AIX-MARSEILLE UNIVERSITÉ
Ecole Doctorale 352 : Physique et Sciences de la Matière
Faculté Des Sciences de Luminy
Centre de Physique des Particules de Marseille

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Physique et Sciences de la Matière
Spécialité : Physiques des Particules et Astroparticules

Thomas CALVET

Search for the production of a Higgs boson in association with top quarks and decaying into a b-quark pair and b-jet identification with the ATLAS experiment at LHC

Thèse soutenue le 08/11/2017 devant le jury :

Elizaveta SHABALINA	University of Göttingen, Germany	Rapporteur
Yves SIROIS	LLR, Ecole Polytechnique Palaiseau, France	Rapporteur
Cristinel DIACONU	CPPM, Marseille, France	Examinateur
Anne-Catherine LE BIHAN	IPHC, Strasbourg, France	Examinateur
Bruno MANSOULIÉ	CEA Saclay, France	Examinateur
Arnaud DUPERRIN	CPPM, Marseille, France	Directeur de thèse
Georges AAD	CPPM, Marseille, France	Co-directeur de thèse

Remerciements

Je n'aurais pu achever cette thèse sans la présence de tout ceux qui m'ont entouré, tant dans le cadre professionnel que personnel. Ces lignes leur sont dédiées.

En tout premier lieu, je tiens à adresser mes remerciements à quiconque a, ou va essayer de lire ce manuscrit. Bonne lecture et bonne chance !

Une thèse ne commence, ni ne finit, sans la présence et le soutien de superviseurs. Je tiens à exprimer mes plus sincères remerciements à Georges Aad et Arnaud Duperrin, pour m'avoir accordé leur confiance et donné l'opportunité de travailler sur ce projet. C'est à leurs savoirs que je dois le mien, et c'est grâce à leurs encouragements et leurs conseils (parfois même tard dans la nuit) que j'ai pu en faire quelque chose. Je les remercie aussi d'avoir eu l'œil sur mon avenir, et de m'avoir poussé à construire un projet sur le long terme. J'espère avoir l'occasion de travailler avec vous à nouveau. Milles fois merci.

Je tiens également à donner une place aux rapporteurs de ce manuscrit, Yves Sirois et Elizaveta Shabalina, ainsi qu'aux membres de mon jury, Christinel Diaconu, Anne-Catherine Le Bihan et Bruno Mansoulié. Je vous remercie de m'avoir lu et écouté. Je vous remercie également d'avoir pris le temps de proposer des corrections pour ce manuscrit, et ainsi me permettre de l'améliorer.

Je suis extrêmement heureux d'avoir pu travailler au Centre de Physique des Particules de Marseille. Il n'est pas si courant de trouver un lieu où on est content d'aller travailler; c'était le cas de ce laboratoire. Je suis d'autant plus heureux d'avoir été un membre du groupe ATLAS, dont les membres sont trop nombreux pour les citer tous ici; j'ai apprécié travailler mais aussi discuter avec vous. Je remercie en particulier Yann Coadou, collaborateur pour le b-tagging et l'analyse $t\bar{t}H(H \to b\bar{b})$, grand connaisseur des BDT (et qu'est ce qu'on en utilise ...), merci pour avoir contribué à cette thèse. Je remercie également Emmanuel Le Guirriec pour son aide et son expertise dans l'informatique et tous les aspect techniques. Enfin, je remercie Laurent Vacavant, notre chef de groupe. Merci pour son soutien, ses conseils avisés et d'avoir gardé un œil sur mon travail. Bon vent !

Evidemment, dans un laboratoire il y a aussi les étudiants. Un grand merci à tous les doctorants du CPPM, collègues et amis. A Venu, Yulia et Asma, qui ont commencé en même temps que moi et avec qui j'ai partagé cette aventure. A Kazuya et Royer, les ainés, pour votre aide et les verres partagés. Et puis à Rima et Robert, qui ont partagé mon bureau ces deux dernières années.

Une thèse est un projet passionnant mais représente aussi beaucoup de travail et un nombre insolent d'heures passées au laboratoire; je n'aurais pas réussi sans la présence de ceux qui m'ont entouré. Je remercie toute ma famille, et en particulier ma mère, Florence, et mon grand-père, Philippe, pour m'avoir soutenu et encouragé à suivre cette voie, et ce depuis bien avant la thèse, je n'en serais pas là sans vous. Je remercie également tous mes amis, la famille "yo", le trio Elise, Christian et Mathieu, et bien d'autres encore, pour m'avoir changé les idées et donné un second souffle quand la fatigue se faisait sentir. Un mot tout particulier pour mes deux chats, Tapioca et Tequila, dont les frimousses ont animé certaines de mes présentations.

Les derniers mots sont réservés à Julie avec qui je partage ma vie. Je n'aurais pas pu aller aussi

loin sans son aide et son soutien (heureusement que nous n'avons pas eu à finir nos thèses en même temps). Je la remercie d'en avoir tant fait pour moi, de m'avoir remonté le moral quand j'étais perdu entre des milliers de lignes, de s'être occupé de moi quand j'en avais besoin, et surtout, je te remercie d'être là.

Abstract

Keywords : LHC, ATLAS, Higgs boson, $t\bar{t}H$, $H \to b\bar{b}$, b-tagging, statistical analysis of data

En Juillet 2012, les expériences ATLAS et CMS ont annoncé la découverte d'une nouvelle particule de masse 125 GeV compatible avec le boson de Higgs prédit par le Modèle Standard. Cependant pour établir la nature de ce boson de Higgs et la comparer aux prédictions du Modèle Standard de la physique des particules, il est nécessaire de mesurer le complage du boson de Higgs au fermions. En particulier le quark top possède le plus fort couplage de Yukawa avec le boson de Higgs. Ce couplage est accessible par le processus de production d'un boson de Higgs en association avec une paire de quarks tops ($t\bar{t}H$). Pour le Run 2 du LHC de nombreuses améliorations sont réalisés et ouvrent l'accès au cannal $t\bar{t}H$: augmentation de l'énergie au centre de mass à 13 TeV, augmentation de la luminosité intégrée à 36.1 fb^{-1}en 2016, améliorations du detecteur avec en particulier l'IBL dont l'impact sur le b-tagging est important. Cette thèse présente la recherche d'événement $t\bar{t}H$ où le boson de Higgs se désintègre en deux quarks b dans les données du Run 2 recueillies en 2015 et 2016 par le détecteur ATLAS. La composition du bruit de fond ainsi que la mesure du signal $t\bar{t}H$ dans les données sont obtenues à partir d'un ajustemement statistique des prédictions aux données. Une attention particulère est portée au bruit de fond $t\bar{t}+$ jets dont originent les plus grandes sources d'incertitudes sur le signal

L'étiquetage des jets issus de quarks b, appelé b-tagging, est primordiale pour l'analyse $t\bar{t}H(H \to b\bar{b})$ dont l'état final contient quatre quarks b. Afin d'améliorer la compréhension des algorithmes de b-tagging pour le Run 2, la définition des jets de saveur b dans les simulations Monte Carlo est revisitée. Les algorithmes standards du b-tagging ne permettant pas la différenciation des jets contenant un ou deux quarks b, une methode spécifique à été développée et est présentée dans cette thèse.

In July 2012, the ATLAS and CMS experiments announced the discovery of a new particle, with a mass about 125 GeV, compatible with the Standard Model Higgs boson. In order to assess if the observed particle is the one predicted by the Standard Model, the couplings if this Higgs boson to fermions have to be measured. In particular, the top quark has the strongest Yukawa coupling to the Higgs boson. The associated production of a Higgs boson with a pair of top quarks ($t\bar{t}H$) gives a direct access to this coupling. The $t\bar{t}H$ process is accessible for the first time in the Run 2 of the LHC thanks to an upgrade of the detector, especially the IBL which improves b-tagging, and the increase of the center of mass energy to 13 TeVand of the integrated luminosity to 36.1 fb^{-1}in 2016. This thesis presents the search for $t\bar{t}H$ events with the Higgs boson decaying to a pair of b-quarks using data collected by the ATLAS detector in 2015 and 2016. The description of the background and the extraction of the $t\bar{t}H$ signal in data are obtained by a statistical matching on predictions to data. In particular the $t\bar{t}+$ jets background is the main limitation to signal sensitivity and is scrutinized.

The identification of jets originating from b-quarks, called b-tagging, is a vital input to the search of $t\bar{t}H(H \to b\bar{b})$ events because of the four b-quarks in the final state. For Run 2 the definition of b-flavoured-jets in Monte Carlo simulations is revisited to improve the understanding of b-tagging algorithms and their performance. Standard b-tagging algorithms do not separate jets originating from a single b-quark from those originating from two b-quarks. Thus a specific method has been developed and is reviewed in this thesis.

Synthèse en français

Introduction

La physique des particules naquit au début du $20^{\text{ème}}$ siècle, expérimentalement par l'observation de rayons cosmiques, et théoriquement avec l'apparition de la Mécanique Quantique. L'effort conjoint des théoriciens ainsi que des expérimentalistes jusqu'à aujourd'hui a permis de révéler l'existence d'un nombre réduit de particules fondamentales et de décrire leurs interactions. Toutes ces connaissances sont rassemblées dans le Modèle Standard (SM) de la physique des particules décrit au chapitre 1. L'un des piliers de cette théorie est le méchanisme de Brout-Englert-Higgs (BEH), également présenté au chapitre 1. Introduit en 1964, il permet d'inclure la masse des particules dans le SM et prédit une nouvelle particule, le boson de Higgs.

La recherche de nouvelle particules fondamentales et les hautes précisions nécessaires aux mesures en physiques des particules requièrent des moyens instrumentaux de hautes performances. Le grand collisioneur de hadron (LHC) est l'accélérateur de particules le plus avancé à ce jour. A ses paramètres de fonctionnement pour le Run 2, commençant en 2015 et toujours en cours lors de la rédaction de ce document en 2017, il permet la collision de deux paquets de $\sim 10^{11}$ protons toutes les 25 ns (50 ns en 2015) correspondant à une énergie au centre de masse de 13 TeV. Plusieurs détecteurs sont installés sur l'anneau de 26.7 km que forme le LHC. Les données sur lesquelles s'appuient les études de cette thèse ont été receuillies par le detecteur ATLAS. Ce gigantesque détecteur cylindrique (44 m de long pour 25 m de diamètre) est conçu pour detecter un maximum de particules afin de satisfaire un vaste programme de recherche, allant des mesures de précision du SM à la recherche de nouvelle physique, et en passant par la recherche du boson de Higgs. Grâce à son excellent fonctionnement au cours du Run 2, il a permit de collecter un grande quantité de données. Pour le travail de cette thèse les données receuillies en 2015 et 2016, correspondant à une luminosité intégrée de 36.1 fb^{-1}, sont utilisées. Le chapitre 2 décrit en détail le LHC et le detecteur ATLAS.

Ce n'est qu'en 2012 (48 ans après sa prédiction) que les expériences ATLAS et CMS au LHC annoncent la découverte d'une particule compatible avec le boson de Higgs du SM. Cependant, plusieurs propriétés clés de la particule observée, telles que les couplages aux quarks lourds de troisième famille (quark tops et b), ne sont pas encore mesurées. La mesure de ces proprétés est nécessaire pour établir l'appartenance du boson de Higgs observé au SM.

Le quark top est la particule connue de plus haute masse, donc de plus fort couplage de Yukawa du boson de Higgs. La production associé du boson de Higgs avec une paire de quark top ($t\bar{t}H$) est le processus le plus favorable à la mesure directe du couplage de Yukawa au quark top au LHC. Cependant, aucune évidence de l'existence de ce cannal n'est trouvée lorsque cette thèse commence. Ce manuscrit présente la recherche d'évènements de production $t\bar{t}H(H \to b\bar{b})$, où le boson de Higgs se désintègre en une paire de quarks b, dont le diagramme de Feynman est présenté figure 0.1. Cette analyse est en particulier limitée par le bruit de fond $t\bar{t}+$ jets (principalement la composante $t\bar{t}+b\bar{b}$ dont les deux quarks additionnels sont des quarks b), dont le diagramme de Feynman est présenté figure 0.1. La séparation de ce bruit de fond du signal, son modèle systematique, et sa description dans les données sont des enjeux majeurs discutés dans ce document.

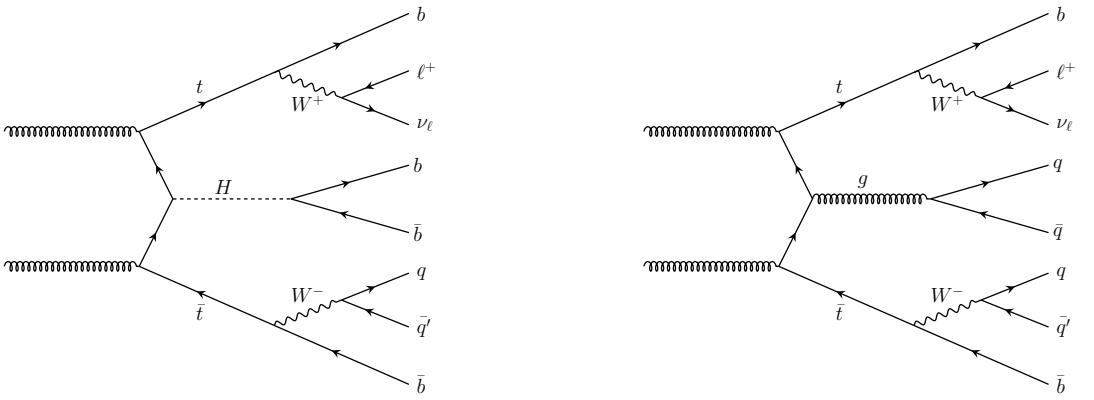


Figure 0.1.: Diagrammes de Feynman des processus (gauche) $t\bar{t}H(H \to b\bar{b})$ et (droite) $t\bar{t}+$ jets.

L'étiquetage des jets issus de quarks b, appelé b-tagging, est un ingrédient majeur pour une large fraction du spectre de recherche dans l'expérience ATLAS. Le b-tagging est en particulier primordial à l'analyse $t\bar{t}H(H \to b\bar{b})$ qui possède quatre quarks b dans son état final. Cette thèse résume le travail effectué pour une meilleure compréhension des performances des algorithmes de b-tagging, ainsi que pour l'amélioration de l'étiquetage des jets issue de la désintégration d'un gluon en deux quarks b.

Etiquetage des jets de saveur b

La recherche d'événements de production $t\bar{t}H(H \to b\bar{b})$ repose grandement sur la capacité à identifier les jets issus des quatre quarks b, les quatre b-jets, dans l'état final. L'étiquetage des b-jets, le b-tagging, présenté au chapitre 3, a pour but de différencier les b-jets des c-jets et $light$-jets (jets issus de quarks c ou de gluons et quarks légers, uds, respectivement). Les algorithmes de b-tagging se basent principalement sur le relativement long temps de vie, de l'ordre de 1.5 ps, des hadrons contenants un quark b, les b-hadrons, par rapport aux autres saveurs de hadrons. Ce long temps de vie se traduit par une grande distance entre le vertex primaire, le PV, (point de collision des protons) et le vertex secondaire de désintégration du b-hadron, le SV. Typiquement, la distance (PS,SV) dans le plan transverse est de $L_{xy} = 4$ mm (voir figure 0.2 gauche) pour un b-hadron de $p_{\mathrm{T}} \sim 50$ GeV, et permet ainsi de reconstruire un SV distinct du PV à partir des traces dans le jet. De plus, les traces de particules chargées issues de la désintégration des b-hadrons prennent leurs origines dans les vertexes secondaires. L'incompatibilité des traces avec le vertex primaire se traduit par de relativement larges paramètres d'impacts, d_0 dans le plan transverse (montré figure 0.2 gauche), et z_0 sur l'axe longitudinal. Les paramètres d'impacts des traces dans les jets, ainsi que les propriétés des SV reconstruits sont utilisés pour définir des variables permettant de différentier les b-jets des c-jets et $light$-jets. Ces variables sont ensuite combinées statistiquement par une méthode dite d'arbre de décision boosté (BDT). La distribution de sortie du BDT est nommée MV2 et est montrée figure 0.2 (droite). Elle est utilisée comme variable discriminante finale pour identifier la saveur des jets. Par exemple, une coupure sur la variable MV2 de 0.8244 permet de conserver 70% des b-jets tout en rejetant 91.7% des c-jets et 99.7% des $light$-jets. Des algorithmes de b-tagging performants et des efficacités bien comprises dans les données sont nécessaires à l'analyse $t\bar{t}H(H \to b\bar{b})$ et sont d'intérêt général pour l'expérience ATLAS (mesures de QCD ou des propriétés du top, recherche SUSY, etc).

