about 10% for 1 TeV muons.

2.2.5 Trigger system

The designed luminosity of the LHC is 10^{34} cm^{-2}s^{-1} and the bunch crossing rate is about 40 MHz. Only maximum amount of data that can be stored from collision is 300 Mbs^{-1}, which corresponds to an event rate of 200 Hz. The event rate of 200 Hz is limited by the offline processing capacity. In 2012, the offline processing

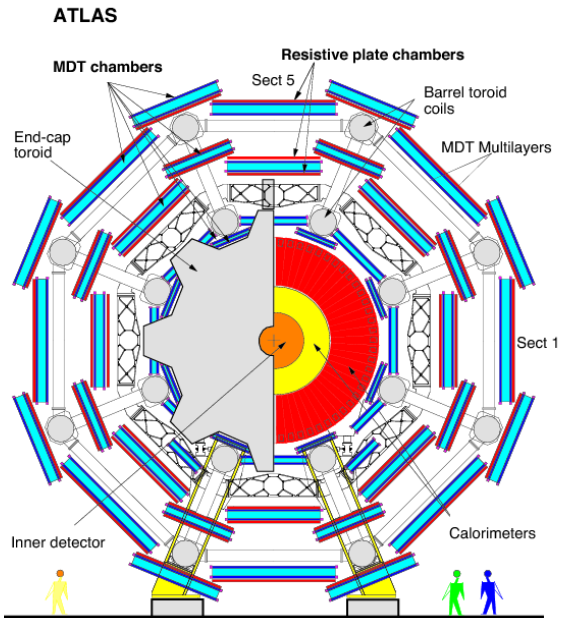


Figure 2.10: Schematic view of the muon spectrometer in the x-y projections. Inner, Middle and Outer chamber stations are denoted BI, BM, BO in the barrel and EI, EM, EO in the end-cap.

capacity was raised to almost 400 Hz, by having a delayed processing of some fraction of the data which won't be available till after the run stops. In order to achieve this challenging goal, the ATLAS has a three level trigger system, shown in Figure 2.11 [18], designed to reduce the event rate to a manageable level. The level 1 trigger consists of hardware while the subsequent two triggers, level 2 and event filter are software based. The combination of the level 2 and event filter are known as the high level trigger (HLT). Each subsequent level refines the decisions made in the previous levels.

2.2.6 The ATLAS simulation and computing

The ATLAS experiment has developed a computing model to allow members of the collaboration all over the world to access the ATLAS data. The main blocks of the model are the Athena [19] software framework, which operates on top of a hierarchical model of computing - the GRID [20]. GRID is used to created a distributed computing framework throughout several facilities in remote locations, that are able to communicate with each other and share tasks. Such facilities are referred to as Tiers. Tier-0 is located at CERN and handles the most unrefined data, referred to as RAW data. Ten worldwide facilities constitute Tier-1, that deals primarily with event reconstruction. Approximately 35 more facilities form Tier-2 which provides the analysis abilities for the ATLAS collaboration. Athena

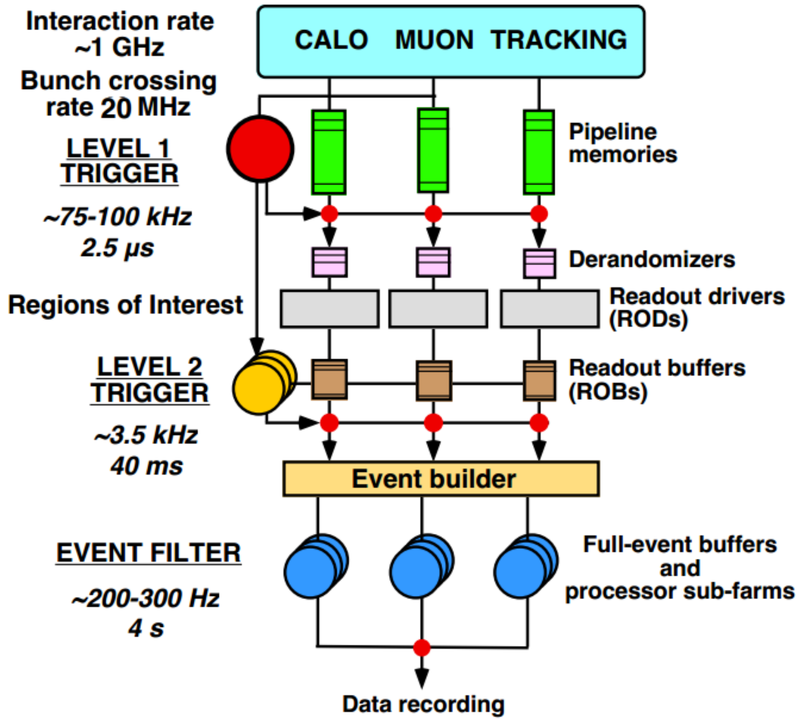


Figure 2.11: The ATLAS trigger system and its integration.

includes software for event simulation, event trigger, event reconstruction and physics analysis tools.

Event simulation [21] is fundamental in an experiment such as the ATLAS. It carries the events from generation to output, in a format which is identical to that of the true detector, allowing for a direct comparison between the real data and the theoretical models. The simulation software chain is generally divided into three steps. The first is the generation of events using a Monte Carlo program. The second step is the simulation of the detector and physics interactions.

In the ATLAS, the simulation is integrated into Athena and uses the Geant4 simulation toolkit [22]. Geant4 basically simulates the entire the ATLAS detector (material, geometry and subsystems, including trigger) and its response to traversing particles. Digitization is the last step, and consists in converting the energy deposited in the sensitive regions of the detector into voltages and currents, for comparison to the readout of the ATLAS detector. The output of this process is referred to as RAW data and is identical to the output of the ATLAS TDAQ system during real data taking.

The complexity of the full simulation (FullSim) of the ATLAS detector has led to the development of fast simulation strategies. These allow for a faster production of the high simulated event statistics needed for physics analyses. ATLFAST-II is a fast simulation framework developed for the ATLAS experiment. It is made of two components: the Fast ATLAS Tracking system (Fatras), for ID and MS simulation, and the Fast Calorimeter Simulation (FastCaloSim), for calorimeter simulation. Optionally, any of these sub-detectors can be simulated with the nominal Geant4, providing flexibility to suit the needs of different physics analyses.

2.2.7 The LHC and ATLAS in Run 2

Figure 2.12: The main 2013-2014 LHC consolidations toward Run 2 data taking.

The LHC Run 2 collisions started in June 2015 with the energy in the center of mass of 13 TeV, the largest energy scale in the world ever before. This high energy leads to a large increase in the production cross sections, ensure the observation and the study of rare process, the higher precision Standard Model measurements and improving the discovery potential for beyond the Standard Model physics like Supersymmetry, dark matter, etc. To adjust this so high collision energy, several upgrades were brought to the LHC and detector machines, and improvements to the online and offline reconstruction and analysis software during the first two years long shut down (2013-2015). Some details of the improvements will be presented in this section.

During the two years long Shut down-1 (LS1) the LHC machine was highly improved to target an energy of 6.5 TeV per proton beam and a bunch spacing of 25 ns. The main consolidations between 2013 and 2014 are illustrated in Figure 2.12 [23]. The Run 2 operation started with the initial beam commissioning in

April 2015. It lasted about two months firstly with test beams with an excellent and improved system performance, such as the beam instrumentation and collimation, injection and beam dump systems, vacuum and magnetic systems and the machine protection. Software and the analysis tools are also highly improved in the meanwhile.

Starting on April 5th, 2015 the first stable beams at 6.5 TeV were circulated in the LHC ring. Just after, a period of scrubbing, or of electron bombardment of the beam pipe surface, was mandatory to highly reduce the electron cloud around the beams. Finally, on June 3rd, 2015 the LHC started the first collisions at $\sqrt{s} = 13$ TeV. The LHC Run 2 physics proton-proton period started with around 30 days of collisions with 50 ns bunch spacing, and ATLAS collected an integrated luminosity about 0.1 fb^{-1}. Following stable runs are with 25 ns bunch spacing after August and the expected integrated luminosity to be collected is about 100 fb^{-1} totally before the next long shut down from 2018 to 2020. Total integral luminosity during the whole 2016 year is shown in Figure 2.13 (left) and the number of interactions per crossing with Run 2 condition is also presented. The average value is about 25 ns as expected and increased considerably compared to the Run 1 condition (Figure 2.4).

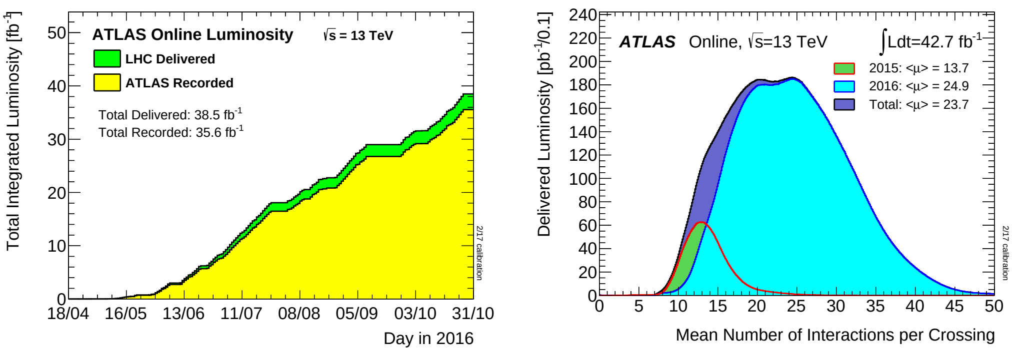


Figure 2.13: Total integrated luminosity in 2016 (left) and number of interactions per crossing with the combined 13 TeV data for 2015 and 2016 [24].

The ATLAS detector was also consolidated and several upgrades were performed during the LS1. It has a new beam-pipe, improved magnet and cryogenic systems. Compared to Run 1, the muon chambers were completed in the [1.1, 1.3] η region and repaired. The dead pixel modules and the calorimeter electronics were also reconditioned. To improve the tracking and the vertex reconstruction performance at high luminosities, a fourth pixel layer or the "Insertable B-layer" (IBL) was added. It is located just near the beam pipe, at a distance R = 3.3 cm, and presents a 3D silicon pixel technology. The purpose of this layer is to ensure a high quality tracking performance and an improved b-tagging at high luminosities for the LHC Run 2 operation.

3 Object reconstruction in the ATLAS

In the ATLAS, all physical objects are reconstructed with dedicated algorithms using the collision information recorded by the detector. All the ATLAS physics analyses always start with the reconstructed and identified objects representing the observed characteristics of the particles coming from the proton-proton collisions and travelling through the detector volume. The algorithms for reconstruction and identification of physics objects have been developed initially on simulated samples or early data, and then commissioned and optimized during the ATLAS running, keeping up with the changing energy and pile-up conditions of the collisions. The three leptons final state of the $t\bar{t}H$ includes leptons, jets and missing transverse energy. It is important to have good view of the details of the objects reconstruction in the ATLAS. Electrons (Section 3.1), muons (Section 3.3), photons (Section 3.2), tau(τ)s (Section 3.4), jets (Section 3.5) and missing transverse energy (Section 3.6) are presented in the following sections. The results described here are clearly the outcome of a well operated the ATLAS detector and a dedicated and committed work of the ATLAS performance and physics groups.

3.1 Electrons

Electrons are employed widely in the ATLAS physical analysis as the main primary signature for many physical processes. They are used in a wide range of physics analyses: from precision standard model measurements to the search for exotic new physics. The electron candidates then can be further selected against background, such as hadrons and background (non-prompt) electrons originating predominantly from photon conversions and heavy flavour hadron decays, using several sets of identification criteria with different levels of background rejection and signal efficiency. Many aspects of the overall design of the ATLAS were driven by the requirements that electrons be well-reconstructed and efficiently identified.

3.1.1 Reconstruction

The signature of an electron in the ATLAS is a reconstructed track in the Inner Detector (ID), associated to a narrow, localized cluster of energy in the electromagnetic (EM) calorimeter. Electron reconstruction in the central region of the ATLAS detector ($|\eta| < 2.47$) is with several steps as following:

- **Seed-cluster reconstruction**
  A sliding window with a size of 3×5 in units of 0.025×0.025, corresponding to the granularity of the EM calorimeter middle layer, in $\eta \times \phi$ space is used to search for electron cluster "seeds" as longitudinal towers with total cluster transverse energy above 2.5 GeV. The clusters are then formed around the seeds using a clustering algorithm [25] which allows for duplicated objects to be removed. The cluster kinematics are reconstructed by using an extended window depending on the cluster position in the calorimeter. The efficiency of this cluster searching ranges from 95% at $E_T = 7$ GeV to more than 99% above $E_T = 15$ GeV.

- **Track reconstruction**
  Track reconstruction is proceeding in two steps: pattern recognition and track fit. The ATLAS pattern recognition uses the pion hypothesis for energy loss because of the interactions with the detector material. This has been complemented with a modified pattern recognition algorithm which allows up to 30% energy loss at each intersection of the track with the detector material to account for possible bremsstrahlung. If a track seed (consisting of three hits in different layers of the silicon detectors) with a transverse momentum larger than 1 GeV can not be successfully extended to a full track of at least seven hits using the pion hypothesis and it falls within one of the EM cluster region of interest, a second attempt is performed with a new pattern recognition using an electron hypothesis that allows for larger energy loss. Track candidates are then fit either with the pion hypothesis or the electron hypothesis (according to the hypothesis used in the pattern recognition), using the ATLAS Global χ^2 Track Fitter [26]. If a track candidate fails the pion hypothesis track fit (for example, due to large energy losses), it is refitted with the electron hypothesis. In this way, a specific electron-oriented algorithm has been integrated into the standard track reconstruction. It improves the performance for electrons and has minimal interference with the main track reconstruction.

- **Electron specific track fit**
  The obtained tracks are then loosely matched to EM clusters by using the distance in η and ϕ between the position of the track, after extrapolation, in the calorimeter middle layer and the cluster barycentre. The matching condition account for energy loss dues to bremsstrahlung and the number of

precision hits in the silicon detector. Tracks which have larger than 4 significant number of precision hits and are loosely associated to electron clusters are refitted by using an optimised Gaussian Sum Filter (GSF) [27], which takes into account the non-linear bremsstrahlung effects.

- **Electron candidate reconstruction**
  The matching of the track candidate to the cluster seed is the final stage of the electron reconstruction procedure. A similar matching as the one described above is repeated for the refit track with more stricter conditions. If several tracks fulfil the matching condition, one track is chosen as a "primary" track. The choice is based on an algorithm using the cluster-track distance R calculated using different momentum hypotheses, the number of pixel hits and the presence of a hit in the first silicon layer [28]. Electron candidates without any associated precision hit tracks are removed and considered to be photons. The electron cluster is then re-formed using 3×7 (5×5) longitudinal towers of cells in the barrel (end-caps) of the EM calorimeter. The energy of the clusters is calibrated to the original electron energy using the multivariate techniques [29] based on simulated MC samples.

3.1.2 Calibration

Due to the detector geometry and material distribution, the correction are needed for the position measurement and energy measurement based on the MC samples. The position (η, ϕ) of the electron cluster is calculated as the energy-barycenter of all cells in the cluster, using the information from the first and second layers of the EM calorimeter. Corrections are considered as the LAr calorimeter accordion geometry and finite granularity. These are applied separately for each layer and for the barrel and end-cap regions of the detector. The η position is estimated in each cell of the cluster as a function of the impact point of the incident particle and is biased toward the center of the cell. As cells can have a finite granularity and the shower energy, which is not fully inside the cell, the η position is computed with respect to the cell center with an η correction. Considering the detector geometry, the correction is computed for several η regions and depends on the distance from the interaction point of the the incident particle to the center of the cell. The measurements are performed for each layer and then combined results are obtained to have the final position (η, ϕ) of the electron cluster.

Although the deposited energy of the electrons are recorded by the ATLAS efficiently, still a small energy is deposited in the non-instrumented section, as the inner detector, cryostat, solenoid, cables, etc. To account for these effects, the energy cells are calibrated. The EM cluster energy is calibrated to the initial electron energy using the special MC samples. A multivariate analysis is used by training on the MC as well as the inter-calibration of the scale of the first and second longitudinal EM calorimeter layers. The overall electron energy resolution measured in the data is adjusted as a function of η, using $Z \to ee$ events. Besides, to account for the worse resolution in data, additional scale factors are applied to MC electron energies to match the data (scale calibration).

The four-momentum of the electrons is computed using information from both the final calibrated energy cluster and the best track matched to the original seed cluster. The energy is given by the final calibrated cluster, while the η and ϕ directions are taken from the corresponding track parameters with respect to the beam-line. For Run 2 analyses, the electron measurements are performed by requiring the track associated with the electron to be compatible with the primary interaction vertex of the hard collision, and the aim is to reduce the background from conversions and secondary particles. The track parameters are calculated in a reference frame where the z-axis is taken along the measured beam-line position. The following conditions are applied together with all the identification operating points considered in this thesis: $d_0/\sigma_{d_0} < 5$ and $\Delta z_0 \sin\theta < 0.5$ mm, where the impact parameter d_0 is the distance of closest approach of the track to the measured beam-line, z_0 is the distance along the beam-line between the point where d_0 is measured and the beam-spot position, and θ is the polar angle of the track. To achieve the compatibility with the primary vertex of the hard collision the Δz_0 between the track and the primary vertex is employed. This vertex is selected from the reconstructed primary vertices (compatible with the beam-line) as the one with the highest sum of transverse momenta of the associated tracks. The σ_{d_0} represents the estimated uncertainty of the d_0 parameter, and θ is the polar angle of the track. The efficiency of these requirements in data and MC is estimated together with the efficiency of the various identification operating points.

3.1.3 Identification

The electrons from W or Z decays are typically isolated, hadron fake if it does not match a true electron or tau or muon, non-isolated if it matches a true electron originating from b or c mesons and background electron if it matches a true electron coming from Dalitz decays or conversions. To determine whether the reconstructed electron candidates are signal-like objects or background-like objects such as hardronic jets or converted photons, electron identification (EID) is applied. The EID uses quantities related to the electron cluster and track measurements including calorimeter shower shapes, information from the transition radiation tracker, track-cluster matching related quantities, track properties, and variables measuring the bremsstrahlung effects for the distinguishing signal from background. The quantities are summarised in Table 3.1 [30].

Towards Run 2, several changes to the input variables used for EID have been introduced. Taking advantage of the IBL, the number of hits in this innermost pixel layer is used for discriminating between electrons and converted photon and

Table 3.1: Definitions of electron discriminating variables.

Type	Description	Name
Hadronic leakage	Ratio of E_T in the first layer of the hadronic calorimeter to E_T of the EM cluster (used over the range $\|\eta\| < 0.8$ or $\|\eta\| > 1.37$)	R_{had1}
	Ratio of E_T in the hadronic calorimeter to E_T of the EM cluster (used over the range $0.8 < \|\eta\| < 1.37$)	R_{had}
Back layer of EM calorimeter	Ratio of the energy in the back layer to the total energy in the EM accordion calorimeter. This variable is only used below 100 GeV because it is known to be inefficient at high energies.	f_3
Middle layer of EM calorimeter	Lateral shower width, $\sqrt{(\Sigma E_i \eta_i^2)/(\Sigma E_i) - ((\Sigma E_i \eta_i)/(\Sigma E_i))^2}$, where E_i is the energy and η_i is the pseudorapidity of cell i and the sum is calculated within a window of 3×5 cells	ω_{stot}
	Ratio of the energy in 3×3 cells over the energy in 3×7 cells centered at the electron cluster position	R_ϕ
	Ratio of the energy in 3×7 cells over the energy in 7×7 cells centered at the electron cluster position	R_η
Strip layer of EM calorimeter	Shower width, $\sqrt{(\Sigma E_i (i - i_{\max})^2)/(\Sigma E_i)}$, where i runs over all strips in a window of $\Delta\eta \times \Delta\phi \approx 0.0625 \times 0.2$, corresponding typically to 20 strips in η, and $i_{\max}$ is the index of the highest-energy strip	w_{stot}
	Ratio of the energy difference between the largest and second largest energy deposits in the cluster over the sum of these energies	E_{ratio}
	Ratio of the energy in the strip layer to the total energy in the EM accordion calorimeter	f_1
Track conditions	Number of hits in the innermost pixel layer; discriminates against photon conversions	n_{Blayer}
	Number of hits in the pixel detector	n_{Pixel}
	Number of total hits in the pixel and SCT detectors	n_{Si}
	Transverse impact parameter with respect to the beam-line	d_0
	Significance of transverse impact parameter defined as the ratio of d_0 and its uncertainty	d_0/σ_{d_0}
	Momentum lost by the track between the perigee and the last measurement point divided by the original momentum	$\Delta p/p$
TRT	Likelihood probability based on transition radiation in the TRT	eProbabilityHT
Track-cluster matching	$\Delta\eta$ between the cluster position in the strip layer and the extrapolated track	$\Delta\eta_1$
	$\Delta\phi$ between the cluster position in the middle layer and the track extrapolated from the perigee	$\Delta\phi_2$
	Defined as $\delta\phi$, but the track momentum is rescaled to the cluster energy before extrapolating the track from the perigee to the middle layer of the calorimeter	$\Delta\phi_{res}$
	Ratio of the cluster energy to the track momentum	E/p

now is the second-to-innermost pixel layer. The change in the TRT gas led to modifications in the detector response and prompted the introduction of a new discriminating variable in the electron identification algorithms. In Run 1, only the fraction of high-threshold hits was used from the TRT as a signature of transition radiation to distinguish electrons from hadrons. In Run 2, a likelihood method based on the TRT high-threshold hits is introduced to compensate for the lower transition radiation absorption probability of the argon. The TRT likelihood method uses the high-threshold probability of each TRT hit to construct a discriminant variable, referred to here as eProbabilityHT. Cut-based method and Likelihood-based (LH) method are used with those discriminate variables to reject the background efficiently, especially for Likelihood-based method is the baseline in Run 2 analysis. However, higher collision energy and more complex pile-up situation compared to Run 1, both methods need re-optimize to adjust to the new condition.

Cut-based method

In Run 1 cut-based method using a set of rectangular cuts on the EID discriminating variables was used. Three sets of cut-based methodology identification menus are defined with increasing background rejection power: *Loose*, *Medium* and *Tight*. Towards Run 2 re-optimize work is done for cut-based method by using the $Z \to ee$, $J/\psi \to ee$ MC samples and $Z \to ee$ events of the first data recorded in 2015. With re-optimized cut-based method, similar performance with Run 2 condition comparing to Run 1 can be achieved. Even under the more complex pile-up condition, the re-optimized menus give a good separation power for the signal-like electron from the background (details in Appendix A, and this is my qualification task done in 2015). The re-optimized cut-based method is used as for cross-checks during the 2015 data taking.

Likelihood based method

Likelihood-based method (LH) is a multi-variate analysis (MVA) technique that simultaneously evaluates several properties of the electron candidates when making a selection decision. The LH method uses the signal and background probability density functions (PDFs) of the discriminating variables (shown in Table 3.1). Based on these PDFs, an overall probability is calculated for the object to be signal or background. The signal and background probabilities for a given electron are then combined into a discriminant $d_{\mathcal{L}}$ on which a requirement is applied:

$$d_{\mathcal{L}} = \frac{\mathcal{L}_S}{\mathcal{L}_S + \mathcal{L}_B}, \qquad \mathcal{L}_{S(B)}(\vec{x}) = \prod_{i=1}^{n} P_{s(b),i}(x_i), \tag{3.1}$$

where $\vec{x}$ is the vector of discriminating variable values and $P_{s,i}(x_i)$ is the value of the signal probability density function of the i^{th} variable evaluated at x_i. In the same way, $P_{b,i}(x_i)$ refers to the background probability function. This allows for a better background rejection for a given signal efficiency than a cut-based algorithm that would use selection criteria sequentially on each variable. In addition to the variables used as input to the LH discriminant, simple selection criteria are used for the variables counting the number of hits on the track.

Three levels of identification operating points are typically provided for EID. In order of increasing background rejection they are *Loose*, *Medium*, and *Tight*. The *Loose*, *Medium*, and *Tight* operating points are defined such that the samples selected by them are subsets of one another. Each operating point uses the same variables to define the LH discriminant, but the selection on this discriminant is different for each operating point. Thus, electrons selected by *Medium* are all selected by *Loose*, and *Tight* electrons are all selected by *Medium*. The distributions of electron shower shapes depend on the amount of material the electrons pass through, and therefore vary with the pseudorapidity of the electron candidates. In addition, significant changes to the shower shapes and track properties are expected with increasing energy. The EID operating points were consequently optimised in several bins in η and E_T. The performance of the LH identification algorithm is illustrated in Figure 3.1 [30]. Depending on the operating point, the signal (background) efficiencies for electron candidates with $E_T = 25$ GeV are in the range from 78 to 90% (0.3 to 0.8%) and increase (decrease) with E_T.

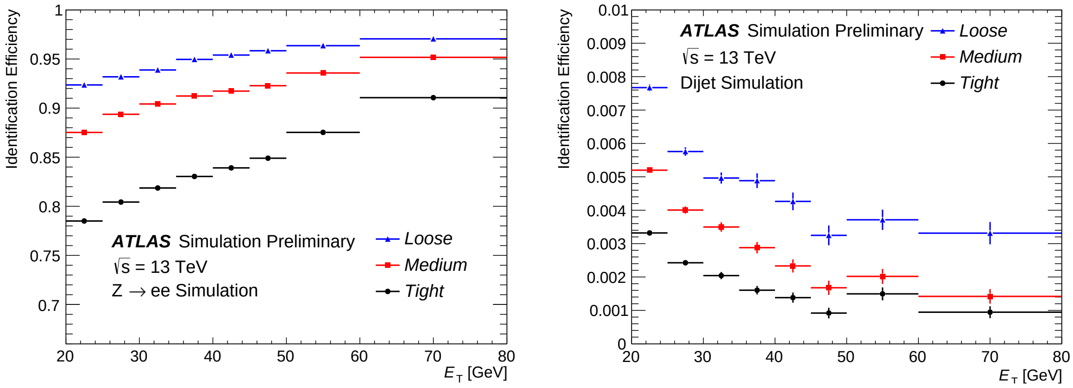


Figure 3.1: The efficiency to identify electrons from $Z \to ee$ decays (left) and the efficiency to identify hadrons as electrons (background rejection, right) estimated using simulated dijet samples. The efficiencies are obtained using MC, and are measured with reconstructed electrons. The candidates are matched to true electron candidates for $Z \to ee$ events. For background rejection studies the electrons matched to true electron candidates are not included. The last bin used for the optimisation of the ID is 45-50 GeV, which explains the signal efficiency increases slightly more in the 50 GeV bin than in others, and the background efficiency increases in this bin as well.

The electron identification performance may be influenced by the parasitic collisions taking place in the same beam crossing (in-time pile-up) or a consecutive bunch crossing (out-of-time pile-up) as the hard pp collision producing the electron candidates. The number of reconstructed primary vertices is indicative of the level of pile-up in each event, with the average number of primary vertices (eight per event) corresponding to an average pile-up of 13.7. Since some shower shape distributions depend on the number of pile-up collisions per bunch crossing, the cut on the LH discriminant value is loosened as a function of the number of primary vertices. This is done to ensure that the LH identification remains efficient at high pile-up condition without drastically increasing the amount of background accepted by the LH selection. The optimisation included simulations with a number of pile-up collisions of is up to 40, covering the range of the pile-up observed in 2015.

Some of the calorimeter variable distributions are different from the typical distributions obtained with $Z \to ee$ at high E_T. Higher energy electrons tend to deposit relatively smaller fractions of their energy in the early layers of the EM calorimeter, and more in the later layers of the EM calorimeter or even in the hadronic calorimeter. *Loose* and *Medium* of LH method were deemed to be loose enough to be robust against these E_T dependent changes. For electron candidates with E_T above 125 GeV, LH *Tight* uses the same discriminant selection as LH *Medium* but adding rectangular cuts on ω_{stot} and E/P, which were found to be particularly effective at discriminating signal from background at high E_T.

3.2 Photons

Reconstruction

The reconstruction of photon is performed separately for converted and unconverted photons. The former is characterized by the existence of at least one track originating from a vertex inside the tracker matched to the electromagnetic cluster, while the later is not. The conversion vertices reconstructed in the inner detector are classified according to the associated number of tracks. If a track is matched in η and ϕ to a reconstructed EM cluster, the object is added to the converted photon collection. The matching is considered to be successful if the impact parameter associated to the track, after extrapolation from its last measurement point to the second EM layer, is inside an (η, ϕ) window of radius 0.05 from the cluster centre. The track is extrapolated to the position corresponding to the expected maximum energy deposit for EM showers. Finally, if a reconstructed electromagnetic cluster can not be associated to a track, then it is considered to be an unconverted photon candidate.

Almost all converted photons are also regarded as electrons. The existence of a track associated to a conversion vertex is used to separate those two categories. If the track associated to the EM cluster coincides with a track originating from a conversion vertex and the electron is treated as an converted photon object. If the track can not be associated to a conversion vertex, the object is classified as an electron. In the case of a initially reconstructed electron with a matched track having only the TRT information (and the usual requirements on energy and track p_T are fulfilled), the candidate is considered to be a converted photon even if no conversion vertex is associated to the EM cluster. Unconverted photons are also recovered: if the reconstructed electron has the best track candidate with only the TRT information or the converted photon candidate condition is not passed and the track p_T is smaller than a typical value of 2 GeV, the candidate is considered to be an unconverted photon. The photon calibration is similar to electrons, and the most important sources of experimental systematic uncertainties are: photon energy scale and photon energy scale resolution.

Identification

Photon identification with high signal efficiency and high background rejection is required for transverse momenta from 10 GeV to the TeV sale to distinguish prompt photons from background photons. Photon identification in the ATLAS is based on a set of cuts on several discriminating variables listed in Table 3.2, characterise 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 one compared to background photons from jets, due to the presence of additional hadrons near the photon candidate in the latter case. Background photons from isolated $\pi^0 \rightarrow \gamma\gamma$ decays, unlike prompt photons, are often characterised by two separate local energy maxima in the finely segmented strips of the first layer, because of the small separation between the two photons. Due to the pile-up in presence of low-E_T activity in the detector this tends to broaden the distributions of the discriminating variables and thus to reduce the separation between prompt and background photon candidates.

A *Loose* and a *Tight* selection are defined. The *Loose* selection is based only on shower shapes in the second layer of the electromagnetic calorimeter and on the energy deposited in the hadronic calorimeter, and is used by the photon triggers. The *Loose* requirements are designed to provide a high prompt-photon identification efficiency with respect to reconstruction. Their efficiency rises from 97% at $E_T^\gamma = 20$ GeV to above 99% for $E_T^\gamma > 40$ GeV for both the converted and unconverted photons. The *Tight* selection adds information from the finely segmented strip layer of the calorimeter, which provides good rejection of hadronic jets where a neutral meson carries most of the jet energy. The *Tight* criteria are separately optimised for unconverted and converted photons to provide a photon identification efficiency of about 85% for photon candidates with transverse energy $E_T >$ 40 GeV. The selection criteria are different in seven intervals of the reconstructed photon pseudorapidity (0.0–0.6, 0.6–0.8, 0.8–1.15, 1.15–1.37, 1.52–1.81, 1.81–2.01, 2.01–2.37) to account for the calorimeter geometry and for different effects on the shower shapes from the material upstream of the calorimeter.

The photon identification criteria were first optimised prior to the start of the data taking in 2010, using simulated samples of prompt photons from γ+jet, diphoton and $H \rightarrow \gamma\gamma$ events and samples of background photons in QCD multi-jet events [31]. To adjust the 8 TeV run in 2012, the identification were re-optimised based on improved simulations in which the values of the shower shape variables are slightly shifted to improve the agreement with the data shower shapes. To cope with the higher pile-up, the criteria on the shower shapes which is more sensitive to pile-up were relaxed while the others were tightened.

For the data taken in 2011, 4.9 fb^{-1} at $\sqrt{s} = 7$ TeV, the efficiency of the cut-based identification algorithm increases from 60–70% at $E_T = 20$ GeV up to 87–95% (90–99%) at $E_T > 100$ GeV for unconverted (converted) photons. With an optimised neural network this efficiency increases from 85–90% at $E_T = 20$ GeV to about 97% (99%) at $E_T > 100$ GeV for unconverted (converted) photon candidates for a similar background rejection. For the data taken in 2012, 20.3 fb^{-1} at $\sqrt{s} =$ 8 TeV, the efficiency of a re-optimised cut-based photon identification algorithm increases from 50–65% (45–55%) for unconverted (converted) photons at $E_T = 10$ GeV to 95–100% at $E_T > 100$ GeV, being larger than 90% for $E_T > 40$ GeV [32].

Category	Description	Name	Loose	Tight
Acceptance	$\|\eta\| < 2.37$, $1.37 < \|\eta\| < 1.52$ excluded	–	$\surd$	$\surd$
Hadronic leakage	Ratio of E_T in the first sampling of the hadronic calorimeter to E_T of the EM cluster (used over the range $\|\eta\| < 0.8$ and $\|\eta\| > 1.37$)	R_{had_1}	$\surd$	$\surd$
	Ratio of E_T in all the hadronic calorimeter to E_T of the EM cluster (used over the range $0.8 < \|\eta\| < 1.37$)	R_{had}	$\surd$	$\surd$
EM Middle layer	Ratio in η of cell energies in 3×7 versus 7×7 cells	R_η	$\surd$	$\surd$
	Lateral width of the shower	ω_{η_2}	$\surd$	$\surd$
	Ratio in ϕ of cell energies in $3{\times}3$ and $3{\times}7$ cells	R_ϕ		$\surd$
EM Strip layer	Shower width for three strips around maximum strip	ω_{s3}		$\surd$
	Total lateral shower width	ω_{stot}		$\surd$
	Fraction of energy outside core of three central strips but within seven strips	F_{side}		$\surd$
	Difference between the energy associated with the second maximum in the strip layer, and the energy reconstructed in the strip with the minimal value found between the first and second maxima	ΔE		$\surd$
	Ratio of the energy difference associated with the largest and second largest energy deposits over the sum of these energies	E_{ratio}		$\surd$

Table 3.2: Variables used for *Loose* and *Tight* photon identification cuts.