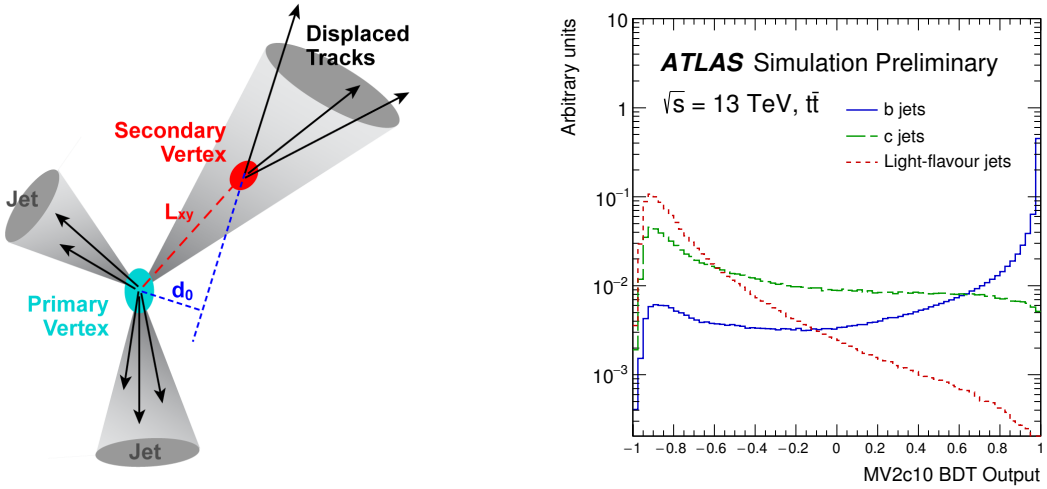


Figure 0.2.: (Gauche) Schéma représentant un b-jets accompagné de deux *light*-jets. (Droite) Discriminant final du b-tagging, MV2, pour la differentiation des b-jets contre les c-jets et *light*-jets.

Fragmentation des quarks b en b-jets

Une bonne compréhension des performances des algorithmes de b-tagging, et des propriétés des b-jets dans les données, nécessite l'étude de la définition des b-jets dans les simulations Monte-Carlo (MC). En particulier, l'étude de la fragmentation des quarks b dans les jets, que j'ai effectuée lors de cette thèse, permet d'établir un lien de parenté entre les particules produites lors des collisions proton-proton et les jets reconstruits après la framentation de ces particules dans le detecteur. Cependant, dans de nombreux cas la définition des b-jets peut être ambigüe. Les produits d'un quark b peuvent se séparer en plusieurs jets, plusieurs quarks b peuvent être a l'origine d'un seul jet, etc.

Dans les simulations MC les b-hadrons sont associés aux jets environants. Les b-jets sont alors définis comme les jets auxquels sont associés au moins un b-hadron. Pour proposer la définition la plus adaptée au b-tagging dans des conditions ambigües, plusieurs algorithmes d'association des b-hadrons aux jets ont été confrontés. L'association dite ΔR [a] définit pour chaque b-hadron un b-jet comme étant le jet le plus proche du b-hadron, s'il se trouve dans un cône de taille $\Delta R(b\text{-hadron,jet}) \leq 0.3$. D'autre part, l'association fantôme (AF) utilise l'algorithme de reconstruction des jets pour identifier les jets issus de b-hadrons. Chaque b-hadron est alors associé au jet minimisant la distance $\Delta R(b\text{-hadron,jet})$ à laquelle est assigné un poids inversement proportionel au p_{T} au carré. Par conséquent, pour deux jets équidistants d'un b-hadron, ce dernier sera associé au jet de plus haut p_{T}.

Dans des événements $t\bar{t}$, qui fournissent un échantillon de b-jet nets, les algorithmes ΔR et AF associent les mêmes jets aux mêmes b-hadrons dans 99% des cas. Cette différence d'identification de 1% induit une variation de 10% sur les efficacités des *light*-jets. L'impact du choix de l'association entre les particules et les jets est, néanmoins, sous-dominant par rapport aux incertitudes mesurées pour les efficacités des *light*-jets (de l'ordre de 50% celon le p_{T}, η et MV2 du jet). Cependant, de plus grandes différences sont observées dans des conditions plus ambigües, en particulier les envi-

[a] Pour l'expérience ATLAS, un système de coordonnées droit est utilisé. Son origine est le point d'intération au centre du détecteur, et son axe z coïncide avec le tube de faisceau. Le plan tranverse au tube de faisceau (x, y) est orienté de sorte que l'axe x pointe vers le centre de l'anneau du LHC. Par égart à la simplicité, les coordonnées cylindriques sont majoritairement utilisées. Les coordonnées ϕ et θ décrivent alors respectivement les angles entre le point considéré et l'axe x dans le plan transverse où l'axe z. La pseudo-rapidité est alors définie à partir de l'angle polaire θ par $\eta = -\log\left(\tan\frac{\theta}{2}\right)$. La distance angulaire ΔR est enfin définie par $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$.

ronements chargés où plusieurs jets entourent les b-hadrons. En effet, dans 89% des cas où un jet est associé à un b-hadron par l'algorithme ΔR et non par l'algorithme AF, un second jet est associé au b-hadron par l'algorithme AF et non par l'algorithme ΔR. Si figure 0.3 (gauche) montre que l'énergie du b-hadron est également répartie entre les deux jets, figure 0.3 (droite) montre que les traces issues de la désintégration du b-hadron sont majoritairement trouvées dans le jet identifié par l'algorithme ΔR. Ceci est du à l'association des traces aux jets, également basé sur un algorithme ΔR. Par conséquent, l'algorithme ΔR est plus cohérent avec les algorithmes du b-tagging (basés sur les traces), et donc est plus à même de représenter les jets pouvant être étiquetés "b" dans les données. Il fut donc choisi comme algorithme par défaut pour les études de b-tagging dans ATLAS. Les détails de mes travaux sur la fragmentation des quarks b, et sur le contenu en traces des b-jets, sont présentés dans le chapitre 3.

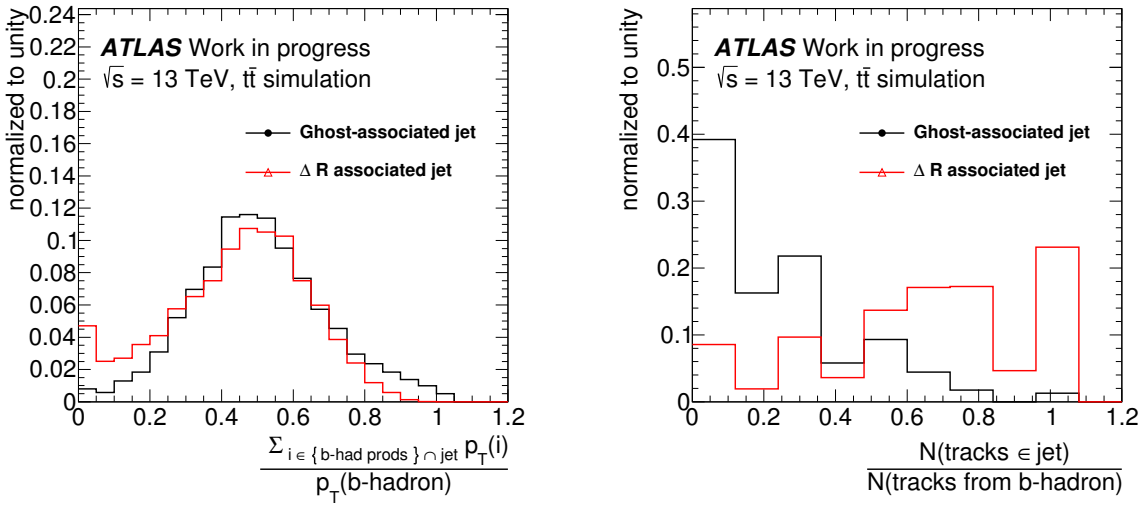


Figure 0.3.: Distributions de la fragmentation d'un hadron b en deux jets, l'un des deux jets étant étiqueté b par association ΔR (le ΔR-associated-truth-jet) et l'autre par association fantôme (ghost-associated-truth-jet). (Gauche) fraction d'énergie du b-hadron retrouvé dans chaqu'un des deux jets. (Droite) nombre de traces issues du hadron b retrouvé dans chaque jets divisé par le nombre total de traces issues du hadron b.

Identification des bb-jets

Les jets issus de deux quarks b (bb-jets), par exemple lors de la désintégration d'un gluon en deux quarks b (gluon splitting) à faibles angles d'ouverture, peuvent aussi être séparés des jets issus d'un seul quark b (single-b-jet). Un algorithme capable de différentier les bb-jets des single-b-jets peut être un apport important pour les recherches, ou mesures, à haut p_{T} dans les environements boostés ainsi que pour les mesures en QCD. Cependant, les algorithmes de base du b-tagging ne sont pas conçus pour la séparation des bb-jets et des single-b-jets.

La reconstruction des multiples vertexes secondaires dans les jets donne des informations importantes pour cette étude. Dans le cas des single-b-jets deux vertexes issus la chaîne de désintégration PV $\rightarrow$ b-hadron $\rightarrow$ c-hadron sont attendus. Dans cette configuration, les deux vertexes se trouvent sur un axe passant proche du PV. Au contraire, dans les bb-jets quatre vertexes sont produits par deux chaînes PV $\rightarrow$ b-hadron $\rightarrow$ c-hadron, et dont les deux principaux vertexes sont ceux de la désintégration des deux b-hadrons. Contrairement aux vertexes des single-b-jets, l'axe formé par ces deux

vertexes ne se prolonge pas au PV.

Mes études montrent que l'utilisation des deux vertexes reconstruits de plus hautes masses est une approximation suffisante pour estimer la position et la cinématique des deux b-hadrons dans les bb-jets. J'ai ensuite utilisé les propriétés des deux vertexes de plus hautes masses dans les jets pour définir des variables discriminantes entre les bb-jets et les autres saveurs de jets, en particulier les single-b-jets. Ces variables sont combinées à des propriétés des jets, telles que le nombre de vertexes reconstruits, dans deux BDT (un axé sur la cinématique des vertexes, l'autre incluant des variables topologiques) dont les distributions de sorties sont appelées MultiSVbb1 et MultiSVbb2 et sont montrées figure 0.4. En ne gardant que 5% des single-b-jets pour 35% de bb-jets conservés, ces algorithmes rejettent jusqu'à sept fois plus de single-b-jets que les algorithmes standards du b-tagging.

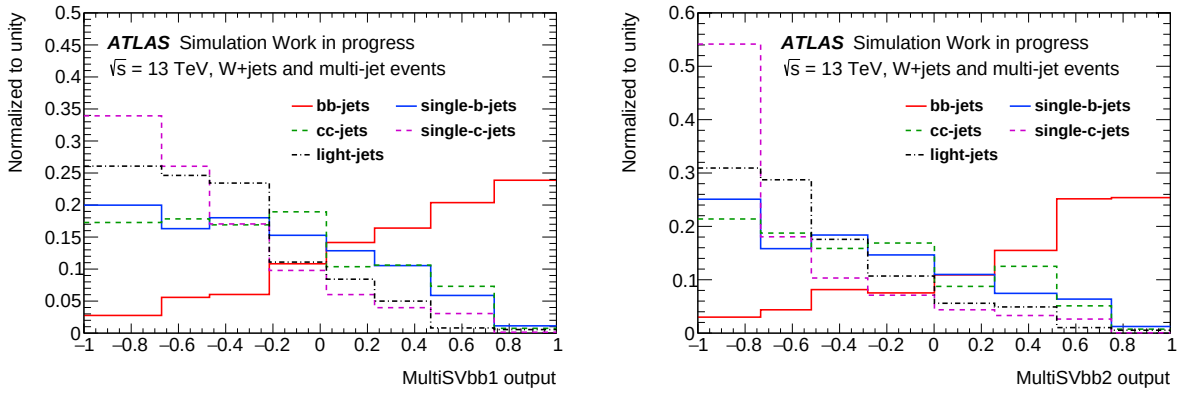


Figure 0.4.: Distributions de sortie des BDTs pour les algorithmes (gauche) MultiSVbb1 et (droite) MultiSVbb2 pour chaque saveur de jet.

Recherche d'évènements de production $t\bar{t}H(H \to b\bar{b})$

Ce n'est que récement, en 2012, qu'une particule compatible avec le boson de Higgs a été découverte par les expériences ATLAS et CMS au LHC. Dès lors, un effort important est dédié aux mesures des propriétés de cette particule (masse, largeur de désintégration, couplages, spin, ...) et à la comparaison des données aux prédictions du SM. Une sélection de ces mesures sont présentées dans le chapitre 1. A ce jour, toutes les mesures effectuées sont compatibles avec les prédictions du SM. En particulier, le couplage de Yukawa au quark top est compatible avec le SM avec une incertitude prédite de $\sim 15\%$. Cette relativement faible incertitude est obtenue par la combinaison de mesures directes[b] et indirectes[c] du couplage de Yukawa au quark top. Cependant, des particules inconnues dans le Modèle Standard peuvent participer aux boucles des mesures indirectes. Dans ce cas, seule la mesure directe provenant de l'analyse $t\bar{t}H$ au Run 1 peut être utilisée. Aucune évidence de production $t\bar{t}H$ n'est trouvé dans les données Run 1. Ce processus contraint donc faiblement le couplage de Yukawa au top et l'incertitude attendue pour ce couplage monte à $\sim 30\%$ sans les contraintes des mesures indirectes.

La recherche d'événements de production $t\bar{t}H(H \to b\bar{b})$ est revisitée pour le Run 2 et est présentée dans les chapitres 4 et 5 en se concentrant sur les résultats produits lors de cette thèse. Le principal

[b] La mesure du couplage de la particule A à la particule B est dite *directe* si elle se base sur un processus dont le digramme de Feynman contient un vertex liant A et B en dehors d'une boucle au premier ordre.

[c] Contrairement à la mesure directe, une mesure indirecte du couplage entre deux particules utilise un processus dont le diagramme de Feynman prédit la production d'une des particules à l'intérieur d'une boucle au premier ordre.

défi pour cette analyse est le bruit de fond $t\bar{t}$ dont la section efficace est plus de 1000 fois supérieure à celle du processus $t\bar{t}H$: 832^{+46}_{-51} pb^{-1}pour la production $t\bar{t}$ contre 507^{+35}_{-50} fb^{-1}pour la production $t\bar{t}H$. Le signal $t\bar{t}H(H \to b\bar{b})$ produit en principe six jets dont quatre sont issus de quarks b. A haut nombre de jets, les événements $t\bar{t}$ avec jets additionels, dits $t\bar{t}+$ jets, sont séparés en trois composantes:

- $\boldsymbol{t\bar{t}}$ **+ light** dont les jets additionels sont issus de saveurs légères. Cette composante se situe principlament à faible nombres de jets étiquetés comme b-jets par le b-tagging (dits b-tagged-jets) et donc contribue peu dans les régions associées au signal. De plus, le large effort mis dans les mesures du bruit de fond $t\bar{t}$ et les ajustements des générateurs d'événements MC permettent à cette composante d'être relativement bien décrite par les simulations.

- $\boldsymbol{t\bar{t}+\geq 1b}$ dont les jets additionels sont issus de quarks b. Cette composante irréductible se trouve principalement dans les régions de signal avec plusieurs b-tagged-jets. De plus, le bruit de fond $t\bar{t}+\geq 1b$ est difficile à prédire théoriquement et souffre de peu de contraintes par des mesures alternatives. De larges incertitudes lui sont donc associées. La composante $t\bar{t}+\geq 1b$ est donc la plus grande source de limitation pour l'analyse $t\bar{t}H(H \to b\bar{b})$.

- $\boldsymbol{t\bar{t}+\geq 1c}$ dont les jets additionels sont issus de quarks c. Le bruit de fond $t\bar{t}+\geq 1c$ se situe typiquement entre les composantes $t\bar{t}+$light et $t\bar{t}+\geq 1b$. Aucune mesure alternative ne contraint les incertitudes sur le $t\bar{t}+\geq 1c$. Il est cependant sous-dominant en comparaison avec le bruit de fond $t\bar{t}+\geq 1b$ dans les régions de signal.

L'analyse $t\bar{t}H(H \to b\bar{b})$ vise l'extraction du bruit fond $t\bar{t}+$jets et du signal $t\bar{t}H(H \to b\bar{b})$ simultanément dans les données par un ajustement statistique aux données, dit ajustement à partir de maintenant. Les événements de production $t\bar{t}H(H \to b\bar{b})$ sont séparés en fonction du nombre de leptons issus de la désintégration des bosons W, eux-mêmes produits par les deux quarks tops. Les cannaux un-lepton et deux-leptons sont traités comme deux analyses distinctes, suivant la même stratégie, et adoptant la même description des bruits de fond principaux. Ces deux cannaux sont ensuite combinés dans l'ajustement final aux données. Mon travail de thèse est focalisé sur le cannal un-lepton et la combinaison finale avec le cannal deux-leptons. De fait, seul le cannal un-lepton est décrit en détail.

En premier lieu, seuls les événements contenants un lepton et au moins cinq jets sont conservés. Pour maximizer la sensibilité à chaque composante du bruit de fond $t\bar{t}+$ jets, les évéments sont catégoriés en fonction du nombre de jets, exactement cinq jets ou au moins six jets, et en fonction du nombre de jets étiquetés b. Onze catégories sont ainsi obtenues, deux régions de contrôle pour chaque composante du bruit de fond $t\bar{t}+$ jets et cinq régions de signal. Une douzième catégorie dit "boosted" est ajoutée. Elle cible le régime à haut p_{T} du boson de Higgs et des quarks tops pour les futures mesures différentielles. Les régions de signal sont dominées par le bruit de fond $t\bar{t}+\geq 1b$ et le rapport du nombre d'événements de signal sur le nombre d'événements de bruit de fond est d'au plus 5.4%.

Afin de séparer le bruit de fond $t\bar{t}+\geq 1b$ et le signal une analyse multi-variée en deux étapes est appliquée aux régions enrichies en signal. Lors de la première étape, dite de reconstruction, plusieurs méthodes utilisent les différences cinématiques et topologiques des états finaux des processus $t\bar{t}+\geq 1b$ et $t\bar{t}H(H \to b\bar{b})$ pour définir des variables discriminantes. L'une de ces méthodes, le BDT de reconstruction, utilise un BDT permettant également de trouver la meilleure correspondance possible entre les jets reconstruits et les quarks de l'état final du processus $t\bar{t}H(H \to b\bar{b})$. Le moment transverse et la masse du boson de Higgs ainsi reconstruit sont montrés figure 0.5 et sont en accord avec les données. Les informations obtenues par ces méthodes sont ensuite combinées dans un BDT entrainé à différentier les événements $t\bar{t}H(H \to b\bar{b})$ et $t\bar{t}+\geq 1b$. La distribution de sortie, nommée "BDT de classification", est utilisée comme discriminant final, et est ajustée aux données dans l'ajustement.

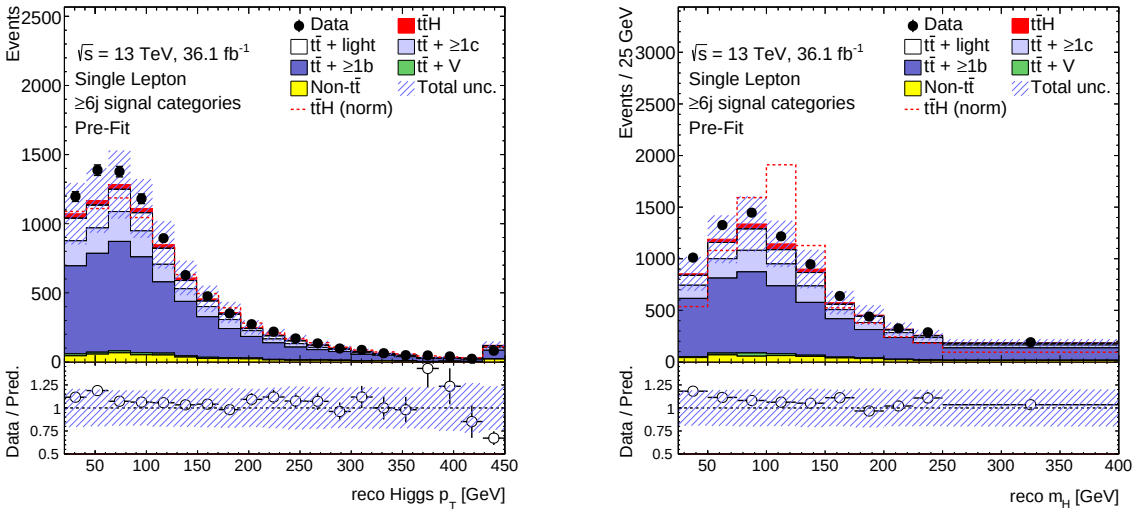


Figure 0.5.: Comparaison des distributions prédites et observées dans les données pour (gauche) le p_{T} et (droite) la masse du candidat au boson de Higgs. Seuls les événements satisfaisants les selections pour les régions de signal sont gardés. La zone hachurée représente l'incertitude prédite totale (systematique et statistique) sans prendre en compte les incertitudes de normalisation des composantes $t\bar{t}+\geq 1b$ et $t\bar{t}+\geq 1c$. Les prédictions sont montrées avant les corrections apportées par l'ajustement.

Reconstruction du système $t\bar{t}H(H \to b\bar{b})$ sur réseau

Lors de cette thèse j'ai développé une nouvelle méthode de reconstruction, présentée pour la première fois dans ce manuscrit et illustrée figure 0.6. Elle utilise le lepton et les jets reconstruits comme les vertexes d'un graphe $\mathcal{G}$ dont tous les vertexes sont connectés par des liens. A chaque lien est ensuite associé un poids représentant la probabilité que les deux particules qu'il connecte aient la même origine. Différents algorithmes sont ensuite appliqués sur le réseaux afin de retrouver le motif originel:

- Le boson de Higgs formé d'une paire de quarks b.
- Le quark top de désintégration semi-leptonique (dit top-leptonique) formé d'une paire {lepton, quark b }.
- Le quark top dont le W se désintègre en quarks (dit top-hadronique) formé d'un triplet de quarks dont un quark b.

Afin de définir le poids de chaque lien un BDT est optimisé pour identifier les paires d'objets (jets ou leptons) provenant de la même particule (le top-leptonique, le top-hadronique ou le boson de Higgs). La variable définie par ce BDT, nommée ici jumelage, est ensuite utilisée pour définir le ratio de participation du lien $i \to j$ au vertex i, noté $\mathrm{rp}(i \to j)$:

$$\mathrm{rp}(i \to j) = \frac{\mathrm{jumelage}(i,j)}{\sum_{z \in \mathcal{G}} \mathrm{jumelage}(i,z)} \tag{0.1}$$

Le ratio de participation définit ainsi la probabilité que le vertex j soit associé au vertex i étant donné les possibles partenaires de i.

J'ai ensuite proposé trois algorithmes, définis au chapitre 4. Le plus prometteur, la résolution sur

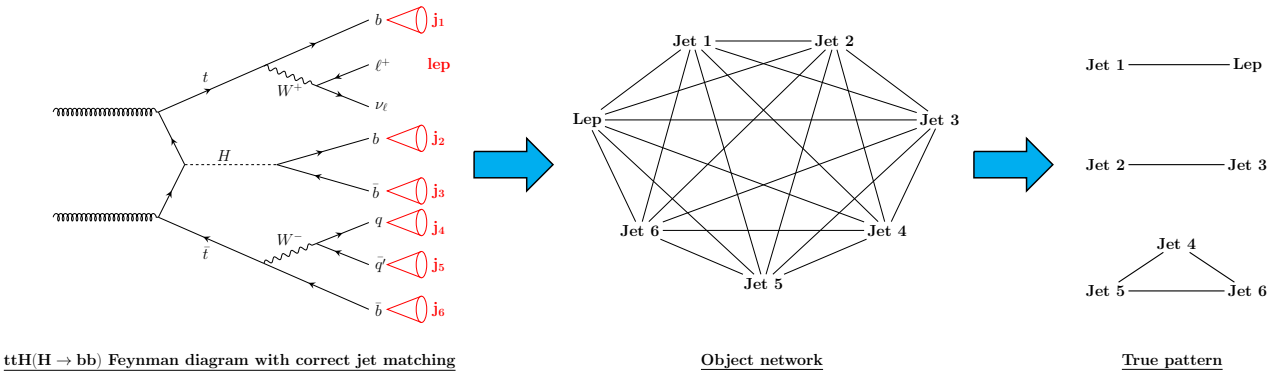


Figure 0.6.: Schéma de la reconstruction sur réseau. (Gauche) Diagramme de Feynman avec les jets associés à chaque parton. (Milieu) diagramme formé par les objets reconstruits. (Droite) Motif originel recherché par les algorithmes d'agglomération et de résolution sur le réseau.

réseau, est basé sur la recherche directe de la combinaison d'un triplet et de deux paires maximizant les ratios de participation. Cette technique proposée pour le futur de l'analyse $t\bar{t}H(H \to b\bar{b})$ donne des performances similaires au BDT de reconstruction. En particulier, les BDTs de reconstruction permettent de retrouver le boson de Higgs dans 31% à 48% des événements. La résolution sur réseau quand à elle reconstruit correctement le boson de Higgs dans 41% des événements. De plus, le chevauchement entre ces deux méthodes est relativement faible: dans seulement 18% des événements le boson de Higgs est reconstruit correctement par les deux méthodes. En conséquence, la reconstruction sur réseau est une nouvelle méthode non seulement compétitive mais aussi complémentaire des méthodes bien établies pour la reconstruction de l'état final du processus $t\bar{t}H(H \to b\bar{b})$.

Analyse statistique des données pour la recheche d'évènements de production $t\bar{t}H(H \to b\bar{b})$

Les résultats de l'analyse $t\bar{t}H(H \to b\bar{b})$ sont obtenus par un ajustement statistique des distributions prédites aux données. La méthode utilisée est l'ajustement par maximum de vraissemblance. Les incertitudes systématiques sont traitées comme des paramètres de nuisance corrigeants les distributions prédites pour coller aux données. De plus, cette technique permet de réduire les incertitudes systematiques (contraintes) en fonction de l'incertitude statistique sur les données. Dans cet ajustement statistique, le signal est paramétré par μ: le ratio entre la section efficace mesurée pour le processus $t\bar{t}H(H \to b\bar{b})$ et la section efficace prédite. Etant donné la complexité de l'analyse (plusieurs couches d'analyses multi-variées dites MVAs) et le peu de mesures alternatives pour les bruits de fond dominants (principalement le processus $t\bar{t}+\geq 1b$), l'élaboration du modèle pour l'ajustement est un point clef de la recherche d'événements de production $t\bar{t}H(H \to b\bar{b})$. J'ai ensuite produit de nombreux ajustements afin de tester et valider le modèle tout en assurant les performances de l'analyse. Le chapitre 5 décrit en détail le modèle utilisé par l'analyse, sa validation et ses performances.

La sensibilité au signal $t\bar{t}H(H \to b\bar{b})$ étant principlament limité par les incertitudes systématiques sur le bruit de fond $t\bar{t}+\geq 1b$, les performances de l'analyse ne peuvent être pleinement estimées qu'après l'ajustement. Pour éviter d'optimiser l'analyse sur les données et de biaiser la mesure du signal, j'ai évalué les performances de l'analyse par un ajustement au jeu de données d'Asimov. Ce dernier est défini par le nombre d'événements prédits dans tous les bins, auxquels des incertitudes de type Poisson sont ensuite associés. Les ajustements aux échantillons de données d'Asimov sont utilisés en

particulier pour choisir le classement des événements, les MVAs, le binning des distributions. Dans la configuration finale, une incertitude de $^{+68\%}_{-65\%}$ sur μ est attendue dans le cannal un-lepton.

Le manque de connaissance sur les bruits de fond prinpaux exige l'utilisation d'un modèle complexe, prenant en compte toutes les sources d'incertitudes possibles, et une description fine de ces bruits de fond. C'est en particulier vrai pour le bruit de fond $t\bar{t}$+$\geq$1b qui domine dans les régions de signal. Les simulations MC pour le bruit de fond $t\bar{t}$ + jets sont générés avec POWHEG+PYTHIA8 (PP8). PP8 donne la meilleure modélisation connue pour le processus $t\bar{t}$ et a été très largement ajusté pour coller aux données 8 TeV et 13 TeV. De nombreux autres échantillons sont produits en variant le générateur MC, les paramètres de désintégration en cascade des particules, etc. Pour chaque variation possible, une incertitude est ajouté au modèle. Pour la composante $t\bar{t}$+$\geq$1b un échantillon suplémentaire est produit. Là où PP8 ne permet d'obtenir deux quarks b additionnels au processus $t\bar{t}$ ($t\bar{t}+b\bar{b}$) que par le parton shower, ce nouvel échantillon extrait le processus $t\bar{t}+b\bar{b}$ directement de l'élément de matrice et au second ordre dans le calcul de l'élément de matrice (NLO) en QCD. Cependant, les données actuelles ne permettent pas d'exclure l'une ou l'autre de ces deux prédictions. Plusieurs modèles sont donc construits. Les deux modèles que j'ai développés et étudiés sont décrits dans ce manuscrit: le modèle de base, utilisé pour la publication en cours, et un second modèle permettant la validation des observations du permier. De plus, de nombreuses incertitudes proviennent de la reconstruction des différents objets présents dans l'état final. En particulier, les incertitudes de reconstruction des jets et sur les efficacités des b-jets, c-jets et $light$-jets auront un impact important sur cette analyse. Mes études du modèle systématique et sa validation sont très largement discutés dans ce manuscrit.

De même que pour les performances, l'étude de la modélisation des bruits de fond doit se faire avant de regarder le signal. J'ai alors réalisé plusieurs ajustements. J'ai réutilisé l'ajustement au jeu de données d'Asimov pour identifier les sources majeures d'incertitudes sur le signal: les incertitudes systematiques sur la composante $t\bar{t}$+$\geq$1b (equivalentes à $\sim 70\%$ de l'incertitude sur μ), les incertitudes statistiques sur le MC (equivalentes à $\sim 45\%$ de l'incertitude sur μ), les incertitudes de b-tagging et sur la reconstruction des jets (equivalentes à $\sim 20\%$ de l'incertitude sur μ chacune). Il permet également l'étude des contraintes les plus importantes sur les incertitudes systematiques. Entre autres, ce manuscrit présente mes études approfondies de l'incertitude systematique ayant le plus grand impact sur la sensibilité au signal.

J'ai également estimé la complétude du modèle par un ajustement à des pseudo-données. Ces dernières sont construites à partir des prédictions, mais en remplaçant la simulation $t\bar{t}$ par un autre échantillon. La différence ainsi obtenue entre les pseudo-données et les prédictions doit être couverte par les incertitudes sur le processus $t\bar{t}$ + jets. Dans le cannal un-lepton $\mu = 0.90^{+0.61}_{-0.58}$ est observé après l'ajustement aux pseudo-données, ce qui est pleinement compatible avec la valeur injectée $\mu = 1$.

L'étape finale de l'analyse est l'ajustement aux données. Dans un premier temps, j'ai confirmé les observations faites dans les précédents ajustements par un ajustement aux données sans inclure les bins sensibles au signal. Les mouvements des incertitudes systematiques pour faire correspondre les données et les prédictions sont égalements étudiés en détail. Pour ce manuscrit, j'ai répété ces études pour les incertitudes systématiques majeures dans l'ajustement final aux données incluant tous les bins. La compatibilité des observations entre tous ces ajustements, et indépendamment de la présence du signal, constitue une preuve majeure de l'absance de biais sur la mesure de μ.

J'ai ensuite combiné le cannal un-lepton au cannal deux-leptons pour obtenir la plus haute sensibilité. Une valeur combinnée de $\mu = 0.84^{+0.64}_{-0.61}$ est observée, correspondant à un excès de signal d'une significance de 1.4σ (pour 1.6σ attendus). Cette valeur de μ est comparée à celles des cannaux un et deux leptons (obtenues en utilisant un ajustement combinné des incertitudes systematiques) figure 0.7.

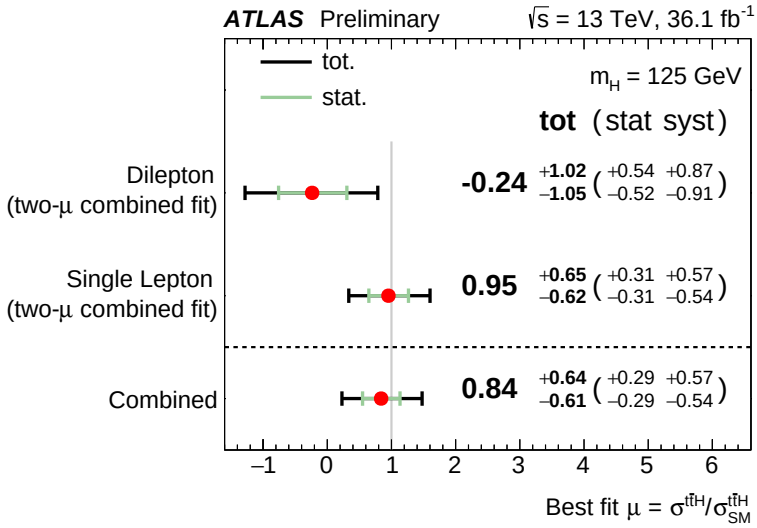


Figure 0.7.: Valeurs de μ pour les cannaux un-lepton, deux-lepton et leur combinaison, observées après l'ajustement statistique des deux cannaux combinés aux données.

Conclusions

La découverte d'une particule compatible avec le boson de Higgs du SM au Run 1 du LHC est une pierre angulaire de la recherche en physique fondamentale. Au début du Run 2, de nombreuses mesures confirment l'appartenance de ce candidat au SM. Cependant, le secteur des couplages aux fermions, en particulier aux quarks, reste peu exploré. En particulier, une mesure directe du couplage du boson de Higgs au quark top (le plus fort couplage de Yukawa) est nécessaire pour contraindre de potentiels effets au delà du SM dans les mesures indirectes. L'observation d'événements de production $t\bar{t}H$ serait le candidat le plus sérieux pour la mesure directe de ce couplage.

L'excellent fonctionnement du LHC et du detecteur ATLAS au Run 2 a permis de collecter 36.1 fb^{-1} de données de collision proton-proton à 13 TeV en 2015 et 2016. Ce document se concentre sur la recherche d'événements de production $t\bar{t}H$ où le boson de Higgs se désintègre en une paire de quarks b, noté $t\bar{t}H(H \to b\bar{b})$, dans ce nouvel échantillon de données. Les performances de cette analyse complexe reposent principalement sur la capacité à séparer et controler le bruit de fond $t\bar{t}$+$\geq$1b. D'un côté, une catégorisation avancée des événements et une série d'analyses multi-variées permettent une forte séparation des processus $t\bar{t}$+$\geq$1b et $t\bar{t}H(H \to b\bar{b})$. D'un autre côté, un modèle comprenant de nombreuses incertitudes systematiques permet de prendre en compte toutes les inconues de la modélisation du bruit de fond $t\bar{t}$+$\geq$1b. Avec cette analyse, une section efficace de $0.84^{+0.64}_{-0.61}$ fois la section efficace prédite par le SM est observée. De plus les sections efficaces supérieures à deux fois celle prédite par le SM sont exclues avec un niveau de confiance de 95%.

Ce manuscrit relate en détail l'analyse $t\bar{t}H(H \to b\bar{b})$ dans le cannal un-lepton et les résultats de la combinaison avec le cannal deux-leptons. Il met en avant le travail effectué dans cette thèse pour la compréhension du modèle, l'élaboration du modèle systematique et la description des données. Ces enjeux majeurs sont ici étudiés au travers de l'analyse des données par ajustement statistique. Cette technique est ensuite utilisée pour extraire le résultat final de l'analyse.