3.3 Muons

Reconstruction

Muon reconstruction is first performed independently in the ID and MS. The information from individual sub-detectors is then combined to form the muon tracks that are used in physics analyses. Muon track candidates are then built by fitting together hits from segments (MDT, RPC, CSC and TGC) in different layers. The algorithm used for this task performs a segment-seeded combinatorial search that starts by using as seeds the segments generated in the middle layers of the detector where more trigger hits are available. 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. 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 [33]. A track candidate is accepted if the χ^2 of the fit satisfies the selection criteria. The calculation of the energy loss in the calorimeter was also improved. An analytic parametrization of the average energy loss is derived from a detailed description of the detector geometry. The final estimate of the energy loss is obtained by combining the analytic parametrization with the energy measured in the calorimeter. This method yields a precision on the mean energy loss of about 30 MeV for 50 GeV muons.

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 (*Loose*, *Medium*, *Tight*, and *High-*p_T) are provided to address the specific needs of different physics analyses. *Loose*, *Medium*, *Tight* are inclusive categories so that muons identified with tighter requirements are also included in the looser categories. Four identification selection criteria can separate low ($4 < p_T < 20$ GeV) and high ($20 < p_T < 100$ GeV) transverse momentum muon candidates.

Muon reconstruction performance has been measured by using 3.2 fb^{-1} of data at $\sqrt{s} = 13$ TeV recorded during the 25 ns run at the LHC in 2015. The muon reconstruction efficiency is close to 99% over most of the pseudorapidity range of $|\eta| < 2.5$ for $p_T > 5$ GeV. The $Z \to \mu\mu$ sample enables the measurement of the efficiency with a precision at the 0.2% level for $p_T > 20$ GeV. The $J/\psi \to \mu\mu$ sample provides a measurement of the reconstruction efficiency between 5 and 20 GeV with a precision better than 1%. The $Z \to \mu\mu$ sample is also used to measure the isolation efficiency for seven isolation working points in the momentum range 10–120 GeV. The isolation efficiency varies between 93% and 100% depending on the selections and on the momentum of the particle, and is well reproduced in the simulation. The muon momentum scale and resolution have been studied in details using $J/\psi \to \mu\mu$ and $Z \to \mu\mu$ decays. These studies are used to correct the simulation to improve the agreement with data and to minimise the systematic uncertainties in physics analyses [33].

3.4 Hadronic taus

In the ATLAS, τ reconstruction and identification concentrates on the hadronic decay modes of a τ lepton. The majority of τ hadronic decays are characterized by one or three charged pions accompanied by neutral pions and neutrinos. τ candidates are seeded by jets formed using the procedure described in Section 3.5. A τ-specific energy calibration, baseline calibration or boosted regression tree (BRT) [34] is applied to the τ candidate 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.

The τ identification algorithm is designed to reject backgrounds from quark-initiated and gluon-initiated jets. The identification uses Boosted Decision Tree (BDT) based method. Three working points labelled *Loose*, *Medium* and *Tight* are provided, and corresponding to different τ identification efficiency values, with the efficiency designed to be independent of p_T. The input variables of 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.

3.5 Jets

Jets are essential objects in the ATLAS physics analysis. They play a fundamental role in many physics measurements. In the searching of the $t\bar{t}H$ production analysis with the multi-lepton final states, jets can be employed to discriminate the signal events from diboson process for instance. Reconstruction of the jets starts with 3D topological clusters built in the ATLAS calorimeters, which are matched to tracks of charged particles measured in the inner detector. The algorithm used is anti-k_t [35] and the energy calibration is performed using MC simulations.

The inputs of jet reconstruction in the ATLAS are locally-calibrated 3D topological clusters (topo-clusters), built from calorimeter cells. Topo-clustering starts by identifying seed cells with energy significance 4σ above noise. The noise here is defined as the sum in quadrature of electronic and pile-up noise. Neighbour cells with energy significance higher than 2 are then added to form the seed clusters. An extra ring of direct neighbour cells is added to the final clusters. After topo-clusters finding, a splitting algorithm is used to further separate clusters, based on local energy maxima within clusters. Individual clusters are calibrated using local properties such as energy density, calorimeter depth, and isolation with respect to nearby clusters. This local cluster weighting calibration (LCW) allows clusters to be classified as electromagnetic or hadronic and uses a dedicated cluster calibration derived from single pion Monte Carlo simulations. Jets are built using the anti-k_t algorithm with radius parameters R=0.4. Jets are calibrated to the particle level in dijet events: a Monte Carlo pile-up offset correction, a Monte Carlo jet energy response correction, and an residual calibration are applied to jets in data only, to account for the differences in response between data and Monte Carlo.

Main background jets come from proton beam collisions with the residual gas in the beam pipe, interactions in the tertiary collimators, cosmic muons overlapping in-time with collision events, calorimeter noise, etc. Four selections are defined with different levels of fake-jets rejection: *Looser*, *Loose*, *Medium* and *Tight*. The most loose one has the highest jet efficiency while the tight one has the highest background rejection. Their definitions are based on the reconstructed energy at the cell level, on the jet energy deposited in the direction of the shower development and on the number of reconstructed tracks matched to the jets. The pile-up offset correction aims at subtracting the extra energy added to jet by additional pp interactions overlapping with the physics events of interest. Another effect of pile-up is to generate the additional fake jets. Such fake jets originating from pile-up fluctuations after the application of the offset correction are rejected by using the Jet Vertex Fraction (JVF) algorithm [36]. JVF calculates the fraction of total track p_T matched to a jet that originates from the hard scatter vertex. Pile-up jets have very small JVF values as most of their tracks originate from additional pile-up vertices.

b-jets tagging

The identification of jets containing b hadrons is an important tool in precision measurements in the top quark sector as well as in the search for the Higgs boson and new phenomena, the suppression of background processes that contain predominantly light-flavour jets using b-tagging is of great use. It may also become critical to achieve an understanding of the flavour structure of any new physics (e.g. supersymmetry) revealed at the LHC. An illustration of the production of a b-jet is shown in Figure 3.2. The decay of a b hadron is suppressed by a Cabibbo–Kobayashi–Maskawa (CKM) [37] factor which results in a longer flight path before decaying in the inner detector compared to charm and light hadrons.

The rate at which a true b-jet is identified defines the b-tagging efficiency for a particular b-tagging algorithm. The true flavour of a jet is defined in simulated data using a spatial ΔR matching between stable hadrons and reconstructed jets. A hierarchy matching is performed, first checking whether a b hadron can be matched, followed by a charm hadron and followed by a τ-lepton. This matching procedure results in a jet being classified respectively as either a b-jet, c-jet, τ-jet or a light-jet.

There are three main categories of b-tagging algorithms commonly used in the ATLAS. There are impact parameter based algorithms (IP2D, IP3D) [38]; inclusive secondary vertex reconstruction algorithms (SV); and decay chain reconstruction algorithms (JetFitter) [39]. These algorithms contribute complementary information and can be combined by using a multivariate function to create a single b-tagging discriminant (MV).

The three types of algorithms provide input to a multivariate classifier. In Run 2

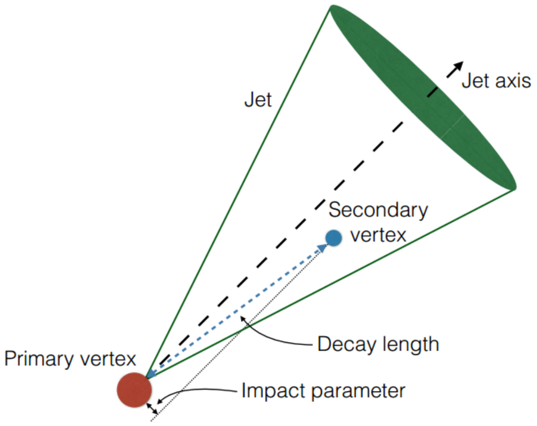


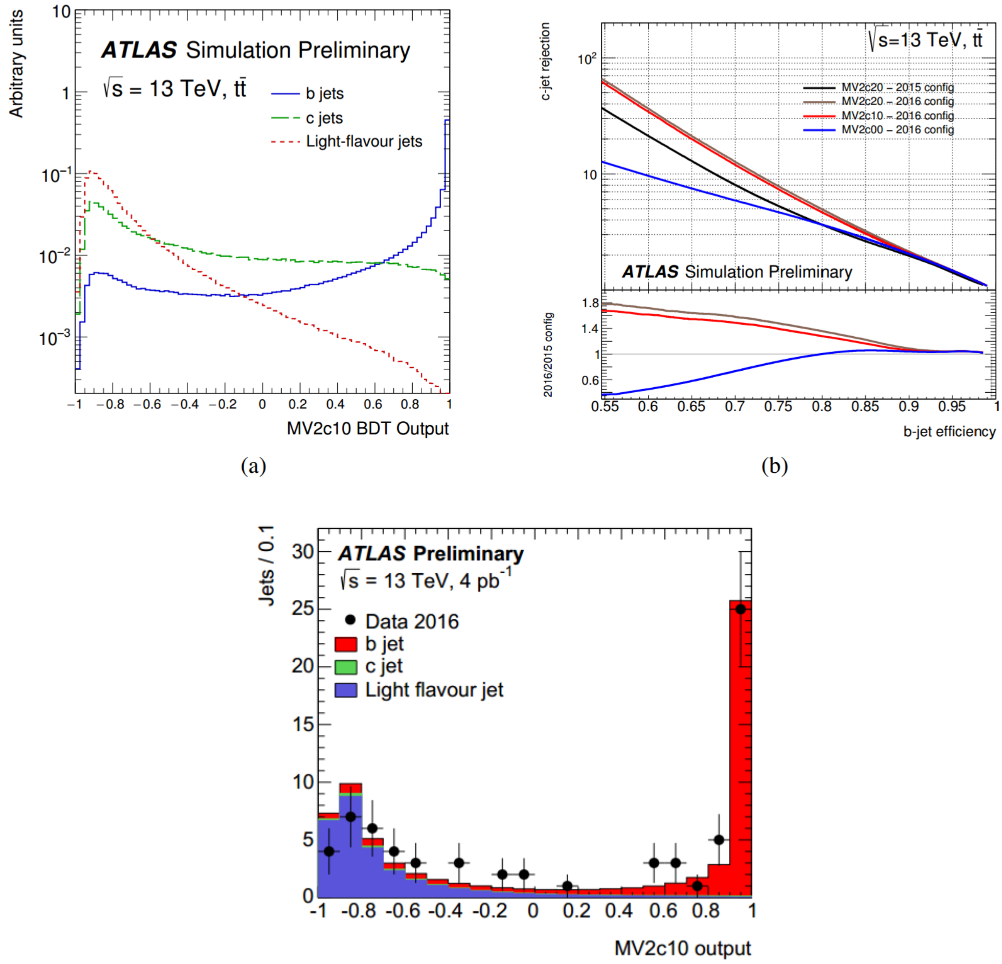
Figure 3.2: Illustration of the production of a b-tagged jet.

this classifier is a boosted decision tree (BDT) trained to discriminate b-jets from light-jets. The background samples of light-jets can contain an admixture of charm and light-jets in order to improve the charm-jet rejection. The discrimination performance is shown in Figure 3.3(a) for the MV2 (MV2c10) algorithm with a 10% charm-jet admixture. The performance between the Run 1 MV algorithm (MV1) and the Run 2 MV algorithm (MV2) has seen an improvement of 10%, primarily driven by the inclusion of the IBL. Between 2015 and 2016, retraining of the b-tagging classifier has further improved the charm-jet rejection by around 40% at the 77% working point. The charm-jet rejection rate as a function of b-tagging efficiency is shown in Figure 3.3(b). Jets which have been selected using a top-pair event selection are shown in Figure 3.3(c) comparing the performance of the b-tagging weight (MV2c10) in simulation and data, where a good agreement can be observed [40].

Benchmark values obtained with MV2 algorithm are shown in Table 3.3 [41]. Several working points (WP) are shown together with the b-tagging efficiency, purity, c quark and light flavor quark rejection factor (RF) together with the tau lepton rejection efficiency. The 70% working point of MV2c10 is chosen in this thesis for the following section when b-tagged jet efficiency mentioned.

WP[%]	b jet efficiency[%]	Purity[%]	c RF	τ RF	light jets RF
60	60.03	99.00	34.54	183.98	1538.78
70	69.97	97.46	12.17	54.72	381.32
77	76.97	95.17	6.21	22.04	134.34

Table 3.3: b-tagging benchmarks of MV2 algorithm.



(a) (b) (c)

Figure 3.3: (a): The performance of the MV2c10 algorithm is shown for different jet flavours [40], (b): a comparison between the MV2c algorithms trained in 2015 and trained in 2016 [40] and (c): the performance of the MV2c10 algorithm compared to 2016 data [42].

3.6 Missing transverse energy

The Missing Transverse Energy (MET) in the ATLAS is reconstructed from cells belonging to topo-clusters and from reconstructed muons. Cells in topo-clusters are calibrated and the calibration of all physics objects in each final state is also propagated to the MET. The soft term of the MET, which consists of topo-clusters not belonging to any reconstructed physics object, is corrected for the effect of pile-up using a track-based technique. The Soft Term Vertex Fraction (STVF) [43] is defined as the ratio of the sum of p_T of all tracks unmatched to jets from the hard-scatter vertex and all tracks unmatched to jets from all vertices in a given event. The soft term of the MET is then rescaled by STVF, event-by-event. The MET performance and systematic uncertainties are established from differences between data and simulations of the MET distribution in $Z \to ll$ and $W \to l\nu$ events. The evolution of the E_T^{miss} resolution is shown for different numbers of jets in Figure 3.4 with the TST E_T^{miss} algorithm as a representative example [44].

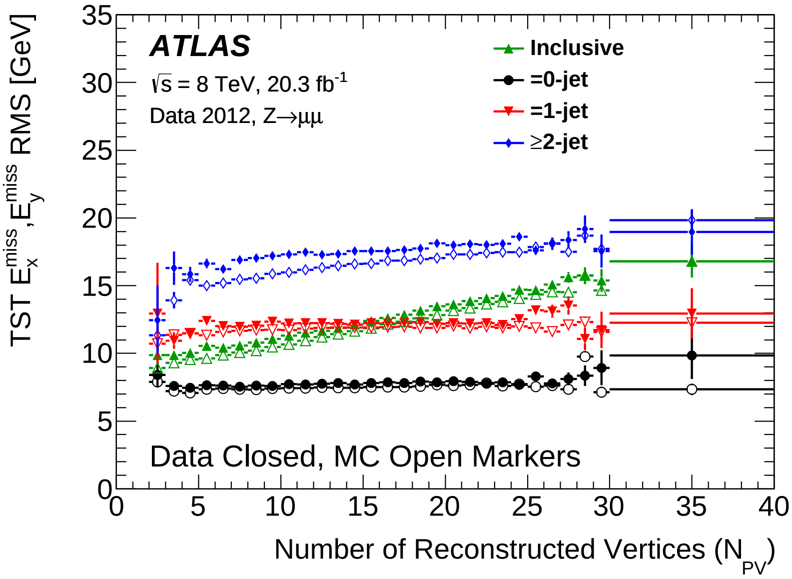


Figure 3.4: The resolution of the combined distribution of E_x^{miss} and E_y^{miss} for the TST E_T^{miss} as a function of NPV for the 0-jet, 1-jet, ≥2-jet, and inclusive $Z \to \mu\mu$ samples. The data (closed markers) and MC simulation (open markers) are overlaid. The jet counting uses the same JVF criterion as the TST E_T^{miss} reconstruction algorithm.

4 Search for the production of the Higgs boson associated with a pair of top quarks

4.1 Introduction

The Higgs boson production associated with a pair of top quarks is one of the most important measurements in Run 2 after Higgs discovery. Details of the $t\bar{t}H$ production and results with Run 1 data have been mentioned in Chapter 1.6. The search for $t\bar{t}H$ production with Run 2 data will be presented in this chapter.

The $t\bar{t}H$ production can be observed through different topologies according to the Higgs and top quark decays. The analysis presented in this thesis is to search for Higgs decays to WW^*, $\tau\tau$, ZZ^*. The top quark and anti-top quark decays to $W^{\pm}b$. Especially for Higgs decays to WW^*, each $W^{\pm}$ boson in the final state decays either leptonically ($\ell = e^{\pm}, \mu^{\pm}, \tau^{\pm}$) with missing energy or hadronically, leading to many topologies. Depending on the number of leptons in the final state of the signal events, topologies could be the 2ℓSS (exactly two light leptons with same sign charges and with hadronic tau veto); 3ℓ (exactly three light leptons and with hadronic tau veto); 4ℓ (exactly four light leptons), etc. The search for $t\bar{t}H$ in the exactly three leptons final state is presented in thesis as it's my main work during the PhD study.

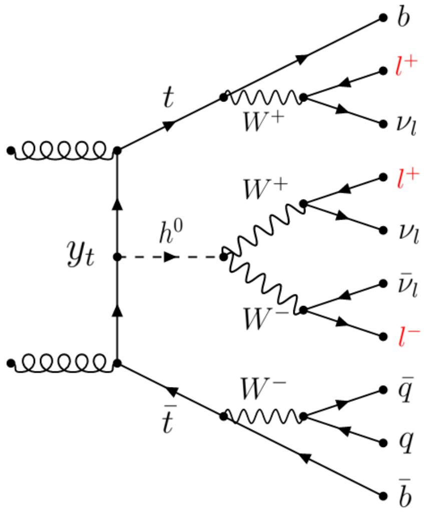


Figure 4.1: Feynman diagram of the $t\bar{t}H$ signature with 3 leptons final state.

The feynman diagram for the final state of 3ℓ with Higgs decays to WW for instance is shown in Figure 4.1. The Higgs decays to WW^* channel makes the dominant contribution compared to others as shown in Table 4.1. The channels associated with hadronic τ is considered separately, denoted τ channel in $t\bar{t}H$ searching. Object definition and the selections used in analysis are mentioned in the following sections.

	3ℓ	4ℓ	2ℓSS
$H \to WW^*$	74%	72%	77%
$H \to ZZ^*$	4%	9%	3%
$H \to \tau\tau$	20%	18%	17%
$H \to$ others	2%	2%	3%

Table 4.1: Contributions from different Higgs decay modes for the 2ℓSS, 3ℓ and 4ℓ channel.

4.2 Data and Monte Carlo samples

4.2.1 Collision data

The analysis presented in this thesis uses 36.1 fb^{-1} of data collected from proton-proton collision recorded by the ATLAS detector at $\sqrt{s} = 13$ TeV during 2015 and 2016. The dataset has been collected with a bunch crossing of 25 ns and verified data quality cuts namely which must be in the recommended Good Run List.

The data is prepared with `xAOD` format and further produced to `DxAOD` format using a dedicated derivation framework. This `xAOD` to `DxAOD` derivation provides a reduction specifically for the $t\bar{t}H$ events with multileptons in the final states. The total size of dataset is reduced to 3.6% for simulated $t\bar{t}$ events and 0.1% for collision dataset. These reductions come from the effects of smart slimming (remove un-needed variables), thinning (remove entire objects from events) and additional skimming for collision dataset only. The skimming means removing the whole event if it is rejected by any of the following selections:

- at least two light leptons passing loose identification criteria with leading lepton $p_T > 15$ GeV and subleading lepton $p_T > 5$ GeV, and $|\eta| < 2.6$.

- at least one light lepton passing loose identification criteria with $p_T > 15$ GeV and $|\eta| < 2.6$, and at least two hadronic taus. The τ jet has to pass `JetBDTSigLoose` requirement with $p_T > 15$ GeV, and its charge must be one and has one or three associated tracks.

4.2.2 Monte Carlo samples

Event generator programs and configurations used for simulating the signal and background processes are summarized in Table 4.2. The Higgs mass is 125 GeV and this is assumed for all.

Productions of $t\bar{t}H$, $t\bar{t}W$ and $t\bar{t}Z$ are generated with a next-to-leading order (NLO) QCD matrix element computed by MadGraph5_aMC@NLO, interfaced to Pythia8 for parton showering and fragmentation into particles. In the case of $t\bar{t}Z$, the inclusive $t\bar{t}\ell\ell$ matrix element is computed, including off-shell Z and γ^* contributions with $m(\ell\ell) > 5$ GeV.

The overall $t\bar{t}H$ cross section is 507.1 fb, which is computed at NLO in QCD [51, 52, 53, 54, 55, 56, 57, 58, 59]. It has uncertainties of $^{+5.8\%}_{-9.2\%}$ from QCD renormalization/factorization scale choice and $\pm 3.6\%$ from parton distribution function uncertainties (including α_s uncertainties).

The cross sections for $t\bar{t}V$ production, including the process $pp \to t\bar{t}\ell^+\ell^- + X$ over the full Z/γ^* mass spectrum, are computed at NLO in QCD and electroweak couplings. These have QCD scale uncertainties $\approx 12\%$ and PDF+α_s uncertainties of $3 - 4\%$. The total cross section used for $t\bar{t}\ell^+\ell^-$ (with $M(\ell^+\ell^-) > 5$ GeV) is 123.7 fb, and 600.8 fb for $t\bar{t}W^\pm$ [60, 57].

$t\bar{t}$ events are generated with Powheg v2.0 and interfaced with Pythia 8 for the parton showering and fragmentation. A14 tune is used for showering. The overall $t\bar{t}$ cross section is 832 pb. Powheg is also used to model other top backgrounds such as single top t-channel, s-channel and Wt-channel.

Diboson processes are generated with Sherpa v2.1.1 at leading order, using CT10 PDF set. The matrix elements consider for the fully leptonic diboson production with up to three additional partons. Processes split into b-filtered and b-vetoed

Process	ME Generator	Parton Shower	PDF	Tune
$t\bar{t}H$	MG5_aMC	Pythia 8	NNPDF 3.0 NLO/2.3 LO	A14[45]
$tHqb$	MG5_aMC	Pythia 8	CT10	A14
tHW	MG5_aMC	Herwig++	CT10/CTEQ6L1	UE-EE-5[46]
$t\bar{t}W$	MG5_aMC	Pythia 8	NNPDF 3.0 NLO/2.3 LO	A14
$t\bar{t}(Z/\gamma^{*})$	MG5_aMC	Pythia 8	NNPDF 3.0 NLO/2.3 LO	A14
$t(Z/\gamma^{*})$	MG5_aMC	Pythia 6	CTEQ6L1	Perugia2012
$tW(Z/\gamma^{*})$	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	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[47, 48]	Pythia 6	CT10/CTEQ6L1	Perugia2012
VV, $qqVV$, VVV	Sherpa 2.1.1[49]	Sherpa	CT10	Sherpa
$Z \to \ell^{+}\ell^{-}$	Sherpa 2.2	Sherpa	NNPDF 3.0 NNLO	Sherpa
$W \to \ell\nu$	Sherpa 2.2	Sherpa	NNPDF 3.0 NNLO	Sherpa

Table 4.2: Configurations of event generations used for 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 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; "Pythia 6" refers to version 6.427; "Pythia 8" refers to version 8.2; "Herwig++" refers to version 2.7. Samples using Pythia 6 and Pythia 8 have heavy flavor hadron decays modelled by EvtGen 1.2.0. All samples include leading-logarithm photon emission, either modelled by the parton shower generator or by PHOTONS [50].

samples, according to having or not truth jets containing a b hadron with transverse momentum above 5 GeV.

All Monte Carlo samples are processed through a complete simulation of the ATLAS detector response based on Geant4. Additional simulated pp collisions generated with Pythia 8 were overlaid to model the effects of both in- and out-of-time pile-up, from additional pp collisions in the same and nearby bunch crossings. The pile-up distribution is reweighed to reflect the mean number of additional interactions observed in data. All simulated events were processed using the same reconstruction algorithms and analysis chain as the data. Simulated events are corrected so that the object reconstruction and identification efficiencies, energy scales and energy resolutions match those determined from data.

4.3 Object definition and basic selection

Many objects like leptons, jets, neutrinos are presented in the 3ℓ final state as shown in the Feynman diagram 4.1. The details of the reconstruction of the common objects are described in Chapter 3. The optimizations on those are done to improve the signal acceptance and signal significance preliminarily. Following sections introduce the detailed criterion on different objects. Firstly at event level, the primary vertex with requirement of the vertex with largest $\sum p_T^2$ of associated tracks with $p_T > 400$ MeV is required in an event. Events with significant noise in the calorimeters or data corruption which is due to software and hardware failures are removed.

4.3.1 Trigger

Because of the different running conditions between 2015 and 2016 (more complex pile-up), the trigger is also different for data collected in each year. Generally, single lepton trigger (single electron, single muon) or dilepton trigger ($ee, e\mu, \mu\mu$) is applied to the events. Lowest p_T threshold, unprescaled single lepton (SLT) and dilepton (DLT) trigger chains used in the current trigger of analysis are presented in Table 4.3 [a] and Table 4.4 for 2015 and 2016 respectively.

	Single lepton triggers (2015)
μ	HLT_mu20_iloose_L1MU15, HLT_mu50
e	HLT_e24_lhmedium_L1EM20VH, HLT_e60_lhmedium, HLT_e120_lhloose
	Dilepton triggers (2015)
$\mu\mu$	HLT_mu18_mu8noL1
ee	HLT_2e12_lhloose_L12EM10VH
$e\mu, \mu e$	HLT_e17_lhloose_mu14

Table 4.3: List of lowest p_T-threshold, unprescaled triggers used for the whole 2015 data taking.

Several studies showed that this logical combination of triggers provides the largest gain in signal acceptance of the same sign lepton pair, allowing the minimum lepton p_T threshold of the trigger-matched leading lepton to be lower than the allowed by the single lepton triggers alone. From the fake estimate side, no trigger bias originating from the online lepton trigger selection is observed to significantly

[a] The "HLT" means the high level trigger which consists of the Level 2 (LVL2) and the Event Filter (EF) trigger. The "mu" is muon and "e" stands for electron. The number after the muon or electron is the cut on the corresponding p_T. The "L1" means the L1 trigger level. The "iloose" is the *Loose* working point of the muon identification and "lhloose", "lhmedium" are the *Loose, Medium* working point of the LH electron identification. The "nod0" means there is no $d0$ requirement which is shown in Table 4.3. The "VH" means the hadronic veto requirement.

	Single lepton triggers (2016)
μ	HLT_mu26_ivarmedium, HLT_mu50
e	HLT_e26_lhtight_nod0_ivarloose, HLT_e60_lhmedium_nod0, HLT_e140_lhloose_nod0
	Dilepton triggers (2016)
$\mu\mu$	HLT_mu22_mu8noL1
ee	HLT_2e17_lhvloose_nod0
$e\mu, \mu e$	HLT_e17_lhloose_nod0_mu14

Table 4.4: List of lowest p_T-threshold, unprescaled triggers used for the whole 2016 data taking.

affect the Matrix Method results when using an `OR` of single lepton and dilepton triggers, hence the proposed choice does not require any additional complexity in the fake estimate procedure. As for data-driven trigger efficiency corrections to be applied to simulated samples, a new tool developed which can provide a per-event correction for the chosen combination of triggers and a given offline lepton selection is used.

4.3.2 Electrons

The electron objects used in thesis are with the requirements as follows. The candidates are with $p_T > 10$ GeV. They are required to satisfy $|\eta_{cluster}| < 2.47$. The electron candidates which are in the transition region between different electromagnetic calorimeter components, $1.37 < |\eta_{cluster}| < 1.52$, are rejected. The *Loose* working point of the likelihood electron identification is employed in the object preselection. No isolation requirement is applied at the preselection level as to increase the statistics of the fake leptons estimate with data driven method. The non-prompt electrons are further reduced with the track requirement of being consistent with originating from the primary vertex and the requirements (shown in Table 4.5) are imposed on the transverse impact parameter significance ($|d_0|/\sigma_{d0}$) and the longitudinal impact parameter ($|\Delta_{Z0} \sin\theta_l|$).

4.3.3 Muons

In the region $|\eta| < 0.1$, where muon spectrometer coverage is reduced, muons are also reconstructed from the inner detector tracks matched to the isolated energy deposits in the calorimeter consistent with the passage of a minimum-ionizing particle. Candidates are required to satisfy $p_T > 10$ GeV and $|\eta| < 2.5$. The transverse impact parameter is sightly tighter than the electrons, while the longitudinal impact parameter selection is the same. No isolation is required in the object preselection.

All the requirements above on lepton candidates are grouped and called Loose level objects. The Tight level objects should satisfy the Tight LH working point for electrons at some cases to reduce the fake and non-prompt electrons. A dedicated isolation variable, `PromptLeptonIso` based on Boosted Decision Tree algorithm is applied to improve the reduction of the background of non-prompt produced in hadron decays. The detailed requirements for Loose and Tight leptons are summarised in Table 4.5.

	Loose		Tight	
	e	μ	e	μ
Non-prompt lepton BDT and loose isolation	-	-	Yes	Yes
Identification	Loose	Loose	Tight	Loose
Charge misID veto BDT	-	-	Yes ($e^{\pm}e^{\pm}$ and $e^{\pm}\mu^{\pm}$)	-
Transverse impact parameter significance $\lvert d_0\rvert/\sigma_{d_0}$	< 5	< 3	< 5	< 3
Longitudinal impact parameter $\lvert\Delta z_0 \sin\theta_\ell\rvert$ (mm)	< 0.5	< 0.5	< 0.5	< 0.5

Table 4.5: Loose and Tight cut definitions for leptons. Selections for tight leptons are applied on top of the selections for loose leptons. IsoLoose is an additional minimum "99% eff" calorimeter and track isolation requirement on top of the `PromptLeptonIso`, and refers to isolation working points designed to be 99% efficient for isolated leptons at all p_T range. The cut on the charge misID is applied in $e^{\pm}e^{\pm}$ and $e^{\pm}\mu^{\pm}$ channels when doing the fake estimation.

4.3.4 Jets and *b*-tagged jets

Jets are reconstructed from calibrated topological clusters built from energy deposits in the calorimeters, using the anti-k_t algorithm with a radius parameter $R = 0.4$ which are described in Section 3.5. Jets with energy contributions arising from noise or detector effects are removed, and only jets satisfying $p_T > 25$ GeV and $|\eta| < 2.5$ are used in this analysis. For jets with $p_T < 60$ GeV and $|\eta| < 2.4$, a jet-track association algorithm is used to confirm that they originates from the selected primary vertex, in order to reject jets arising from pile-up collisions. The average efficiency of this association is 92% per jet.

Jets containing b-hadrons are identified (b-tagging) via a multivariate discriminant method. The working point used in this search corresponds to an average efficiency of 70% (explained in Section 3.5) for b-tagged jets with $p_T > 20$ GeV and $|\eta| < 2.5$. The expected rejection factors against light and c-jets are 380 and 12, respectively.

4.3.5 Overlap removal

A procedure called overlap removal is designed to avoid double counting objects and to remove leptons originating from hadron decays. The overlap removal is done in the following order: any electron candidate within $\Delta R = 0.1$ of another electron candidate with higher p_T is removed; any electron candidate within $\Delta R = 0.1$ of a muon candidate is removed; any jet within $\Delta R = 0.3$ of an electron candidate is removed; if a muon candidate and a jet lie within $\Delta R = \min(0.4,\ 0.04 + 10$ [GeV]$/p_T$(muon)) of each other, the jet is kept and the muon is removed; any τ_{had} candidate within $\Delta R = 0.2$ of an electron or a muon candidate is removed; any jet within $\Delta R = 0.3$ of a τ_{had} is removed in events with two light leptons. The full overlap removal procedure are listed in Table 4.6.

Keep	Remove	Cone size (ΔR)
electron	electron (low p_T)	0.1
muon	electron	0.1
electron	jet	0.3
jet	muon	$\min(0.4,\ 0.04 + 10[\text{GeV}]/p_T(\text{muon}))$
electron	τ	0.2
muon	τ	0.2
τ	jet	0.3

Table 4.6: Summary of the overlap removal procedure between electrons, muons, hadronically decaying τs, and jets.

4.3.6 Event pre-selection (The 3ℓ pre-MVA selection)

Depending on the number of leptons in the final state, $t\bar{t}H$ multi-lepton channel is complex, like the 2ℓSS, 3ℓ, 4ℓ, 2ℓSS1τ, etc. In this thesis, only the three leptons final state is considered. The orthogonality of events entering various channels is ensured in the definition of the channels. In the three leptons final state, the total charge of the three leptons in the events must be ± 1 , no τ candidate is allowed. Among these three leptons, the lepton with opposite charge compared to other two is called "lepton 0" (Lep 0). The one in the same sign pair with the smallest $\Delta R(\ell, \ell_0)$, which means the one has closest distance to the "lepton 0", is designed to be "lepton 1" (Lep 1) and remaining one is "lepton 2" (Lep 2). Besides the Tight requirement listed in Table 4.5 on "lepton 1" and "lepton 2", the $p_T > 15$ GeV for both leptons is also required. As "lepton 0" is always the prompt one, the additional requirement is looser than "lepton 1" and "lepton 2", such as $p_T > 10$ GeV, *Loose* isolation and *Loose* lepton identification.