Comme indiqué plus haut, un second enjeu majeur de l'analyse $t\bar{t}H(H \to b\bar{b})$ est la séparation des processus $t\bar{t}H(H \to b\bar{b})$ et $t\bar{t}$+$\geq$1b. La forte séparation obtenue dans l'analyse Run 2 est en particulier due aux techniques de reconstruction, utilisées pour définir l'origine la plus probable des jets reconstruits. Dans ce manuscrit j'ai proposé une nouvelle méthode, basée sur un réseau d'objets reconstruits,

est proposée. Elle permet déjà d'obtenir des performances similaires aux méthodes bien établies tout en leur étant complémentaire. Elle apporte donc une nouvelle source potentielle d'améliorations pour les itérations futures de l'analyse.

Le b-tagging (l'étiquetage des jets originants de quarks b, dits b-jets) est un atout d'intérêt général pour l'expérience ATLAS. Il est en particulier un ingrédient majeur de la recherche d'événements de production $t\bar{t}H(H \to b\bar{b})$, dont l'état final comprend quatre quarks b. Ce manuscrit établit l'importance de la compréhension de la définition des b-jets dans les simulations pour décrire avec précision les événements chargés. En particulier, mes études de la fragmentation des quarks b dans les jets montrent que l'association des jets aux hadrons par l'algorithme ΔR est plus appropriée aux études de b-tagging que l'association fantôme. Cette étude a notamment amené la décision d'utiliser l'algorithme ΔR comme algorithme par défaut pour les études de b-tagging dans l'expérience ATLAS.

Par ailleur, les algorithmes standards du b-tagging ne sont pas conçus pour séparer les jets issus de multiples quarks b, des jets issus de quarks b isolés. La méthode que j'ai développé et étudié permet de séparer les bb-jets (contenants deux b-hadrons) des single-b-jets (contenants un seul b-hadron). Pour une efficacité typique de 35% sur les bb-jets, ces algorithmes ne gardent que 5% des single-b-jets. Cette méthode rejette donc sept fois plus de single-b-jets que les algorithmes standards de b-tagging.

Contents

Introduction

The Standard Model of particle physics describes the fundamental constituents of ordinary matter and their interactions with great precision. The first component of this model is developed at the end of the first half of the 20^{th} century. After about seventy years of experiments, most of the Standard Model predictions are observed with a very high precision and no sign of Beyond Standard Model processes are found. In particular, the discovery of a Higgs boson compatible with the Standard Model in 2012 by the ATLAS and CMS collaboration, forty years after its prediction by R. Brout, F. Englert and P.W. Higgs, brings the last piece of the Standard Model particle puzzle. However, some of the Higgs boson key properties are not yet measured and the nature of the Higgs boson can only be established after comparing these properties to the Standard Model predictions.

The top quark is the known elementary particle with the highest mass and its Yukawa coupling to the Higgs boson is the largest within the Standard Model. Constraints on the top quark Yukawa coupling can be extracted from Run 1 data of the Large Hadron Collider (LHC). However, these measurements depend on the contribution of the top quark to the production and decay of the Higgs boson inside a particle loop and assume that no beyond standard model particle participates to the loop. The associated production of a Higgs boson with a pair of top quarks ($t\bar{t}H$) is the most favorable channel for an access at tree level of the top Yukawa coupling. An observation of this process would thus allow to test Beyond Standard Model effects that could participate to indirect measurements.

This thesis presents the search for $t\bar{t}H$ events where the Higgs boson decays into a pair of b-quarks, $t\bar{t}H(H \to b\bar{b})$, using 36.1 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV recorded by the ATLAS detector at LHC in 2015 and 2016. A full review of the single lepton channel, with exactly one electron or muon from a W-boson decay in the final state, to which I am one of the leading contributors is presented. The results of the combination of the single lepton and di-lepton channels are also shown. The main limitation of the $t\bar{t}H(H \to b\bar{b})$ analysis is the $t\bar{t}$ + jets background, in particular the $t\bar{t}$+$\geq 1b$ component. Multi-variate techniques are necessary to separate this background from the signal. In particular, the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ final state provides valuable discriminating variables for the separation of the signal and the backgrounds. I developed a novel technique for the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ system based on complex networks formalism. It is presented in this report and compared to the well established method which uses a Boosted Decision Tree.

The predicted distributions of the background and of the signal are statistically matched to data using a profile likelihood fit which I developed in the single lepton channel. This thesis describes the input model to the fit, including my studies of the MC modelling of data, of systematic uncertainties, and of the results of the fit. A particular attention is given to the $t\bar{t}$+$\geq 1b$ background model which is the leading source of uncertainties on the signal. The $t\bar{t}$+$\geq 1b$ background is poorly constrained by current measurements and presents large theoretical uncertainties. A great effort is made in Run 2 to improve the $t\bar{t}$+$\geq 1b$ model which was adapted from the Run 1 analysis after dedicated studies that I helped developing and validating. In order to increase the robustness of the analysis, I proposed a second model using alternative predictions and systematic uncertainties and I confronted it to the main model.

The identification of jets originating from the fragmentation of b-quarks, called b-tagging, is a vital input to the search of $t\bar{t}H(H \to b\bar{b})$ events because of the presence of four b-quarks in the final state.

A precise understanding of the performance of the b-tagging algorithms requires a good knowledge of b-jets in Monte Carlo simulations. Finding the origin of a jet is an ambiguous process. My work to study the definition of the flavour of a jet in various event topologies are presented. In particular, I studied the matching of jets to hadrons looking at the fragmentation of b-quarks into jets, the provenance of the jet constituents and the association of tracks to jets. Jets are then categorized depending on these quantities and their b-tagging performance are inspected.

The large amount of data delivered by the Run 2 of the LHC gives access to new decay topologies. In particular, boosted $H \to bb$ topologies, where the two b-quarks are merged into a single jet, could be observed. However, such searches have a large background coming from the gluon splitting to two b-quarks at low opening angles ($g \to b\bar{b}$) which has large uncertainties. Moreover, standard b-tagging algorithms do not provide information on the number of b-quarks from which a jet originates. This thesis describes a technique developed to differentiate $g \to b\bar{b}$ initiated jets from b-jets using reconstructed secondary vertices in jets. It focuses on the studies I performed to develop, understand and improve the performance of a $g \to b\bar{b}$ identification algorithm with Run 2 conditions at the LHC.

This thesis is organized as follows. The basic principles of the Standard Model, in particular the Brout-Englert-Higgs mechanism, are presented in chapter 1. This chapter also reviews the properties of the Higgs boson measured in the Run 1 of the LHC. Chapter 2 gives an overview of the experimental setup of the LHC and of the ATLAS detector. The definition of b-jets for b-tagging studies in Run 2 is described in chapter 3 together with the description of the identification of $g \to b\bar{b}$-initiated jets. Finally chapter 4 and 5 present the search for the $t\bar{t}H(H \to b\bar{b})$ events in 13 TeV data collected with the ATLAS detector in 2015 and 2016.

1. The Standard Model of Particle Physics

The 20^{th} century began with major changes in our interpretation of the universe and its content. While Einstein's *General Relativity* [1] was proposed to understand gravity and the history of our universe, *Quantum Mechanics* [2] appeared as the theory of objects at the atomic scale. However Quantum Mechanics failed in explaining interactions involving the creations and annihilation of particles, that are observed when particles become relativistic ($E \geq m_0 c^2$). The *Relativistic Quantum Theory of Fields* was built to solve this problem and used as a standing stone for the understanding of the particle world.

A new framework is required to re-think the way particle interactions are described. The *matter fields* and the *vectors of the interactions* are included as representation of the *gauge group*. The invariance of the Lagrangian under this group introduces the couplings between various representations. Gauge theories have produced a large number of predictions verified experimentally and allowed to explain the constant flow of phenomena provided by experiments.

The parallel effort of theorists and of experimentalist allowed in less than a hundred years to understand three of the four known forces at the particle level: the electromagnetism, the weak-interaction and the strong-interaction. Section 1.1 reviews the models describing these interactions. The *Standard Model* of particle physics (SM) [3] gathers all the knowledge related to these 3 forces in a single theory. The content of this model is presented in section 1.2 together with the encoding of the three forces in a single theory. The SM is now set in stone by several experiments that validate most of its predictions with an unprecedented precision. The latest achievement being the discovery of a *Higgs boson* compatible with the standard model in 2012 [4, 5]. The introduction of this new boson in the SM is explained in section 1.2 and a review of our knowledge on the observed Higgs boson is presented in section 1.3.

1.1. The success of gauge theories

Quantum field theory aims at describing particles and their interaction. As its name suggests, particles are interpreted as fields. They are divided into two categories:

- **Half-integer spin** particles: they obey Fermi statistics and are thus referred to as *fermions* (Ψ). All known fermions that are elementary particles are of spin $1/2$.
- **Integer spin** particles: they obey Bose statistics and are thus referred to as *bosons*. They are further more separated in spin 0 particles which represent scalar fields Φ and spin 1 particles which represent vector fields V^μ. Other categories are unnecessary since up to now no fundamental particle is observed with a spin larger than 1.

The *Lagrangian* formalism is the most convenient to build quantum field theories. The Lagrangian must be invariant under the symmetries that are found in nature. The first example of a gauge invariant theory is quantum electrodynamics (1.1.1). This theory being successful, it has been used as a reference to model the weak (1.1.2) and the strong interaction (1.1.3).

1.1.1. An abelian gauge theory: quantum electrodynamics

Quantum Electrodynamics (QED) was built in the 30's with a great contribution of P. Dirac [6] and E. Fermi [7] and completed in the 40's by F. Dyson [8, 9], R.P. Feynman [10–12], J. Schwinger [13, 14] and S.I. Tomonaga [15–18]. This section uses the context of QED to explain the principles of quantum field theories and gauge theories. QED aims at describing the electromagnetic interaction between charged fermions and a massless vector boson, the *photon*. The Lagrangian of this theory thus starts with two terms, the free motion of spinor fields described by the Dirac equation and the free motion of photons described by Maxwell's equations:

$$\mathcal{L}_0^{\text{em}} = \overline{\Psi}\left(i\partial\!\!\!/ - m\right)\Psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \quad \text{with} \quad \partial\!\!\!/ = \gamma^\mu\partial_\mu \tag{1.1}$$

where γ^μ are the four Dirac matrices, $\overline{\Psi} = \Psi^\dagger\gamma^0$ is the anti-spinor of Ψ and $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$ is the electromagnetic tensor built out of the electromagnetic field four-vector A^μ.

Maxwell's electromagnetism is invariant under the transformation $A_\mu(x) \to A'_\mu(x) = A_\mu(x) + \partial_\mu\lambda(x)$, with $\lambda(x)$ a function of x with values in $\mathbb{R}$. In the context of gauge theories this transformation is interpreted as the transformation of iA_μ the generator of the $u(1) = i\mathbb{R}$ Lie algebra. The group $U(1)_{\text{em}}$ of 1×1 matrices satisfying $U^\dagger U = 1$, i.e. of the form $U = e^{iA_\mu(x)}$ is thus defined as the gauge group of electromagnetism. To couple to the gauge group fermions are assumed to live in the *fundamental representation* of $U(1)_{\text{em}}$ which implies that they transform as $\Psi(x) \to \Psi'(x) = e^{ieQ\lambda(x)}\Psi(x)$ for particles of charge Qe. Finally, one can show that replacing the partial derivative with the covariant derivative $\partial_\mu \to D_\mu \equiv \partial_\mu - iQeA_\mu$ ensures the gauge invariance of the Lagrangian. It also introduces a new term in the Lagrangian $\mathcal{L}_{\text{int}}^{\text{em}} = Qe\overline{\Psi}\gamma^\mu A_\mu\Psi$ which represents the interaction between the fermions and the bosons. In quantum field theory the term $\mathcal{L}_{\text{int}}^{\text{em}}$ is interpreted as follows:

1. Choose an initial state $|\Psi(t=-\infty)\rangle = |i\rangle$ and a final state $|\Psi(t=+\infty)\rangle = |f\rangle$ in the Fock space of quantum field theory which accounts for multiple particles as tensor products, i.e. multiple realization, of harmonic oscillators. Then the *transition amplitude* between these two states is defined by:

$$\langle f|\, S\,|i\rangle \quad , \quad S = T\left[\exp\left(i\int_{-\infty}^{+\infty} d^4x L_{\text{int}}^{\text{em}}\right)\right] \tag{1.2}$$

   where T is the time ordering operator. This is easily interpreted as a path integral [19] probability between two states. It is in general more convenient to consider $\langle f|\, S-1\,|i\rangle$ to remove contributions from the trivial interaction.

2. Dirac fields are represented as operators annihilating fermions and creating anti-fermions:

$$\Psi(x) = \int \frac{d^3k}{(2\pi)^3}\frac{m}{2k^0}\sum_{\text{spin}=1,2}\left[b_{\text{spin}}(k)u_{\text{spin}}(k)e^{-ikx} + d^\dagger_{\text{spin}}(k)v_{\text{spin}}(k)e^{ikx}\right] \tag{1.3}$$

   where $b, b^\dagger$ and $d, d^\dagger$ are the *creation* and *annihilation operators* for fermions and anti-fermions respectively, satisfying $\left[b_s(k), b_r^\dagger(p)\right] = \delta_{sr}(2\pi)^3 2k^0\delta(\vec{k}-\vec{p})$ and u, v the solutions of the free motion of particles and anti-particles. Similarly bosonic fields are defined by:

$$A_\mu(x) = \int \frac{d^3k}{(2\pi)^3}\frac{m}{2k^0}\sum_{\text{pol}=1}^{3}\left[a_{\text{pol}}(k)\epsilon_\mu(k,\text{pol})e^{-ikx} + a^\dagger_{\text{pol}}(k)\epsilon_\mu(k,\text{pol})e^{ikx}\right] \tag{1.4}$$

   with $a, a^\dagger$ the creation and annihilation operators for bosons and $\epsilon_\mu(k,\text{pol})$ the polarization

quadri-vector.

3. It is then possible to sketch the procedure assuming that the Lagrangian is normal ordered (i.e. that particles are created before being destroyed). The transition amplitude of a photon with momentum k and polarization r $|i\rangle = |\gamma, k, r\rangle = a_r^\dagger(k)\,|0\rangle$ producing a pair electron-positron with momentum p_1, p_2 and spins s_1, s_2 $|f\rangle = |e^+, p_1, s_1; e^-, p_2, s_2\rangle = b_{s_1}^\dagger(p_1) d_{s_2}^\dagger(p_2)\,|0\rangle$ at a specific point x in space-time at *leading order (LO)* is obtained from the LO Taylor expansion of the exponential in eq 1.2:

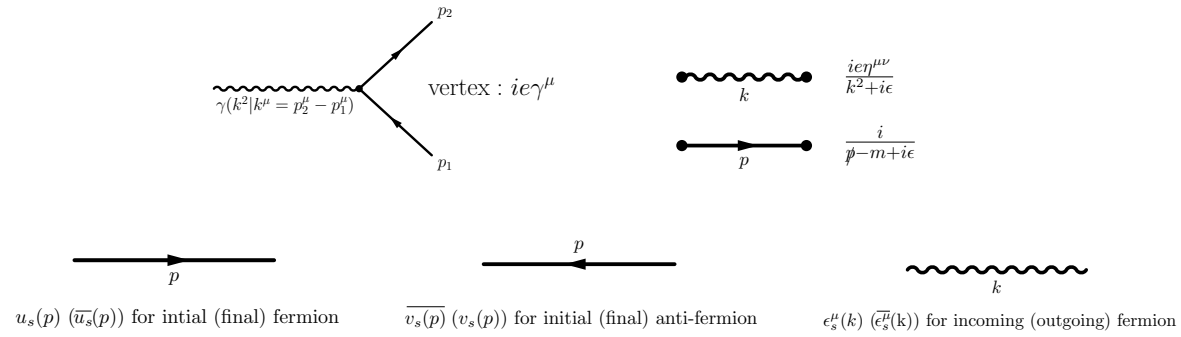
$$\langle f|\, S-1\,|i\rangle = i\int d^4x\, \langle 0|\, b_{s_2}(p_2) d_{s_1}(p_1) \overline{\Psi}(x)\gamma^\mu \Psi(x) A_\mu(x) a_r^\dagger(k)\,|0\rangle \tag{1.5}$$

Noticing that a, b, d operators live in different realizations and thus commute, and that a creation operator applied on $|0\rangle$ gives 0, then one can not have any creation operators coming from A_μ ($a\,|0\rangle = 0 \Leftrightarrow \langle 0|\, a^\dagger = 0$) nor annihilation operators coming from Ψ and $\overline{\Psi}$. One is left with terms $a_{r'}(k') a_r^\dagger(k)$ for the photon and $b_{s_1}(p_1) b_{s_1'}^\dagger(p_1')$, $d_{s_2}(p_2) d_{s_2'}^\dagger(p_2')$ for fermions. This terms are resolved introducing the commutators which bring Kronecker δ that allow to perform the integration over momentum in spinors and $A_\mu(x)$. This last step ensures that the fermionic and bosonic operators create or destroy the particles states including the correct momentum and spin information. Finally one gets:

$$\langle f|\, S-1\,|i\rangle = i\int d^4x e^{i(p_1+p_2-k)x}\, \overline{u_{s_2}}(p_2)\gamma^\mu v_{s_1}(p_1)\epsilon_{\mu r}(k) \tag{1.6}$$

The integration over space time gives $(2\pi)^4\delta^{(4)}(p_1^\mu + p_2^\mu - k^\mu)$ which ensures the four-momentum conservation. The other terms form the so-called *matrix element* of the process $-iM$.

4. The general amplitude for any process in obtained generalizing the previous steps with the *Dyson formula* and *Wick's theorem* [1]. They lead to the *Feynman rules* that allow to compute the matrix element of a process considering all diagrams with the correct final and initial states. The Feynman rules [20] of QED are listed below in terms of initial or final particles, propagator and vertices. The initial and final particles (bottom line) represent respectively the initial and final state of the interaction. A vertex (top-left) refers to the point of interaction between particles. The propagators (top-right) are virtual particles exchanged between two vertices and thus mediating the interaction from one vertex to the other.



The full Lagrangian of QED $\mathcal{L}_{\text{em}} = \mathcal{L}_0^{\text{em}} + \mathcal{L}_{\text{int}}^{\text{em}}$ does not contain any mass term for the gauge boson. The inclusion of a term $\mathcal{L}_m^\gamma \sim \frac{1}{2}m^2 A^\mu A_\mu$ is indeed forbidden by its non-gauge-invariance. This guarantees that the photon is a massless particle in this model.

This theory is one of the best achievements in physics with several predictions confirmed experimentally up to very high precision. The agreement between prediction and measurement of the anomalous moment of the electron [21] at a relative level of 10^{-12} and the hydrogen hyperfine structure [22] characterization are two examples of the numerous successes of QED.

1.1.2. The weak interaction.

Originally, the model for *weak interaction* is not a gauge theory but is rather an extrapolation of QED in the context of observed weak decays.

In 1934 Fermi develops the first weak theory, the *4-point interaction model*, to explain β decays of neutrons [23]. The matrix element of the β decay in this model is directly extrapolated from known QED processes (see fig 1.1) with G_F the *Fermi constant* as coupling constant and reads as follows:

$$M = \frac{G_F}{\sqrt{2}} \left[u_\nu \, \gamma^\mu \, \overline{u_e} \right] \cdot \left[\overline{u_p} \gamma_\mu u_n \right] \tag{1.7}$$

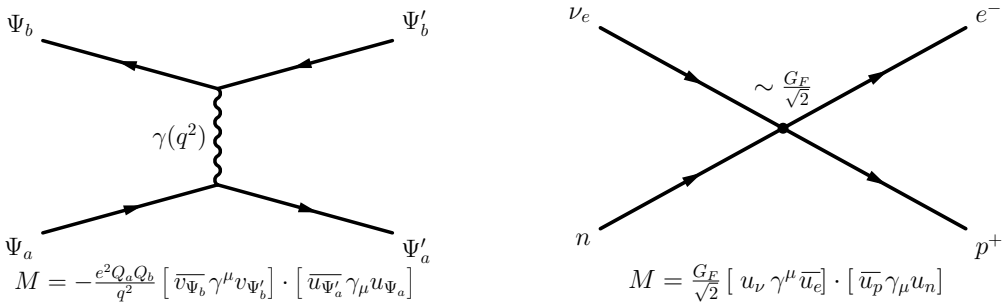


Figure 1.1.: Feynman diagrams for a QED scattering process (left) and Fermi's 4-point interaction (right). The matrix element of the weak interaction is directly extrapolated from QED Feynman rules, replacing the electromagnetic coupling by the Fermi constant G_F.

The *violation of the parity* proposed by C.N Yang and T.D Lee in 1956 [24] and simultaneously discovered by two independent experiments in 1957 [25, 26] is an important input to the theoretical modelling of the weak interaction that granted the Nobel prize to C.N Yang and T.D Lee in 1957.

The explanation of the parity violation requires some more details on the spinor fields. Spinor fields are 4-vectors in the spinor space. Their equation of motion without interactions is given by the Dirac equation [27] $(i\partial\!\!\!/ - m)\,\Psi = 0$. This equation gives four solutions for the spinor fields which can be interpreted as the up- and down- spin states of a solution with positive energy, the fermion, and of a solution with negative energy, the anti-fermion.

The operators P_L and P_R defined in eq 1.8, allow to decompose the spinor representation in two irreducible representations of dimension two. The representation of spinors in terms of these two components is known as the Weyl representation.

$$P_R = \frac{1}{2}\left(1 + \gamma^5\right) \;, \quad P_L = \frac{1}{2}\left(1 - \gamma^5\right) \;, \quad \text{with} \;\; \gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3 \tag{1.8}$$

The projections $\Psi_L = P_L\Psi$ and $\Psi_R = P_R\Psi$ are called the *chiral left-* and *right-handed* components. One can further show that the anti-particle of the left-handed (resp. right-handed) spinor correspond

to the right-chiral-projection (resp. left-chiral-projection) of the anti-spinor:

$$\overline{\Psi_L} = (P_L \Psi)^\dagger \gamma^0 = \overline{\Psi} P_R \quad , \quad \text{using} \quad \{P_L \, , \, \gamma^\mu\} = 0 \tag{1.9}$$

The Parity operator acts on Weyl spinor as an exchange of the left- and right-handed components. Thus the parity violation denotes the breaking of the parity symmetry between left- and right-handed spinors. In 1958 R. Feynman and M. Gell-Mann [28, 29] on one side, E.C.G Sudarshan and R.E. Marshak [30–32] on the other, develop an effective field theory based on vector and axial bosons (the $V - A$ *interaction*). This theory extends the 4-point interaction model by including the P_L operator in the Matrix element to respect the parity violation:

$$u_\nu \, \gamma^\mu \, \overline{u_e} \quad \rightarrow \quad u_\nu \, \gamma^\mu P_L \, \overline{u_e} \tag{1.10}$$

Both the 4-point interaction and its extension the $V - A$ interaction model do not involve any propagators and are thus non-renormalizable[a]. The standard model described in more details in 1.2.3 completes the description of the weak interaction. It introduces two charged bosons W^+, W^- as propagators and the new matrix element reads as:

$$M = -\frac{g_w^2}{2\left(q^2 - m_W^2\right)} \left[\overline{v_{\Psi_b}} \gamma^\mu P_L v_{\Psi'_b} \right] \cdot \left[\overline{u_{\Psi'_a}} \gamma_\mu P_L u_{\Psi_a} \right] \tag{1.11}$$

Figure 1.2 shows a comparison of the weak interaction matrix element with the one of Fermi's model. The mass of the charged bosons can then be directly derived as proportional to the Fermi constant $\frac{G_F}{\sqrt{2}} = \frac{{g_w}^2}{8{m_w}^2}$.

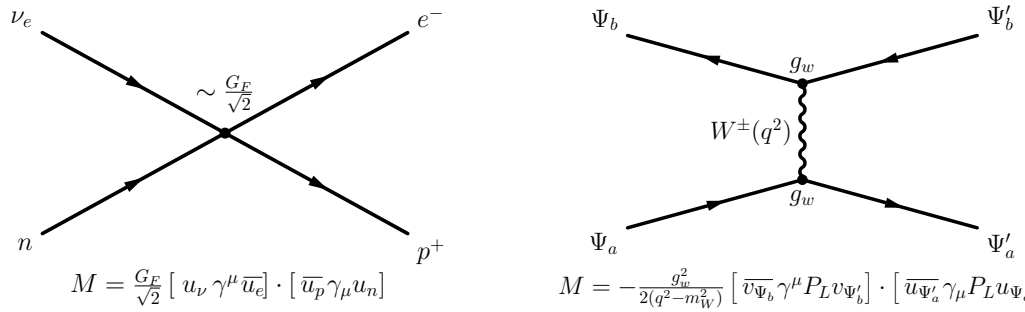


Figure 1.2.: Feynman diagrams for Fermi's 4-point interaction (a) and a weakly interacting process (b). The matrix element is modified to include a massive propagator containing the weak coupling constant g_w and the left-handed projections to account for the parity violation.

The $SU(2)$ structure of weak interaction in the standard model also produces a neutral vector boson Z^0. Unlike charged vector bosons, the Z^0 is subject to a soft chiral asymmetry, i.e. it couples both to left- and right-handed particles but with different strengths. This last piece of the weak interaction is validated by the discovery of weak decays through neutral current in 1973 at the Gargamelle experiment in CERN [33].

[a] In perturbative quantum field theories, such as the Standard Model, the renormalisation procedure avoids divergence at low or high energy scales. In particular, it modifies the propagator (proportional to $\frac{1}{p^2}$ or $\frac{1}{p^2 - m}$) to avoid the divergence when $p \rightarrow 0$ or $p \rightarrow m$. Thus a theory that can not be renormalizable (for example if there exist no propagator) can have infra-red or ultra-violet divergences which are unphysical.

1.1.3. The strong interaction.

The quark model finds its origin in the search for a classification of the hadron masses and decay amplitudes led by Y. Ne'eman [34] and M. Gell-Mann [35] in 1961. This structure is identified, 3 years later [36], as an underlying $SU(3)$ symmetry of hadron constituents that are then referred to as *quarks* [37] in 1972.

The concept of non-abelian gauge theories is introduced by C.N. Yang and R.L. Mills [38] and is extended to the group $SU(3)_C$ of hadron constituents to build the *Quantum Chromodynamics (QCD)*. To ensure their interaction in QCD, quarks are introduced as living in the fundamental representation of $SU(3)_C$ "à la QED". Three *color charges* are thus associated to quarks, red, green and blue, and the action of the $SU(3)_C$ group on quarks is defined as a rotation in the color space $q_i \to q'_i = U_{ij}q_j$, $U_{ij} \in SU(3)_C$ (from now on we assume that there is an implicit sum of all repeated indices).

The Lie algebra $su(3)_C$ of $SU(3)_C$ in its representation **3** is described by the set of eight traceless and hermitian matrices $iT^A_{\mathbf{3}}$ which generate the $SU(3)_C$ color rotations:

$$U \in SU(3)_C \Rightarrow U = e^{i\alpha^A T^A_{\mathbf{3}}} \quad \text{with} \quad T^A_{\mathbf{3}} = \frac{1}{2}\lambda^A \quad , \quad \lambda^A : \text{Gell-Mann matrices} \tag{1.12}$$

with the condition $Tr(T^A_{\mathbf{3}}T^B_{\mathbf{3}}) = \frac{1}{2}\delta^{AB}$.

As for $U(1)_{\text{em}}$ each generator of $su(3)_C$ is associated to a spacetime 4-vector $G^a_\mu(x)$ that is commonly named a *gluon*. The field strength tensor for gluons is defined by :

$$G^A_{\mu\nu} = \partial_\mu G^A_\nu - \partial_\nu G^A_\mu + g_s f^{ABC} G^B_\mu G^C_\nu \tag{1.13}$$

where f^{abc} are the structure constants : $\left[T^a_{\mathbf{3}}, T^b_{\mathbf{3}}\right] = if^{abc}T^c_{\mathbf{3}}$.

The QCD Lagrangian is built similarly to the QED one:

$$\mathcal{L}_{\text{QCD}} = \overline{\Psi}^c \left(i\not{D} - m\right)\Psi^c - \frac{1}{4}G^{c\mu\nu}G^c_{\mu\nu} \tag{1.14}$$

here the index c refers to the color of the quark and the covariant derivative is defined by $D_\mu\Psi^c \equiv \partial_\mu\Psi^c - ig_s G^A_\mu \left(T^A_{\mathbf{3}}\right)^c_{\ d}\Psi^d$.

As in QED, the introduction of the covariant derivative provides an interaction term between gluons and quarks. The non-abelian structure of $SU(3)_C$ (non-vanishing commutators) also induces new terms in the Lagrangian such as $\frac{g_s}{4}f^{ABC}G^{B\mu}G^{C\mu}G^A_{\mu\nu}$ which represent 3- and 4-points *self-interactions* of gluons which are not possible in QED for the photon. A review of the motivations for this octet structure can be found in [39].

An important consequence of the self-interaction of gluons is to compensate the screening effect at short distances. While QED gets stronger and stronger at short distances, this anti-screening effect causes an *asymptotic freedom* of QCD [40, 41]. The experimental verification of this effect (measurement of the running coupling constant [42, 43], etc) granted the Nobel prize to D.J. Gross, H.D. Politzer, F. Wilczek in 2004. On the other side of the spectrum, the quark *confinement* [44], which states that quarks are glued together at low energies, ensures that quarks are trapped in color singlets, the hadrons.

1.2. The Standard Model

The Standard Model of particle physics unifies the weak and electromagnetic interactions and includes the strong interaction in a single theory. First introduced in the sixties by S. Glashow, A. Salam to define the electroweak sector [45, 46] the compatibility with the strong sector is proved by S. Weinberg in 1973 [47] right after the proposition of QCD. Almost all experimental results are compatible with the SM predictions [48].

The SM is based on the invariance under the gauge groups $U(1)_Y \otimes SU(2)_L \otimes SU(3)_C$ where Y stands for hypercharge, L for left-handed and C for color. The *hypercharge* is defined by the Gell-Mann-Nishijima formula $Q = I^3 + \frac{Y}{2}$ [49, 50], where Q is the electric charge and I^3 the weak isospin. It generates a $U(1)_Y$ invariance in the hypercharge space as the electric charge generates a $U(1)_{\text{em}}$ invariance in QED. After a review of the SM content, this section will explain how the $U(1)_Y \otimes SU(2)_L$ invariance is mapped to the $U(1)_{\text{em}}$ invariance and the weak sector.

The R. Brout, F. Englert and P. Higgs (BEH) mechanism [51–53] sketched at the end of this section brings the missing piece of the Standard Model.

1.2.1. Symmetries and gauge bosons

As seen before, the generators of the Lie algebra of the symmetry group define the gauge vector bosons of the theory. The Standard Model strong sector is the QCD model. Thus the first set of generators is the eight gluons of the $SU(3)_C$ Lie Algebra.

The strategy used to describe $SU(3)_C$ can also be applied to the non abelian $SU(2)_L$ group. The Lie algebra $su(2)_L$ of the $SU(2)_L$ group in its representation **2** is fully described by the very well known Pauli matrices σ^A. For convenience, the $su(2)_L$ algebra is often represented by the matrices $T_{\mathbf{2}}^A = \frac{1}{2}\sigma^A$. The $T_{\mathbf{2}}^A$ matrices follow the normalization $Tr(T_{\mathbf{2}}^A T_{\mathbf{2}}^A) = \frac{1}{2}\delta^{AB}$ and their structure constants ϵ^{abc} are defined by :

$$\left[T_{\mathbf{2}}^A, T_{\mathbf{2}}^B\right] = i\epsilon^{ABC} T_{\mathbf{2}}^C \quad \Rightarrow \quad \epsilon^{ABC} = \epsilon^{CAB} = -\epsilon^{BAC}\ , \ \ \epsilon^{123} = 1 \tag{1.15}$$

The Standard Model is thus enriched with the 3 vector bosons of the $SU(2)_L$ Lie algebra $W^\mu(x)$.

Finally the $U(1)_Y$ group is described by a phase factor $e^{i\omega T}$ and generates a single vector boson B_μ. The hypercharge is the equivalent of the electric charge for the electromagnetic interaction as it represents the strength of the interaction of a particle with the $U(1)_Y$ group.

The gauge Bosons and their field strength are summarized in the following table 1.1.

Gauge group	Bosons	index range	Field strength
$SU(3)_C$	G_μ^A	$A = 1, ..., 8$	$G_{\mu\nu}^A = \partial_\mu G_\nu^A - \partial_\nu G_\mu^A + g_s f^{ABC} G_\mu^B G_\nu^C$
$SU(2)_L$	W_μ^a	$a = 1, 2, 3$	$W_{\mu\nu}^a = \partial_\mu W_\nu^a - \partial_\nu W_\mu^a + g\epsilon^{abc} W_\mu^b W_\nu^c$
$U(1)_Y$	B_μ	$\emptyset$	$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$

Table 1.1.: Gauge bosons of the standard model summary and their field strength.

1.2.2. Matter content of the Standard Model: fermions

In the Standard Model, ordinary matter is composed of 12 spin $1/2$ fundamental particles. They are classified in two categories: 6 *leptons* which interact only with the electro-weak sector, and 6 quarks which interact with both the electro-weak and the strong sectors.

Leptons

When the first pieces for the description of the weak and electromagnetic interactions appear, only two charged leptons are discovered, the electron (1897) and the muon (1936) and the existence of neutral and massless leptons, the neutrinos, is postulated. The discovery of the first neutrinos in 1958 and of the τ in 1974 complete the panorama of leptons.

Leptons are sensitive to the electro-weak sector. Therefore, they are organized as left-handed doublets (fundamental representation of $SU(2)_L$) and right-handed singlets (trivial representation of $SU(2)_L$) of the weak interaction. Table-1.2 summarizes the leptonic content of SM and their representation under the gauge groups.

Leptons	Representation $(U(1)_Y, SU(2)_L, SU(3)_C)$	Weak isospin
$\begin{pmatrix}\nu_e\\ e\end{pmatrix}_L, \begin{pmatrix}\nu_\mu\\ \mu\end{pmatrix}_L, \begin{pmatrix}\nu_\tau\\ \tau\end{pmatrix}_L$	$(-1,\ 2,\ 1)$	$\begin{pmatrix}1/2\\ -1/2\end{pmatrix}$
$e_R\ ,\ \mu_R\ ,\ \tau_R$	$(-2,\ 1,\ 1)$	0

Table 1.2.: Leptonic content of matter and the representations of leptons under the three groups of SM.