The 3ℓ pre-MVA selection
Three light leptons with $p_T > 10$ GeV; sum of light lepton charges $\pm$ 1
Two same-charge leptons must be Tight and have $p_T > 15$ GeV
The opposite-charge lepton must be loose, isolated and pass the non-prompt BDT
Zero medium τ_{had} candidates
$m_{\ell^+\ell^-} > 12$ GeV and $\lvert m_{\ell^+\ell^-} - 91.2 \text{ GeV}\rvert > 10$ GeV for all SFOC pairs
$\lvert m_{\ell\ell\ell} - 91.2 \text{ GeV } \rvert > 10$ GeV
$N_{jets} \geq 2$ and $N_{b-jets} \geq 1$

Table 4.7: The 3ℓ pre-MVA selection criteria applied in analysis. Same-flavour, opposite-charge lepton pairs are referred to as SFOC pairs.

To remove the leptons from quarkonium decays, all same flavour $\ell^+\ell^-$ pairs must satisfy $m_{\ell^+\ell^-} > 12$ GeV. The potential backgrounds with Z decays to $\ell\ell\gamma^{(*)}$ or $\ell\ell\ell'(\ell')$, where one lepton has very low momentum and is not reconstructed, the three lepton invariant mass should satisfy $\lvert m_{\ell\ell\ell} - 91.2 \text{ GeV } \rvert > 10$ GeV. The jets and b-tagged jets requirement is at least 2 jets and at least 1 b-tagged jets.

An additional cut to reject the $t\bar{t}Z$ process is used in the training stage of the 3ℓ analysis, where all same flavour lepton pairs in the event should satisfy $\lvert m_{\ell^+\ell^-} - 91.2 \text{ GeV}\rvert > 10$ GeV. There is no need to consider the charge flip issue in the 3ℓ as the study shows that no charge flip events exist in this channel after basic selection.

As the selections are applied before the Multi-variate analysis stage, it's also called the 3ℓ pre-MVA selection. A summary of the 3ℓ pre-MVA selection is shown in Table 4.7. The expected event yields with the 3ℓ pre-MVA selection for different processes are shown in Table 4.8.

Process	Expected yields
$t\bar{t}H$	22.5 $\pm$ 0.4
$t\bar{t}W$	42.9 $\pm$ 0.6
$t\bar{t}Z$	40.4 $\pm$ 0.5
VV	19.7 $\pm$ 2.4
$t\bar{t}$	52.4 $\pm$ 7.2
Rare	22.1 $\pm$ 0.6
Total	200.0 $\pm$ 7.7

Table 4.8: The expected event yields for signal and background processes after the 3ℓ pre-MVA selection.

4.4 Background estimation

Several physics processes can produce the similar signal signature. Good estimation of those backgrounds is the key issue for this analysis. The background in this analysis are sorted into two categories:

- **Irreducible background**
  Events which can lead to the same final state as the signal and the final state do indeed contain the three prompt charged leptons. $t\bar{t}W$, $t\bar{t}Z$ and di-boson processes are in this category.
- **Reducible background**
  Mostly events in which a non-prompt lepton or fake lepton is selected as prompt lepton are called the reducible background. These processes can lead to a final state which is compatible with signal with the misreconstructed object. The main process are $t\bar{t}$ and Z+jets. The data-driven method (Matrix Method) is used to estimate the reducible background in Section 4.4.2.

As a reference, Table 4.9 shows the expected and observed event yields in the 3ℓ channel at 8 TeV using the full Run 1 20.3 fb^{-1} dataset [61]. The dominant backgrounds are the non-prompt (fake leptons) and $t\bar{t}V$ process. The contribution from the VV is small. Although many conditions (energy scale, signal region definition, etc.) are different between Run 1 and Run 2, a brief picture of the background composition can be drawn according to the table.

Category	Non-prompt	$t\bar{t}W$	$t\bar{t}Z$	VV	Expected bkg.	$t\bar{t}H$	Observed
3ℓ	3.2 ± 0.7	2.3 ± 0.7	3.9 ± 0.8	0.86 ± 0.55	11.4 ± 2.3	2.34 ± 0.35	18

Table 4.9: Expected and observed event yields in the 3ℓ with the 20.3 fb^{-1} dataset at 8 TeV.

4.4.1 Fake leptons

Due to the limitation of the detector, the objects are not reconstructed perfectly. As one of the mis-reconstructed objects, fake leptons are common issue in the analysis. The main origins of the fake leptons are: photon conversion, light flavour jets and heavy flavour jets.

Fake electrons from conversion are produced when a photon interacting with detector material and split into an electron-positron pair. When this happens in the inner detector, a track is left and the track will be combined with the EM cluster from the electron and positron, and this may lead a reconstructed electron. In principle, the fake electrons from photon conversion can be removed by requiring the hits in the very first layer of the tracker, but random calorimeter clusters can still be combined with conversion tracks to form fake electrons. According to the description of jet object reconstruction in Section 3.5, the algorithm employs the track information in the inner detector and energy in the calorimeter. If a jet is reconstructed with a large fraction of EM energy and there is a random track that matches the calorimeter cluster, this would lead to an electron.

The fake muons from photon conversion are expected to be negligible. If a charged hadron with a lifetime sufficient to go through the whole detector, whilst the track of this hadron in the ID might not match that expected from a muon. But there are chances that matches can be found with unrelated hits and a fake muon can be reconstructed in this way.

A dedicated study of the fake lepton origin is performed in 3ℓ for both fake electrons and muons by using the truth information with MC. The study of the origin composition of the fake leptons in the 3ℓ pre-MVA region shows that the heavy flavour fakes (fake leptons which are from b and c meson) is dominant for both fake electrons and fake muons; large contribution comes from photon conversion fakes for fake electrons but no contribution is from this for fake muons as expected.

The data-driven method is introduced in this analysis to have a good estimation of the fake leptons. The following section will discuss the Matrix Method which is used widely in many analyses.

4.4.2 The Matrix Method

The Matrix Method (MM) is a data-driven technique to have the estimation of the fake contamination in the analysis. The estimation based on MC samples can not describe the fake objects well as the MC is not reliable all the time. Data-driven technique can employ the information from real collision data and get rid of the dependence on MC. The MM is a widely used data-driven method in the ATLAS. To have a good estimation of the fakes background, the MM is employed in the three leptons final state.

Generally, the MM is used to estimate the fake electrons and muons in a tight (denoted with T) region by using the loose object information in an anti-tight (denoted with $\bar{T}$) region. The basic strategy of the MM can be explained with a simplified case where only one lepton is considered. The number of the events with a tight (referred to N^T) lepton and the one with a lepton which fails the selection ($N^{\bar{T}}$) can be expressed in the terms of efficiencies and inefficiencies for the baseline loose real (N^r) or fake (N^f) leptons to pass the tight selection via two equations:

$$
\begin{aligned}
N^T &= \varepsilon_r \ N^r + \varepsilon_f \ N^f, \\
N^{\bar{T}} &= \bar{\varepsilon_r} \ N^r + \bar{\varepsilon_f} \ N^f,
\end{aligned} \tag{4.1}
$$

where ε_r (ε_f) represents the efficiency for a real (fake) lepton to pass tight selection, and $\bar{\varepsilon}_r \equiv (1-\varepsilon r)$ ($\bar{\varepsilon}_f \equiv (1-\varepsilon_f)$) represents the probability for a real (fake) lepton to fail tight but still pass baseline selection, or in a matrix form:

$$\begin{pmatrix} N^T \\ N^{\bar{T}} \end{pmatrix} = \begin{pmatrix} \varepsilon_r & \varepsilon_f \\ \bar{\varepsilon}_r & \bar{\varepsilon}_f \end{pmatrix} \begin{pmatrix} N^r \\ N^f \end{pmatrix}. \tag{4.2}$$

Then reversing the matrix can give the relation between the number of real and fake leptons events, given the observed number of tight and anti-tight lepton events:

$$\begin{pmatrix} N^r \\ N^f \end{pmatrix} = \frac{1}{\varepsilon_r - \varepsilon_f} \begin{pmatrix} \bar{\varepsilon}_f & -\varepsilon_f \\ -\bar{\varepsilon}_r & \varepsilon_r \end{pmatrix} \begin{pmatrix} N^T \\ N^{\bar{T}} \end{pmatrix}. \tag{4.3}$$

Thus the number of the fake leptons in the tight region can be obtained:

$$N_f^T = \frac{\varepsilon_r \varepsilon_f}{\varepsilon_r - \varepsilon_f} N^{\bar{T}} + \frac{(-\bar{\varepsilon}_r)\varepsilon_f}{\varepsilon_r - \varepsilon_f} N^T. \tag{4.4}$$

For a simplified example above, fake leptons can be estimated with a 2 × 2 matrix. A similar matrix method can be developed in the 3ℓ as well but with an 8 × 8 matrix. Besides building the complex matrix, the most difficulty for the 3ℓ is to find a reliable definition of the real lepton and fake lepton enriched control region. This is not as easy as the case in the tight and anti-tight region mentioned above. However, a simplified Matrix Method based on a 4 × 4 matrix (which is already employed in the $t\bar{t}H$ 2ℓSS final state) can be used and this simplified Matrix Method can solve those difficulties with negligible effect.

To complete this simplified Matrix Method in the 3ℓ, a premise is needed: "Lep 0" which is with opposite sign is with a much lower possibility to be the fake lepton. $t\bar{t}$ MC is used to check this premise with the 3ℓ pre-MVA selection. Following check is to see the real and fake raw number and corresponding percentage in the $t\bar{t}$ MC. Figure 4.2 shows the possibility of "Lep 0" to be the fake one in the 3ℓ pre-MVA region with $t\bar{t}$ MC. Overall only 1% "Lep 0" would be the fake one in the 3ℓ pre-MVA region.

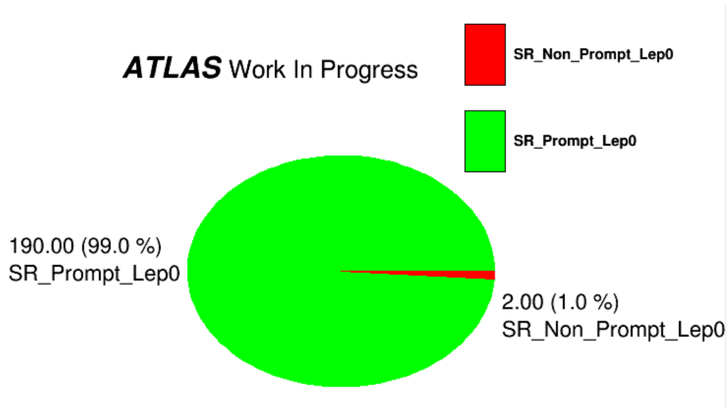


Figure 4.2: The raw number of event and the corresponding percentage for the real and the fake lepton events in $t\bar{t}$ MC.

Generally, most "Lep 0" are the prompt leptons and very few would be the fake objects for both electron and muon. The situation "Lep 0" to be the fake would be neglected. Thus three leptons final state is much similar to the $t\bar{t}H$ two leptons with same sign charge channel, and the main different is to estimate the same sign lepton pair, "Lep 1" and "Lep 2" in the 3ℓ. The advantages of this simplified Matrix Method in the 3ℓ are: the simple Matrix Method which is the same one used in the 2ℓSS; no need to redefine the real and fake enriched control region in the 3ℓ; the real efficiency and fake rate which are used in matrix are the same as the measured results from the 2ℓSS.

Since the assumption of "Lep 0" works well in the 3ℓ final states, the same sign lepton pair which pass the baseline selection could be fakes. The real and fake efficiency can be measured by using the two leptons events with the 2ℓ selection. Then the Matrix Method is applied to the same sign lepton pair in three leptons events during the application stage with the 3ℓ selection.

Depending on whether each lepton pass the Tight requirement (the definition of Loose and Tight is in Table 4.5), each i-th events can be categorised into any of four orthogonal (*sidebands*) region:

- TT_i: Events with both leptons are under Tight requirement (total events is denoted as N^{TT}).

- $T\bar{T}_i$: Events with leading leptons is Tight and subleading lepton is anti-Tight (total events is denoted as $N^{T\bar{T}}$).

- $\bar{T}T_i$: Events with leading lepton is anti-Tight and subleading lepton is Tight (total events is denoted as $N^{\bar{T}T}$).

- $\bar{T}\bar{T}_i$: Events with both leptons are under anti-Tight requirement (total events is denoted as $N^{\bar{T}\bar{T}}$).

The 4 × 4 matrix is defined to map the total number of such events into the total number of events in four di-lepton regions characterized by different real and fake lepton composition :

- rr_i: Events with both real leptons (total events is denoted as N^{rr}).

- rf_i: Events with leading leptons is real and subleading lepton is fake (total events is denoted as N^{rf}).

- fr_i: Events with leading lepton is fake and subleading lepton is real (total events is denoted as N^{fr}).

- ff_i: Events with both fake leptons (total events is denoted as N^{ff}).

The matrix is as follows:

$$\begin{pmatrix} N^{TT} \\ N^{T\bar{T}} \\ N^{\bar{T}T} \\ N^{\bar{T}\bar{T}} \end{pmatrix} = \begin{pmatrix} \varepsilon_{r,1}\varepsilon_{r,2} & \varepsilon_{r,1}\varepsilon_{f,2} & \varepsilon_{f,1}\varepsilon_{r,2} & \varepsilon_{f,1}\varepsilon_{f,2} \\ \varepsilon_{r,1}\bar{\varepsilon}_{r,2} & \varepsilon_{r,1}\bar{\varepsilon}_{f,2} & \varepsilon_{f,1}\bar{\varepsilon}_{r,2} & \varepsilon_{f,1}\bar{\varepsilon}_{f,2} \\ \bar{\varepsilon}_{r,1}\varepsilon_{r,2} & \bar{\varepsilon}_{r,1}\varepsilon_{f,2} & \bar{\varepsilon}_{f,1}\varepsilon_{r,2} & \bar{\varepsilon}_{f,1}\varepsilon_{f,2} \\ \bar{\varepsilon}_{r,1}\bar{\varepsilon}_{r,2} & \bar{\varepsilon}_{r,1}\bar{\varepsilon}_{f,2} & \bar{\varepsilon}_{f,1}\bar{\varepsilon}_{r,2} & \bar{\varepsilon}_{f,1}\bar{\varepsilon}_{f,2} \end{pmatrix} \begin{pmatrix} N^{rr} \\ N^{rf} \\ N^{fr} \\ N^{ff} \end{pmatrix}, \tag{4.5}$$

where the indexes for ε_r and ε_f are ordered in terms of lepton p_T .

The number of fakes in the signal region as a function of observables can be obtained by reverting the matrix:

$$\begin{pmatrix} N^{rr} \\ N^{rf} \\ N^{fr} \\ N^{ff} \end{pmatrix} = \begin{pmatrix} \varepsilon_{r,1}\varepsilon_{r,2} & \varepsilon_{r,1}\varepsilon_{f,2} & \varepsilon_{f,1}\varepsilon_{r,2} & \varepsilon_{f,1}\varepsilon_{f,2} \\ \varepsilon_{r,1}\bar{\varepsilon}_{r,2} & \varepsilon_{r,1}\bar{\varepsilon}_{f,2} & \varepsilon_{f,1}\bar{\varepsilon}_{r,2} & \varepsilon_{f,1}\bar{\varepsilon}_{f,2} \\ \bar{\varepsilon}_{r,1}\varepsilon_{r,2} & \bar{\varepsilon}_{r,1}\varepsilon_{f,2} & \bar{\varepsilon}_{f,1}\varepsilon_{r,2} & \bar{\varepsilon}_{f,1}\varepsilon_{f,2} \\ \bar{\varepsilon}_{r,1}\bar{\varepsilon}_{r,2} & \bar{\varepsilon}_{r,1}\bar{\varepsilon}_{f,2} & \bar{\varepsilon}_{f,1}\bar{\varepsilon}_{r,2} & \bar{\varepsilon}_{f,1}\bar{\varepsilon}_{f,2} \end{pmatrix}^{-1} \begin{pmatrix} N^{TT} \\ N^{T\bar{T}} \\ N^{\bar{T}T} \\ N^{\bar{T}\bar{T}} \end{pmatrix}, \tag{4.6}$$

where the final number of fakes in the signal region N^f_{TT}, namely the total number of TT events with at least one fake lepton, can be obtained with equation:

$$N^f_{TT} = N^{rf}_{TT} + N^{fr}_{TT} + N^{ff}_{TT} = \varepsilon_{r,1}\varepsilon_{r,2}N^{rf} + \varepsilon_{f,1}\varepsilon_{r,2}N^{fr} + \varepsilon_{f,1}\varepsilon_{f,2}N^{ff}. \tag{4.7}$$

According to the equation above, the main work is to measure the real and fake efficiency as the input of the matrix.

This method deals with the $t\bar{t}$ background because the most signal lepton of $t\bar{t}$ are the fakes in the $t\bar{t}H$ 3ℓ final state. The simulated $t\bar{t}$ sample with at least one W decaying leptonically is used for the real and fake measurement. A dedicated $t\bar{t}$ sample generated with a dileptonic filter is used for the 3ℓ case to enhance the statistics in the closure test stage. Additional photon radiation, which can give rise to extra leptons (mostly electron) through material interactions, is simulated inclusively in these samples. In order to improve statistics for those events, another sample specifically targeting $t\bar{t}$ production with an additional prompt photon radiated by any of the top quarks ($t\bar{t}\gamma$), has been generated and used. An overlap removal procedure which is based on distance with respect to any lepton in the event is applied to avoid the double counting of events between $t\bar{t}$ and $t\bar{t}\gamma$.

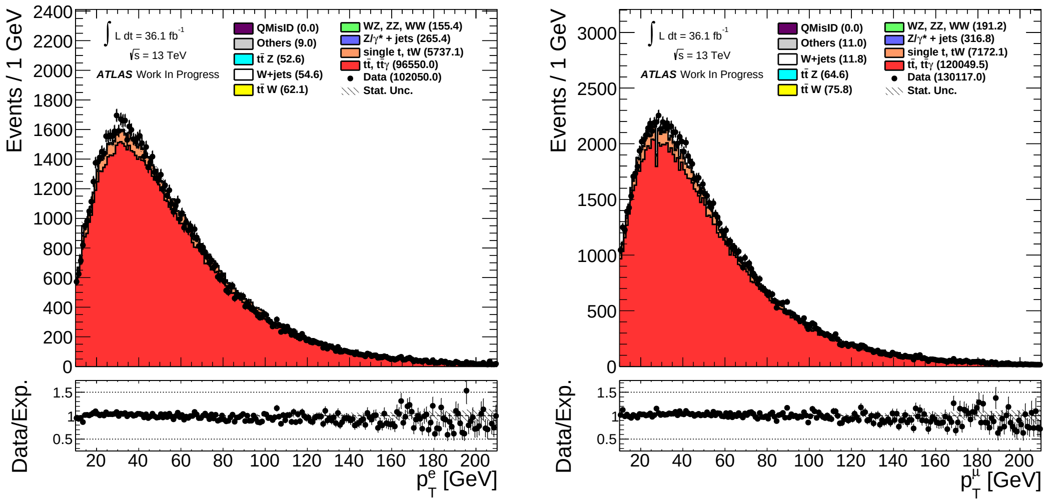
4.4.2.1 Real and fake lepton efficiency measurement

Two control regions, real and fake enriched, are designed to measure the efficiency of the real and fake lepton to pass the Tight selection requirements (mentioned in Table 4.5). Those regions are with sufficiently large statistics and can represent the kinematics and background composition in the signal region. The di-lepton requirement ensures orthogonality to the 3ℓ signal region. At least one b-tagged jet is required to avoid potentially large changes in the fake composition. The efficiencies have been factorised in bins of p_T to get a fake prediction dependent on the lepton kinematics. A $(N^{Jet}_{b-tagged}, p^{\ell}_{T})$ two dimensional parametrisation is used for electron fake rate as it can improve fakes modelling in events with high b-tagged jet multiplicity. The two dimensional parametrisation of distance of the muon and its closet jets and p_T $(\Delta R(\mu, jet), p^{\ell}_{T})$ is designed for fake muon rate to have a good fake modelling.

Real lepton control region and real lepton efficiency measurement

The real lepton control region is designed to be enriched with prompt leptons from $t\bar{t}$ di-lepton decays by requiring the presence of two opposite-sign charge, opposite flavour leptons. Table 4.10 shows the definition of the real control region where ε_r is performed. The prompt lepton purity achieved in this region is very high and is shown in Figure 4.3.

A standard Tag and Probe method is used to measure the real lepton efficiency for both electrons and muons. Each event in the CR is Tagged by requiring at least one of the leptons to pass the Tight selection and be trigger-matched. The other one, which is unbiased by the Tight and trigger selection is the Probe candidate for the efficiency measurement. In case two tag candidates are found, both leptons are considered as valid probes and used for the measurement. This procedure follows the one used in CP groups to measure the electron and muon efficiencies for calibration.



(a) Real electron CR composition

(b) Real muon CR composition

Figure 4.3: Real lepton enriched control region composition for p_T distribution of Tight electrons (left) and Tight muons.

The efficiency for prompt leptons is defined on pre-event basis as the ratio of numerator events where the probe passes Tight selection and denominator events where no additional selection requirement other than the baseline Loose one (Loose = Tight + anti-Tight) is applied on the probe. The subtraction of background events is applied to both numerator and denominator:

$$\varepsilon_r(e,\mu)_i = \frac{N_i^{T(e,\mu)} - N_{bkg}^{T(e,\mu)}{}_i}{N_i^{L(e,\mu)} - N_{bkg}^{L(e,\mu)}{}_i}, \tag{4.8}$$

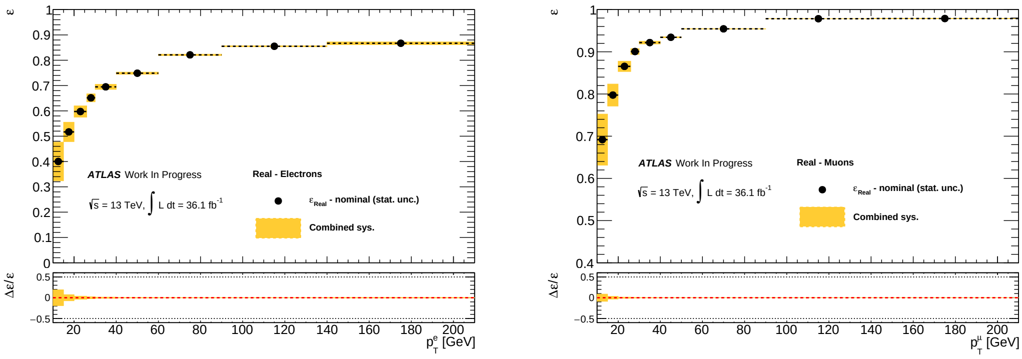
where the background to be subtracted accounts for the contamination of OS

	Real lepton enriched CR
N_{jets}	2, 3
$N_{b-taggeg\ jets}$	≥ 1 (MV2c10, 70% eff.)
N_ℓ	2
lepton charge	OS
lepton flavour	$e\mu$, μe
p_T^ℓ	≥ 10 GeV
No. of trigger-matched ℓ	≥ 1

Table 4.10: Definition of the control region used for measuring the real lepton efficiency. The same region is used to measure both ε_r^e and ε_r^μ.

events with a fake lepton, and mostly comes from the top and single-Top. The index i is the p_T binning chosen for parametrising the efficiency.

As the contribution of those fakes are small, estimated from simulation samples directly, and a 30% systematic uncertainty is added on their normalization. Figure 4.4 shows the measured real electron and muon efficiency in data with systematic uncertainty coming from fake background subtraction in data.



(a) $\varepsilon_r(e)$ (b) $\varepsilon_r(\mu)$

Figure 4.4: Real efficiencies for electrons and muons from CR in data. The systematic uncertainty from the OS fakes normalisation uncertainty are show in orange band.

Fake control region and fake rate measurement

It is difficult to define good control region for fake study like real lepton control region. Actually, only the control region dominated by $t\bar{t}$ semileptonic events can be defined with one real lepton and one non-prompt lepton or converted photon. To address this ambiguity, a modified Tag and Probe method is designed to measure the fake efficiency for electrons and muons separately.

- **Fake rate $\varepsilon_f(e)$ of electrons**

In the opposite flavour same sign dilepton events when a tight muon firing the single muon trigger is found, the muon has a high probability of being the prompt lepton of the pair, and the remaining electron as a suitable unbiased probe candidate to measure the fake efficiency. Additional benefit by using opposite flavour events is that the reduction of the amount of the charge misID events to be subtracted which can be large in same-sign electron-electron region, hence lead to a large uncertainties on the efficiency measurement.

The definition of the fake electron control region is listed in Table 4.11. The check on fake origin in fake CR as a function of the number of b-tagged jets is shown in Figure 4.5. There is a large difference in the fake electron origin

	Fake electron enriched CR
N_{jets}	2, 3
$N_{b-tag\ jets}$	≥ 1 (`MV2c10`, 70% eff.)
N_{ℓ}	2
lepton charge	SS
lepton flavour	$e\mu$, μe
p_T^{ℓ}	≥ 10 GeV

Table 4.11: Definition of the control region used for measuring the fake electron efficiency. Note that we require the muon to *tag* the event, with the electron being used as selection-unbiased *probe* for the efficiency measurement.

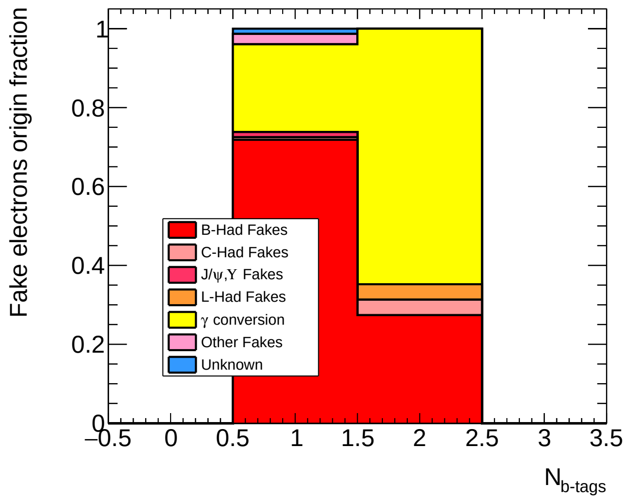


Figure 4.5: Fake electron origin fraction for $t\bar{t} + t\bar{t}\gamma$ events, in the dileptonic OF CR where electron fake rates are measured, as a function of $N_{b-tagged}$.

fraction between 1 b-tagged jet and 2 b-tagged jets. The photon conversion fakes in 1 b-tagged jet bin is 20% compared to 60% in the 2 b-tagged jets bin. One dimensional parametrisation of the fake rate depending on p_T only together with at least one b-tagged jets requirement can lead to a poor modelling of the b-tagged jet distribution in data. Thus, a two dimensional $(N^{Jet}_{b-tagged}, p_T^{\ell})$ parametrisation of the electron fake rate is done. The efficiency of the fake electron is similar to the Equation 4.8:

$$\varepsilon_f(e)_{i,j} = \frac{N^{T(e)}_{i,j} - N^{T(e)}_{bkg\ \ i,j}}{N^{L(e)}_{i,j} - N^{L(e)}_{bkg\ \ i,j}}. \tag{4.9}$$

The double index refers to the $(N^{Jet}_{b-tagged}, p^{l}_{T})$ bin. The subtracted background includes the event with prompt SS lepton pair, charge misID events and $t\bar{t}$ semileptonic events with the mis-assigned probe electron. Overall the fake lepton purity is good, the $t\bar{t}$ event where the probe electron happens to be incorrectly matched to the truth prompt lepton in the pair is overall small for only about 10% of the total background events to be subtracted at the numerator.

Additional check on the truth origin of fake leptons is performed in the 2ℓSS CR, 2ℓSS SR and 3ℓ SR (with the 3ℓ pre-MVA selection). Figure 4.6 shows the truth origin of fake electrons and fake muons for those three regions. The ratio is defined as follows: fake leptons from specific source over the total fakes. The majority of fakes are the non-prompt leptons from heavy flavoured particles decays in b, c quark according to Figure 4.6. The fraction of fake muons which from heavy flavour decays is stable in the 2ℓSS CR and 3ℓ SR, the difference is only about 15%.

As shown in Figure 4.6, the electron fakes from heavy flavour decay dominate in the 2ℓSS CR (about 60%), but there is a large fraction of fake electron coming from photon conversion (about 30%). The fraction of the photon conversion fakes is larger (about 50%) in the 3ℓ SR. Those photons can originate from the ISR/FSR, a π^0 decay product or prompt photons radiated off the top or the anti-top quark. Then interacting with the detector material to produce a pair of opposite sign electrons of which only one get reconstructed. A dedicated correction is applied to cover this difference in the both 2ℓSS SR and 3ℓ SR.

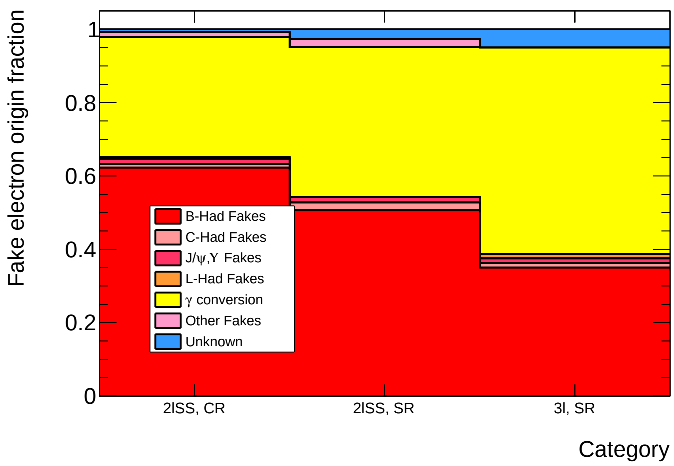


Figure 4.6: Origin fraction of fake electrons for $t\bar{t}$ + $t\bar{t}\gamma$ events in the dileptonic CR, 2ℓSS SR and 3ℓ SR.

As shown in Figure 4.6 in the 3ℓ SR, the fraction of the fake electrons which are from photon conversion is about 50%. The difference between the 2ℓSS CR and 3ℓ SR is non-negligible. Subsequently the fake rate which is measured in the 2ℓSS CR can not describe the situation in the 3ℓ SR precisely. To solve this, a dedicated scale factor is designed to rescale the fraction of fake electrons from photon conversions in the 2ℓSS CR to be the same level in the 3ℓ SR.

The dedicated factor α is applied during the fake rate measurement to adjust the difference between the same sign fake enrich control region and the 3ℓ SR. The definition of factor α is:

$$\alpha = \frac{(1-f_x)\varepsilon_f^{non-PhConv} + f_x \varepsilon_f^{PhConv}}{(1-f_{CR})\varepsilon_f^{non-PhConv} + f_{CR}\varepsilon_f^{PhConv}} - 1, \tag{4.10}$$

where f_x and f_{CR} is the fraction of the fake leptons which are from photon conversion in dedicated region (the 3ℓ pre-MVA region in this thesis) and fake enrich control region, $\varepsilon_f^{non-PhConv}$ and ε_f^{PhConv} is the fake rate of leptons for non-photon conversion and photon conversion origin respectively. The measurement of α is performed for electron channel and opposite flavour channel in the fake enrich control region and 3ℓ pre-MVA region (only targeting at same sign lepton pair). The $\varepsilon_f^{non-PhConv}$ (ε_f^{HF} in the table) and ε_f^{PhConv} are derived from fake enrich control region and measurements of all corresponding variables are shown in Table 4.12, the "X" in the table represents electrons or muons, and the impact on all related variables by varying fake fraction is also shown in the table to have the uncertainty on this correction factor. The uncertainties on conversion fraction (f, f_{CR}) is: 40% taken as conservative estimate from the 2015 data/MC agreement for material description in Ref. [62]; 50% from heavy flavour fake efficiency ($\varepsilon_f^{non-PhConv}$); 50% from conversion fakes efficiency (ε_f^{PhConv}). The final measured results concerning the uncertainties:

$$\begin{aligned} \alpha_{Xee} &= 0.57^{+0.11}_{-0.12}(f)^{+0.17}_{-0.13}(\varepsilon_f^{HF})^{+0.10}_{-0.22}(\varepsilon_f^{PhConv}), \\ \alpha_{Xe\mu(\mu e)} &= 0.07 \pm 0.01(f) \pm 0.02(\varepsilon_f^{HF})^{+0.01}_{-0.03}(\varepsilon_f^{PhConv}). \end{aligned} \tag{4.11}$$

Then the fake rate in the interested region would be $\varepsilon_f^x = (1+\alpha)\varepsilon_f^{CR}$. Effective rescaling of the fake efficiencies (symmetrizing uncertainties) are:

$$\begin{aligned} \varepsilon_f^{Xee}/\varepsilon_f^{CR} &= (1+\alpha_{Xee}) = 1.57 \pm 0.25, \\ \varepsilon_f^{Xe\mu(\mu e)}/\varepsilon_f^{CR} &= (1+\alpha_{Xe\mu(\mu e)}) = 1.07 \pm 0.04. \end{aligned} \tag{4.12}$$

Generally this correction describes the fakes situation in the 3ℓ very well, and the closure test which to be introduced in Section 4.4.2.2 shows a good agreement between MC and the Matrix Method prediction.