The absence of right-handed spinors for neutrino lead to the absence of neutrino masses in the Standard Model. Indeed, in the Weyl representation the mass term of the Lagrangian is replaced by $m\bar{\Psi}\Psi \to m\overline{\Psi_R}\Psi_L + m\overline{\Psi_L}\Psi_R$ (this is shown using the projection properties of P_R and P_L). The absence of right-handed spinor naturally cancels the mass term in the Lagrangian (see also 1.2.4). Even though the Super-Kamiokande collaboration has found neutrino oscillation proving that neutrinos have masses. Neutrino masses are supposed to be very small and thus negligible in the context of this thesis but extensions of the standard model can be built to include neutrino masses.

Quarks

Unlike leptons, quarks are sensitive to both the electroweak and strong sectors. In their early sixties proposal, Gell-Mann and Zweig postulated the existence of the up-, down-, and strange-quarks, and their anti-quarks (see discussion in 1.1.3). After its validation in 1968 at the SLAC experiment, this model is extended by the introduction of the newly discovered charm-quark in 1974 at SLAC and Brookhaven, bottom-quark in 1977 at Fermilab, and top-quark in 1995 at Fermilab.

The coupling of quarks to the weak sector imposes the classification of their left-handed component in isospin doublets of observed pair decays:

$$\begin{pmatrix}u\\ d'\end{pmatrix}_L, \quad \begin{pmatrix}c\\ s'\end{pmatrix}_L, \quad \begin{pmatrix}t\\ b'\end{pmatrix}_L$$

where for down type quarks, the notation q' stands for the mass eigenstate of quarks which is connected to its color eigenstate q via the *Cabibbo-Kobayashi-Maskawa (CKM)matrix* [54, 55]:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_{\mathrm{CKM}} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{1.16}$$

The quark mixing induced by this matrix is motivated by the observation of $K^-(us) \to \mu\nu$ decays. This process involves the color eigenstate of the u and s quarks, which in absence of quark mixing belongs to two different generation. Thus the K^- would not be allowed to decay weakly. The quark mixing ensures that the mass eigenstates decaying weakly are linear combinations of the color eigenstates and preserves QCD properties. The CKM matrix parameters have been measured in data and looks as follows [56]:

$$\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} = \begin{pmatrix} 0.97427 \pm 0.00015 & 0.22534 \pm 0.00065 & 0.00351^{+0.00015}_{-0.00014} \\ 0.22520 \pm 0.00065 & 0.97344 \pm 0.00016 & 0.0412^{+0.0011}_{-0.0005} \\ 0.00867^{+0.00029}_{-0.00031} & 0.0404^{+0.0011}_{-0.0005} & 0.999146^{+0.000021}_{-0.000046} \end{pmatrix}$$

The electromagnetic interaction of quarks is assured by the charge they carry. For quark triplets (hadrons) and doublets (meson) to carry their known charges ($\pm e$) the charge of up- and down-type quarks is set to $2/3$ and $-1/3$ respectively.

On top of this composite structure quarks are sensitive to the strong interaction. Their color eigenstates thus belong to the representation **3** of the $SU(3)_C$ group and complete the representation of quarks.

1.2.3. Recovering the weak-interaction and QED

This section is restricted to the electro-weak sector of the Standard Model. As announced previously the Standard Model electro-weak sector can be mapped to the QED and weak-interaction models.

The interaction term of the Lagrangian for the coupling of fermions to the gauge bosons is extracted by requiring the invariance of the Lagrangian under the $U(1)_Y$ and the $SU(2)_L$ groups and is shown in the following:

$$\mathcal{L}_{\mathrm{int}}^{\mathrm{U(1)_Y \otimes SU(2)_L}} = -g_w \overline{\Psi}_L W_\mu^a T_{\mathbf{2}}^a \gamma^\mu \Psi_L - g \frac{Y}{2} \overline{\Psi} B_\mu \gamma^\mu \Psi \tag{1.17}$$

The comparison of the interaction term presented in equation 1.11 to the expansion of the first term over the a index probes the charged vector bosons of the weak interaction:

$$W_\mu^\pm = \frac{1}{\sqrt{2}} \left(W_\mu^1 \mp i W_\mu^2 \right) \tag{1.18}$$

The neutral vector bosons are then extracted from the remaining terms of the Lagrangian performing a rotation in the (I^3, Y) plane, where the isospin I^3 is the third projection of the spin:

$$Z_\mu = W_\mu^3 \cos\theta_W - B_\mu \sin\theta_W \tag{1.19}$$

$$A_\mu = W_\mu^3 \sin\theta_W + B_\mu \cos\theta_W \tag{1.20}$$

The angle of the rotation θ_W is known as the *Weinberg angle*. It links together the coupling constants of the electro-weak sector of Standard Model to the coupling constant of QED $g_w \sin\theta_W = g \cos\theta_W = \alpha_{\mathrm{em}}$ and is measured in data with a high precision $\sin\theta_W = 0.22295 \pm 0.00028$.

1.2.4. The electro-weak $SU(2)_L \otimes U(1)_Y$ symmetry breaking and the BEH mechanism

The Standard Model is successful in defining the strong, weak and electromagnetism interactions for elementary particles. However, in the formalism described up to now, no mass term for fermions is allowed which is in contradiction with their observed masses. Indeed, the mass term in the Weyl representation $m\bar{\Psi}\Psi = m\overline{\Psi_R}\Psi_L + m\overline{\Psi_L}\Psi_R$ is not invariant under the $U(1)_Y \otimes SU(2)_L$ symmetry. Moreover this theory is based on 4 massless bosons while the weak interaction described in section 1.1.2 needs 3 bosons (Z^0, W^+, W^-) with relatively high masses.

The R. Brout, F. Englert and P. Higgs mechanism is introduced in the Standard Model to solve this problem. Following the previous statement the BEH mechanism proposes that the $U(1)_Y \otimes SU(2)_L \otimes SU(3)_C$ symmetry is spontaneously broken to $U(1)_{\text{em}} \otimes SU(3)_C$. It will ensure that the $W^{\pm}$ and Z^0 bosons can have a mass term in the Lagrangian while keeping the photon and gluons mass-less.

To achieve this goal a complex scalar field is introduced, the Higgs field. The interaction with the $SU(2)_L$ bosons and with fermions requires a doublet structure of this scalar field:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \begin{pmatrix} \phi_1^+ + i\phi_2^+ \\ \phi_1^0 + i\phi_2^0 \end{pmatrix} \tag{1.21}$$

The Higgs field being a scalar field, it follows the Klein-Gordon equation of motion. Its Lagrangian is:

$$\mathcal{L}_{\text{Higgs}} = D^\mu \phi^\dagger D_\mu \phi - V(\phi) \tag{1.22}$$

with the potential defined as:

$$V(\phi) = \frac{1}{2}\mu^2 \phi^\dagger \phi + \frac{1}{4}\lambda \left(\phi^\dagger \phi\right)^2 \tag{1.23}$$

the most general gauge invariant potential. The unitarity requires that the free parameters μ^2 and λ are real, and to obtain a finite minimal value of the potential (vacuum stability) the condition $\lambda > 0$ is imposed. Finally $\mu^2 < 0$ is required and the potential adopts the so called "Mexican hat" shape illustrated in figure 1.3. The *spontaneous symmetry breaking* is illustrated by the fact that nature chooses a specific value for the vacuum and breaks the symmetry $SU(2)$ of the vacuum.

The vacuum is chosen to be:

$$\phi = \begin{pmatrix} 0 \\ v \end{pmatrix} \quad , \quad v = \sqrt{\frac{-\mu^2}{\lambda}} \text{ the vacuum expectation value (v.e.v.)} \tag{1.24}$$

This state is not invariant under the $SU(2)_L$ and the $U(1)_Y$ group symmetries but is invariant under the action of the $U(1)_{\text{em}}$ group.

In order to generate the physical states one has to expand the Higgs field around the minimum of its potential. This is done in a perturbative approach. A real field $H(x)$, the Higgs boson, is added to the vacuum state:

$$\phi = \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{1.25}$$

For a general perturbation one should also add a term $e^{i\frac{G^a(x)}{2v}}$ where G^a are the Nambu-Goldstone bosons that account for the three broken generators. However a clever fixing of the $SU(2)$ gauge (i.e. a choice of the position on the vacuum circle) allows to suppress this term without any loss of generality. Developing $\mathcal{L}_{\text{Higgs}}$ and replacing the W_μ^a and B_μ bosons by the weak bosons as done in

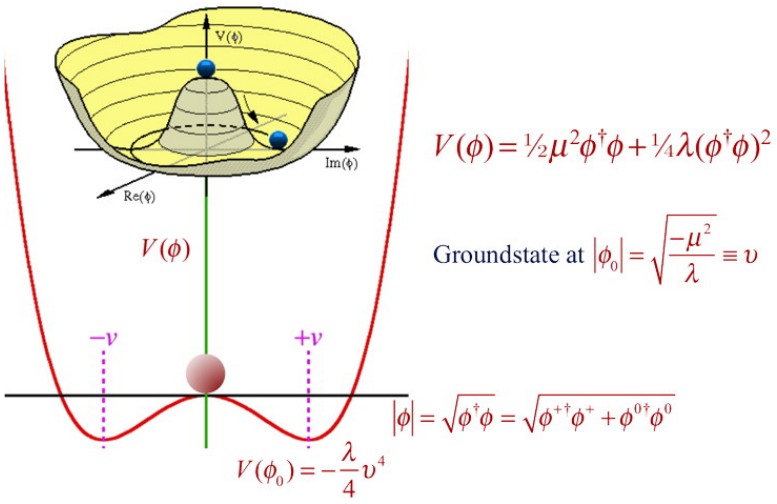


Figure 1.3.: The potential of the Higgs boson $V(\phi)$ for a single scalar field.

section 1.2.3 one finds:

$$\begin{aligned}\mathcal{L}_{\text{Higgs}} = & \frac{1}{2}\,\partial^\mu H \partial_\mu H \\ & + \frac{1}{2}\frac{g_w^2 v^2}{4\cos^2(\theta_W)} Z^\mu Z_\mu + \frac{g_w^2 v^2}{4} W^{-\mu} W_\mu^+ \\ & + \left(\frac{2}{v}H + \frac{1}{v^2}H^2\right)\left(\frac{1}{2}\frac{g_w^2 v^2}{4\cos^2(\theta_W)} Z^\mu Z_\mu + \frac{g_w^2 v^2}{4} W^{-\mu} W_\mu^+\right) \\ & - \mu^2\frac{(v+H)^2}{2} - \lambda\frac{(v+H)^4}{4}\end{aligned} \tag{1.26}$$

The second line of this equation shows the mass terms of gauge bosons appearing through the BEH mechanism. Then the third line presents the possible interaction of the Higgs boson with the weak vector bosons, and the fourth line introduce the mass term of the Higgs boson $m_H = v\sqrt{\lambda}$ and the self coupling of the Higgs boson.

It is important to quote that the mass term of the $W^\pm$ boson fixes the Higgs boson v.e.v. even though the SM does not fix the value of m_W, thus the BEH mechanism only adds one extra free parameter to the SM, the Higgs boson mass. The value of the Higgs boson v.e.v. can be extracted from the relation seen in section 1.1.2 between m_W and the precisely measured Fermi constant:

$$\left.\begin{aligned} m_W &= \frac{g_w v}{2} \\ \frac{G_F}{\sqrt{2}} &= \frac{g_w^2}{8 m_w^2} \end{aligned}\right\} \Rightarrow \quad v = \left(\sqrt{2}G_F\right)^{\frac{1}{2}} \simeq 246.2 \text{ GeV} \tag{1.27}$$

Fermion masses are finally added to the theory as Yukawa Lagrangian interactions:

$$\mathcal{L}_{\text{Yukawa}} = -C_f\,\overline{\Psi_L}\,\phi\,\Psi_R + \text{h.c.} \tag{1.28}$$

The same procedure as for bosons allows to build the mass terms for "weak-isospin-down-type" fermions (d, s, b, e, μ, τ) and their interaction with the Higgs boson. The coupling constant is then proportional to the mass of the fermion : $C_f = \sqrt{2}\frac{m_f}{v}$.

To include the mass of "weak-isospin-up-type" fermions (u, c, t) one has to repeat the procedure starting from a doublet $\phi' = i\sigma_2\phi$ which introduces the neutral component as the "weak-isospin-up-type" scalar field. The exact same results can then be derived and one finds:

$$\mathcal{L}_{\text{Yukawa}} = -\frac{v\, C_f}{\sqrt{2}}\,\overline{\Psi_L}\,\Psi_R - \frac{C_f}{\sqrt{2}}\,\overline{\Psi_L}\,\Psi_R H + \text{h.c.}\ , \quad C_f = \sqrt{2}\frac{m_f}{v} \tag{1.29}$$

1.3. The Higgs boson searches and its discovery

The importance of the Higgs mechanism is clear from section 1.2.4. A huge effort is done to search for the Higgs boson in several experiments (LEP [57], Tevatron [58] and LHC) and yet it remained undiscovered for several decades. The discovery of a particle compatible with the Standard Model Higgs boson in July 2012 by the ATLAS and CMS experiments at LHC is an important milestone in the history of physics. This section reviews the discovery of this Higgs boson and the measurement of some of its properties at the LHC experiments.

1.3.1. Higgs boson production and decay modes

A review of the different production and decay modes of the Higgs boson in a proton-proton (noted pp in what follows) collisions is presented here. All the results presented here can be found in [59].

The production modes considered in pp collisions, such as at the LHC, are listed here in decreasing order of production cross-section:

- **Gluon fusion**: The gluon fusion (noted $gg \to H$ or simply ggH), which leading order diagram is shown in figure 1.4(a), is the leading contribution to SM Higgs boson production at LHC due to the overwhelming presence of gluons in pp collisions. The top- and bottom-quarks are the main contributors to the quark loop and contributions from other quarks are negligible for current searches.
- **Vector boson fusion**: The leading order diagram for vector boson fusion (VBF or also noted qqH) is shown in figure 1.4(b). Two quarks produce a vector boson V ($W^{\pm}$ or Z^0) and their fusion produces a Higgs boson. The presence of diagrams with a vertex connecting the bosons to the Higgs boson without being in a loop is referred to as direct coupling. The direct coupling of the Higgs boson to the vector bosons in VBF allows a direct measurement of the coupling of the Higgs bosons to vector bosons in addition to the bosonic decays of the Higgs boson.

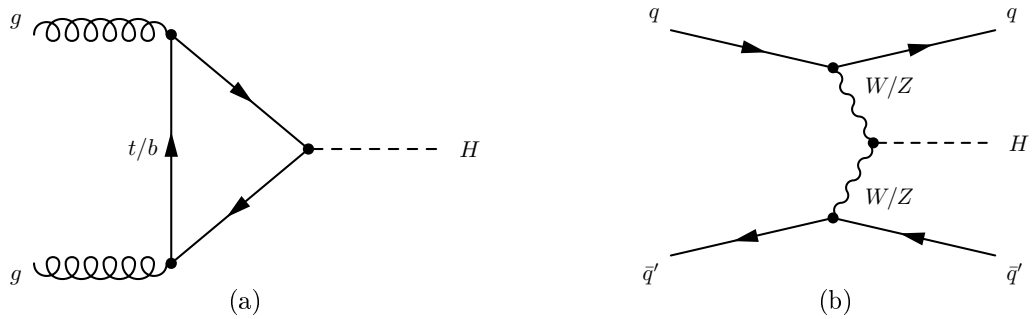


Figure 1.4.: Leading order diagrams for the gluon fusion (left) and vector boson fusion (right) initiated SM Higgs boson production.

- **Higgs-strahlung**: The Higgs-strahlung, or associated production of Higgs bosons with vector bosons (referred to as VH processes), leading order Feynman diagrams for qq and gg initiated processes are shown in figure 1.5. These production modes are privileged processes to study $H \to bb$ since they benefit from the leptonic decays of the additional vector bosons to reduce the multi-jet background.

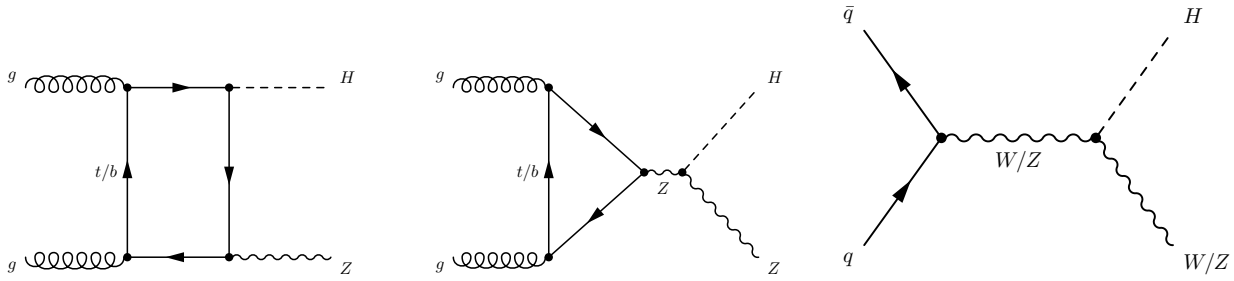


Figure 1.5.: Leading order diagrams for the production of a Higgs boson in association with a vector boson.

- **Associated production of the Higgs boson with top-quarks**: Figure 1.6 shows a set of Feynman diagrams for the production of a Higgs boson in association with top quarks. These diagrams involve direct coupling of the Higgs boson to the top quarks. Thus these are privileged production modes for the study of the Yukawa coupling of the Higgs boson to top quarks which is the highest Yukawa coupling in the SM. In particular, the $t\bar{t}H$ production (upper diagram) is the preferred channel for the measurement of this coupling as it has a higher cross-section than the tH processes (bottom diagrams). However, the tH processes are still important as they are sensitive to the sign of the coupling via beyond SM effects. In the case of $tHb+\text{j}$ (where j stands for one jet) production (bottom left and bottom center diagrams in figure 1.6), two production modes of the Higgs boson are involving coupling to both W-bosons and top-quarks. Since the final state is the same they can not be separated and the coupling to top-quarks can not be directly accessed.

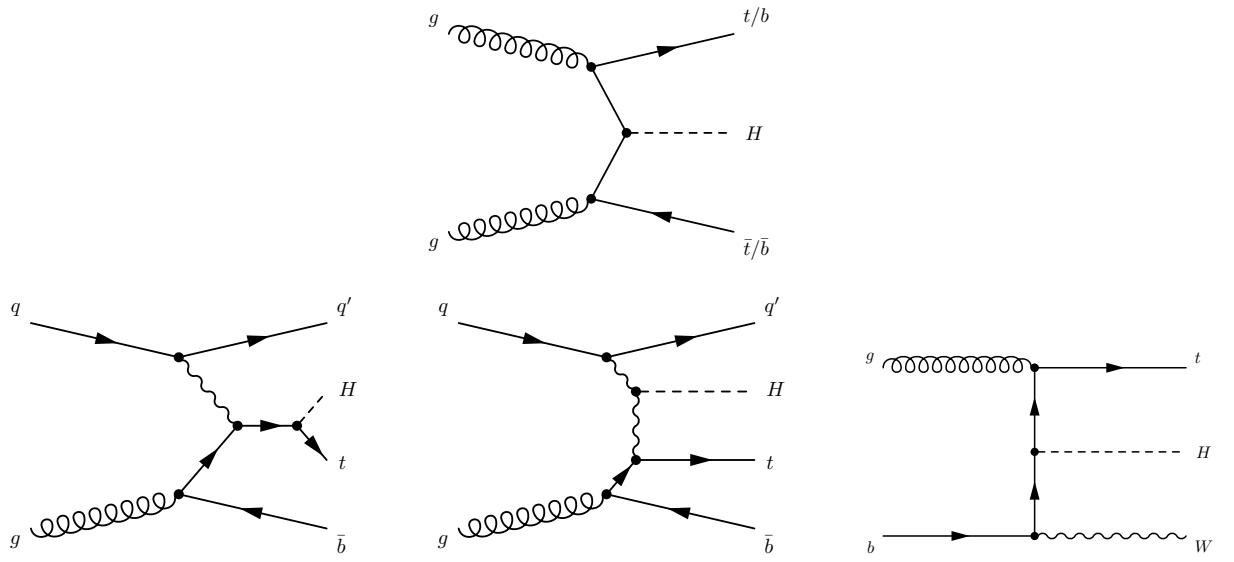


Figure 1.6.: Leading order diagrams for the production of a Higgs boson in association with top quarks.

Figure 1.7 summarizes the main production modes cross-sections as a function of the square root of the center-of-mass energy $\sqrt{s}$. The Branching ratios of Higgs decays modes are presented as a function of the Higgs boson mass in fig 1.8. At the measured mass of 125.09 GeV the Higgs boson mainly decays to a $b\bar{b}$ pair (58%). The gain from the relative high rate of $H \to b\bar{b}$ events is however balanced by the relatively larger difficulty to identify b-jets compared to leptons and photons in the detectors and large backgrounds.

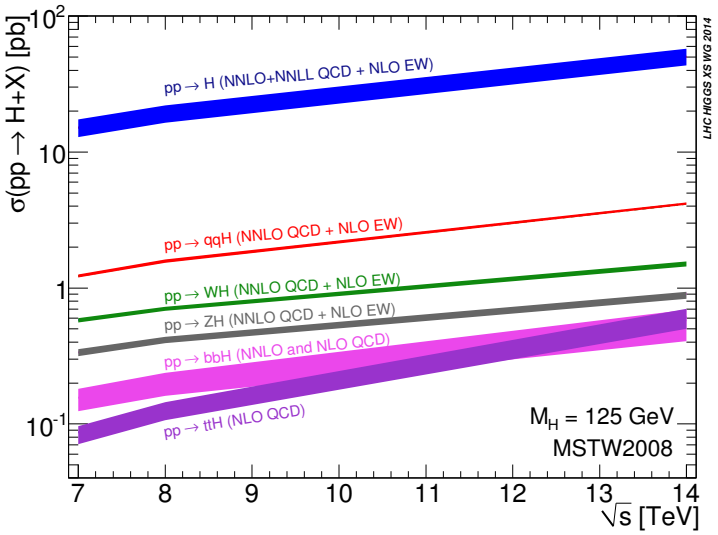


Figure 1.7.: The production cross-section of the SM Higgs boson as a function of the pp collisions center-of-mass energy for a Higgs boson of mass 125 GeV [60].

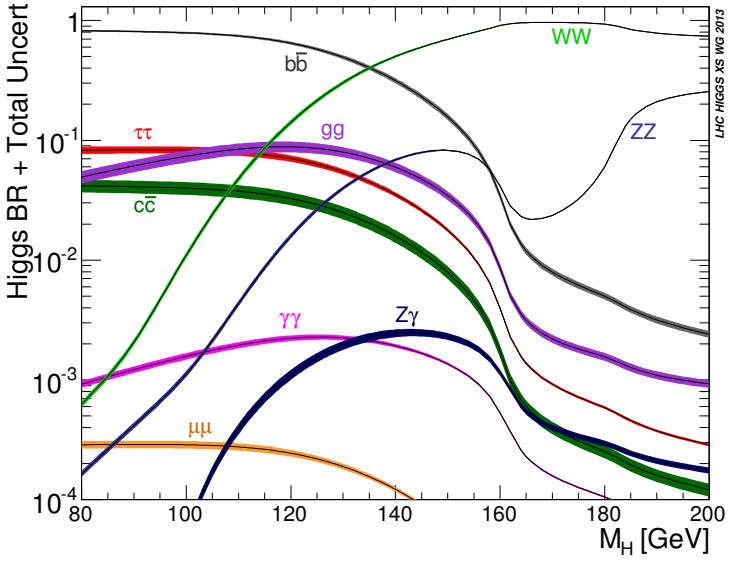


Figure 1.8.: The branching ratios of the SM Higgs boson to various decay modes as a function of its mass [59].

1.3.2. The Higgs boson discovery at LHC

The LHC started to produce pp collision in March 2010 at the never reached before center of mass energy of $\sqrt{s} = 7$ TeV. It provided collisions at this energy until the end of 2011. Then the LHC delivered $\sqrt{s} = 8$ TeV pp collisions from April 2012 to December 2012. This whole period is referred to as LHC-Run 1 and a total integrated luminosity of 5.5 fb^{-1} at $\sqrt{s} = 7$ TeV and 22.8 fb^{-1} at $\sqrt{s} = 8$ TeV is recorded.

It is in this period that the LHC achieves one of its major goals and announces the discovery of a SM-like Higgs Boson. Higgs boson searches are first performed in all decay modes of the Higgs boson. However bosonic decay modes of the Higgs boson provide better signal sensitivity compared to fermionic final states, leading to a focus on the former topologies for the initial searches. Figure 1.9 shows the comparison between data recorded in early 2012 by the ATLAS detector and SM predictions for the $H \to \gamma\gamma$ (1.9(b)) and $H \to ZZ^* \to 4l$ (1.9(a)) channels [4]. These famous two bumps at $\sim$ 125 GeV indicate the presence of the new boson compatible with the SM. The combined discrimination of the background hypothesis, known as significance, by these two channels is shown in figure 1.10 and outstands the 5σ threshold for a process to be announced as a *discovered*. In parallel the searches done by the CMS collaboration in the same channels are equally successful [5]. These two independent observations of the same Higgs boson demonstrate the validity of each single discovery.

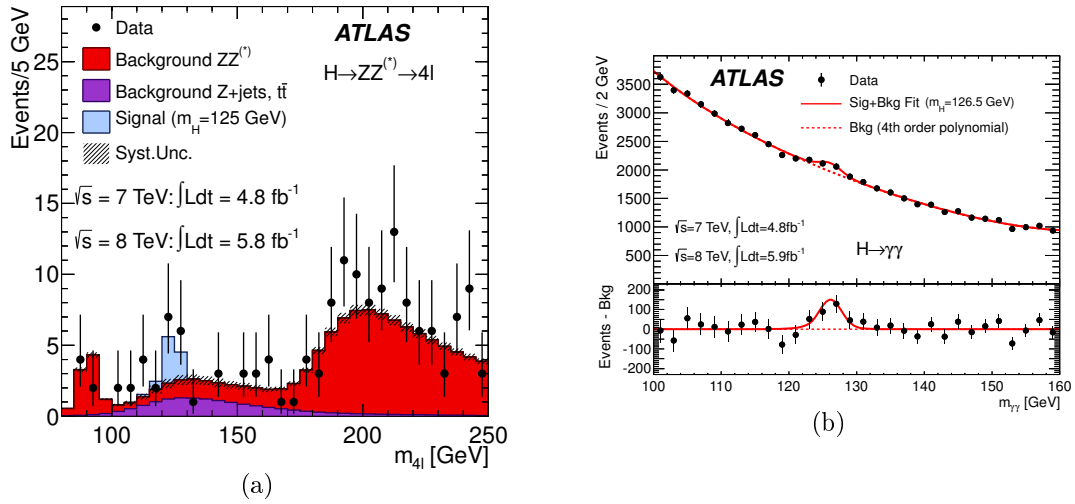


Figure 1.9.: The distribution of the four-lepton invariant mass m_{4l} for the selected candidates in $H \to ZZ^* \to 4l$ events (left) and di-photon invariant mass for the selected candidates in $H \to \gamma\gamma$ events (right), compared to the background expectation, for the combination of the $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV collected data at the LHC. The signal expectation for a SM Higgs boson with $m_H = 125$ GeV is also shown for m_{4l} [4].

At the end of Run 1 the ggH production and the $H \to ZZ^*$, $H \to WW^*$ and $H \to \gamma\gamma$ decay modes are observed by both the ATLAS and CMS collaborations. In addition a combination of the analyses performed within the ATLAS collaboration with the analyses performed within the CMS collaboration is done [61]. This combination improves our knowledge of the Higgs boson production and decay modes and allowed the observation of the VBF production and $H \to \tau\tau$, as shown in table 1.3.

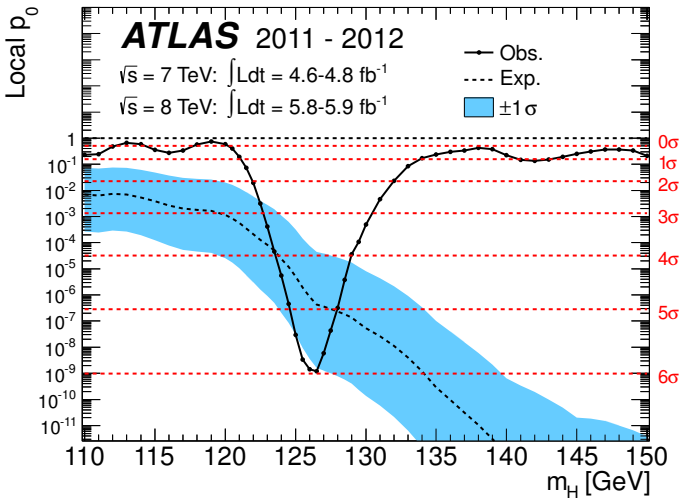


Figure 1.10.: The observed (solid) local p_0 as a function of the Higgs boson mass (m_H) in the low mass range. The dashed curve shows the expected local p_0 under the hypothesis of a SM Higgs boson signal with its $\pm 1\sigma$ band. The horizontal dashed lines indicate the p-values corresponding to significances of 1 to 6 σ [4].

Production process	Measured significance (σ)	Expected significance (σ)
VBF	5.4	4.6
WH	2.4	2.7
ZH	2.3	2.9
VH	3.5	4.2
ttH	4.4	2.0
$H \to \tau\tau$	5.5	5.0
$H \to b\bar{b}$	2.6	3.7

Table 1.3.: Measured and expected significances for the observation of Higgs boson production processes and decay channels for the combination of ATLAS and CMS. The ggH production process and the $H \to ZZ^*$, $H \to W^+W^-$, and $H \to \gamma\gamma$ decay channels, have already been clearly observed and thus are not included. All results are obtained constraining the decay branching fractions to their SM values when considering the production processes, and constraining the production cross sections to their SM values when studying the decays [61].

1.3.3. Properties of this newly found particle

This newly discovered Higgs boson is compatible with the SM Higgs boson introduced in section 1.2.4. However beyond SM (BSM) models are not excluded and precise measurements of this observed particle properties are necessary to discriminate between the various hypotheses.

The Higgs mass and couplings:

The Higgs boson mass is a free parameter of the SM. Measuring precisely the mass of the Higgs boson is necessary to determine the branching ratios of the Higgs boson and the cross section of the Higgs boson production modes at LHC.

This measurement is done in the context of $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ decays, where as seen in section 1.3.2, the Higgs mass peak is narrow and gives a high experimental resolution of a few GeV. The combined measurement of the Higgs mass in the ATLAS and CMS collaborations with the full Run 1 dataset has been performed [62–64]. Figure 1.11 shows the measured Higgs mass in the different channels and their successive combinations towards the final result:

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

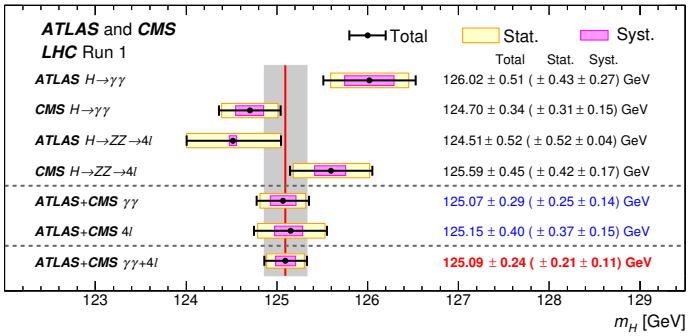


Figure 1.11.: Summary of Higgs boson mass measurements from the individual analyses of ATLAS and CMS and from the combined analysis for data collected in the Run 1 of the LHC. The systematic (narrower, magenta-shaded bands), statistical (wider, yellow-shaded bands), and total (black error bars) uncertainties are indicated. The (red) vertical line and corresponding (gray) shaded column indicate the central value and the total uncertainty of the combined measurement, respectively [61].

Once the Higgs mass is fixed it is possible to compute the cross sections and branching ratios of the various Higgs boson production and decay modes, and thus of the couplings. The coupling modifiers κ_i are expressed as ratios of cross-sections or branching ratios to the standard model ones ${\kappa_i}^2 = \frac{\sigma^i}{\sigma^i_{\text{SM}}}$ or ${\kappa_i}^2 = \frac{\Gamma^i}{\Gamma^i_{\text{SM}}}$ where i denotes the production or decay mode. Figure 1.12 shows the constraint from the combined ATLAS and CMS data [61] on the *global fermionic and bosonic coupling modifiers*: κ_F for the coupling to fermions and κ_V for the coupling to bosons. κ_F and κ_V are obtained assuming that the coupling modifiers of the Higgs boson to the $W^{\pm}$ and Z^0 are the same $\kappa_V = \kappa_W = \kappa_Z$ and the coupling modifiers to the top-, bottom-quarks and the τ are the same $\kappa_F = \kappa_t = \kappa_b = \kappa_\tau$. Coupling modifiers of individual channels are also shown assuming all couplings κ^f_F, $f = t, b, \tau$ and κ^f_V, $f = W, Z$ uncorrelated. All the results are in agreement with the SM prediction within one standard deviation.

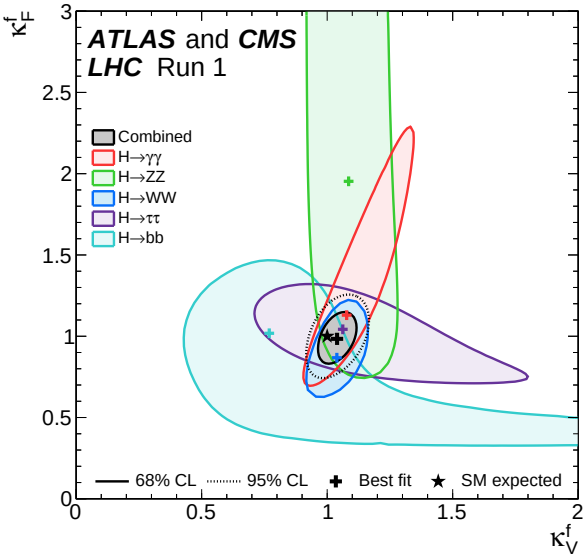


Figure 1.12.: Negative log-likelihood contours at 68% and 95% CL in the (κ_F^f , κ_V^f) plane for the combination of ATLAS and CMS and for the individual decay channels, as well as for their combination (κ_F versus κ_V shown in black), without any assumption about the sign of the coupling modifiers [61].