The fake rate of the electrons in fakes CR with data is shown in Figure 4.8. One dimensional efficiency represents a projection for inclusive $N_{b-tagged}^{Jet} \geq 1$ events of the $(N_{b-tagged}^{Jet}, p_T^l)$ efficiency which is also shown in the figure and used as input in the MM. p_T bin is set [20, 210+] GeV as to have a better statistics. For the systematics uncertainty on the efficiency, a theory uncertainty of 14% is set because of the theoretical cross section, QCD scale and PDF uncertainties ($t\bar{t}W, t\bar{t}Z$). 50% uncertainty is set on the VV, and 30% uncertainty is set on the remaining

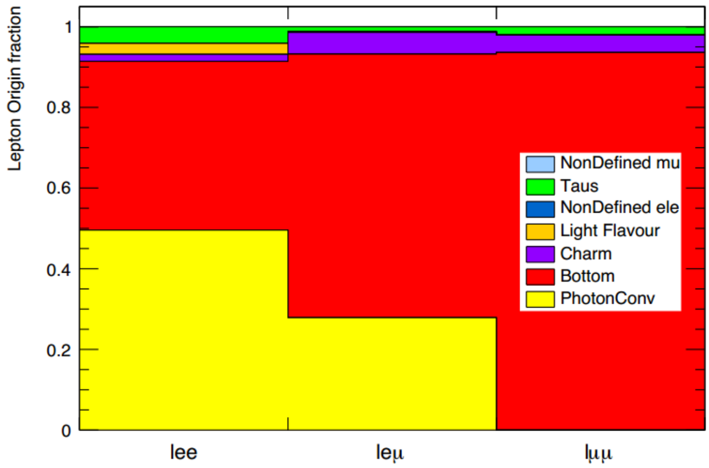
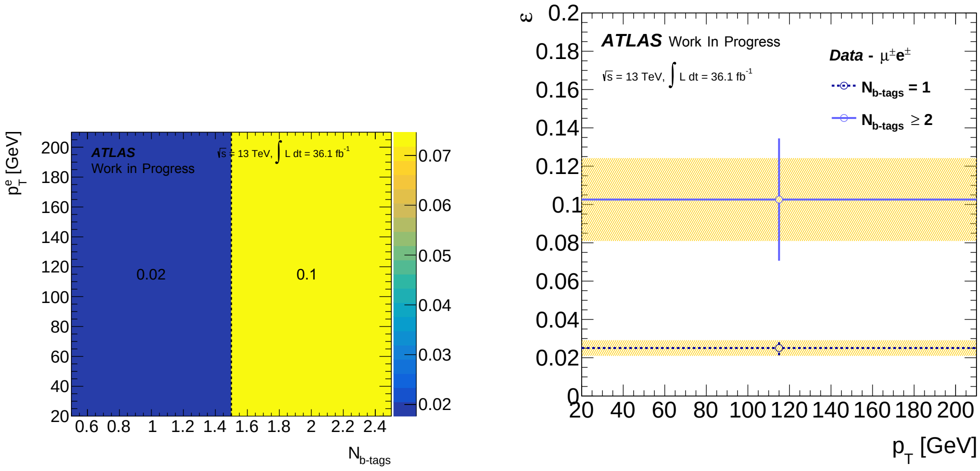


Figure 4.7: Origin fraction of fake electrons and fake muons for $t\bar{t}$ + $t\bar{t}\gamma$ events in the 3ℓ by flavoured channel.

	Nominal	$f_{CR} - 40\%$	$f_{CR} + 40\%$	$\varepsilon_f^{HF} - 50\%$	$\varepsilon_f^{HF} + 50\%$	ε_f^{PhConv}-50%	ε_f^{PhConv}+50%
				Xee			
$f_x(3\ell\text{SR})$	0.498	0.299	0.697	0.498	0.498	0.498	0.498
f_{CR}	0.25	0.15	0.35	0.25	0.25	0.25	0.25
ε_f^{HF}	0.009	0.009	0.009	0.005	0.014	0.009	0.009
ε_f^{PhConv}	0.057	0.057	0.057	0.057	0.057	0.029	0.086
α_{Xee}	0.56	0.44	0.65	0.74	0.44	0.35	0.67
				$Xe\mu$ **or** $X\mu e$			
$f_x(3\ell\text{SR})$	0.279	0.167	0.391	0.279	0.279	0.279	0.279
f_{CR}	0.25	0.15	0.35	0.25	0.25	0.25	0.25
ε_f^{HF}	0.009	0.009	0.009	0.005	0.014	0.009	0.009
ε_f^{PhConv}	0.057	0.057	0.057	0.057	0.057	0.029	0.086
$\alpha_{Xe\mu(\mu e)}$	0.07	0.06	0.08	0.09	0.05	0.04	0.08

Table 4.12: The impact of varying fake fraction on correction factor α.

processes to be subtracted. The total uncertainty of the data-driven charge misID is also considered. All the systematics uncertainties related to the detector are neglected as those uncertainties can be covered by the size of the background normalization uncertainties.



(a) $\varepsilon_f(e)$, $(N^{Jet}_{b-tagged}, p^{\ell}_T)$ parametrisation (b) $\varepsilon_f(e)$, p_T projection for $N^{Jet}_{b-tagged}$ slices

Figure 4.8: Fake rate for electrons from CR in data. 4.8(a) shows the two-dimensional $(N^{Jet}_{b-tagged}, p^{\ell}_T)$ map, and 4.8(b) represents the fake rate projection over p_T for each $N^{Jet}_{b-tagged}$ slice. The orange bands represent the total systematic uncertainty on the fake rate. Please note that the efficiencies in 4.8(b) have been rescaled by the α_{ee} factor.

- **Fake rate $\varepsilon_f(\mu)$ of muons**

Unlike using the opposite flavour SS events to measure fake electron rate, a Tag and Probe based on $\mu\mu$ events is adopted in fake muon rate measurement, assuming the subleading muon is more likely to be the fake in the lepton pair, the subleading one is chosen as probe when both muons are tight and fire the trigger. According to Figure 4.9, the fake muons from heavy flavour decay dominate in the 2ℓSS CR (about 90%). No fake contributions from photon conversion in the both 2ℓSS CR and 3ℓ SR for fake muons. About 20% fake muons in the 3ℓ SR are from "Unknown", and actually those muons are the prompts which come from virtual photon conversions.

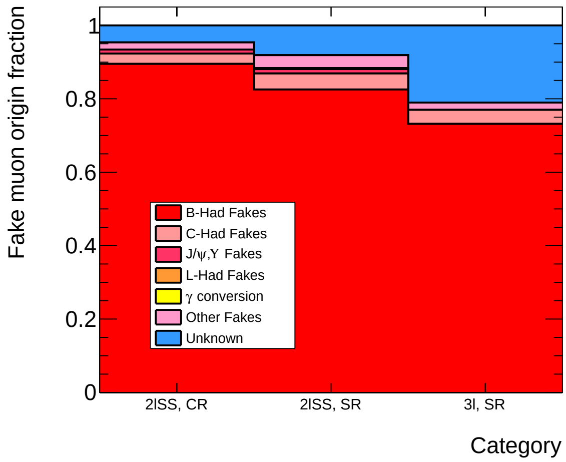


Figure 4.9: Origin fraction of fake muons for $t\bar{t}$ + $t\bar{t}\gamma$ events in the dileptonic CR, 2ℓSS SR and 3ℓ SR.

Poor modelling of the low $\Delta R(\mu, jet)$ is observed when using one dimensional p_T parametrisation of the muon fake rate. A two dimensional measurement on $(\Delta R(\mu, jet), p_T^\ell)$ for muon fake rate is designed to have a better modelling of the fakes. The binning in $\Delta R(\mu, jet)$ is [0.0, 1.0, 5.0] and the p_T binning is [20.0, 50.0, 210+]. The fake rate for muons is:

$$\varepsilon_f(\mu)_{i,j} = \frac{N_{i,j}^{T(\mu)} - N_{bkg\ \ i,j}^{T(\mu)}}{N_{i,j}^{L(\mu)} - N_{bkg\ \ i,j}^{L(\mu)}}, \tag{4.13}$$

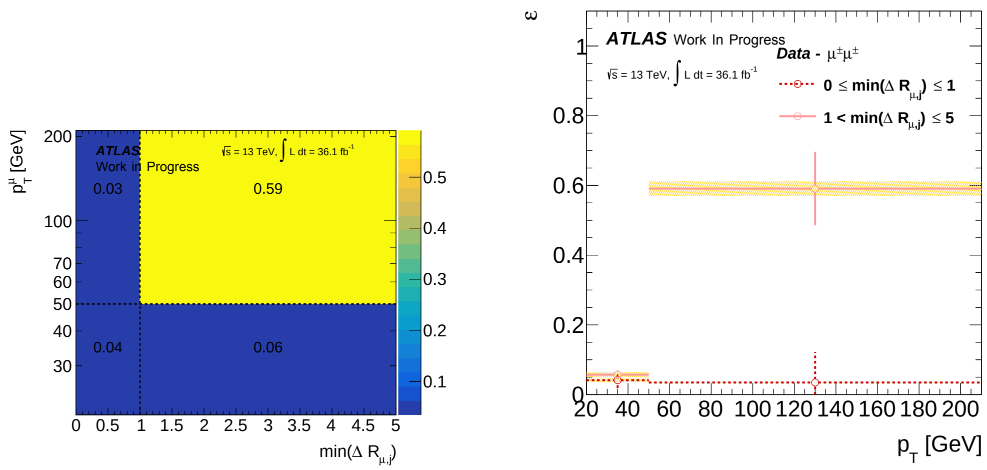
the background to be subtracted comes from the prompt same sign lepton pairs and $t\bar{t}$ semileptonic events with mis-assigned probe muons. Table 4.13 summaries the definition of the fake muons control region which is used to measure the fake rate of muons.

The assumption of using p_T ranking as criterion to solve the ambiguity of both muons being tight and trigger matched does not have a significant impact on the

	Fake muon enriched CR
N_{jets}	2, 3
$N_{b-tag\ jets}$	≥ 1 (`MV2c10`, 70% eff.)
N_{ℓ}	2
lepton charge	SS
lepton flavour	$\mu\mu$
p_T^{ℓ}	≥ 10 GeV

Table 4.13: Definition of the control region used for measuring the fake muon efficiency.

muon fake estimate. Overall the fake muons purity is good and $t\bar{t}$ events where the probe is the prompt lepton in the pair is only 16% of the total process to be subtracted from data in the numerator. Figure 4.10 shows the fake rate of the muons after background subtracting in data. The systematics uncertainties of the fake rate is similar to the electron case, but uncertainties from charge misID is neglected.

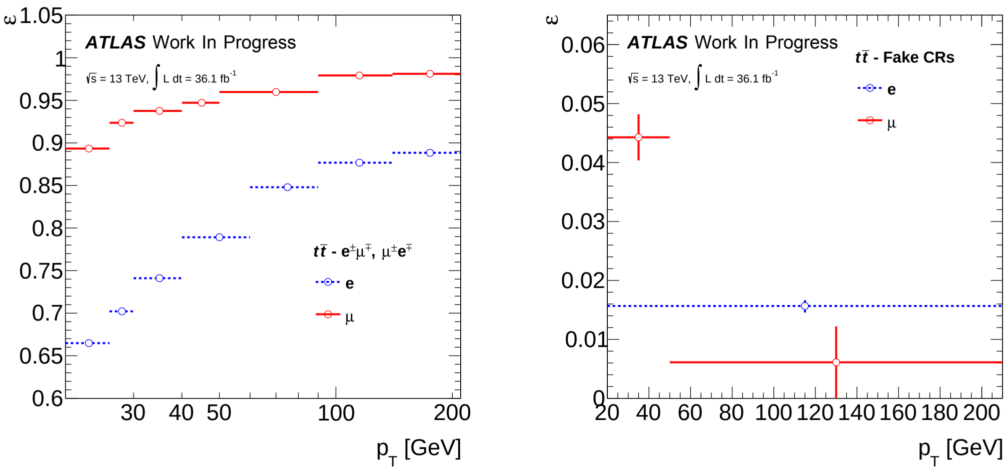


(a) $\varepsilon_f(\mu)$, $(\Delta R(\mu, jet), p_T^{\ell})$ parametrisation (b) $\varepsilon_f(\mu)$, p_T projection for $\Delta R(\mu, jet)$ slices

Figure 4.10: Fake rate for muons from CR in data. 4.10(a) shows the two-dimensional $(\Delta R(\mu, jet), p_T^{\ell})$ map, and 4.10(b) represents the fake rate projection over p_T for each $\Delta R(\mu, jet)$ slice. The orange bands represent the total systematic uncertainty on the fake rate.

4.4.2.2 Closure test

The Matrix Method is based on the assumption that the fake composition is stable between CR where the real or fake efficiency measured and SR where efficiency are applied. In fact the fake sources are different in the CR and the SR, and also the efficiency of passing the Tight selection may be different. Thus it is necessary to check the assumption, as well as to ensure no additional bias from the method, and the Matrix Method performs correctly. A so-called closure test has been performed in the 3ℓ pre-MVA region. Generally, it's done by comparing the fake event yields from MC and the Matrix Method in the 3ℓ pre-MVA region, and events are normalized to the luminosity of 36.1 fb^{-1}. The corresponding real and fake efficiencies are measured by using the MC in the related CRs. Figure 4.11 shows the real and fake efficiencies measured in $t\bar{t} + t\bar{t}\gamma$ events, where a truth requirement of exactly two prompt leptons is applied in the real enriched CR. The probe lepton is required to be a non-prompt in fake enriched CRs. Truth charge flip events have been vetoed in both real and fake CRs. Such efficiencies are used to compute the Matrix Method weights for the same $t\bar{t} + t\bar{t}\gamma$ sample to derive the fake yields.



(a) ε_r (b) ε_f

Figure 4.11: Electron, muon real (4.11(a)) and fake (4.11(b)) efficiencies as measured in $t\bar{t}+t\bar{t}\gamma$ MC simulation. In the fake case, the plots are inclusive projections over p_T of the 2D efficiency maps previously described.

To check the robustness of the method and the additional effects other than one cited, the closure test is done in three channels by lepton flavour (ee, $\mu\mu$, and opposite flavour (OF)) in the SR of 3ℓ. Any non-closure observed will also account for missing non-trivial efficiency parametrisations and binning effects. Finally the observed non-closure is the one to be quoted as an additional nuisance parameter into the final fit.

Due to the high fraction of the photon conversion fakes, $t\bar{t}\gamma$ sample is added when doing the closure test. Besides this $t\bar{t}\gamma$ sample, the di-lepton filter $t\bar{t}$ sample is used in the 3ℓ case instead of the non-allhadronic one which is used for the efficiency measurement to increase the statistics of the 3ℓ events. The event selection for the closure test is the 3ℓ pre-MVA selection. The expected prediction is obtained from $t\bar{t} + t\bar{t}\gamma$ MC sample with two Tight leptons (the same sign lepton pair in the 3ℓ channel), and then requiring at least one truth-matched fake lepton either coming from a photon conversion or heavy flavour decay. The truth charge flip lepton events are vetoed. The same truth selection is applied to the Matrix Method as well.

The non-closure (ζ) is obtained from the ratio of the Matrix Method (N_{MM}) and pure MC ($N_{t\bar{t},t\bar{t}\gamma}$) prediction, and the relative bias between the two estimates is defined as:

$$\begin{aligned} \zeta &= \frac{N_{MM} - N_{t\bar{t},t\bar{t}\gamma}}{N_{MM}}, \\ \sigma_\zeta = \sigma_{SF} &= \sqrt{\frac{1}{N_{MM}^2} \cdot \sigma^2_{N_{t\bar{t},t\bar{t}\gamma}} + \frac{N^2_{t\bar{t},t\bar{t}\gamma}}{N^4_{MM}} \cdot \sigma^2_{N_{MM}}}. \end{aligned} \tag{4.14}$$

The error on the Matrix Method estimate $\sigma_{N_{MM}}$ is a combination of the statistical uncertainty driven by the size of the $TT, T\bar{T}, \bar{T}T, \bar{T}\bar{T}$ sidebands, and the uncertainty on the measured efficiencies. The above formula for the non-closure uncertainty (σ_ζ) holds in case $\sigma_{N_{MM}}$ and $\sigma_{N t\bar{t},t\bar{t}\gamma}$ are not correlated. This is true in the first order, since the contribution of the TT sideband, which is the only one not independent between the pure MC event set and the MM set has a small contribution to the total MM fake event yields compared to the sidebands with anti-tight leptons.

The 3ℓ pre-MVA selection is applied to both MC and the Matrix Method to get the fake event yields for closure test. The two tight leptons with same sign and at least one fake from photon conversion or heavy flavour decay are required. The truth selection is applied to remove the charge flip and keep the pure fakes from MC. Figure 4.12 shows the variables' distribution comparison between simulation prediction and the Matrix Method predicted fakes in inclusive channel. Table 4.14 shows the fake event yields comparison between MC and MM in flavoured channel with the 3ℓ pre-MVA selection. The agreement is good considering the low statistics and large uncertainties.

4.4.2.3 Fake results in data

The measured real efficiency and fake rate which are mentioned in Section 4.4.2.1 are as the input to the Matrix Method to get the fake leptons in the 3ℓ pre-MVA region by using the data. The uncertainties on the efficiencies have been propa-

Fake Yields	$\mu\mu$	ee	OF
MM	16.4 $\pm$ 4.1	11.1 $\pm$ 3.3	24.9 $\pm$5.0
Expected number	19.1 $\pm$ 4.3	11.0 $\pm$ 3.3	27.4 $\pm$5.2
Non-closure	-0.14 $\pm$ 33%	0.01 $\pm$ 43%	-0.1 $\pm$ 26%

Table 4.14: Expected fake event yields from MC and from the Matrix Method with $t\bar{t}$ + $t\bar{t}\gamma$ in the 3ℓ pre-MVA region.

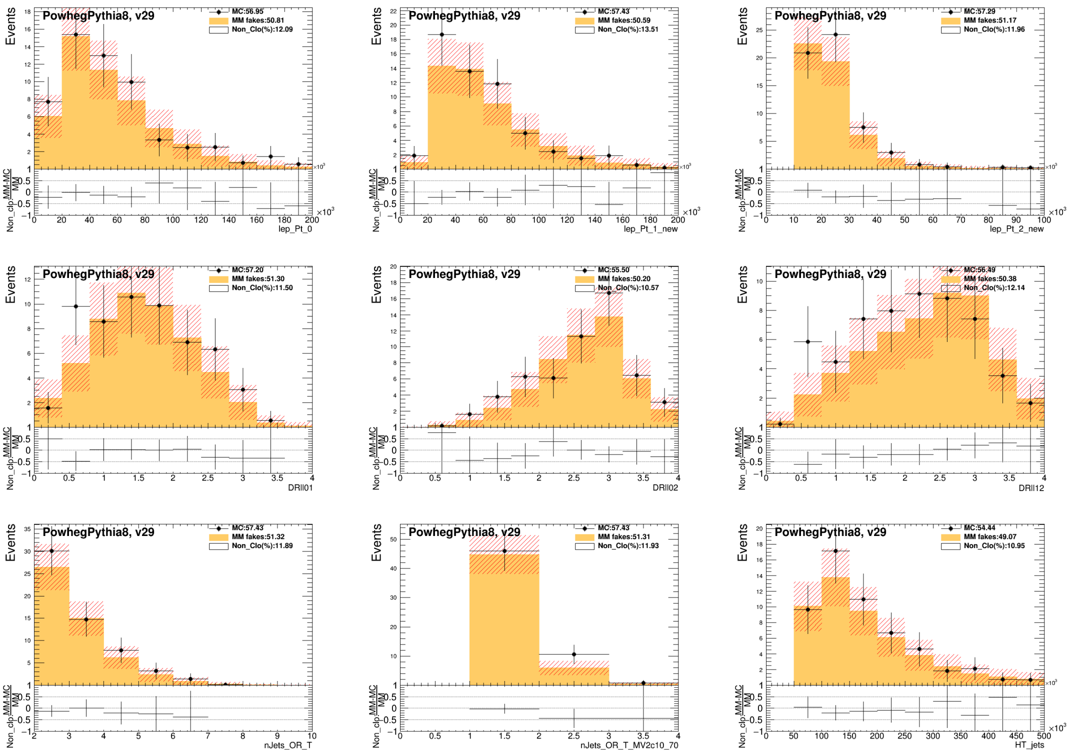


Figure 4.12: Closure of the p_T of leptons, distance between leptons (DRll) and number of jets and b-tagged jets between $t\bar{t} + t\bar{t}\gamma$ simulated events and MM fake events.

gated to the event weights. As for the statistical uncertainty on the efficiencies and the systematic uncertainty from charge misID subtraction, each bin of the efficiency parametrisation has been varied independently from others. The background theory uncertainty variation is applied simultaneously across the efficiency bins. The final shape uncertainty on a variable of interest is taken as the sum in quadrature of all the uncertainties contributing to each bin of the distribution.

Table 4.15 shows the fakes event yields by using the Matrix Method with data in ee, $\mu\mu$ and OF in the 3ℓ pre-MVA region with statistical and systematic uncertainties.

	$\mu\mu$	ee	OF
Fake Yields(MM)	35.3 ± 5.9	24.0 ± 4.9	62.0 ±7.9
MC prediction	16.4 ± 4.1	11.1 ± 3.3	24.9 ±5.0
MM/MC-prediction ratio	2.15	2.18	2.49

Table 4.15: Expected fake event yields from MC and from the Matrix Method with data in the 3ℓ pre-MVA region.

4.5 Background suppressing using the multi-variate analysis technique

The multi-variate analysis (MVA) technique has been widely employed in the ATLAS experiment in many analyses. For the analysis in this thesis, the Boosted Decision Trees (BDT) is optimized to further reject the "reducible" fake background which is dominantly from the $t\bar{t}$ and irreducible background from $t\bar{t}W$ and $t\bar{t}Z$. Before the MVA optimization, the 3ℓ pre-MVA selection shown in the Section 4.3.6 is required and corresponding events yields are shown in Table 4.8. As shown in Table 4.8, the dominant contributions are from non-prompt background and $t\bar{t}V$. To suppress those backgrounds, a two dimensional (2D) BDT analysis strategy is designed to enhance the signal events in the pre-MVA region.

4.5.1 Modelling of the MVA input variables

A set of variables are used as the input for the MVA. The definition of those variables is shown in Table 4.16. The comparison between data and MC for those input variables is shown in Figure 4.14. The agreement is good and those inputs show the promising discriminate power.

Variable	Definition
$t\bar{t}H$ vs $t\bar{t}V$ BTDG	
$P_T(\ell_0, \ell_1)$	Transverse momentum of the lepton pair ℓ_0, ℓ_1
$m(\ell_0, \ell_1)$	Invariant mass of the leptons ℓ_0, ℓ_1
$m(\ell_1, \ell_2)$	Invariant mass of the leptons ℓ_1, ℓ_2
$best\_Z\_m(\ell\ell)$	The closest invariant mass between two leptons to the Z boson pole mass
$\Sigma\Delta R(\ell_0, \ell_1, \ell_2)$	Scalar sum of the ΔR between (ℓ_0, ℓ_1), (ℓ_0, ℓ_2), (ℓ_1, ℓ_2)
$\Delta\eta(\ell_1, \ell_2)$	Pseudo-rapidity difference between ℓ_1, ℓ_2
m_{eff}	The effective mass defined as the sum of muons, electrons, jets transverse momentum and E_T^{miss}
$t\bar{t}H$ vs Non-prompt BTDG	
n_{jets}	Number of jets
$n_{\text{jets}} + 10 \cdot n_{\text{b-jets}}$	Multiplicity of jets and multiplied by 10 times the multiplicity of b-tagged jets
P_t^{l2}	Third lepton Pt
$m(\ell_1, \ell_2)$	Invariant mass of the lepton pair ℓ_1, ℓ_2
$\Sigma\Delta R(\ell_0, \ell_1, \ell_2)$	Scalar sum of the ΔR between (ℓ_0, ℓ_1), (ℓ_0, ℓ_2), (ℓ_1, ℓ_2)
$\Delta\eta(\ell_1, \ell_2)$	Pseudo-rapidity difference between ℓ_1, ℓ_2
m_{eff}	The effective mass defined as the sum of muons, electrons, jets transverse momentum and E_T^{miss}

Table 4.16: List of input variables for the BDT.

4.5.2 Performance of the BDT

The BDTG (Gradient boosting algorithm) is trained on sets of MC samples for the signal, irreducible background and data-riven sample for the non-prompt background with total 10 variables which are mentioned in Section 4.5.1. Especially for the reducible background, the data-driven result, the Matrix Method estimation is used as the good description of the non-prompt background. The closure test of the Matrix Method shows the good performance of the method, and the predicted non-prompt could be employed in the BDTG training especially for the shape modelling. The cross training strategy is employed as the good performance for the low statistics analysis. The main idea is that using the half of the total events for training and rests are used for testing, and events are chosen by requiring odd or even event number.

The separation of the BDTG for signal and background events is illustrated in Figure 4.13. No overtraining issue is observed for both. The input variables in the top ranking are:

- $t\bar{t}V$: $P_T(\ell_0, \ell_1)$, $m(\ell_0, \ell_1)$, $\Sigma\Delta R(\ell_0, \ell_1, \ell_2)$ and m_{eff}.
- Non-prompt: n_{jets}, $\Delta\eta(\ell_1, \ell_2)$, $m(\ell_1, \ell_2)$, m_{eff} and $\Sigma\Delta R(\ell_0, \ell_1, \ell_2)$.

Figure 4.13(c) and Figure 4.13(d) show the background rejection versus signal efficiency (Receiver Operating Characteristic curve, the ROC curve) obtained with odd or even events. Similar performance are obtained for the both BDTGs demonstrating by using odd or even events.

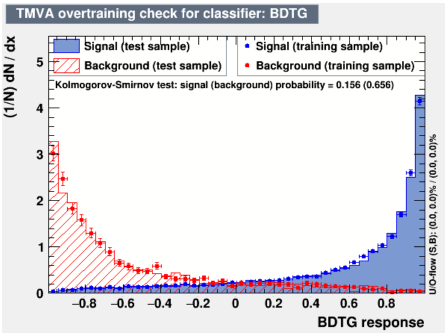
The full shape of the BDTGs are used to maximize the signal sensitivity. Two BDTGs are combined into one final discriminant, called $BDTG_{Combination}$ and is defined as:

$$BDTG_{Combination} = (\text{BDTG\_Non-prompt} + a \times \text{BDTG\_}t\bar{t}V)/(1 + a), \quad (4.15)$$

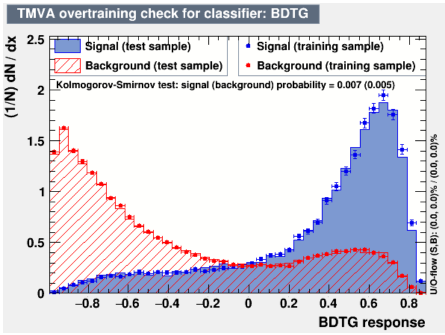
here a is defined as the slope in the BDTG_Non-prompt vs BDTG_$t\bar{t}V$ plane as shown in Figure 4.13(e).

The studies have been done to find a good combination of those two BDTGs, finally parameter a with value 1 is the best choice. The discriminant BDTG is then splitted in 6 bins which has been optimised to separate signal and background using the TransfoD function called auto-binning in ***ttHFitter***. This iterative algorithm defines the bins of the BDTG distribution according to two free parameters b and s, whose sum is constrained to the number of bins. Several tries to improve on the signal sensitivity are attempted by scanning 2 parameters: b and s. The Asimov fit of the $t\bar{t}H$ signal strength including all statistical and systematic uncertainties is used for this study. The best sensitivity is obtained with $a = 1$, $b = 6$ and $s = 0$. This corresponds to a situation where the signal is constant in each bins of the BDTG discriminant. As a result the lowest bins are more populated by non-prompt background, the central bins more populated by $t\bar{t}V$ and the highest bins more pure by $t\bar{t}H$. The final discriminant is therefore defined as :

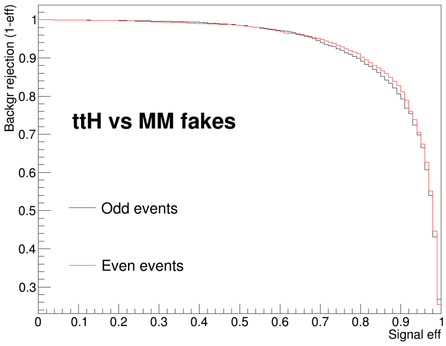
$$BDTG_{Combination} = (\text{BDTG\_Non-prompt} + \text{BDTG\_}t\bar{t}V)/2 \quad (4.16)$$



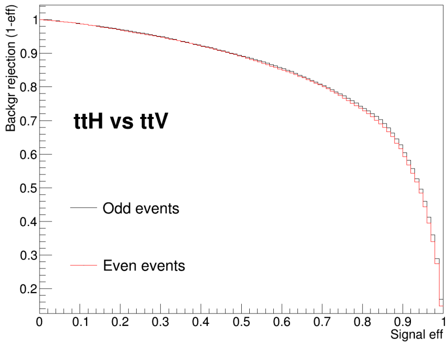
(a) BDTG response to non-prompt background



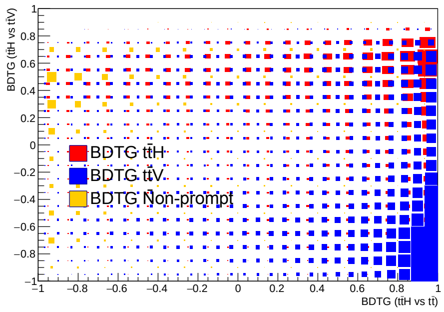
(b) BDTG response to ttV



(c) The ROC curve to non-prompt background



(d) The ROC curve for ttV



(e) BDTG_Non-prompt vs BDTG_ttV plane

Figure 4.13: Distribution of the BDTG non-prompt response and the BDTG $t\bar{t}V$ response. The background rejection versus signal efficiency (ROC) for the both BDTGs and BDTG_Non-prompt vs BDTG_$t\bar{t}V$ plane.

with TransfoD (6,0) configuration.

The modelling of this final BDT discriminant is good as shown in Figure 4.14, the $BDTG_{Combination}$.

As $t\bar{t}V$ is one of the main backgrounds, it is necessary to check its modelling with the BDTG output. A set of cuts on the BDTG output is applied to select a relative pure $t\bar{t}V$ region. A detailed cut and count method is designed to find the $t\bar{t}V$ CR and it is mentioned in Section 4.6. Figure 4.15 shows the BDTG input variables distribution comparison between data and prediction in the $t\bar{t}V$ region based on the BDTG output. The purity of $t\bar{t}V$ and variables' modelling is fine. The validation of the BDTG output is good.

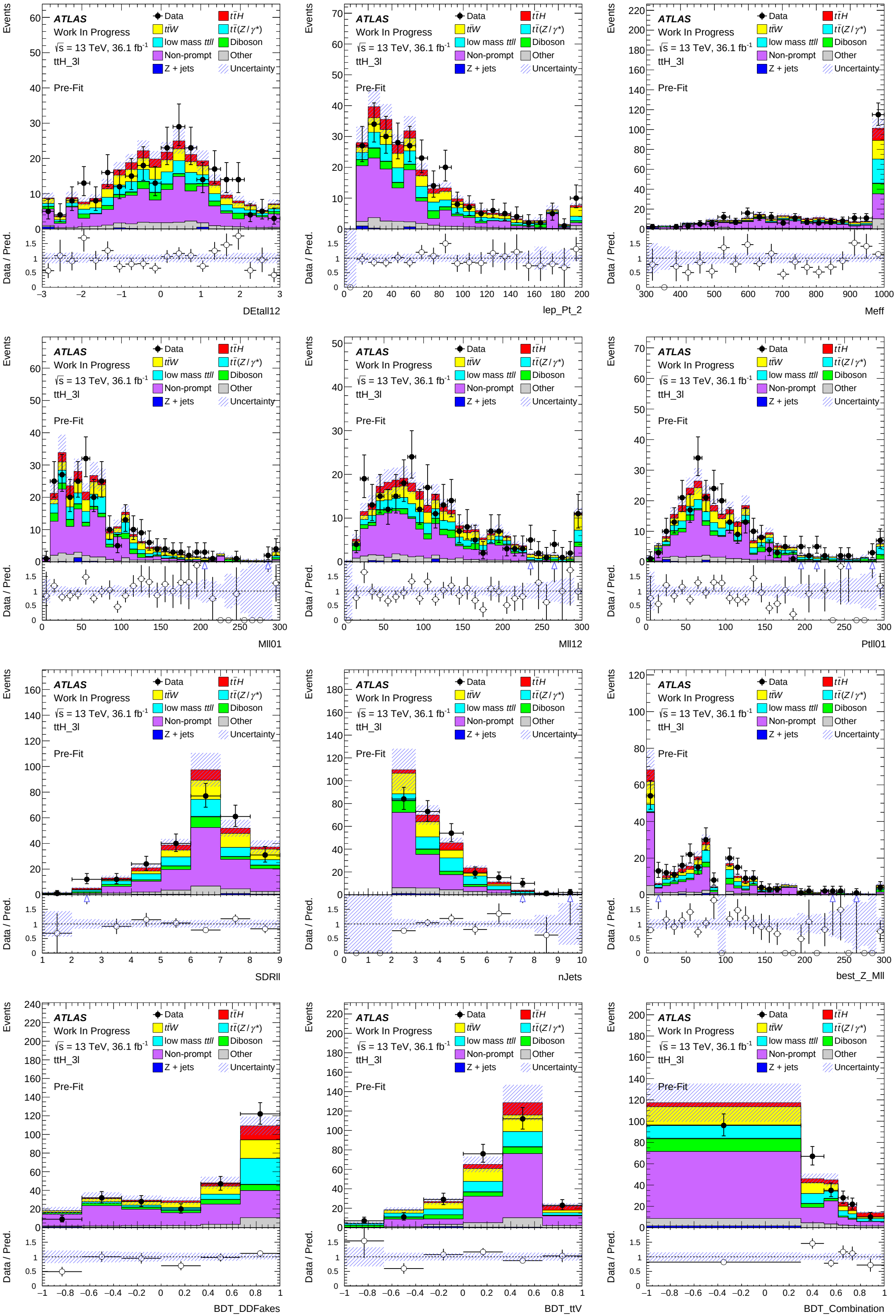


Figure 4.14: Pre-fit distributions of the MVA input and output variables in the pre-MVA region.

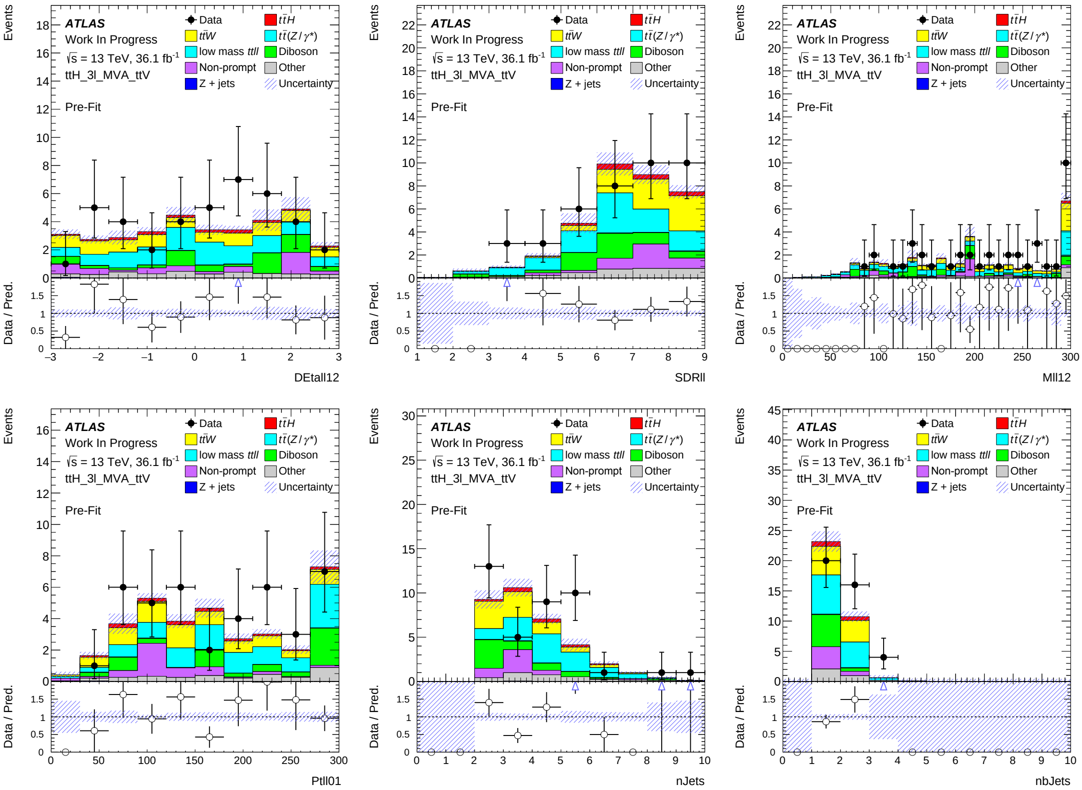


Figure 4.15: Pre-fit distributions of the MVA input variables in the $t\bar{t}V$ control region.

4.6 Validation and control regions

The description of data by various Monte Carlo samples (MC) is tested in several validation regions close to the signal region in the event topology. The validation regions under study are defined in Table 4.17. A selection of those comparison plots is shown in this section. The $t\bar{t}Z$ and $t\bar{t}W$ backgrounds are modelled from MC simulation in the current plots and can be used as a cross check in the analysis.

VR	Selection
$t\bar{t}Z$	3ℓ lepton and jet selection require at least one OS SF pair within 10 GeV of $m_Z = 91.2$ GeV and requiring 4 jets and 2 b-tagged jets
$t\bar{t}W$	2ℓ lepton selection, leading and subleading lepton $p_t > 30$ GeV require ≥ 2 b-tagged jets and 3 or 4 reconstructed jets H_T(jets) > 220 GeV in ee and $e\mu$ channel M_{ee} not within [75,105] GeV and $E_T^{miss} > 50$ GeV in ee channel both leptons with positive charge

Table 4.17: Description of the validation regions being designed for $t\bar{t}Z$ and $t\bar{t}W$ background.

4.6.1 $t\bar{t}W$

As $t\bar{t}W$ is the dominant irreducible background and it has not been observed yet in ATLAS, it is important to build a validation region to make sure that the MC prediction matches the data in $t\bar{t}W$ enriched region. Two VRs are built based from pre-MVA region: a region defined with cut and count using simple kinematic variables (Table 4.17) and a region based on the BDTG_Non-prompt vs BDTG_$t\bar{t}V$ plane (shown in Figure 4.15). In both cases, the purity is above 40% but modelling of $t\bar{t}W$ region based on MVA output is better. In both cases, a fit to the strength of $t\bar{t}W$ gives an agreement with the MC prediction at better than 1σ level.

Distributions of the number of electrons and the number of jets are shown in Figure 4.16. A good agreement between prediction and data is observed and the tests reinforce the confidence that $t\bar{t}W$ is well modelled by the Monte Carlo prediction.

4.6.2 $t\bar{t}Z$

The Run 1 definition uses an inverted Z-veto cut in the 3ℓ signal region selection, which means selecting events in Z mass window with at least 4 jets and 1 b-tagged

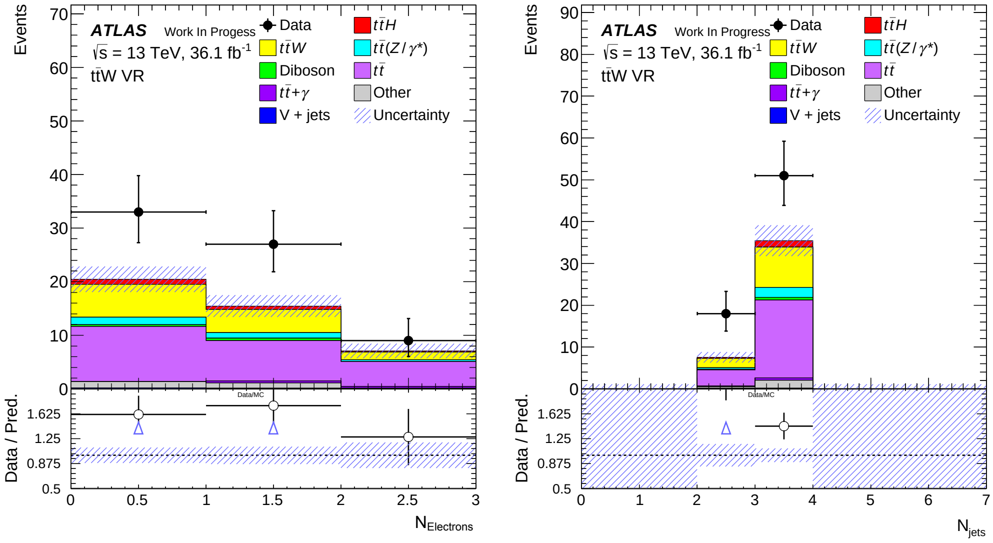


Figure 4.16: Distributions of the number of electrons and number of jets for the $t\bar{t}W$ cut-based validation region.

jet or at least 3 jets and 2 b-tagged jets. $t\bar{t}Z$ purity (especially against WZ events) can be improved by requiring that events must have at least 4 jets and 2 b-tagged jets. Overall $t\bar{t}Z$ purity is about 85%. Comparisons between data and MC are shown in Figure 4.17.

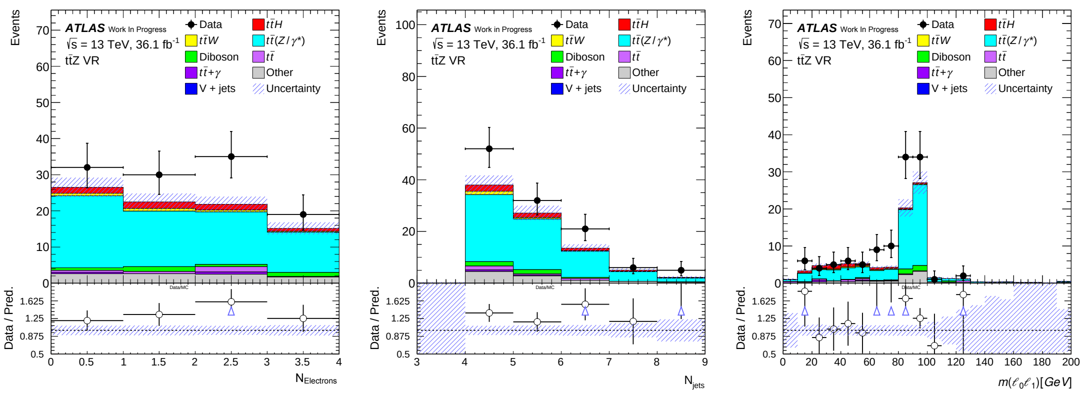


Figure 4.17: Distributions of the number of electrons, number of jets and invariant mass of Lep0 and Lep1 for $t\bar{t}Z$ cut-based validation region and errors include statistical and systematic uncertainties.

4.7 Uncertainties

Sets of uncertainties which are related to the analysis in this thesis are presented. A brief summary on the uncertainties is discussed here, mainly covering experimental systematics, signal and background modelling theoretical systematics. The sources of systematic uncertainty considered in this analysis are summarised in Table 4.21. They impact the estimated signal and background rates, the migration of events between categories and/or the shape of the discriminants used in the final fit.

The uncertainty in the combined 2015+2016 integrated luminosity is 2.1%. It is derived, a methodology similar to that detailed in Ref. [63], from a calibration of the luminosity scale by using x-y beam-separation scans performed in August 2015 and May 2016.

The uncertainties in physics objects which are related to the reconstruction and identification of light leptons and hadronic τ leptons, to the reconstruction and b-tagging of jets, and to the reconstruction of E_T^{miss} are all considered. The impact of the b-tagging efficiency uncertainty is weaker, as the b-tagged jet multiplicity in all selected events is typically one or two. The impact of the uncertainty of jets containing either charm jets or τ_{had} is significant and, due to the calibration procedure applied, is taken as fully correlated between the two jet flavours. Uncertainties in lepton reconstruction, identification and trigger efficiencies have negligible impact.

The systematic uncertainties associated with the generation of signal and background processes are due to uncertainties in the assumed cross sections and the acceptance for each process in each category and bin used in the final fit. The most important uncertainty arising from theoretical predictions is on the assumed SM cross sections and the modelling of the acceptance for $t\bar{t}H$, $t\bar{t}Z$ and $t\bar{t}W$ production.

4.7.1 Experimental systematics

Many quantities used in thesis are subject to the experimental systematic uncertainties. Each systematic effect is evaluated individually by using the given uncertainties on an event by event basis. These uncertainties are related to the trigger efficiency, lepton reconstruction and identification, jet calibration, continuous b-tagging and the global event activity. The experimental systematic treatments are evaluated by ATLAS performance groups and are used in this analysis as an overall reweighting or rescaling of the object energy and momentum. The list of the systematics uncertainties to be considered and included is summarised in Table 4.18 along with their type, description and name of systematics in the workspace and status of inclusion in the analysis.

The application row in Table 4.18 indicates the methodology of inclusion of the systematic in the analysis: overall event reweight or as a data/MC determined scale factor of the transverse momentum. If no explicit indication is provided the rescaling approach is adopted. The systematic associated to Jet Vertex Tagger, taking into account various information to quantifying the fraction of the track transverse momentum associated to a jet from the hard scattering interaction, requires a particular treatment. This systematic uncertainty is determined from the variation of the corresponding cut. The effect of the various systematics is evaluated assuming a positive and negative 1σ variation around the nominal value of the interested quantity to evaluate the effects on the results yields.

The summary of experimental systematics for the b-tagged jets in the analysis is shown in Table 4.20.

Type	Description	Systematics Name	Application	Analysis
		Experimental Systematics on Leptons		
		Trigger		
Scale Factors	Trigger Efficiency	lepSFTrigTight_MU(EL)_SF_Trigger_STAT(SYST)	Event Weight	$\surd$
		Muons		
Efficiencies	Reconstruction and Identification	lepSFObjTight_MU_SF_ID_STAT(SYST)	Event Weight	$\surd$
	Isolation	lepSFObjTight_MU_SF_Isol_STAT(SYST)	Event Weight	$\surd$
	Track To Vertex Association	lepSFObjTight_MU_SF_TTVA_STAT(SYST)	Event Weight	$\surd$
p_T Scale	p_T Scale	MUONS_SCALE	p_T Correction	$\surd$
Resolution	Inner Detector Energy Resolution	MUONS_ID	p_T Correction	$\surd$
	Muon Spectrometer Energy Resolution	MUONS_MS	p_T Correction	$\surd$
		Electrons		
Efficiencies	Reconstruction	lepSFObjTight_EL_SF_ID	Event Weight	$\surd$
	Identification	lepSFObjTight_EL_SF_Reco	Event Weight	
	Isolation	lepSFObjTight_EL_SF_Isol	Event Weight	$\surd$
Scale Factor	Energy Scale	EG_SCALE_ALL	Energy Correction	$\surd$
Resolution	Energy Resolution	EG_RESOLUTION_ALL	Energy Correction	$\surd$
		Hadronic Taus		
Efficiencies	Reconstruction	tauSFLoose_TAU_SF_RECO_TOTAL	Event Weight	$\surd$
	Identification BDT	tauSFTight_TAU_SF_JETID_TOTAL	Event Weight	$\surd$
	Electron Veto BDT	tauSFTight_TAU_SF_ELEOLR_TOTAL	Event Weight	$\surd$
Scale Factor	p_T Scale	TAUS_TRUEHADTAU_SME_TES_MODEL	p_T Correction	$\surd$
		TAUS_TRUEHADTAU_SME_TES_DETECTOR	p_T Correction	$\surd$
		TAUS_TRUEHADTAU_SME_TES_INSITU	p_T Correction	$\surd$

Table 4.18: Summary of experimental systematics in the analysis for muons, electrons and hadronic tau objects. From left: type, description, name of systematics in the code, mode of application and status of inclusion in the analysis. The mode of application indicates the systematic evaluation: overall event re-weighting (Event Weight) or rescaling (*e.g.* p_T Correction).

4.7.2 Signal and backgroung modelling theoretical systematics

The uncertainties concerning on the cross section, parton shower and generator variation for the main background relying on the MC prediction are summarised in Table 4.19. The systematic uncertainties related to the data driven fakes are discussed in the Section 4.4.2.

Process (default)	X-section [%]	Generator (alternative)	Parton Shower (alternative)	Scale uncertainty
$t\bar{t}H$ (aMC@NLO+Pythia8)	QCD Scale: $^{+5.8}_{-9.2}$ PDF($+\alpha_S$): $\pm$ 3.6		(aMC@NLO+Herwig++)	event-by-event weight
$t\bar{t}Z$ (aMC@NLO+Pythia8)	QCD Scale: $^{+9.6}_{-11.3}$ PDF($+\alpha_S$): $\pm$ 4	(Sherpa)		event-by-event weight event-by-event weight
$t\bar{t}W$ (aMC@NLO+Pythia8)	QCD Scale: $^{+12.9}_{-11.5}$ PDF($+\alpha_S$): $\pm$ 3.4	(Sherpa)		event-by-event weight event-by-event weight
Diboson (Sherpa 2.2.1)	$\pm$ 50			event-by-event weight event-by-event weight

Table 4.19: Summary of theoretical uncertainties for $t\bar{t}H$, $t\bar{t}V$ and diboson MC predictions.

Experimental Systematics on *b*-tagged jets			
Type	Origin	systematics Name	Analysis
		b-tags	
Scale Factors	MV2c10 b-tagger efficiency on b originated jets in bins of η	MC2c10_70_EventWeight_B0-5	$\surd$
	MC2c10 b-tagger efficiency on c originated jets in bins of η	MC2c10_70_EventWeight_C0-3	$\surd$
	MC2c10 b-tagger efficiency on light flavoured originated jets in bins of η and p_T	MC2c10_70_EventWeight_Light0-11	$\surd$
	MC2c10 b-tagger extrapolation efficiency	MC2c10_70_EventWeight_extrapolation MC2c10_70_EventWeight_extrapolation_from_charm	$\surd$ $\surd$

Table 4.20: Summary of experimental systematics for the *b*-tagged jets in the analysis, using the MC2c10 tagging algorithm used at the 70% Working Point. All of the b-tagging related systematics are applied as event weights. From left: type, description, name of systematic in the code and status of inclusion in the analysis.

Table 4.21: Sources of systematic uncertainties considered in the analysis. "N" means that the uncertainty is taken as normalisation-only for all processes and channels affected, whereas "S" denotes systematics that are considered shape-only in all processes and channels. "SN" means that the uncertainty is taken on both shape and normalisation. Some of the systematic uncertainties are split into several components("Comp"), as indicated by the number in the rightmost column.

Systematic uncertainty	Type	Comp
Luminosity	N	1
Pile-Up reweighting	SN	1
Physics Objects		
Electron	SN	6
Muon	SN	15
Jet energy scale and resolution	SN	28
Jet vertex fraction	SN	1
Jet flavour tagging	SN	126
E_T^{miss}	SN	3
Total (Experimental)	–	181
Data-driven non-prompt/fake leptons and charge misassignment		
Control region statistics	SN	38
Light lepton efficiencies	SN	22
Non-prompt light lepton estimates: non-closure	N	5
γ-conversion fraction	N	5
Electron charge misassignment	SN	1
Total (Data driven reducible background)	–	71
$t\bar{t}H$ **modelling**		
Cross section (QCD, PDF)	N	2
QCD scale	S	3
Parton shower	SN	1
Higgs branching ratio	N	4
Shower tune	SN	1
$t\bar{t}W$ **modelling**		
Cross section (QCD, PDF)	N	2
QCD scale	S	3
Generator	SN	1
Shower tune	SN	1
$t\bar{t}Z$ **modelling**		
Cross section (QCD, PDF)	N	2
QCD scale	S	3
Generator	SN	1
Shower tune	SN	1
Other background modelling		
Cross section	N	15
QCD Scale	SN	1
Total (Signal and background modelling)	–	41
Total (Overall)	–	293

4.8 Statistical treatment

The simultaneous fit is used to drive the signal significance and quantify the agreement between the observed data and MC prediction in the signal region for the $t\bar{t}H$ three leptons final state.

A statistical analysis tool called ***ttHFitter*** [64] is employed. It is implemented with ***HistFitter*** [65] and the ***RooFit/RooStat*** [66]. The method is based on profile-likelihood ratio test (more details are in the Section 4.8.1). The signal and backgrounsd are described by binned probability density function (PDF) defined in the statistical independent regions and implemented as histograms.

In this thesis, the combined BDTG shape is used as the input for the fitting to have shape information. In the final fit configuration, the signal strength $\mu_{t\bar{t}H}$ is the only parameter of interest (POI) (scaling factors which are not constrained, used to adjust the relative contribution of the main background and signal components) and is extracted from the shape fit on combined BDTG with 6 autobins configuration. The considered nuisance parameter (NP)(scaling factors with external constrained used to model all statistical and systematic uncertainties) are listed in Table 4.21.

4.8.1 Likelihood function

The likelihood - $\mathcal{L}(\mu, \theta)$ is a function of the signal strength μ and of set of nuisance parameters $\theta = \theta_a, \theta_b, ...$, which represent the uncertainties. The signal strength μ ($\mu_{t\bar{t}H} = \frac{\sigma \times BR}{(\sigma \times BR)_{SM}}$) is the parameter of interest that needs to be determined and is a free parameter of the fit. $\mu = 0$ corresponds to a background only hypothesis and $\mu = 1$ refers to the SM signal and background hypothesis. The real value of these parameters are unknown and are estimated by maximizing the likelihood in the fit. Generally the likelihood function describes the analysis with a product of the Poisson terms for numbers of events in bins and Gaussian term for systematic uncertainties as follows:

$$\begin{aligned}
\mathcal{L} &= P_{SR} \times P_{CR} \times C_{Syst} \\
&= \prod_{i=1}^{N_{SR}} \prod_{j=1}^{N_{bins}} P(n_{SR}^{ij} | \lambda_{SR}(\mu_{sig}, \vec{b}, \vec{\theta})) \times \prod_{i=1}^{N_{CR}} \prod_{j=1}^{N_{bins}} P(n_{CR}^{ij} | \lambda_{CR}(\mu_{sig}, \vec{b}, \vec{\theta})) \\
&\times C_{Syst}(\vec{\theta}).
\end{aligned} \tag{4.17}$$

The first two items are the Poisson items for the numbers of observed events in the signal region (n_{SR}^{ij}) and the numbers of events in each control region (n_{CR}^{ij}). The index i and j are accounting for the number of signal region or control regions and numbers of considered bins of the discriminant variables, respectively. The expectation value ($\lambda_{SR,CR}$) for Poisson is written as:

$$\lambda_{SR,CR}(\mu_{sig}, \vec{b}, \vec{\theta}) = \mu_{sig} n^{ij}_{sig}(\vec{\theta}) + n^{ij}_{bkg}(\vec{\theta}, \vec{b}), \tag{4.18}$$

where μ_{sig}: the free parameter signal strength; $\vec{b} = \sum_i b_i$ are the background normalization factors for the background sources i; $\vec{\theta} = \theta_1, ..., \theta_N$ are the nuisance parameters used to parametrize the systematic uncertainties like the luminosity, etc. The n^{ij}_{sig} and n^{ij}_{bkg} are the number of signal events and background events in the i signal (control) region and bin index j.

The different sources of the systematics are included using the PDF $C_{Syst}(\vec{\theta})$, which is the product of the probability distribution describing each source of systematic uncertainty. Generally the PDF is a Gaussian with a width equal to 1:

$$C_{Syst}(\vec{\theta}) = \prod_{k=1}^{N_\theta} G(\theta_j). \tag{4.19}$$

where k is the number of the nuisance parameters. The impact of nuisance parameters on the expectation value is described by the $\lambda_{SR,CR}$.

4.8.2 Hypothesis testing

We can not get the information on the level of agreement between data and the background only or signal plus background hypothesis through the estimation of POI $\mu_{t\bar{t}H}$. Therefore, the hypothesis testing [67] used in the analysis is based on profiled likelihood ratio, $\tilde{\lambda}(\mu)$, defined as:

$$\tilde{\lambda}(\mu) = \begin{cases} \dfrac{\mathcal{L}(\mu, \hat{\hat{\theta}}(\mu))}{\mathcal{L}(\hat{\mu}\hat{\theta}(\mu)}, & \hat{\mu} \geq 0; \\ \dfrac{\mathcal{L}(\mu, \hat{\hat{\theta}}(\mu))}{\mathcal{L}(0, \hat{\theta}(0)}, & \hat{\mu} < 0. \end{cases} \tag{4.20}$$

The numerator maximize the likelihood for a specific value of μ and denominator is the unconditional maximized likelihood when $\mu \geq 0$. The $\mu < 0$ is refer to avoiding complications in the computations in the case of deficit of signal-like events. Depending on the tested hypothesis, different test statistics $\tilde{t}_\mu$ are defined to test a particular of μ. For the observed data, the test statistics is called $\tilde{t}^{obs}_\mu$. The sampling distribution of the test statistics is denoted by $f(\tilde{t}_\mu|\mu, \hat{\hat{\theta}}(\mu))$. The p-value for a given observation, under a particular hypothesis, is the probability for an equal or more extreme outcome than the observed, under the assumed hypothesis:

$$p_\mu = \int_{\tilde{t}^{obs}_\mu}^{\infty} f(\tilde{t}_\mu|\mu, \hat{\hat{\theta}}(\mu)) d\tilde{t}_\mu. \tag{4.21}$$

p-value with a small value presents evidence against the test hypothesis and is often converted into the normal significance Z value.

The test statistics $\tilde{q}_0$ is defined as follows in case of any excess in data:

$$\tilde{q}_0 = \begin{cases} -2ln\tilde{\lambda}(0), & \hat{\mu} > 0; \\ 0, & \hat{\mu} \leq 0. \end{cases} \tag{4.22}$$

To test the compatibility of the data with the background only hypothesis ($\mu = 0$):

$$p_0 = \int_{\tilde{q}_0^{obs}}^{\infty} f(\tilde{q}_0|0, \hat{\hat{\theta}}(0))d\tilde{q}_0, \tag{4.23}$$

where p_0 is the probability the background only hypothesis leads to a test statistics equal to or larger than observed. A discovery is declared if the background only hypothesis is rejected at the 5σ level.

Besides a test statistics $\tilde{q}_\mu$:

$$\tilde{q}_\mu = \begin{cases} -2ln\tilde{\lambda}(\mu), & \hat{\mu} \geq \mu; \\ 0, & \hat{\mu} < \mu. \end{cases} \tag{4.24}$$

is used to for setting upper limits in the μ. The Confidence Level of signal (CL_s) is employed to set these limits.

$$p_\mu = \int_{\tilde{q}_\mu^{obs}}^{\infty} f(\tilde{q}_\mu|\mu, \hat{\hat{\theta}}(\mu))d\tilde{q}_\mu, \quad p_b = \int_{\infty}^{\tilde{q}_0^{obs}} f(\tilde{q}_0|0, \hat{\hat{\theta}}(0))d\tilde{q}_0, \tag{4.25}$$

where p_b is the p_μ under the background-only hypothesis. The 95% CL upper limits on μ are given for

$$CL_s(\mu) = \frac{p_\mu}{1 - p_b} \leq 5\%. \tag{4.26}$$

To compute p-value related to a certain hypothesis, the full distribution of the test statistics $\tilde{t}_\mu$ needs to be determined. By constructing pseudo-experiments of the hypothesis and calculating the test statistics $\tilde{t}_\mu$. An approximation of the $f(\tilde{t}_\mu|\mu, \hat{\hat{\theta}}(\mu))$ can be obtained in the limit of large statistics by using the Wald's approximation [68]. A "Asimov" dataset is used to test the signal plus background hypothesis to get the expected significance in the thesis. The "Asimov" dataset is an artificial dataset which can replace the ensemble testing performed with MC pseudo-experiment with single dataset. The dataset is defined in a way to return exactly true value for each estimated parameter and is also used to get the error bands on the median expected limit on μ. This dataset can study the constraints on nuisance parameters which are obtained with expected data and statistical uncertainties.

To summarise, the profiled likelihood method can give a picture of some of the systematic uncertainties by fitting the data and a good agreement between data and MC can be achieved through the shift (pull) of a given systematic uncertainty. Nuisances parameters can reveal correlations among themselves during the likelihood maximisation including further reduction of the effect of the total systematic uncertainties.

The binning of the templates used in the fit model is a subtle parameter of the fitting. The choice is from a compromise between the best separation for the classifier shape and available amount of events in data and in MC. Extremely narrow bins could cause a lack of background events in the signal enriched side of classifier distribution. On the other hand, a large width binning could reduce the discriminating power of the classifier by mixing the shape between signal and background. In this thesis, an auto-binning configuration of ***ttHFitter*** is adopted with the optimized parameters b and s (mentioned in Section 4.5.2) to have good separation and balance between signal and background as well as to avoid the low statistics bins during the fit.

5 Results with 13 TeV Run 2 data

5.1 Asimov fit and expected significance

The ***ttHFitter*** based on profile-likelihood is used to drive the final fit results as well as the $\mu_{t\bar{t}H}$ evaluation which is performed at this stage of the analysis. The binned combined BDT output distributions are taken as the input to the fit. All the related systematic uncertainties mentioned in Section 4.7 are considered. Data-driven fakes, $t\bar{t}V$ and diboson background are constrained simultaneously by the fit. Expected results are performed with Asimov dataset first and then the full dataset is used for the final fit. The pre-fit event yields are shown in Table 5.1, and Table 5.2 gives a summary on expected value on $t\bar{t}H$ signal strength and expected significance for the 3ℓ channel, by fitting on Asimov dataset. The results include statistical and systematic uncertainties discussed in previous sections.

The ranking plots are shown in Figure 5.1 and the correlation between NPs is provided in Figure 5.2.

The nuisance parameters corresponding to systematic uncertainties are all centred on zero and the normalisation scale factors are all centred around 1 as expected. For what concerns the analysis sensitivity, the leading source of systematic uncertainties are those that show a high level of correlation with the signal strength.

The impact of a nuisance parameter on the fit $\mu_{t\bar{t}H}$ is calculated by the fixing the corresponding nuisance parameter at a level of $\theta \pm \sigma_\theta$ and perform the fit again. θ is the fitted value of the nuisance parameter and σ_θ is the post-fit uncertainty. The difference between the default and modified $\mu, \delta\mu$, means the effect on the μ of this particular systematic uncertainty.

Especially in the 3ℓ channel, the high ranking NPs are those which are related to the fakes estimation, such as the non-closure of the fakes in the 3ℓ, statistics of the fake muon and also the difference between the 2ℓ and 3ℓ because of the photon conversion fakes. Generally, there is no over-constrain issue in the 3ℓ channel.

5.2 Data fit

The fit has been performed to the data in one signal region and the results are presented in this section.

Pre-fit expected event yields of the 3ℓ channel	
Main background	
$t\bar{t}W$	42.9 ± 7.75
$t\bar{t}Z$	40.4 ± 4.72
VV	19.7 ± 14.3
Non-prompt	117 ± 31.9
Rare background	
Single top t-chan	$5 \times 10^{-5} \pm 3.51 \times 10^{-6}$
Single top s-chan	$5 \times 10^{-5} \pm 3.51 \times 10^{-6}$
Wt	0.266 ± 0.609
Three top	0.448 ± 0.225
Four top	2.45 ± 1.23
$t\bar{t}WW$	2.76 ± 0.381
low mass $ttll$	1.09 ± 0.157
$Z + jets$	1.91 ± 1.85
tZ	$5 \times 10^{-5} \pm 2.51 \times 10^{-5}$
WtZ	3.59 ± 1.83
rare top	7.62 ± 3.83
VVV	0.593 ± 0.317
VH	0.691 ± 1.15
$tHjb\,(H \to WW)$	0.367 ± 0.0536
$tHjb\,(H \to ZZ)$	0.0456 ± 0.0133
$tHjb\,(H \to \tau\tau)$	0.0479 ± 0.0116
$tHjb\,(H \to \text{Others})$	0.0549 ± 0.0163
$WtH\,(H \to WW)$	0.855 ± 0.105
$WtH\,(H \to ZZ)$	0.0875 ± 0.0192
$WtH\,(H \to \tau\tau)$	0.271 ± 0.0495
$WtH\,(H \to \text{Others})$	0.0153 ± 0.00870
Signal	
$t\bar{t}H\,(H \to WW)$	17.7 ± 1.78
$t\bar{t}H\,(H \to ZZ)$	1.14 ± 0.124
$t\bar{t}H\,(H \to \tau\tau)$	3.75 ± 0.468
$t\bar{t}H\,(H \to \text{Others})$	0.492 ± 0.412
Total	266 ± 34.3

Table 5.1: Pre-fit expected event yields of the 3ℓ channel.

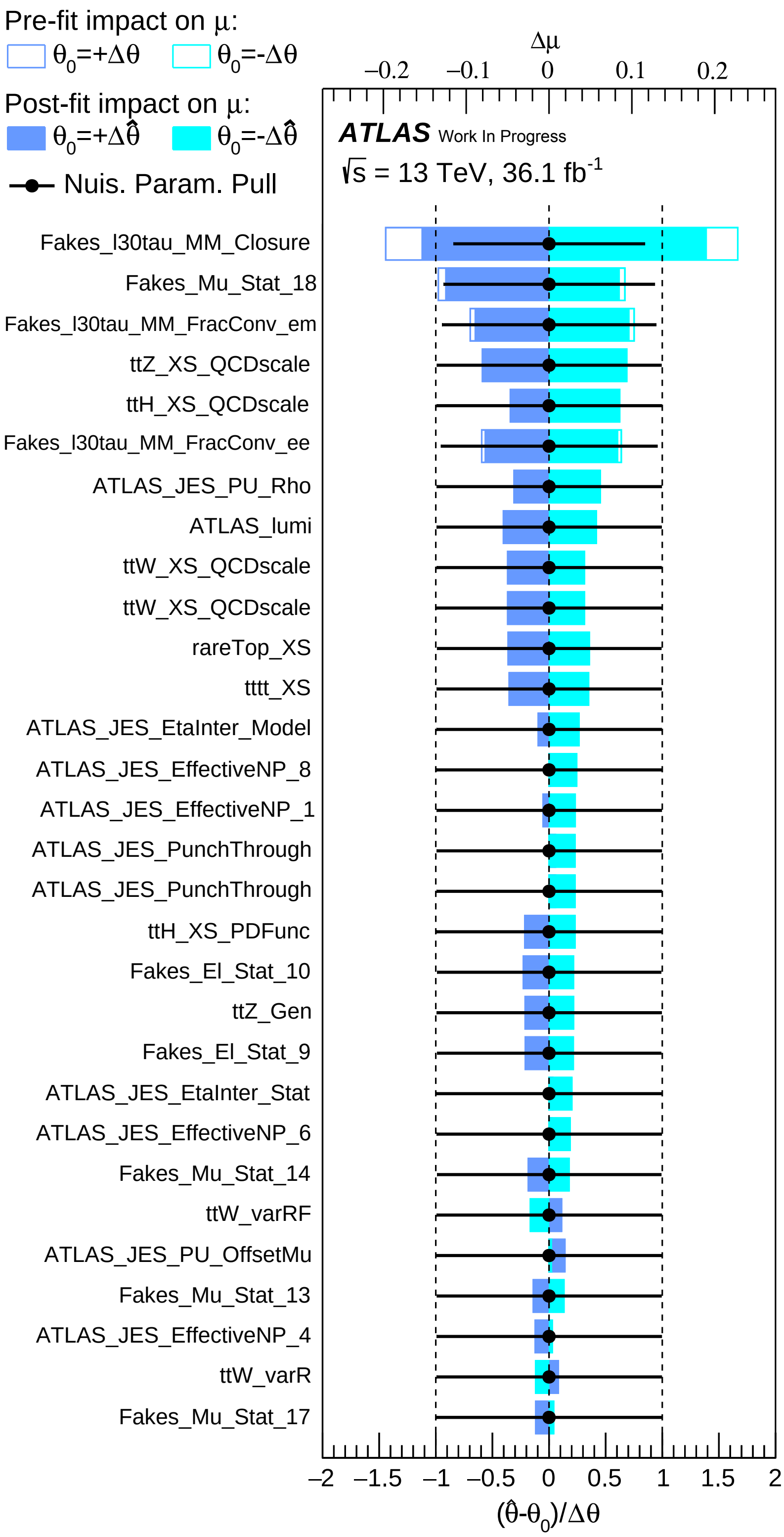


Figure 5.1: Ranking plot of the nuisance parameters from the Asimov dataset fitting.

	μ with total uncertainty	Expected significance (σ)
3ℓ	$1^{+0.8}_{-0.7}$	1.5 ± 1.2

Table 5.2: Summary on expected signal strength with all uncertainties and expected significance from the Asimov fit.

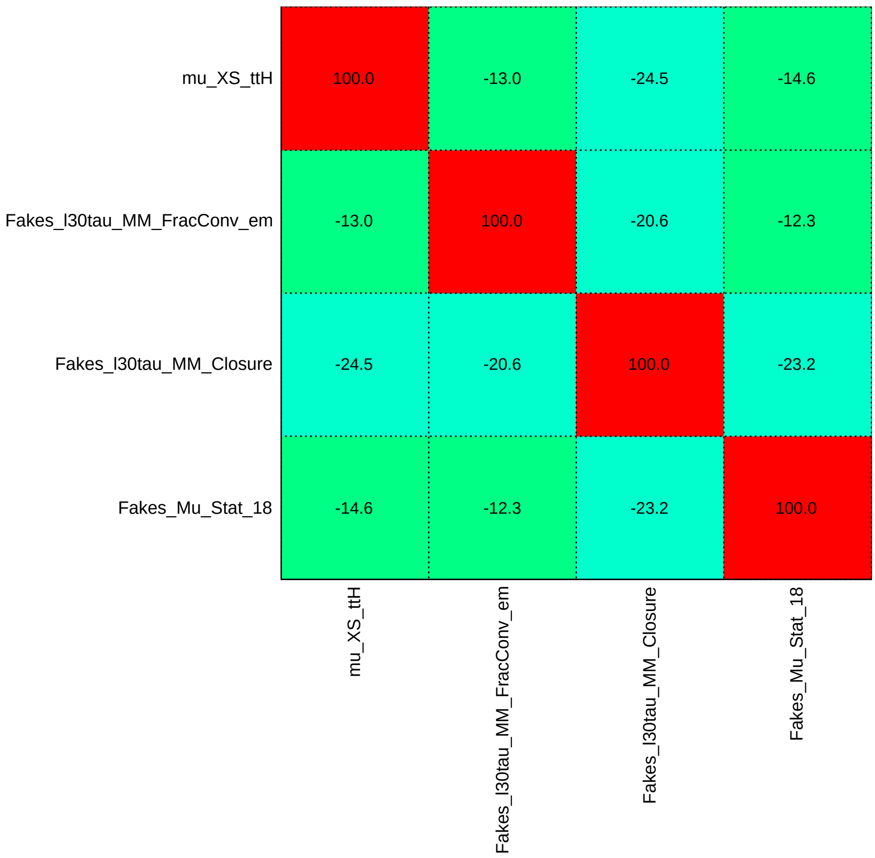


Figure 5.2: Correlation matrix of the nuisance parameters for Asimov dataset fitting.

Table 5.3 shows the event yields comparison between MC and real data after the fit. Table 5.4 shows the observed signal strength and observed significance with respect to the null hypothesis by using the total Lumi 36.1 fb^{-1} data with the total uncertainties (which including the statistical uncertainty and systematic uncertainty). The results show that the observed significance with respect to the null hypothesis has reached 2.18 standard deviations which is much better than the Run 1 measurement in the 3ℓ, but still large uncertainties need to be considered at the same time.

The distribution of combined BDT after fitting to the data is shown in Figure 5.3. The relative uncertainties decrease significantly due to the constraints provided by the data.

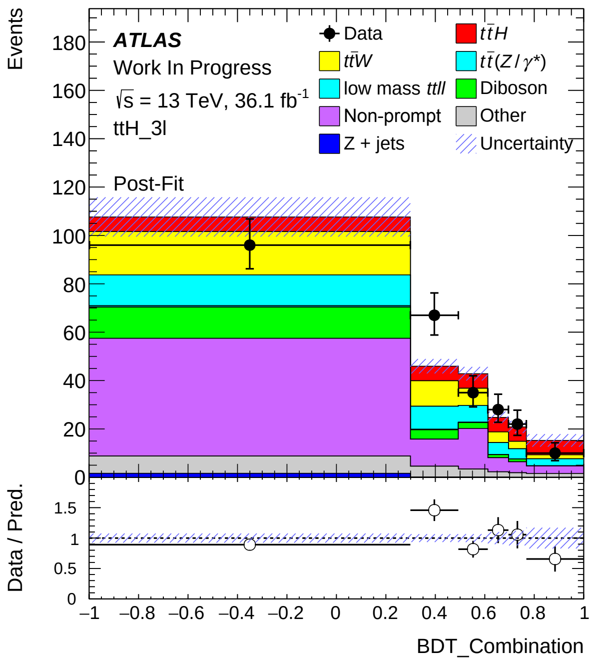


Figure 5.3: The combined BDT distribution after data fitting.

The Figure 5.4 shows the ranking of the NP after data fitting. The largest impact on the signal strength fitting comes from the fakes estimation, the uncertainties due to the non-closure of the fake estimation.

Post-fit event yields of the 3ℓ channel	
Main background	
$t\bar{t}W$	44.7 ± 11.2
$t\bar{t}Z$	41.2 ± 4.64
VV	21.6 ± 20.5
Non-prompt	90.2 ± 17.7
Rare background	
Single top t-chan	$6.01 \times 10^{-5} \pm 4.06 \times 10^{-6}$
Single top s-chan	$6.01 \times 10^{-5} \pm 4.06 \times 10^{-6}$
Wt	0.260 ± 0.664
Three top	0.451 ± 0.225
Four top	2.48 ± 1.23
$t\bar{t}WW$	2.76 ± 0.379
low mass $ttll$	1.11 ± 0.155
$Z+$jets	1.86 ± 1.83
tZ	$6.01 \times 10^{-5} \pm 3.00 \times 10^{-5}$
WtZ	3.70 ± 1.84
rare top	7.94 ± 3.88
VVV	0.603 ± 0.318
VH	0.695 ± 0.873
$tHjb\,(H \to WW)$	0.369 ± 0.0534
$tHjb\,(H \to ZZ)$	0.0461 ± 0.0134
$tHjb\,(H \to \tau\tau)$	0.0483 ± 0.0116
$tHjb\,(H \to \text{Others})$	0.0553 ± 0.0163
$WtH\,(H \to WW)$	0.858 ± 0.105
$WtH\,(H \to ZZ)$	0.0878 ± 0.0193
$WtH\,(H \to \tau\tau)$	0.271 ± 0.0503
$WtH\,(H \to \text{Others})$	0.0154 ± 0.00870
Signal	
$t\bar{t}H\,(H \to WW)$	27.5 ± 13.3
$t\bar{t}H\,(H \to ZZ)$	1.77 ± 0.866
$t\bar{t}H\,(H \to \tau\tau)$	5.85 ± 2.85
$t\bar{t}H\,(H \to \text{Others})$	0.824 ± 0.748
Total	257 ± 19.4
Data	258

Table 5.3: Post-fit event yields of the 3ℓ channel.

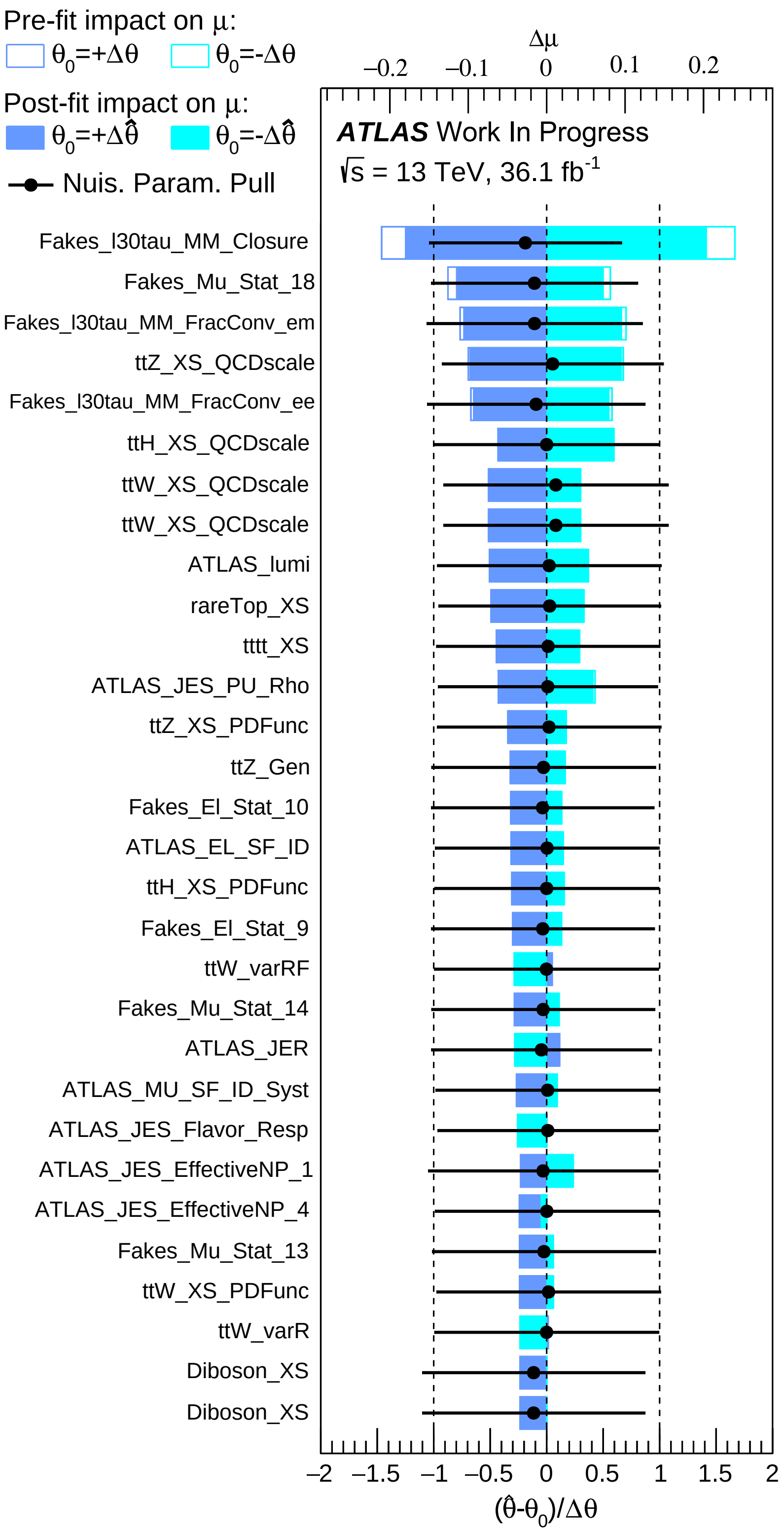


Figure 5.4: Ranking plot of the nuisance parameters to the fit.

	μ with total uncertainty	Observed Significance (σ)
3ℓ	$1.5^{+0.8}_{-0.7}$	2.2 ± 1.4

Table 5.4: Summary on observed signal strength with all uncertainties and the corresponding significance with respect to the null hypothesis from data.

5.3 Interpretation

The results shown in the above section with the two dimensional BDT cut in a single signal region show the promising outcomes compared to Run 1 $t\bar{t}H$ three leptons final state study. The higher cross section and advanced analysis strategy bring the significant improvements. Especially the optimised analysis in the 3ℓ channel presented in this thesis can have a better statistics of the signal region and good description of both MC and the data-driven estimation.

The ATLAS has released the approved results with 36.1 fb^{-1} data in the multi-lepton channel including the 3ℓ. Especially in the 3ℓ channel, a multi-class analysis strategy which is based on the multi-variate analysis as well is adopted. The most difference between the method used in this thesis and the multi-class method is the categorization procedure applied on final BDT output to have a pure signal and control region. The multi-class analysis has one signal region and the other four control regions which cover the $t\bar{t}W$, $t\bar{t}Z$, non-prompt and diboson process. All those regions are based on a 5D BDT training output.

The observed (expected) best-fit value of $\mu_{t\bar{t}H}$, combining all multilepton channels, is $1.56\,^{+0.30}_{-0.29}$ (stat) $^{+0.39}_{-0.30}$ (syst) $=1.56\,^{+0.49}_{-0.42}$ ($1.00\,^{+0.29}_{-0.28}$ (stat) $^{+0.32}_{-0.27}$ (syst) $=1.00\,^{+0.43}_{-0.39}$). The best-fit value of $\mu_{t\bar{t}H}$ for each individual channel and the combination of all channels are shown in Figure 5.5 and Table 5.5. The consistency of the fitted signal strengths between the seven channels is 34%. The observed (expected) significance associated with respect to the background hypothesis is 4.1σ (2.7σ). This excess provides evidence for $t\bar{t}H$ production in the multilepton final states. The observed significance associated with the SM expectation $\mu_{t\bar{t}H} = 1$ is 1.4σ.

Table 5.5: Observed and expected best-fit values of the signal strength $\mu_{t\bar{t}H}$ and associated significance with respect to the null hypothesis. The expected values are shown for the pre-fit background estimates.

Channel	Best fit $\mu_{t\bar{t}H}$ (observed)	Best fit $\mu_{t\bar{t}H}$ (expected)	Observed (expected) significance
2ℓ	$1.51\,^{+0.43}_{-0.41}$ (stat.) $^{+0.50}_{-0.41}$ (syst.)	$1.00\,^{+0.43}_{-0.40}$ (stat.) $^{+0.47}_{-0.40}$ (syst.)	2.8σ (1.8σ)
3ℓ	$1.76\,^{+0.61}_{-0.57}$ (stat.) $^{+0.60}_{-0.50}$ (syst.)	$1.00\,^{+0.57}_{-0.53}$ (stat.) $^{+0.51}_{-0.44}$ (syst.)	2.4σ (1.5σ)
4ℓ	$-0.51\,^{+1.35}_{-0.81}$ (stat.) $^{+0.29}_{-0.29}$ (syst.)	$1.00\,^{+1.75}_{-1.15}$ (stat.) $^{+0.43}_{-0.22}$ (syst.)	- (0.8σ)
1ℓ2tau	$-0.58\,^{+1.07}_{-0.86}$ (stat.) $^{+1.15}_{-1.33}$ (syst.)	$1.00\,^{+1.31}_{-1.10}$ (stat.) $^{+1.45}_{-1.29}$ (syst.)	- (0.6σ)
2ℓ1tau	$3.68\,^{+1.47}_{-1.25}$ (stat.) $^{+0.99}_{-0.53}$ (syst.)	$1.00\,^{+1.13}_{-0.91}$ (stat.) $^{+0.52}_{-0.31}$ (syst.)	3.5σ (1.1σ)
2ℓOS1tau	$1.74\,^{+1.56}_{-1.45}$ (stat.) $^{+1.38}_{-1.16}$ (syst.)	$1.00\,^{+1.53}_{-1.42}$ (stat.) $^{+1.29}_{-1.15}$ (syst.)	0.9σ (0.5σ)
3ℓ1tau	$1.61\,^{+1.71}_{-1.27}$ (stat.) $^{+0.64}_{-0.19}$ (syst.)	$1.00\,^{+1.56}_{-1.12}$ (stat.) $^{+0.49}_{-0.23}$ (syst.)	1.3σ (0.9σ)
Combined	$1.56\,^{+0.30}_{-0.29}$ (stat.) $^{+0.39}_{-0.30}$ (syst.)	$1.00\,^{+0.29}_{-0.28}$ (stat.) $^{+0.32}_{-0.27}$ (syst.)	4.1σ (2.8σ)

The combined fit on Asimov data with $\mu_{ttH} = 1$ of the no-tau channels 2ℓSS, 3ℓ and 4ℓ gives an expected significance of 2.3 and errors on μ_{ttH} of $^{+0.49}_{-0.45}$. For the 4ℓ, $1\ell2\tau$, 2ℓOS1τ and $3\ell1\tau$ channels, the uncertainties on μ are mainly statistical, while

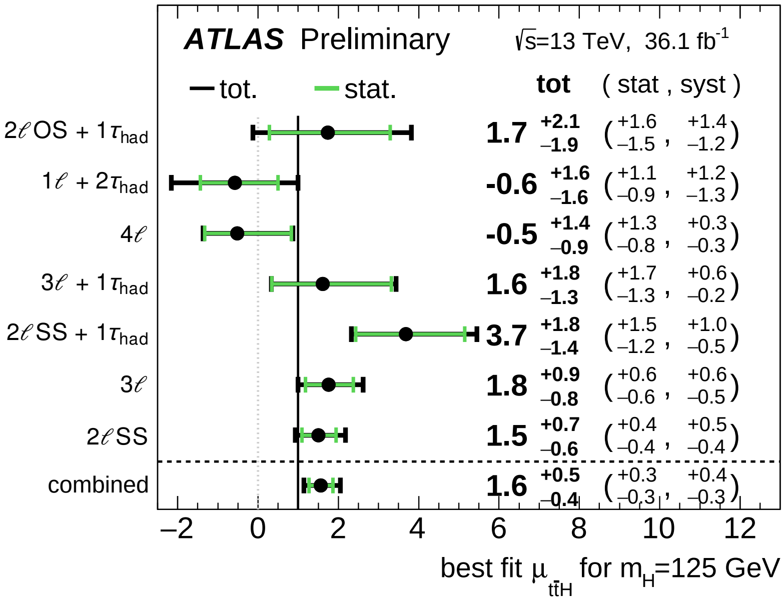


Figure 5.5: Observed best fit values of the $t\bar{t}H$ signal strength $\mu_{t\bar{t}H}$ and their uncertainties by final state category and combined. The SM prediction is $\mu_{t\bar{t}H} = 1$.

the statistical and systematic uncertainties are of comparable size for the 2ℓ, 3ℓ and $2\ell 1\tau$ channels. The uncertainties with highest impact on $\Delta\mu_{ttH}$ are currently $t\bar{t}H$ and $t\bar{t}W$ cross-section QCD scale, jet energy scale, pile-up subtraction, JES flavour composition and several Fake rate statistical uncertainties.

The results presented in the thesis are compatible with the above ATLAS approval results. Especially in the 3ℓ, the best-fit value of $\mu_{t\bar{t}H}$ is 1.76 $^{+0.61}_{-0.57}$ (stat) $^{+0.60}_{-0.50}$ (syst), and the corresponding observed significance is 2.4σ in the ATLAS approved results. The non-prompt background estimation with the Matrix Method in this thesis is adopted in the ATLAS approval results. Besides, the improvement (MVA strategy and the treatment of the uncertainty) can be seen for the current multi-variate analysis used in the thesis.

Conclusion

Higgs boson is discovered in 2012 by the ATLAS and CMS. The last piece of the Standard Model is found, and the structure of the Standard Model is almost completed. But there are still many questions to be answered, and the Yukawa coupling of the Higgs boson to the top quark is one of them. The top Yukawa coupling influences the evolution of the effective Higgs potential with the energy: any tension between the values allowed by the Standard Model and the observation would demand for new physics to solve the inconsistency. Precise measurements of its properties will confirm its nature, and any significant deviations from the Standard Model prediction will represent a clear sign of new physics. The associated production of the Higgs boson with a top quark pair allows for a direct measurement and is the main topic of this dissertation.

The work presented in this thesis focuses on a search for the Higgs boson production associated with a pair of top quarks in three leptons final state. The multi-variate technical (BDT) is employed to have a good discriminate power between signal and background. One of the main backgrounds, fake leptons, the Matrix Method is used to estimate its contribution. Generally, good modelling of the fake background and better separation with BDT output are achieved. Final results are obtained with a simultaneous fit on a combined 2D BDT.

For a Higgs boson mass of 125 GeV, an excess of events over the expected background from the other Standard Model processes is found with an observed significance of 2.2 standard deviations, compared to an expectation of 1.5 standard deviations. The best fit for the $t\bar{t}H$ production cross section, assuming a Higgs boson mass of 125 GeV, is $1.5^{+0.8}_{-0.7}$ times the SM expectation, and is consistent with the value of the Yukawa coupling to top quarks in the Standard Model. The results are compatible with the ATLAS approval results by using the same dataset in three leptons final state and improvements can be introduced into the analysis presented in this thesis.

Compared to the Run 1 results, a great improvement is obtained for both the significance and uncertainty on the signal strength measurement. The improvements come from the high centre of mass energy which brings the higher cross section, the upgrades of the ATLAS which bring lots of improvements, and also the advanced analysis strategy using. The promising result shows the evidence of the $t\bar{t}H$ production but more data are need to have the more precise measurement.

Uncertainties on the final measurements are still large, and can be re-optimized to reduce the impact.

After the restart of the LHC in 2015, larger statistics data brings further possibilities to validate the predictions of the SM and to search for new physics. During the current upgrades, the ATLAS detector is making upgrades to the limited subsystems. The LHC will shut down for next upgrades at the end of 2018 and the ATLAS will make improvements to the trigger system in order to deal with higher trigger rates. After restarting in 2019, the LHC will operate for three years at $\sqrt{s} = 14$ TeV and collect 300 fb^{-1} of data. The final LHC shut down will be in 2023, during which time the ATLAS will install a new, higher granularity, more radiation-hard inner tracker. A new triggering scheme, as well as improvements to the muon and calorimeter electronics, will be made. The upgraded High-Luminosity LHC (HL-LHC) will start in 2025, after which the collider is expected to collect 3000 fb^{-1} of data. The new dataset will allow for more precise Higgs measurements to have a further validation of the predictions by SM. With more and more colliding data and advanced technical analysis methods, the picture of the physics will be more clear or the discovery of new physics, which all need time to search for.

Bibliography

[1] ATLAS Collaboration. "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] (cit. on p. 38).

[2] CMS Collaboration. "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] (cit. on p. 38).

[3] ATLAS Collaboration. "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] (cit. on p. 38).

[4] ATLAS Collaboration. *Measurement of fiducial, differential and production cross sections in the $H \to \gamma\gamma$ decay channel with 13.3 fb^{-1} of 13 TeV proton-proton collision data with the ATLAS detector*. Tech. rep. ATLAS-CONF-2016-067. Geneva: CERN, 2016. URL: http://cds.cern.ch/record/2206210 (cit. on p. 38).

[5] ATLAS Collaboration. "Standard Model Total Production Cross Section Measurement." In: (2016). URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CombinedSummaryPlots/SM/index.html (cit. on p. 38).

[6] ATLAS Collaboration. "Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment." In: *Eur. Phys. J.* C76.1 (2016), p. 6. DOI: 10.1140/epjc/s10052-015-3769-y (cit. on p. 42).

[7] CMS Collaboration. "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV." In: *Eur. Phys. J.* C75.5 (2015), p. 212. DOI: 10.1140/epjc/s10052-015-3351-7. arXiv: 1412.8662 [hep-ex] (cit. on p. 42).

[8] ATLAS Collaboration. "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV." In: *JHEP* 08 (2016), p. 045. DOI: 10.1007/JHEP08(2016)045. arXiv: 1606.02266 [hep-ex] (cit. on pp. 43, 44).

[9] De Melis and Cinzia. "The CERN accelerator complex. Complexe des accélérateurs du CERN." In: (July 2016). URL: https://cds.cern.ch/record/2197559 (cit. on p. 45).

[10] Jean-Luc Caron. "Cross section of LHC dipole.. Dipole LHC: coupe transversale." May 1998. URL: https://cds.cern.ch/record/841539 (cit. on p. 47).

[11] LHC. "LHC Luminosity in 2011, 2012." In: (2017). URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResults (cit. on p. 49).

[12] Joao Pequenao. "Computer generated image of the whole ATLAS detector." In: (Mar. 2008). URL: https://cds.cern.ch/record/1095924 (cit. on p. 50).

[13] D. Parchet. "Axes Detecteur, Machine, Genie Cevil. technical drawing." In: (Dec. 1998). URL: http://www.hep.lu.se/atlas/thesis/egede/thesis-node39.html (cit. on p. 52).

[14] Joao Pequenao. "Computer generated image of the ATLAS inner detector." In: (Mar. 2008). URL: http://cds.cern.ch/record/1095926 (cit. on p. 53).

[15] Joao Pequenao. "Computer generated image of the ATLAS Colarimeter." In: (Mar. 2008). URL: http://cds.cern.ch/record/1095927 (cit. on p. 54).

[16] Joao Pequenao. "Event Cross Section in a computer generated image of the ATLAS detector." In: (Mar. 2008). URL: https://cds.cern.ch/record/1096081 (cit. on p. 54).

[17] ATLAS Collaboration. "Commissioning of the ATLAS Muon Spectrometer with Cosmic Rays." In: *Eur. Phys. J. C* 70.arXiv:1006.4384, CERN-PH-EP-2010-045 (Aug. 2010), 875–916. 58 p. URL: http://cds.cern.ch/record/1275998 (cit. on p. 56).

[18] ATLAS Collaboration. "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]. URL: http://inspirehep.net/record/931205/files/arXiv:1110.1910.pdf (cit. on p. 57).

[19] ATLAS Collaboration. *ATLAS Computing: technical design report*. Technical Design Report ATLAS. Geneva: CERN, 2005. URL: https://cds.cern.ch/record/837738 (cit. on p. 57).

[20] LHC. "Worldwide LHC Computing Grid." In: (2017). URL: http://wlcg.web.cern.ch/ (cit. on p. 57).

[21] ATLAS Collaboration. "The ATLAS Simulation Infrastructure." In: *The European Physical Journal C* 70.3 (2010), pp. 823–874. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-010-1429-9. URL: http://dx.doi.org/10.1140/epjc/s10052-010-1429-9 (cit. on p. 58).

[22] S. Agostinelli and etc. "Geant4 – a simulation toolkit." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250 –303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8. URL: http://www.sciencedirect.com/science/article/pii/S0168900203013688 (cit. on p. 58).

[23] Graphic Design. "LHC consolidations 2013-14." Feb. 2013. URL: http://cds.cern.ch/record/1516031 (cit. on p. 59).

[24] LHC. "LHC Luminosity in 2015, 2016." In: (2017). URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2 (cit. on p. 60).

[25] W Lampl, S Laplace, D Lelas, et al. *Calorimeter Clustering Algorithms: Description and Performance*. Tech. rep. ATL-LARG-PUB-2008-002. ATL-COM-LARG-2008-003. Geneva: CERN, Apr. 2008. URL: https://cds.cern.ch/record/1099735 (cit. on p. 63).

[26] T G Cornelissen, M Elsing, I Gavrilenko, et al. "The global Chi2 χ_2 track fitter in ATLAS." In: *Journal of Physics: Conference Series* 119.3 (2008), p. 032013. URL: http://stacks.iop.org/1742-6596/119/i=3/a=032013 (cit. on p. 63).

[27] ATLAS Collaboration. *Improved electron reconstruction in ATLAS using the Gaussian Sum Filter-based model for bremsstrahlung*. Tech. rep. ATLAS-CONF-2012-047. Geneva: CERN, May 2012. URL: http://cds.cern.ch/record/1449796 (cit. on p. 64).

[28] ATLAS Collaboration. *Electron efficiency measurements with the ATLAS detector using the 2012 LHC proton-proton collision data*. Tech. rep. ATLAS-CONF-2014-032. Geneva: CERN, June 2014. URL: http://cds.cern.ch/record/1706245 (cit. on p. 64).

[29] ATLAS Collaboration. "Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data." In: *Eur. Phys. J.* C74.10 (2014), p. 3071. DOI: 10.1140/epjc/s10052-014-3071-4. arXiv: 1407.5063 [hep-ex] (cit. on p. 64).

[30] ATLAS Collaboration. *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data*. Tech. rep. ATLAS-COM-CONF-2016-028. Geneva: CERN, Mar. 2016. URL: https://cds.cern.ch/record/2142831 (cit. on pp. 65, 68).

[31] ATLAS Collaboration. *Expected photon performance in the ATLAS experiment*. Tech. rep. ATL-PHYS-PUB-2011-007. Geneva: CERN, Apr. 2011. URL: http://cds.cern.ch/record/1345329 (cit. on p. 70).

[32] ATLAS Collaboration. "Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run-1 data." In: *Eur. Phys. J.* C76.12 (2016), p. 666. DOI: 10.1140/epjc/s10052-016-4507-9. arXiv: 1606.01813 [hep-ex] (cit. on p. 71).

[33] ATLAS Collaboration. "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] (cit. on p. 72).

[34] ATLAS Collaboration. *Measurement of the tau lepton reconstruction and identification performance in the ATLAS experiment using pp collisions at $\sqrt{s} = 13\ TeV$*. Tech. rep. ATLAS-CONF-2017-029. Geneva: CERN, May 2017. URL: http://cds.cern.ch/record/2261772 (cit. on p. 73).

[35] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The Anti-k(t) jet clustering algorithm." In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph] (cit. on p. 73).

[36] ATLAS Collaboration. *ATLAS Insertable B-Layer Technical Design Report*. Tech. rep. CERN-LHCC-2010-013. ATLAS-TDR-19. 2010. URL: https://cds.cern.ch/record/1291633 (cit. on p. 74).

[37] H. Pérez R., P. Kielanowski, and S. R. Juárez W. "Cabibbo–Kobayashi–Maskawa matrix: rephasing invariants and parameterizations." In: *J. Phys. Conf. Ser.* 485 (2014), p. 012058. DOI: 10.1088/1742-6596/485/1/012058. arXiv: 1209.5812 [hep-ph] (cit. on p. 74).

[38] ATLAS Collaboration. *Flavour tagging algorithms and performance at the ATLAS experiment*. Tech. rep. ATL-PHYS-PROC-2016-105. Geneva: CERN, 2016. URL: https://cds.cern.ch/record/2209195 (cit. on p. 74).

[39] Ovidiu Miu. "JetFitter vertex reconstruction performance in the ATLAS experiment." In: (2016). URL: http://cds.cern.ch/record/2210576 (cit. on p. 74).

[40] ATLAS Collaboration. *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*. Tech. rep. ATL-PHYS-PUB-2016-012. Geneva: CERN, June 2016. URL: http://cds.cern.ch/record/2160731 (cit. on pp. 75, 76).

[41] *b-tagging Benchmarks for taggers MV2c10 on DC14 8 TeV Samples (Feb 2014)*. Tech. rep. URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/BTaggingBenchmarksRelease17 (cit. on p. 75).

[42] ATLAS Collaboration. *ATLAS Collaboration, b-tagging performance plots in a $t\bar{t}$ dominated sample from early 2016 ATLAS data*. Tech. rep. URL: http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2016-001 (cit. on p. 76).

[43] ATLAS Collaboration. "ATLAS jet and missing-ET reconstruction, calibration, and performance." In: *Nucl. Part. Phys. Proc.* 273-275 (2016), pp. 1121–1126. DOI: 10.1016/j.nuclphysbps.2015.09.176 (cit. on p. 76).

[44] ATLAS Collaboration. "Performance of algorithms that reconstruct missing transverse momentum in $\sqrt{s} = 8$ TeV proton-proton collisions in the ATLAS detector." In: *Eur. Phys. J.* C77.4 (2017), p. 241. DOI: 10.1140/epjc/s10052-017-4780-2. arXiv: 1609.09324 [hep-ex] (cit. on p. 77).

[45] 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] (cit. on p. 81).

[46] 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] (cit. on p. 81).

[47] Simone Alioli, Paolo Nason, Carlo Oleari, et al. "NLO single-top production matched with shower in POWHEG: s- and t-channel contributions." In: *JHEP* 09 (2009). [Erratum: JHEP02,011(2010)], p. 111. DOI: 10.1007/JHEP02(2010)011,10.1088/1126-6708/2009/09/111. arXiv: 0907.4076 [hep-ph] (cit. on p. 81).

[48] Emanuele Re. "Single-top Wt-channel production matched with parton showers using the POWHEG method." In: *Eur. Phys. J.* C71 (2011), p. 1547. DOI: 10.1140/epjc/s10052-011-1547-z. arXiv: 1009.2450 [hep-ph] (cit. on p. 81).

[49] T. Gleisberg, Stefan. Hoeche, F. Krauss, 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] (cit. on p. 81).

[50] 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] (cit. on p. 81).

[51] W. Beenakker, S. Dittmaier, M. Kramer, et al. “Higgs radiation off top quarks at the Tevatron and the LHC.” In: *Phys. Rev. Lett.* 87 (2001), p. 201805. DOI: 10.1103/PhysRevLett.87.201805. arXiv: hep-ph/0107081 [hep-ph] (cit. on p. 80).

[52] W. Beenakker, S. Dittmaier, M. Kramer, et al. “NLO QCD corrections to t anti-t H production in hadron collisions.” In: *Nucl. Phys. B* 653 (2003), p. 151. DOI: 10.1016/S0550-3213(03)00044-0. arXiv: hep-ph/0211352 [hep-ph] (cit. on p. 80).

[53] S. Dawson, L. H. Orr, L. Reina, et al. “Associated top quark Higgs boson production at the LHC.” In: *Phys. Rev. D* 67 (2003), p. 071503. DOI: 10.1103/PhysRevD.67.071503. arXiv: hep-ph/0211438 [hep-ph] (cit. on p. 80).

[54] S. Dawson, C. Jackson, L. H. Orr, et al. “Associated Higgs production with top quarks at the large hadron collider: NLO QCD corrections.” In: *Phys. Rev. D* 68 (2003), p. 034022. DOI: 10.1103/PhysRevD.68.034022. arXiv: hep-ph/0305087 [hep-ph] (cit. on p. 80).

[55] Yu Zhang, Wen-Gan Ma, Ren-You Zhang, et al. “QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider.” In: *Phys. Lett. B* 738 (2014), p. 1. DOI: 10.1016/j.physletb.2014.09.022. arXiv: 1407.1110 [hep-ph] (cit. on p. 80).

[56] S. Frixione, V. Hirschi, D. Pagani, et al. “Weak corrections to Higgs hadroproduction in association with a top-quark pair.” In: *JHEP* 09 (2014), p. 065. DOI: 10.1007/JHEP09(2014)065. arXiv: 1407.0823 [hep-ph] (cit. on p. 80).

[57] S. Frixione, V. Hirschi, D. Pagani, et al. “Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons.” In: *JHEP* 06 (2015), p. 184. DOI: 10.1007/JHEP06(2015)184. arXiv: 1504.03446 [hep-ph] (cit. on p. 80).

[58] S. Heinemeyer et al. “Handbook of LHC Higgs Cross Sections: 3. Higgs Properties.” In: (2013). DOI: 10.5170/CERN-2013-004. arXiv: 1307.1347 [hep-ph] (cit. on p. 80).

[59] LHC Higgs Cross Section Working Group. 2016. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageAt13TeV (cit. on p. 80).

[60] 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] (cit. on p. 80).

[61] ATLAS Collaboration. "Search for the associated production of the Higgs boson with a top quark pair in multilepton final states with the ATLAS detector." In: *Phys. Lett.* B749 (2015), pp. 519–541. DOI: 10.1016/j.physletb.2015.07.079. arXiv: 1506.05988 [hep-ex] (cit. on p. 87).

[62] ATLAS Collaboration. "Study of the material of the ATLAS inner detector for Run 2 of the LHC." In: (2017). arXiv: 1707.02826 [hep-ex] (cit. on p. 97).

[63] ATLAS Collaboration. "Luminosity determination in pp collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector at the LHC." In: *Eur. Phys. J.* C76.12 (2016), p. 653. DOI: 10.1140/epjc/s10052-016-4466-1. arXiv: 1608.03953 [hep-ex] (cit. on p. 114).

[64] ATLAS Collaboration. "ttHFitter information." In: (2017). URL: https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/TtHFitter (cit. on p. 118).

[65] ATLAS Collaboration. "HistFitter information." In: (2017). URL: http://histfitter.web.cern.ch/histfitter/ (cit. on p. 118).

[66] ATLAS Collaboration. "Roofit information." In: (2017). URL: http://roofit.sourceforge.net/intro.html (cit. on p. 118).

[67] Glen Cowan, Kyle Cranmer, Eilam Gross, et al. "Asymptotic formulae for likelihood-based tests of new physics." In: *Eur. Phys. J.* C71 (2011). [Erratum: Eur. Phys. J.C73,2501(2013)], p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0, 10.1140/epjc/s10052-013-2501-z. arXiv: 1007.1727 [physics.data-an] (cit. on p. 119).

[68] Abraham Wald. "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large." In: *Transactions of the American Mathematical Society* 54.3 (1943), pp. 426–482. ISSN: 00029947. URL: http://www.jstor.org/stable/1990256 (cit. on p. 120).

Appendix

A The optimization of the cut-based electron identification in the ATLAS towards Run 2

A.1 Introduction and motivation

The following part will introduce a optimization study of the cut-based electron identification towards Run 2 condition at low transverse energy (5-20 GeV). The background at lower energy is much higher due to photons and light pions and this makes it necessary to optimize cut-based electron identification in this low energy region because of the new situation in Run 2 (higher centre mass of energy, more complex pile-up condition and etc.). Signal, background objects and Monte Carlo samples are presented in Section A.2. Section A.3 shows the performance of the LikelihoodPCA method while following sections are dedicated to the cut-based optimization, changes in the TRT and their effects on cut-based electron identification and the pile-up dependence. The last sections are the results and conclusions.

A.2 Monte Carlo samples

The process of optimization begins with monitoring of different signal and background signatures. The proper signal samples are selected to study "good" quality electrons, and a comparison is made with suitable background samples as well. Variables are ranked depending on their signal-background discrimination power. The signal MC samples used for the studies are $J/\psi \to ee$ and $Z \to ee$ events with only good electrons. A variety of $J/\psi \to ee$ with different truth p_T cuts are available for the studies. $J/\psi e3e3$, $J/\psi e3e8$ and $J/\psi e3e13$ have truth p_T cuts of 3 GeV on the first electron and 3 GeV, 8 GeV, and 13 GeV on the second electron respectively. The $Z \to ee$ has a dilepton filter which means that there is a p_T cut of 15 GeV on both electrons.

The minimum bias MC sample is used as background. The statistics in the p_T bin 15-20 GeV is very low for this sample. The JF17 sample is added in this bin to cover this issue. The JF17 sample has a truth p_T cut at 17 GeV and hence cannot be used for $p_T < 17$ GeV. Any true electrons in these background samples are removed for the optimization studies. To further select good quality electrons, a loose isolation cut is applied. This causes an efficiency changes about 5%.

The kinematics of the electron, p_T and η, determine the distributions of the variables included in the cut-based menu. This justifies a binning in those two kinematic variables. The menu would be more effective if this binning is finer. But a lack of statistics in finer bins introduces statistical artefacts and affects the quality of the optimization. This also makes the truth cuts in the MC samples more dangerous since the kinematics, and hence the variables themselves are biased. For example, as mentioned above, the $Z \rightarrow ee$ dilepton filter sample has truth p_T cuts at 15 GeV each. Thus, these cannot be used for bins with p_T less than 15 GeV. Beside the statistical reasons, the shapes of variables, like R_η, and the isolation variables, are significantly different at high and low p_T. This makes it necessary to optimize the low p_T range independently.

A.3 Performance of LikelihoodPCA

Performance of the LikelihoodPCA method with MC samples is presented in this section. LikelihoodPCA is selected due to its good performance as a multi-variate analysis method. Thus it can provide a baseline for the cut-based method. MVA LikelihoodPCA method is as a reference here and main goal is to check the input variables' performance and get a overview of MVA method, and one can also get a better understanding of cut-based method by comparing different MVA methods' performance.

Electron ID uses many discriminating variables to select good electrons generally. So the MVA can take advantage of the characteristics of these variables to employ them to distinguish between signal and background more efficiently. This test will be a cross check for the cut-based method and can be a good reference for choosing the working point at the same time.
Following are the variables which are used as input of the MVA (definitions are listed in Table 3.1):

- f_3, w_{stot}, $T_{ratio}(DEmaxs1)$, R_{Had}, R_η, R_ϕ, $W_{\eta 2}$, eOverP, $\Delta\Phi$.

Distributions of the variables are shown in Figure A.1.

According to Figure A.2, a high correlation can be seen between $DEmaxs1$ and w_{stot}, and this will be considered for forming groups of variables which are needed to be optimized independently during the optimization process. Figure A.3 shows clearly that the LikelihoodPCA can give a very good performance by using those variables.

A.4 Cut-based optimisation methodology

Preparation

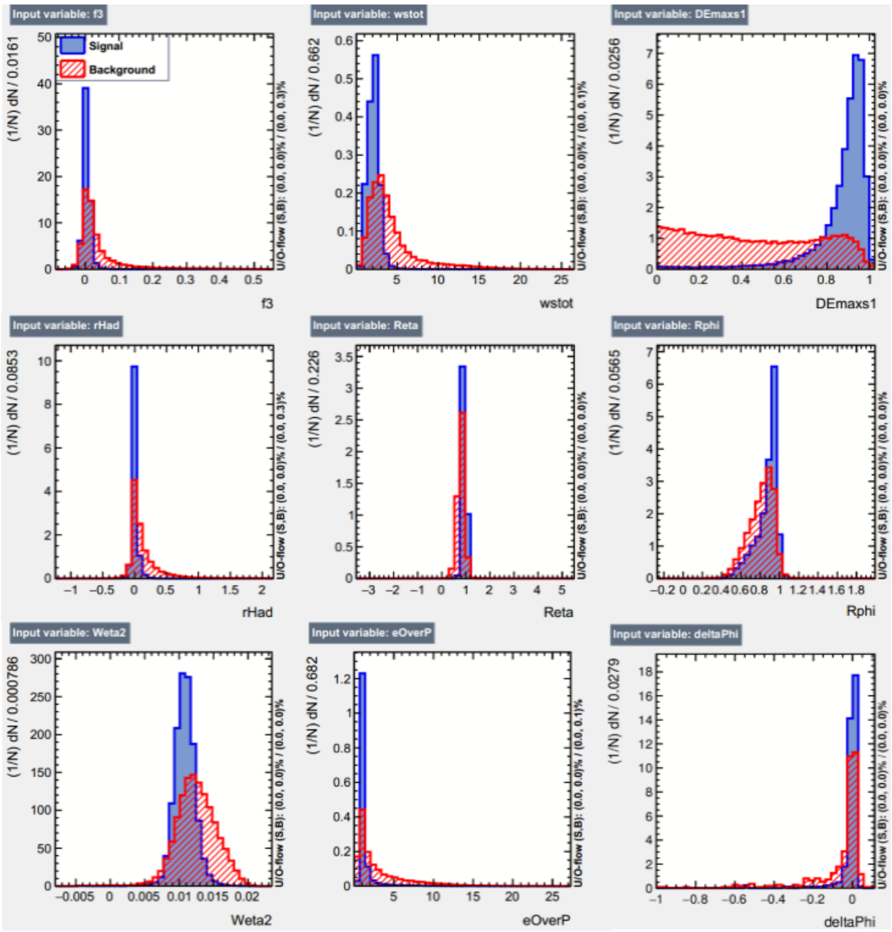
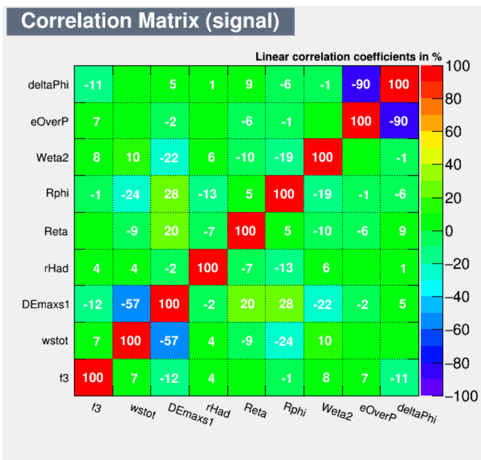
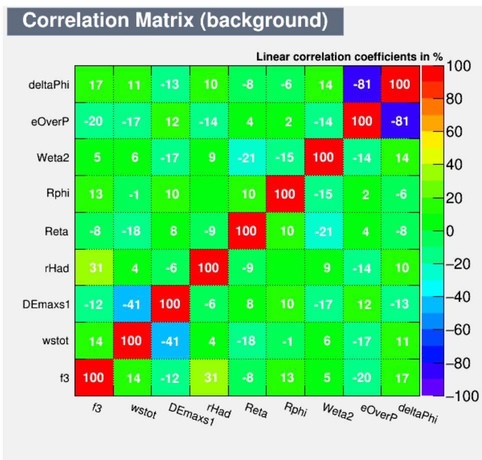


Figure A.1: Variables distribution for signal and background.



(a) Correlation for signal.



(b) Correlation background.

Figure A.2: Correlation Matrix of input variables for signal and background.

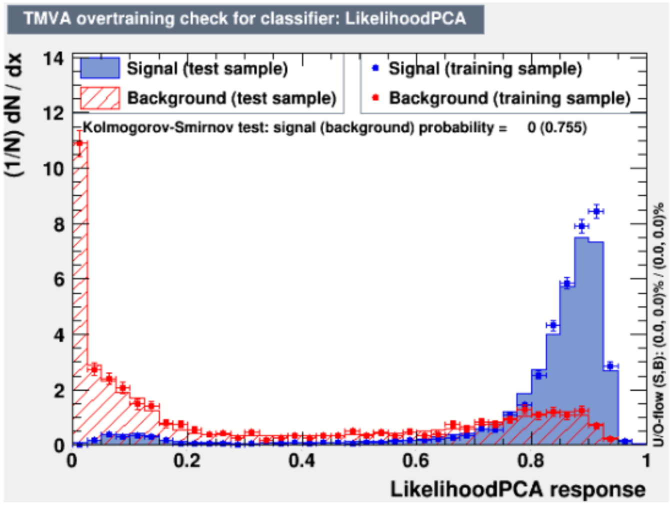


Figure A.3: LikelihoodPCA response.

Nine variables mentioned above are considered in the cut-based menu due to their separation power. Especially for the menus used in the cut-based electron identification, the cuts for different discriminate variables are processed in p_T and η bins. To simplify the optimization, the p_T and η binning used in this study is the same as the cut-based EID menus of 2012. So optimization is processed in ten η bins: 0-0.1, 0.1-0.6, 0.6-0.8, 0.8-1.15, 1.15-1.37 1.37-1.52, 1.52-1.81, 1.81-2.01, 2.01-2.47 and three p_T bins: 5-10 GeV, 10-15 GeV, and 15-20 GeV.

Results from the bin 20-30 GeV are used as a reference. One thing to mention, for the η bin 2.01-2.37 and 2.37-2.47 which used in the old menu, these two bins are merged into one due to the lack of statistics. Eventually, the same cut is used in both bins.

Pile-up dependence and pre-cleaning

When choosing the proper discriminate variables, one should consider not only the correlation between the variables (shown in Figure A.2), but also the pile-up dependence of each cut used in electron ID menu.

Because of the complex pile-up (μ) condition in Run 2 compared to Run 1 it's necessary to avoid the pile-up dependence of input variables which leads to the bias in the final results as function of pile-up. The check of the pile-up dependence for those input variables is processed and this can be studied in two ways:

- Applying only one cut, and measuring the slope of the efficiency vs μ curve.
- Applying all but one cut, measuring the efficiency, and taking the difference between the slopes of efficiency vs μ when all cuts are applied and when all cuts but one are applied.

Group1	$f_3, w_{stot}, T_{ratio}(DEmaxs1), R_{Had}$
Group2	$R_\eta, R_\phi, W_{\eta 2}$
Group3	$eOverP, \Delta\Phi$

Table A.1: The input variables of each group.

The above studies are performed in three p_T bins (5-10 GeV, 10-15 GeV, 15-20 GeV) of interest while also the bin (20-30 GeV) as reference. The results reflect the fact that different variables could be sensitive in difference p_T ranges. The variables deemed sensitive here are treated with extra care. Once the pile-up sensitive variables are studied, a pre-cleaning process is applied on some variables to avoid having cuts in the tails[a]. An attempt is made to remove a lot of the background events and remove as few of the signal events at the same time.

For each variable preliminary cuts are applied such that efficiency of the background decreases by 10% for each cut while keeping the decrease in signal efficiency less than 0.1%. After this is done, the global decrease in the signal efficiency is under $1\% - 2\%$. After applying the tail cuts, 10% background events are rejected and only 1% signal are lost totally.

Variables in groups

Concerning the variable correlations which are show in Figure A.2, the input variables are sorted into groups shown in Table A.1.

The variables used for the Loose, Medium, and Tight working points are the same as those used in the high p_T regime. The TRT PID variable, eProbabilityHT, is treated separately and will be mentioned later; this is justified by the fact that its linear correlation with other variables is negligible.

Efficiency target

The performance of the cut-based EID menus of 2012 is used as a benchmark here. For each p_T bin, the same signal efficiency as the 2012 menu is used for the three working points, Loose, Medium and Tight. In addition, a flatness in the efficiency as a function of η is forced by hand. Table A.2 shows the target signal efficiency (ϵ_0) used in the optimization process.

Optimization flow

Signal and background events are pre-selected firstly. Furthermore, all cuts from the 2012 menu are applied except the variables which need to be optimized (eg.,

[a]Cut values could jump a lot with negligible changes in efficiency if this is not taken care of as unexpected.

Menus	5-10(GeV)	10-15(GeV)	15-20(GeV)
Loose	0.88	0.88	0.88
Medium	0.75	0.75	0.80
Tight	0.60	0.65	0.75

Table A.2: Target efficiency for each menu in different p_T bins.

group 1 variables).A primary selection efficiency (ϵ_1) can be obtained after this step. Then those optimized variables will be the input variables for cut-based method. The CutBased method which is implemented in the TMVA toolkit is chosen for the optimization work. The general optimization work flow is shown in Figure A.4.

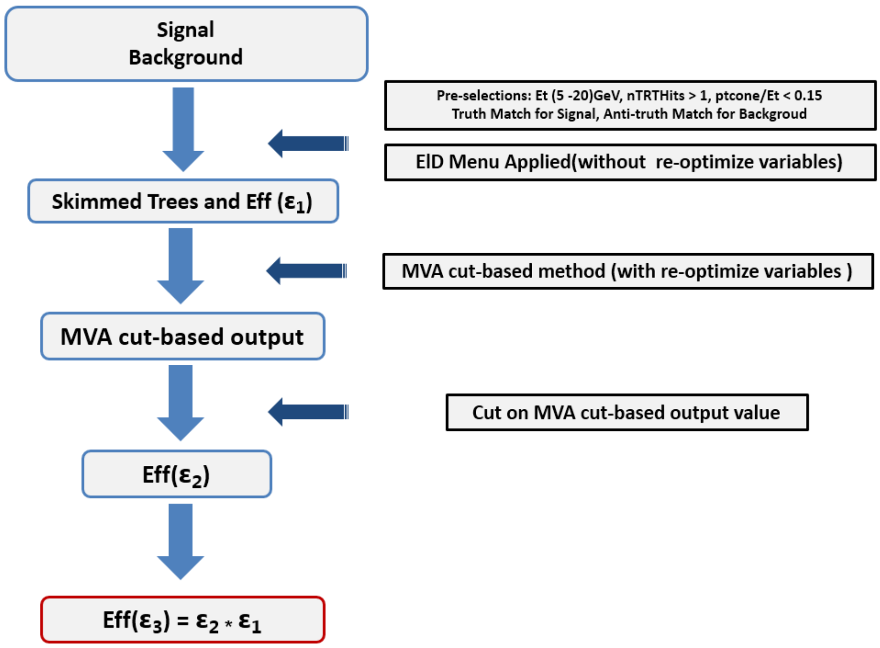


Figure A.4: Optimization work flow.

After MVA cut-based method training and testing, a new signal efficiency (ϵ_2) can be obtained. The total signal efficiency (ϵ_3) can be calculated with equation $\epsilon_3 = \epsilon_1 * \epsilon_2$ which is the target efficiency in the table above. The values of all cuts corresponding to the signal efficiency (ϵ_2) can be found in the MVA output.

The variables are grouped into three and the optimization chain can be explained like Group-1 $\rightarrow$ Group-2 $\rightarrow$ Group-3. It means that during the Group-2 optimization process, cut-based method will use the new first Group variables' cut value which are obtained from Group-1 process and so on. To give more optimization room for each group, a proper efficiency target tuning is needed for each

group. Signal efficiencies are set differently for Groups to leave a balanced optimization room. And after whole optimizing process, targeted signal efficiency will be achieved as expected. An example of Tight menu optimization which contains all three group variables:
1).Efficiency with pre-cleaning: ϵ_C.
2).G1 optimization: preparing samples with pre-cleaning and target efficiency: $\epsilon_{G1} = \epsilon_0 + 0.66 * (\epsilon_C - \epsilon_0)(\epsilon_0$ is the target efficiency above).
3).G2 optimization: preparing samples with pre-cleaning+G1 and target efficiency: $\epsilon_{G2} = \epsilon_0 + 0.33 * (\epsilon_C - \epsilon_{G1})$.
4).G3 optimization: preparing samples with pre-cleaning+G1+G2 and target efficiency: $\epsilon_{G3} = \epsilon_0$.

The CutBased method in TMVA is chosen for *Loose* and *Medium* menu optimization when doing the training and testing. The CutsGA method in TMVA is chosen for *Tight* menu as variables' cut value with two sides can be obtained. Variables' cut value corresponding to the signal efficiency will be chosen as new cuts value of the menu. Such as G1 variables' cuts value are the values corresponding to ϵ_{G1} in MVA cut-based method, etc.

Smoothing and monotonicity process

After inspecting the cut values in different η bins in any given p_T bin, it is found that some cut values changed a lot from one η-bin to the next. This is not desired, so a post-optimization cuts smoothing procedure is applied to avoid these jumps without loss in efficiency or the flatness. The procedure is as follows:

- Move cuts in small steps. The step value is defined using the maximum and minimum values of the cut among all η bins in a given p_T bin (call this C_{max} and C_{min}). The step size is then defined as $\frac{C_{max}-C_{min}}{200}$.
- This change in cut values is stopped as soon as the efficiency reduces by 1%.

But it is observed that this resulted in a global decrease of about 4%. To get around this, an extra condition is applied on the efficiency by forcing it to depend on the working point efficiency by requiring the following relation:

$$\epsilon_T = \epsilon_{T,MAX}\left(1 - C\epsilon_{WP}^2\right), \tag{A.1}$$

where $\epsilon_{T,MAX} = 0.01$ and $C = 0.05$. The results are as desired which are shown in Figure A.5.

Figures A.6, A.7, A.8 show the comparison of *Loose*, *Medium*, *Tight menu* with/without the cut smoothing. And also this process affects very little for the whole performance with many check tests.

Besides this cut smoothing process, inclusiveness of menus should be ensured which means *Loose* menu is looser than *Medium* and *Medium* is looser than *Tight*.

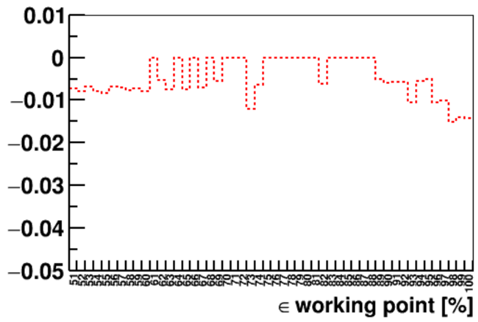


Figure A.5: Efficiency lost because of smoothing.

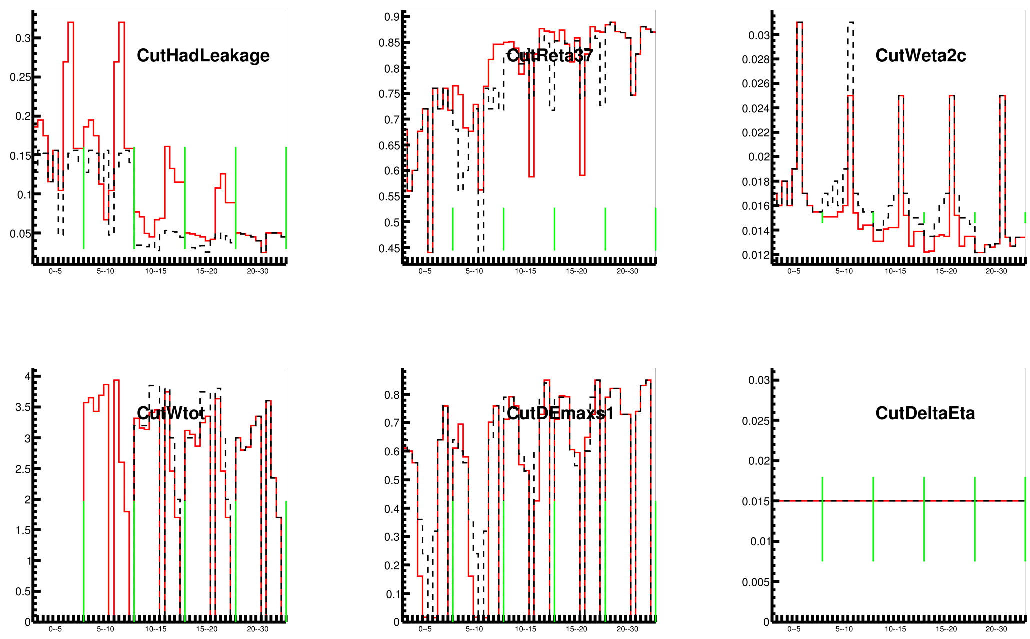


Figure A.6: Cuts value comparison of *Loose* menu with (Red)/without (Black) smoothing.

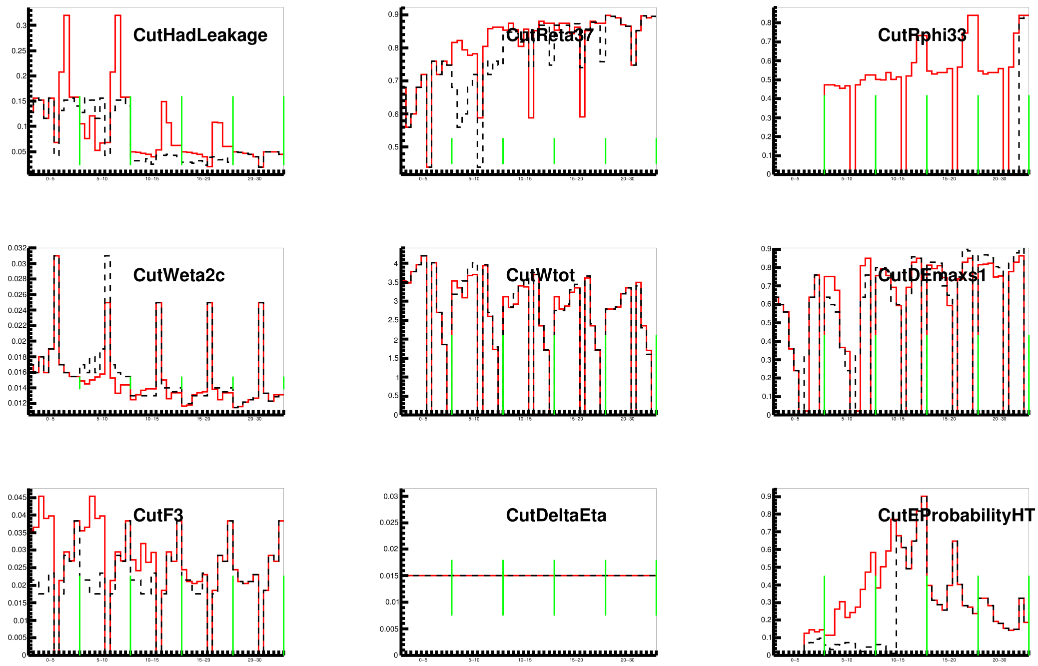


Figure A.7: Cuts value comparison of *Medium* menu with (Red)/without (Black) smoothing.

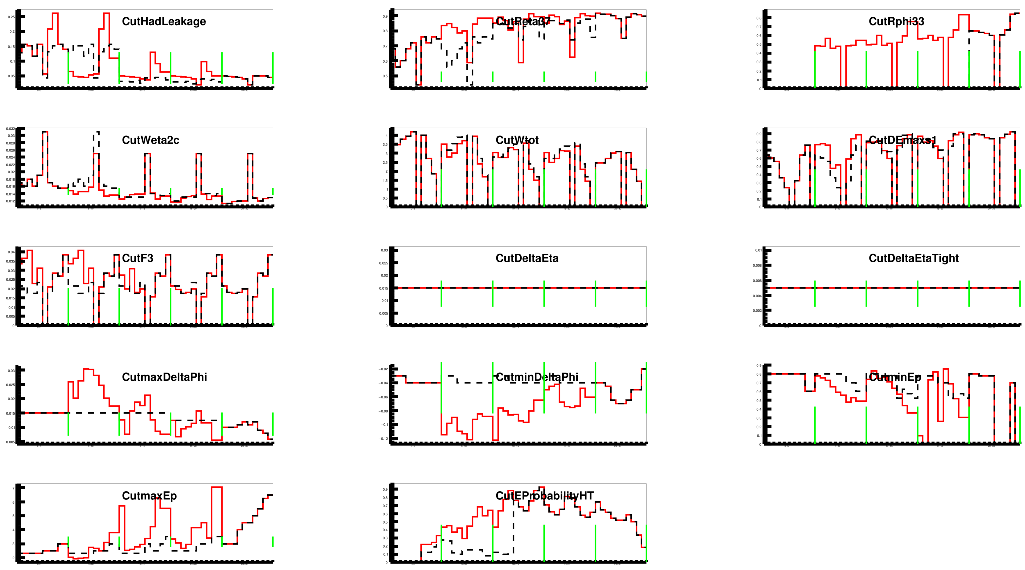


Figure A.8: Cuts value comparison of *Tight* menu with (Red)/without (Black) smoothing.

This is necessary after doing smooth process as this will affect final performance a lot. Monotonicity of cut values as a function of p_T is also taken into consideration.

A.5 Coping with the changes in the TRT

During Run 1, the gas in the TRT used to be only Xenon, which is very expensive. This was problematic because there were major leaks in the TRT. A cheaper solution was to use a Xenon-Argon mixture instead. This had major effects on electron identification because of the TRT-related variables. Two new scenarios called the Baseline scenario and the Pessimistic scenario are simulated based on the amount of Argon being used in the Xe-Ar mixture. The electron ID includes cuts on the total number of TRT hits, and the fraction of high threshold Xenon hits (defined as $f_{HT} = \frac{\text{HT Xe hits}}{\text{Total of TRT hits}}$). The comparison between different scenarios is shown in Figure A.9.

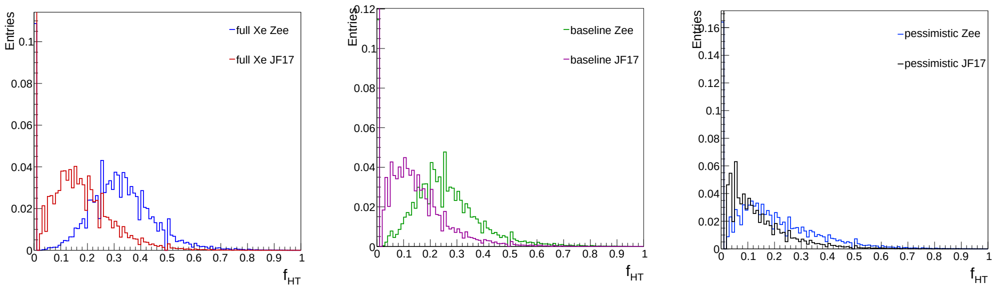


Figure A.9: The effect of changing the TRT gas on the fraction of the high threshold TRT hits (right: full Xe, middle: baseline, right: pessimistic).

It is clear that the discrimination power of the variable is significantly reduced in shifting from full Xe to the baseline Xe-Ar scenario or the pessimistic Xe-Ar scenario. To cope with this, a new variable, called eProbabilityHT, which also uses the high threshold hits information is introduced. The eProbabilityHT is a likelihood-type variable defined as,

$$p^{e,\pi} = \Pi p_{HT}^{e,\pi} \times \Pi \left(1 - p_{HT}^{e,\pi}\right), \tag{A.2}$$

$$p_{final}^{e,\pi} = \frac{p^{e,\pi}}{p^e + p^\pi}. \tag{A.3}$$

The main purpose of this variable is to distinguish electrons from hadrons, especially pions and hence, plays an important role in electron ID. The cut-based menu has to be re-tuned to replace f_{HT} by eProbabilityHT. To do this, it is important to understand the distribution for the signal and background. This is shown in Figure A.10 for the baseline and pessimistic scenarios. These are inclusive in p_T and η. The true electrons tend to have values closer to 1, which is clear in the plots.

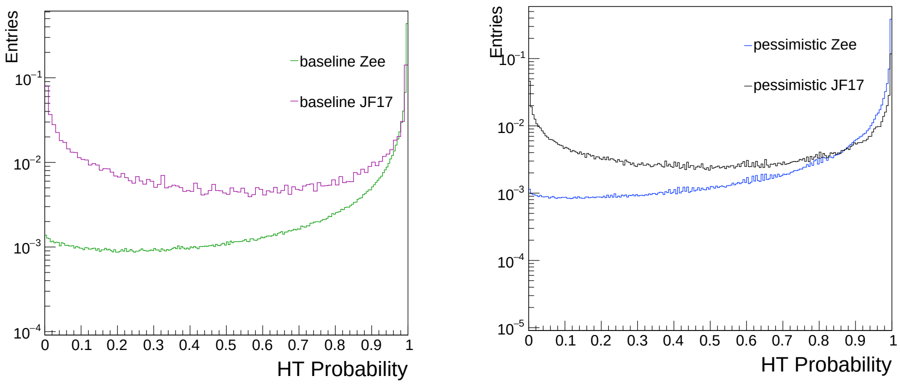


Figure A.10: Distribution of eProbabilityHT for the signal and background.

After checking the distribution, it is necessary to replace f_{HT} by eProbabilityHT. This is done in two steps:

- The effect of removing the f_{HT} cut on the efficiency is evaluated in all the p_T and η bins in which electron ID is optimized.
- The eProbabilityHT distributions in these bins are scanned and cuts are proposed to cause the same effect in the efficiency as f_{HT}.

The fact that the procedure works to give the same efficiencies is shown in Figure A.11.

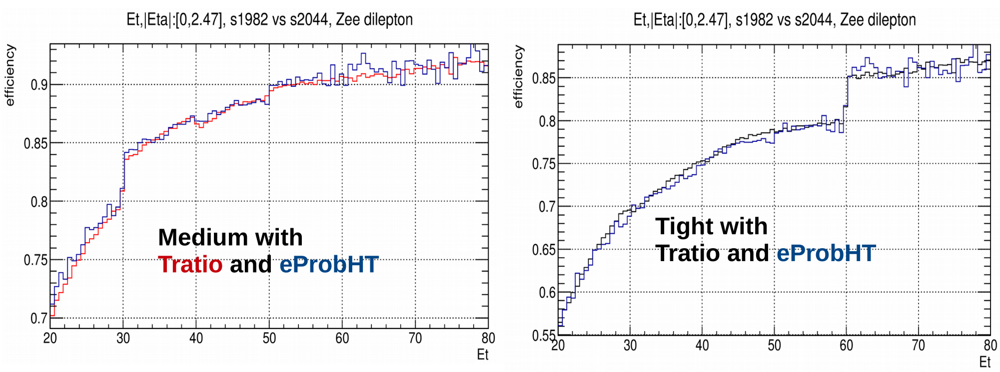


Figure A.11: T_{ratio} and eProbabilityHT comparison with different scenarios in signal events.

It should be noted that T_{ratio} is used with a sample in which the TRT only contained Xe whereas eProbabilityHT is used with the sample which contained the Xe-Ar mixture. This is distinguished by the s-tags; s1982 corresponds to the full Xe samples while s2044 corresponds to the baseline Xe-Ar mixture. A major feature of the TRT variables is that they are very pile-up sensitive and need to be treated with special care. A positive aspect of the swap of variables is that the efficiency is more stable with respect to μ and is shown in Figure **??**.

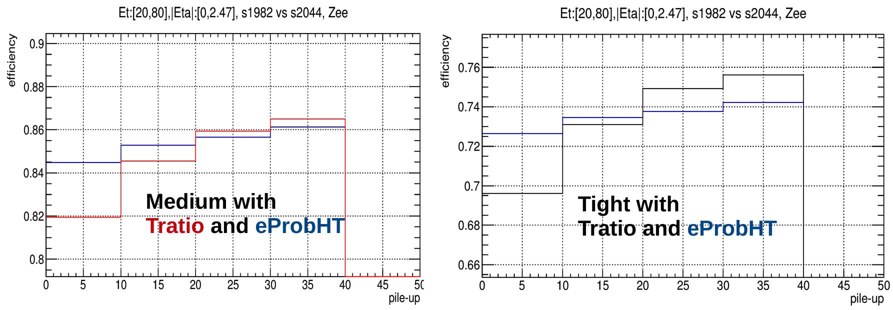


Figure A.12: T_{ratio} and eProbabilityHT comparison with different scenarios in signal events as function of pile-up.

A comparison of the performance of the menus is made between the baseline and pessimistic scenarios. The performance is similar, which is not expected given the difference in the eProbabilityHT distribution in the two scenarios. The following plots show the efficiencies as a function of the μ for signal and background for the two scenarios. A major feature of the TRT variables is that they are very pile-up sensitive and need to be treated with special care. A positive aspect of the swap of variables is that the efficiency is more stable with respect to μ. This is shown in Figure A.13.

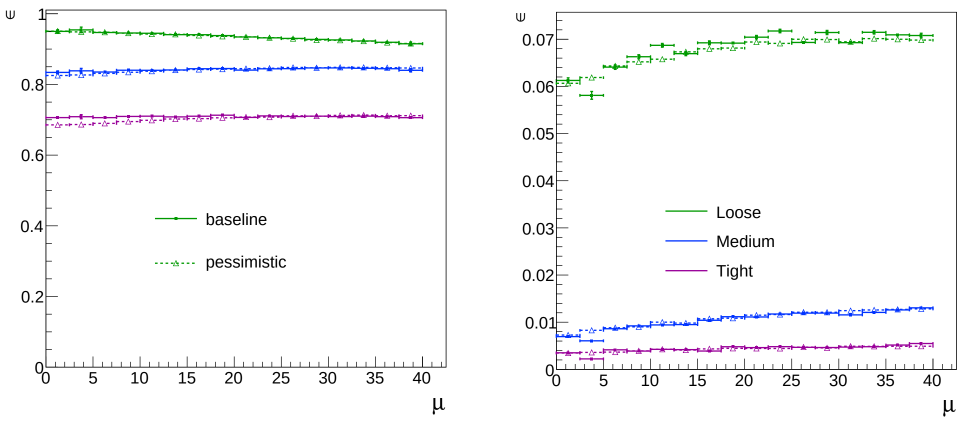


Figure A.13: $|\eta| < 2.47$ and $E_T > 5$ GeV: Efficiency vs μ for signal (left) and background (right).

The efficiencies of the eProbabilityHT cut alone in the $-2 < \eta < 2$ range for signal and background are shown in Figure A.14[b].

[b]The loose menu does not have the eProbabilityHT cut and the plots corresponding to the loose menu represent the complete loose menu applied.

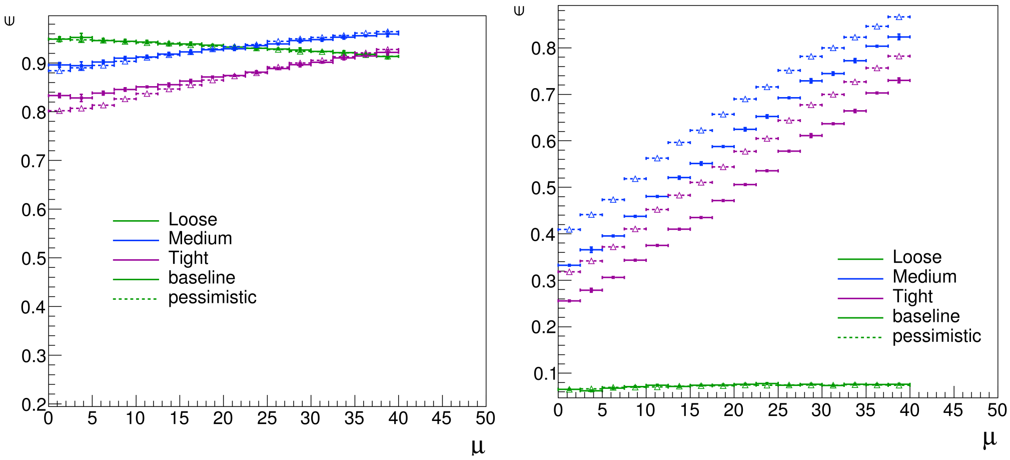


Figure A.14: $|\eta| < 2$: Efficiency vs μ for signal (left) and background (right).

The differences can be further attributed mostly to the $-1 < \eta < 1$ region as shown in Figure A.15.

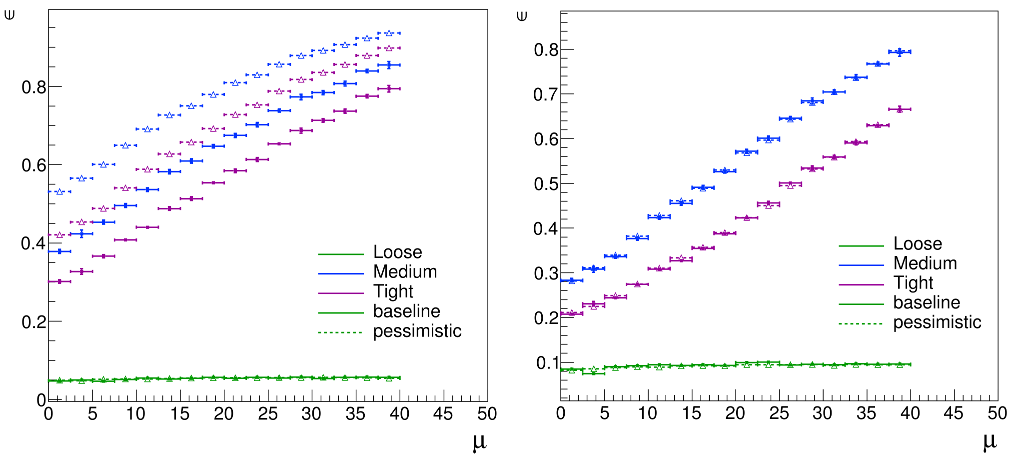


Figure A.15: $|\eta| < 1$ (left) and $|\eta| > 1$ (right): Efficiency vs μ for background.

The bin-wise cuts for eProbabilityHT and the efficiency comparison between the T_{Ratio} cut and the eProbabilityHT cut can be seen in Figure A.16.
The monotonicity that is enforced while tuning with f_{HT} is destroyed during this swap of variables. The modelling of eProbabilityHT has slightly changed. This causes effects on the pile-up sensitivities.

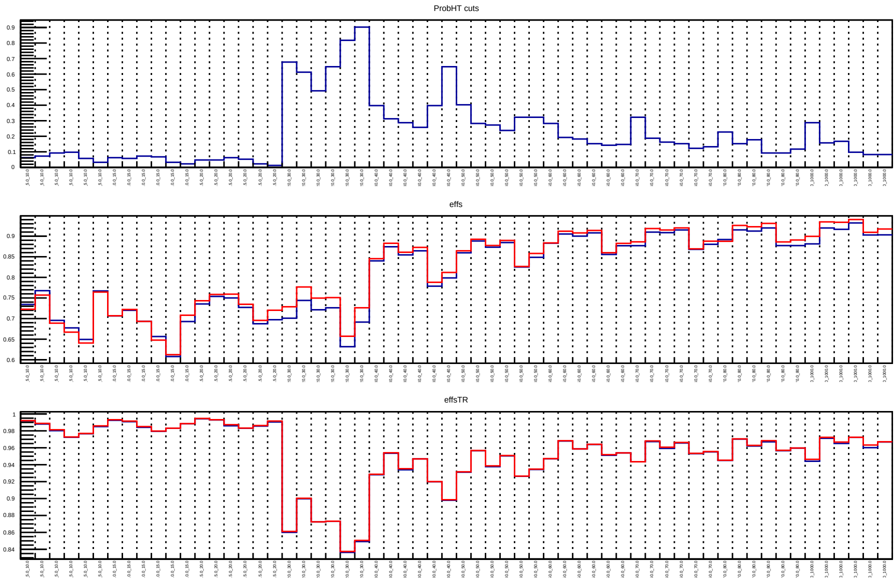


Figure A.16: Top: cut values for eProbabilityHT; middle: total efficiency comparison; bottom: comparison of efficiency of T_{Ratio} and eProbabilityHT cuts.

A.6 Results

Figure A.17 shows the efficiency comparison between cut-based EID menus of 2012 (EID 2012) (benchmark) and cut-based EID menus of 2015 (EID 2015) as function of η for the good electrons and background electrons. The re-optimised menu EID 2015 can have the good robustness with respect to the high pile-up situation. Figures A.18, A.19 and A.20 show efficiency comparison between EID 2012 and EID 2015 as function of pile-up for the background in full η region and $0 < p_T < 20$ GeV for *Loose*, *Medium* and *Tight* menu. Still EID 2015 menu can have higher background rejection and flat efficiency as function of pile-up compared to EID 2012.

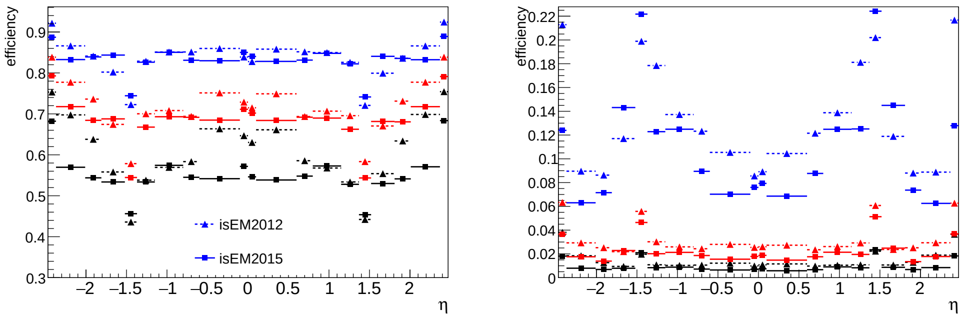


Figure A.17: Signal (left) and background (right): Efficiency vs η comparison between EID 2012 menu and new tuned EID 2015 menu.

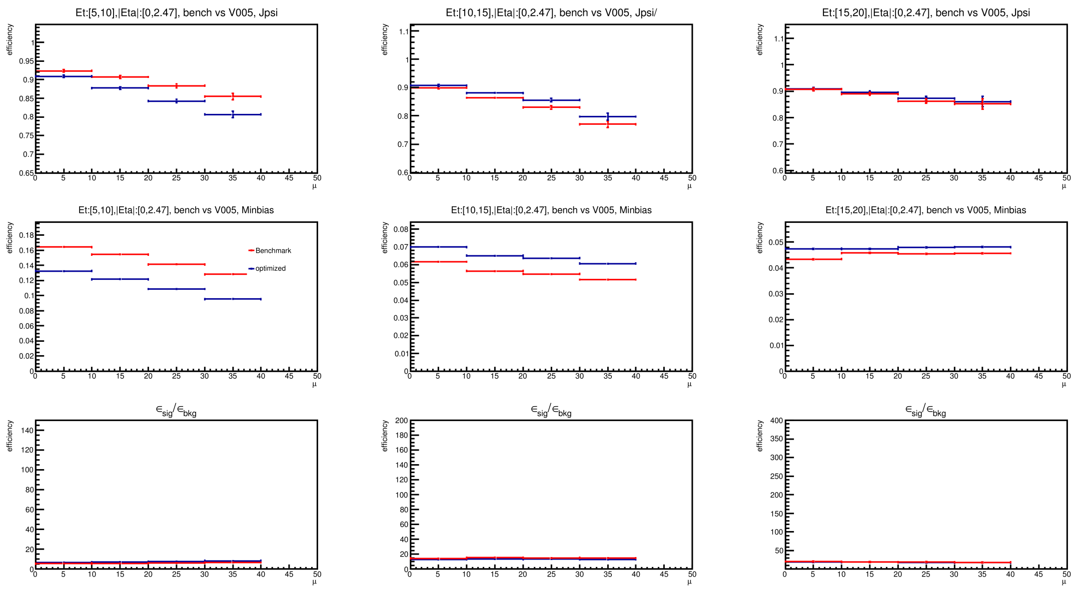


Figure A.18: Efficiency vs pile-up of *Loose* menu.

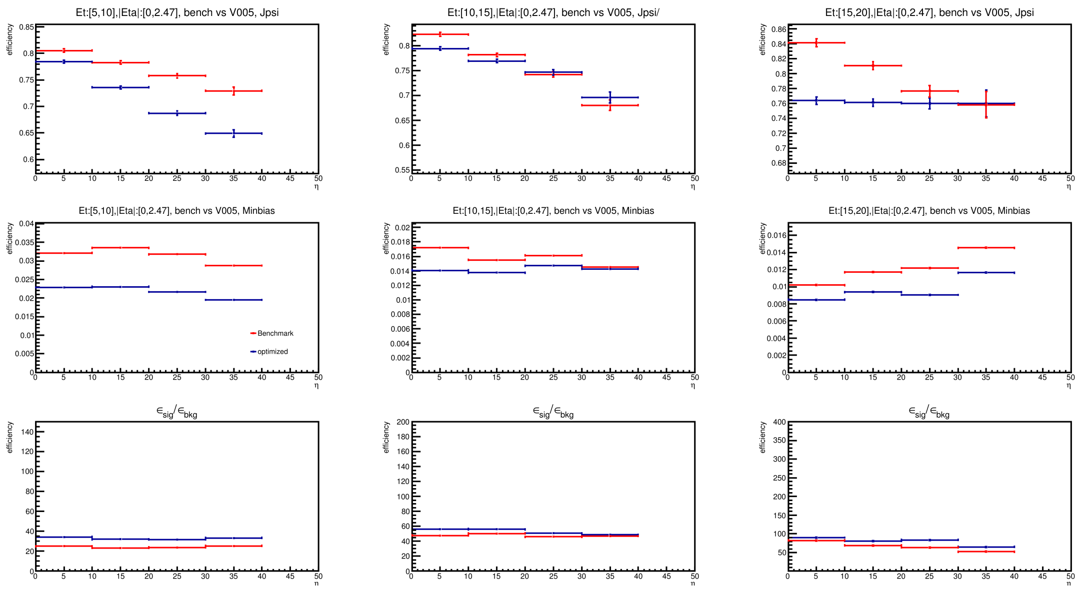


Figure A.19: Efficiency vs pile-up of *Medium* menu.

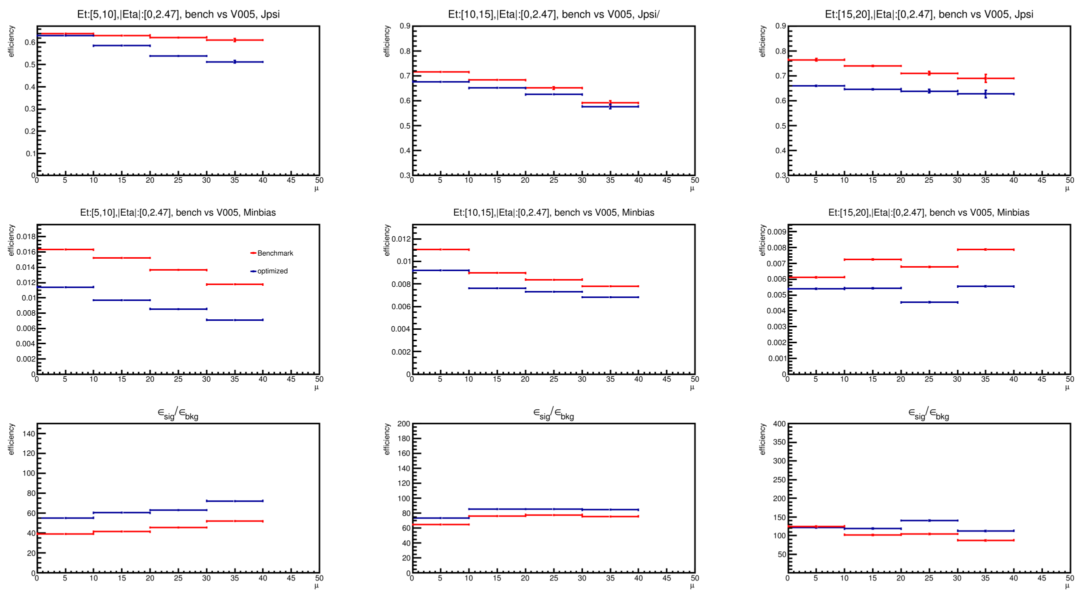


Figure A.20: Efficiency vs pile-up of *Tight* menu.

Figures A.21, A.22 and A.23 show the menu performance comparison between EID 2012 menu and new re-tuned menu for *Loose*, *Medium* and *Tight*. Overall, the re-tuned menu can have similar signal efficiency but with a better background rejection and keep a flatness of the efficiency as function of pile-up in the meanwhile.

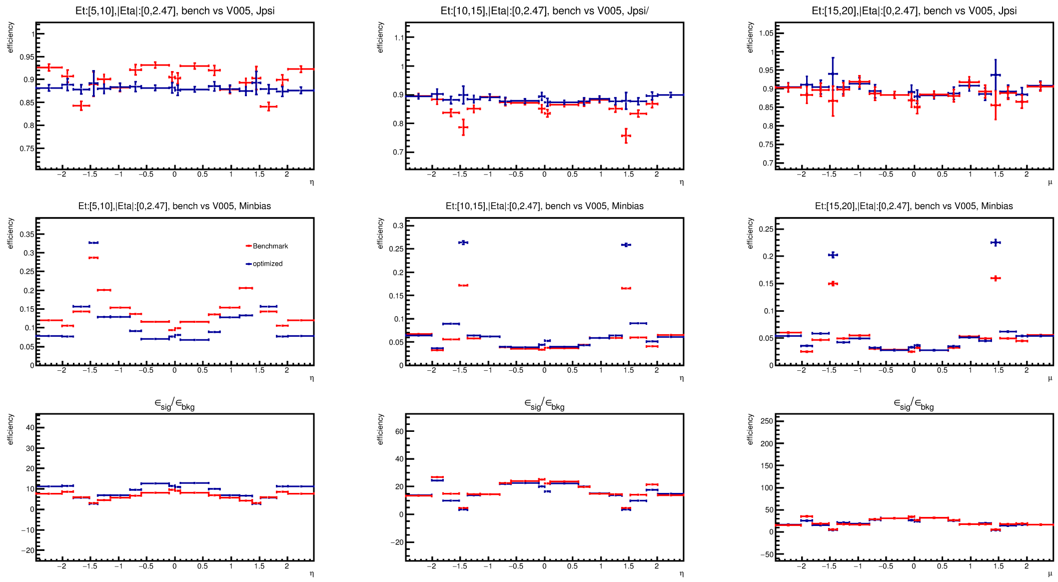


Figure A.21: *Loose* menu performance comparison between EID 2012 menu and new re-tuned menu.

Additional check on the efficiency as function of E_T is shown in Figure A.24. No strange things and obvious bias can be seen.

A.7 Conclusion

The new re-tuned menus towards Run 2 condition keep the good performance as 2012 menus towards Run 1 and have many improvements at the same time, such as flat signal efficiency in full eta range and dealing with more complex pile-up condition well, higher background rejection compared to the EID 2012 menu. The re-tuned cut-based electron identification menu achieves the goals as expected overall.

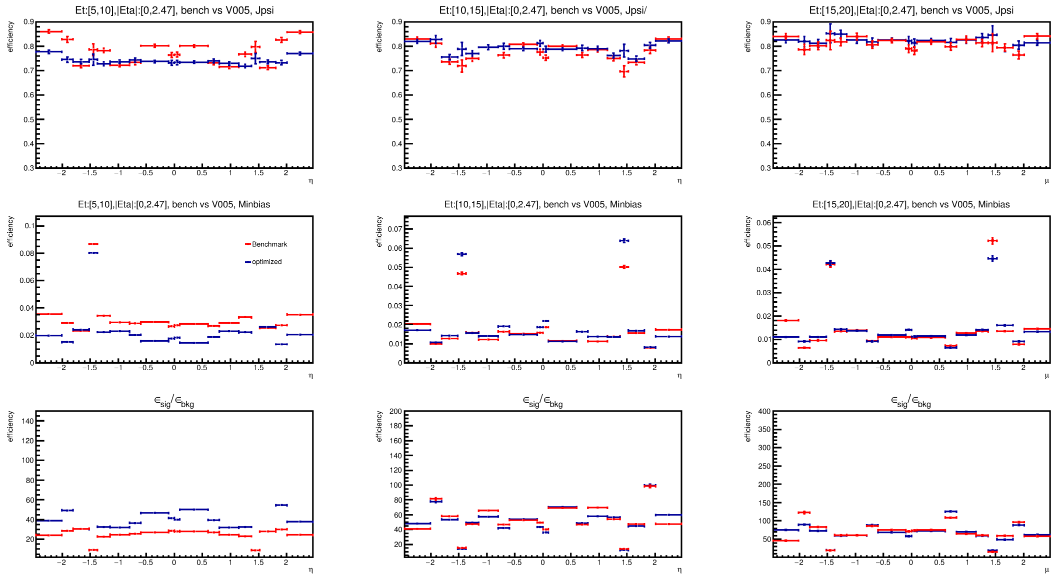


Figure A.22: *Medium* menu performance comparison between EID 2012 menu and new re-tuned menu.

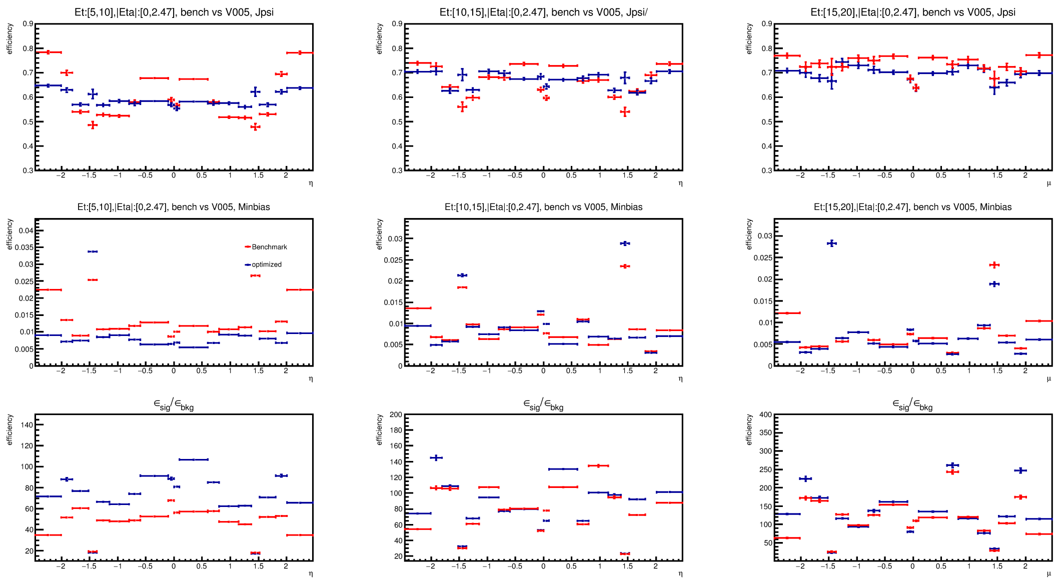


Figure A.23: *Tight* menu performance comparison between EID 2012 menu and new re-tuned menu.

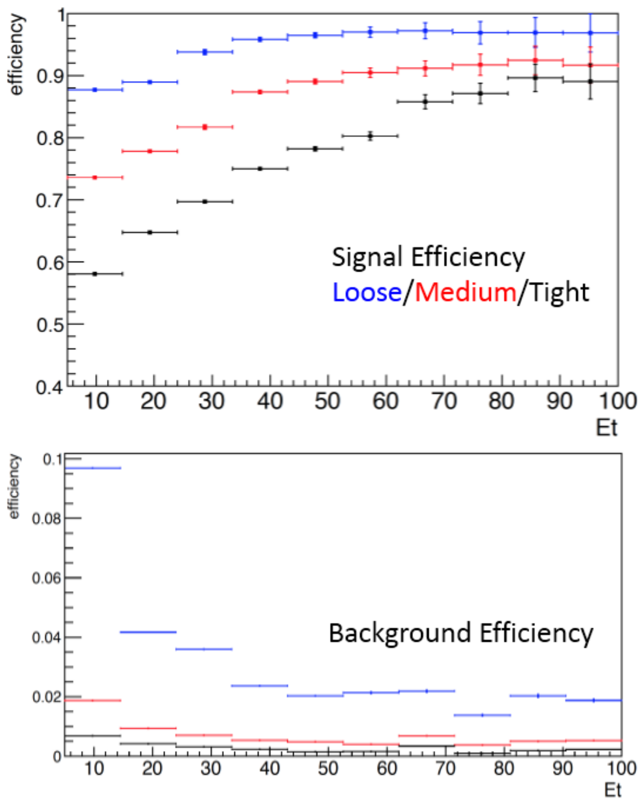


Figure A.24: Efficiency versus E_T for signal and background.

Abstract

English version

The production of the Higgs boson associated with a pair of top quarks is one of the most important Higgs boson productions still not observed by far. Therefore, its discovery is one of the most challenging searches after the Higgs discovery: not only it will be the first time we observe the existence of this Higgs production mode but moreover it will allow to measure the Yukawa coupling of top to the Higgs. This measurement can probe the basics of the Standard Model (SM) but also can search for any hints of new physics beyond the SM prediction. This makes this production search and measurement one of the most important analyses in post Higgs discovery era.

The ATLAS (A Toroidal LHC Apparatus) collaboration is operating since 2009 at the Large Hadron Collider (LHC), at CERN, in Geneva. During the first data taking campaign, so-called Run 1, and with corresponding to an integrated luminosities per experiment of 5 fb^{-1} at 7 TeV center of mass energy (in 2011) and 20 fb^{-1} at 8 TeV (in 2012), several Higgs boson decay channels were considered to measure the Higgs Yukawa coupling to the top quark within $t\bar{t}H$ production through multileptons final states. The ATLAS experiment produced individual combinations which were then converted into a combined significances with the CMS collaboration combinations to get a $t\bar{t}H$ observation significance, statistical of 4.4σ, for a 2.0σ expected from the Standard Model, corresponding to a measured excess of 2.3σ.

After a two years shutdown period, the LHC is now running at 13 TeV since 2015 and the cross section of $t\bar{t}H$ increases by a factor of four compared to the 8 TeV run. The two years upgrade of the ATLAS detector brought substantial improvements to the physics analysis, especially for the SM measurement such as $t\bar{t}H$ production and new physics searching.

After having introduced the theoretical and experimental framework, the work presented in this thesis focuses on the search for the Higgs boson production associated with a pair of top quarks in multilepton final states and more particularly those with three leptons. Multi-variable analysis techniques are introduced and employed to have a good discriminate power between signal and background. For one of the main backgrounds, fake leptons, a Matrix Method is used to estimate its contribution. Generally, good modelling of the fake leptons background and better separation are achieved with a Boosted Decision Tree (BDT) distribution. Final results are obtained with a fit on a combined 2D BDT distribution.

For a Higgs boson mass of 125 GeV with 36.1 fb^{-1} at 13 TeV data recorded by ATLAS, an excess of events over the expected background from other Standard Model processes is found with an observed significance of 2.18 standard deviations, compared to an expectation of 1.45 standard deviations. The best fit for the $t\bar{t}H$ production cross section is $1.54^{+0.81}_{-0.74}$ times the SM expectation, and is consistent with the value of the Yukawa coupling to top quarks in the Standard Model. The results are compatible with the latest ATLAS approved results obtained with a further optimized background treatment.

Comparing to the Run 1 results, a great improvement has been obtained on the significance and the uncertainty of the signal strength measurement. This results from a higher centre of mass energy hence a higher cross section, but also from the ATLAS performance upgrades as well as a more advanced analysis strategy. Still large uncertainties on the final measurements remains and further optimization of the analysis and higher statistics will be needed to get a more precise results in the future.

Français version

La production du boson de Higgs associée à une paire de quarks top est une des plus importantes productions non encore observée à ce jour. Sa découverte est donc une des recherches les plus ambitieuses après la découverte du Higgs : non seulement cela sera la première fois que ce mode sera observé mais surtout il permettra la mesure du couplage de Yukawa du Higgs avec le quark top. Le résultat de cette mesure pourra sonder les bases du Modèle Standard (SM) mais aussi rechercher des traces de nouvelle physique au-delà du Modèle Standard. Cela fait de cette recherche et de la mesure de cette production une des analyses les plus importantes depuis la découverte du Higgs. La collaboration ATLAS (A Toroidal LHC Apparatus) opère depuis 2009 le détecteur généraliste ATLAS auprès du grand collisionneur de Hadron (LHC), à Genève. Durant la première campagne de prise de données appelée Run 1 et correspondant à une luminosité intégrée de 5 fb^{-1} à 7 TeV d'énergie dans le centre de masse (2011) et de 20 fb^{-1} à 8 TeV (2012), plusieurs canaux de désintégration du boson de Higgs ont été considérés pour mesurer le couplage de Yukawa du Higgs avec le quark top dans les productions $t\bar{t}H$ avec des états finaux multi-leptoniques. L'expérience ATLAS a ainsi produit une mesure combinée qui a ensuite été associée avec la mesure combinée obtenue par l'expérience CMS pour produire une significance globale observée de 4.4σ, pour 2.0σ prédite correspondant à un excès mesuré de 2.3σ.

Après une période d'arrêt de deux ans, le LHC tourne maintenant à 13 TeV depuis 2015 et la section efficace $t\bar{t}H$ a augmentée d'un facteur quatre par rapport au run à 8 TeV permettant des mesures plus précises voire des découvertes. Les deux années de mise à niveau du détecteur ATLAS ont apportées des améliorations substantielles pour les analyses de physique, et plus particulièrement pour les mesures du SM comme la production $t\bar{t}H$ et la recherche de nouvelle physique.

Après avoir introduit le cadre théorique et expérimental, le travail présenté dans cette thèse se concentre ensuite sur la recherche et l'étude de la production du boson de Higgs associée à une paire de quarks top dans des états finaux semi-leptoniques et plus particulièrement ceux à trois leptons. Des techniques d'analyse multi-variables sont introduites et employées pour obtenir une bonne discrimination entre signal et bruit de fond. Pour un des principaux bruits de fond, celui de faux leptons, une méthode matricielle est présentée et utilisée pour estimer sa contribution. Une bonne modélisation du bruit de fond de faux leptons et une meilleure séparation du signal est obtenue avec une distribution d'arbre de décision stimulé (BDT). Les résultats finaux sont obtenus par un ajustement sur les distributions BDT combinées à deux dimensions.

Pour un boson de Higgs de 125 GeV de masse, un excès d'événements par rapport au bruit de fond attendu des autres processus du Modèle Standard est obtenu avec une significance observée de 2.18 déviations standard, à comparer à une estimation attendue de 1.45σ. Le meilleur ajustement pour la section efficace de production $t\bar{t}H$ est de $1.54^{+0.81}_{-0.74}$ fois l'estimation du SM et est en accord avec la valeur du couplage de Yukawa du Higgs au quark top dans le Modèle Standard. Ces résultats sont compatibles avec les derniers résultats officiels d'ATLAS obtenues avec une optimisation plus poussée du traitement des bruits de fond.

En comparant avec les résultats du Run 1, une amélioration importante a été obtenue sur la significance et l'incertitude sur la mesure de la force du signal. Cela provient non seulement de l'augmentation de l'énergie de collision et donc de la section efficace mais aussi de performances améliorées du détecteur ainsi que d'une stratégie d'analyse optimisée. Néanmoins, les incertitudes sur la mesure finale demeurent importantes et une optimisation plus poussée avec une augmentation de la statistique sera nécessaire pour obtenir un résultat plus précis dans le futur.