The Higgs boson spin and parity:

The Higgs boson is introduced in the standard model as a spin 0 and CP even particle ($J^P = 0^+$) but other models can generate other types of Higgs bosons. To discriminate between these representations precise measurements of the Higgs spin and parity are necessary. These measurements are based on the kinematic properties of the $H \to \gamma\gamma$, $H \to ZZ^* \to 4l$ and $H \to WW^* \to l\nu l\nu$ which differ depending on J^P. These measurements are presented in [65] following the prescriptions of [66] and the result of this analysis is shown in figure 1.13. The $J^P = 0^+$ nature of the Higgs boson is confronted to the alternative hypotheses $J^P = 0^-, 1^+, 1^-, 2^+_m$. The rejection of the spin 1 and 2 hypotheses at respective confidence levels higher than 99.7% and 99.9% is an evidence of the spin 0 nature of the Higgs boson, and thus of the compatibility of SM with the ATLAS data. This analysis also shows a preference for the even parity predicted by the SM.

The Higgs boson width Γ_H:

The Higgs boson *width* Γ_H corresponds to the total decay rate of a particle and is materialized by the width of the Higgs boson mass peak. This parameter is sensitive to new physics as BSM introduce new massive particles that would couple to the Higgs boson and enlarge its width.

The Higgs boson width, of a few MeV in the SM, is way below the mass resolution of the detector as one can see in figure 1.9. This issue is solved using the asymmetry in the *off-shell* and *on-shell* productions of the Higgs boson in $gg \to H \to ZZ$ events (see [67, 68] for the introduction of the width term and [69] for the theoretical motivation of the experimental setup). In fact, the propagator of the Higgs introduces a dependence on Γ_H of such events :

$$\frac{d\sigma_{gg\to H\to ZZ}}{dm^2_{ZZ}} \sim \frac{g^2_{ggH} g^2_{HZZ}}{(m^2_{ZZ} - m^2_H)^2 + m^2_H \Gamma^2_H} \tag{1.30}$$

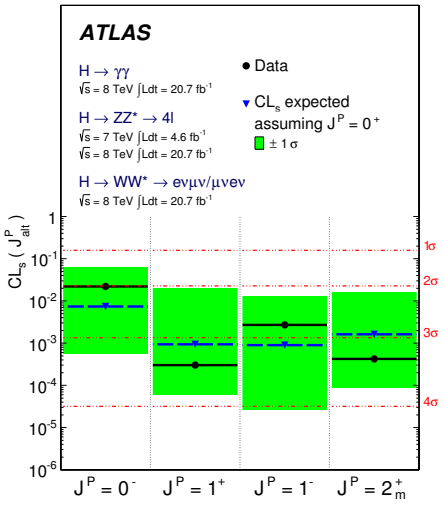


Figure 1.13.: Expected (blue triangles/dashed lines) and observed (black circles/solid lines) confidence level CL_s for alternative spin–parity hypotheses assuming a $J^P = 0^+$ signal. The green band represents the 68% $\mathrm{CL}_s(J^P_{\mathrm{alt}})$ expected exclusion range for the same signal assumption. On the right y-axis, the corresponding numbers of Gaussian standard deviations are given, using the one-sided convention [65].

with g_{ggH} and g_{HZZ} the Higgs boson couplings to the gluon and Z pairs respectively and m_{ZZ} is the invariant mass of the ZZ pair. The on-shell Higgs boson production is found by assuming $m_{ZZ} = m_H$ which cancels the first term of the denominator while one can choose the Higgs boson to be sufficiently off-shell to have $m_{ZZ} >> m_H$ changing the denominator to m^2_{ZZ}. In the end one obtains:

$$\sigma^{\mathrm{on-shell}}_{gg\to H\to ZZ} \sim \frac{g^2_{ggH} g^2_{HZZ}}{m^2_H \Gamma^2_H} \quad , \quad \sigma^{\mathrm{off-shell}}_{gg\to H\to ZZ} \sim \frac{g^2_{ggH} g^2_{HZZ}}{m^2_{ZZ}} \tag{1.31}$$

This analysis has been performed by CMS [70] and then reproduced by ATLAS [71] with enhanced interpretation of theoretical limits. The measured value of the Higgs boson width as a function of negative-log-likelihood is shown in figure 1.14 together with the limit of its ratio with the standard model expectation. With these analyses the upper limit on Γ_H has been reduced to a few tens of MeV which is two orders of magnitude lower than the direct measurements.

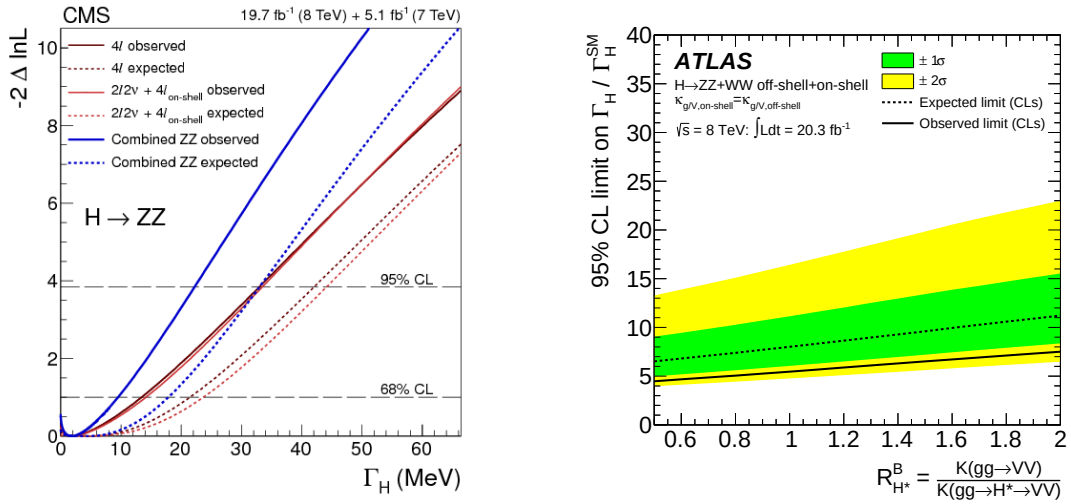


Figure 1.14.: (left) Scan of the negative log-likelihood as a function of Γ_H in the different analysis channels (red and black) and for the combined fit (blue) in the CMS analysis [70]. (right) Observed and expected combined 95% CL upper limit on $\Gamma_H/\Gamma_H^{\mathrm{SM}}$ as a function of coupling ratio [71] in the ATLAS analysis. The upper limits are calculated from the CL_s method, with the SM values as the alternative hypothesis. The green (yellow) bands represent the 68% (95%) confidence intervals for the CL_s expected limit.

2. The ATLAS detector for the LHC experiment

The Large Hadron Collider (LHC) [72, 73], described in section 2.1, is the most advanced circular particle accelerator. It provides proton-proton collisions at the highest center of mass energy ever achieved at a very high rate. Nominally two bunches of $\sim 10^{11}$ protons at 7 TeV cross every 25 ns. This makes the LHC a privileged environment for modern experimental study of particle physics — Higgs boson and BSM searches, precise measurement of SM processes — which requires both high energy events and a large amount of data.

The research work presented in this thesis is based on the data recorded by the ATLAS detector placed at one of the collision points on the LHC ring. A description of the ATLAS detector is presented in section 2.2. Sections 2.3 and 2.4 describe respectively the simulation of data collisions and the object reconstruction within the ATLAS experiment.

2.1. The Large Hadron Collider

The LHC is a 26.7 km long ring installed 100 m below the surface at the French-Swiss border at CERN (Geneva). The LHC tunnel was originally built between 1984 and 1989 to host the Large Electron-Positron (LEP) collider. In the early 2000, the LEP programme came to its end and the LHC ring building started. In 2008 the LHC was ready to provide data. However an incident in a connection between two magnets damaged the ring and the first data-taking with high energy collisions was postponed to 2010 when the Run 1 started. As explained in section 1.3.2 the Run 1 provided data until 2012. From 2012 to 2015 the LHC and the detectors underwent an upgrade to reach higher center of mass energy and higher luminosity. The Run 2 started mid 2015. This thesis is based on data recorded in 2015 and 2016 during Run 2.

2.1.1. A proton-proton collider

Proton-proton collision events are chosen to marry high energies and large amount of data. Indeed e^+e^- circular colliders suffer from a large loss of energy due to synchrotron radiation and proton-antiproton collisions cannot offer a large amount of data due to the difficulty to produce antiprotons. However pp collisions come with a set of difficulties. The *hard scatter* (interaction of interest) occurs between constituents of the protons, namely quarks (q) and gluons (g) which are inclusively referred to as *partons*. At the LHC gg initiated processes are favored rather than $q\bar{q}$ or qg initiated processes due to the parton dynamics inside protons. Partons carry only a fraction of the proton energy following the parton distribution function. If this phenomenon allows to scan a larger range of $\sqrt{s}$, it involves non perturbative QCD. Therefore it requires input from other experiments which come with their uncertainties. Moreover, on top of the hard scatter, the remaining partons in the protons can interact generating an underlying event. Such events are badly described by existing models. An other challenge in pp collisions is the overwhelming production of gluons and quarks (observed as multi-jet events) due to the large QCD coupling. These events are an important source of background events for a large fraction of analysis of the LHC physics programme and should be suppressed.

Figure 2.1 shows the production cross-section of several of the main SM processes as a function of the pp center of mass energy. These plots also illustrate that pp collisions products are dominated by multi-jet events as mentioned before. The relatively large cross section of top-quark production modes makes the LHC the first top-quark factory for precise measurements of the top-quark properties. In addition the operating energy of the LHC rises the Higgs boson production rate to an accessible value, making the discovery of the Higgs boson possible. Figure 2.2 summarizes the cross sections of SM processes measured with the ATLAS detector.

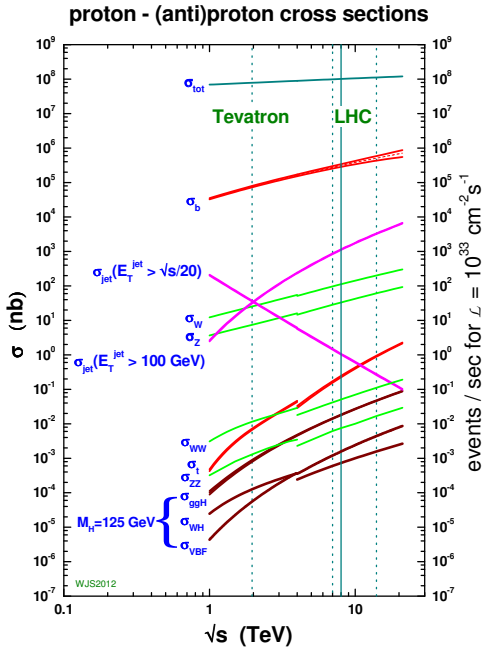


Figure 2.1.: Expected cross sections for a few typical SM processes in proton-(anti)proton collisions as a function of the center of mass energy [74].

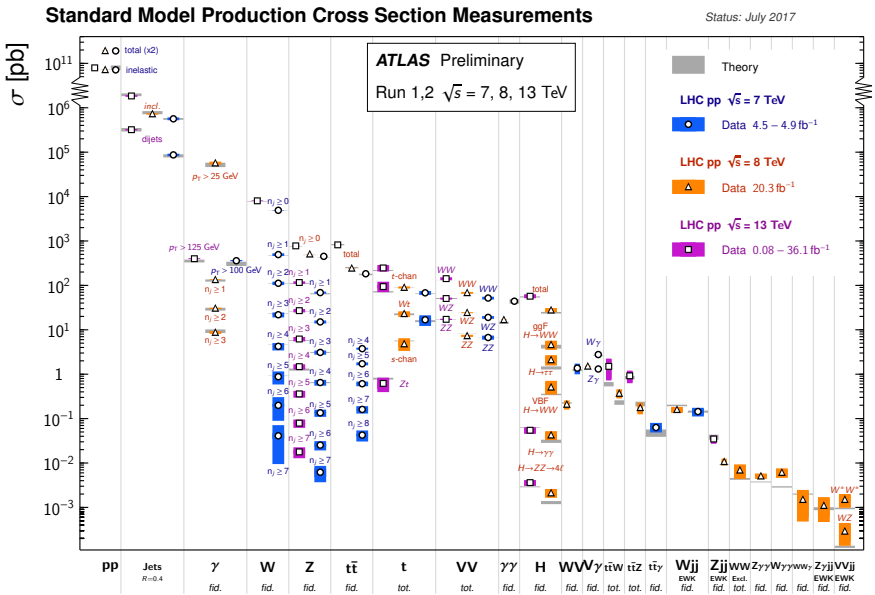


Figure 2.2.: Summary of measured cross sections for Standard Model processes confronted to expected values at $\sqrt{s} = 7, 8, 13$ TeV with the ATLAS detector [75].

2.1.2. The LHC setup

The LHC is the final element of the accelerating chain at CERN. Hydrogen atoms are first ionized and the obtained protons are injected in accelerators from past experiments present at CERN. The accelerator chain is shown in figure 2.3 and reads as follow:

1. Protons are injected in LINAC II and linearly accelerated to an energy up to 50 MeV, i.e. a third of the speed light (c).
2. The BOOSTER proton synchrotron rises their energy to 1.7 GeV which corresponds to $\sim 0.916c$.
3. In the Proton Synchrotron (PS) protons are accelerated to 26 GeV or $0.999c$.
4. Then the Super Proton Synchrotron (SPS) delivers proton beams at an energy of 450 GeV.
5. After 4 min and 20 s the LHC is filled and gives the final shape (size, spacing, ...) and stabilizes proton beams which energy is increased to $3.5, 4$ or 6.5 TeV (depending on the operating year) in 20 min for the highest energy.

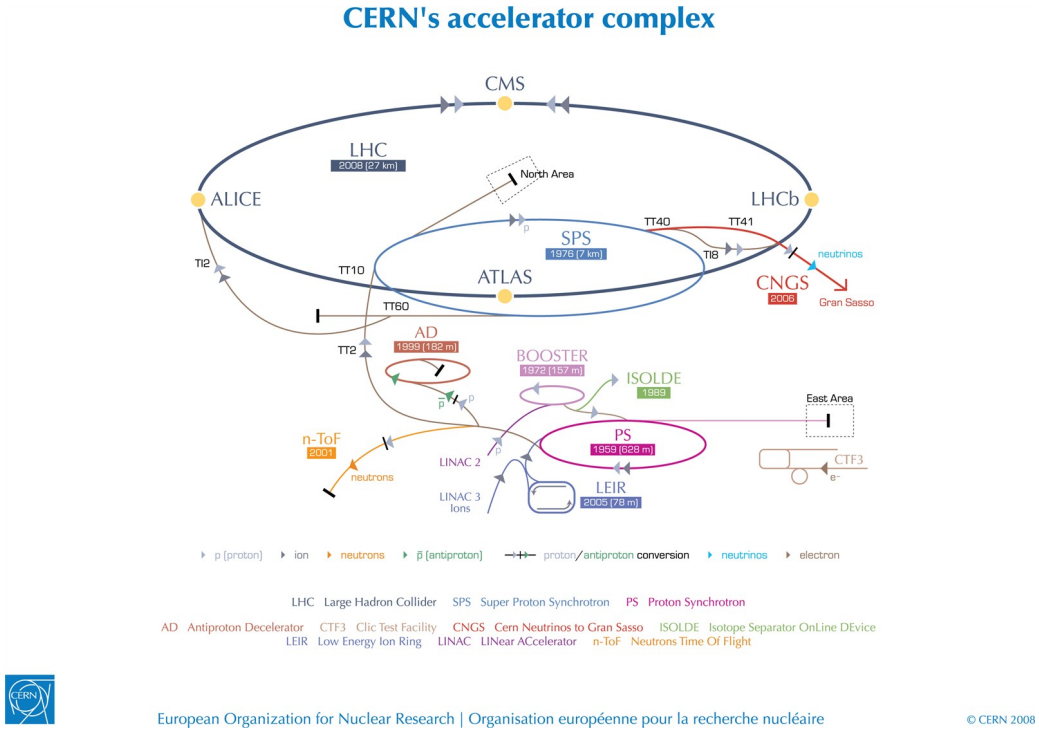


Figure 2.3.: Schematic view of the LHC accelerator chain [76].

At the LHC, the proton beams are divided in up to 2808 bunches containing approximately 10^{11} protons. Nominally the bunches are 25 ns or 50 ns apart (depending on the data taking period). This distance is called bunch-spacing. However the injection scheme from the SPS necessitates the grouping of bunches into bunch-trains with additional time spacing imposed between bunch-trains. Bunches being composed of 10^{11} protons each bunch-crossing leads to multiple pp interactions which are referred to as pile-up. The interaction of interest, usually the interaction of highest energy, is called the hard-scatter.

Although the LHC is mainly designed for pp collisions it can also perform heavy lead ions collisions

at an energy of 2.76 TeV per nucleon. With hundreds of protons and neutrons colliding for each ion collision, a plasma of quarks and gluons is formed. It allows to study the behavior of matter in similar conditions as of the very early universe ($\sim 10^{-6}$ s).

The LHC accelerator is divided in eight straight sections and arcs which are required to accelerate and bend two counter rotating beams.

To maintain the protons in the beam pipe a total of 9600 super-conducting magnets made of Niobium and Titanium are installed and kept at 1.9 K with super-fluid helium. 1232 of them are the dipole magnets shown in figure 2.4 which bend the path of protons in the arcs. At a current of 11 kA they deliver a magnetic field as high as 8.3 T. To correct for imperfections at the extremities of the magnetic field, they are coupled with sextupole, octupole and decapole magnets. Then 858 quadrupole magnets delivering a nominal gradient of 223 T/m and 241 T/m are used to focus the beam. They are installed by pairs all over the LHC ring, the first magnet controlling the width and the second the height of the beam.

In the straight sections radio-frequency chambers deliver 2 MV to generate an electric field of 5 MV/m oscillating at 400 MHz which accelerates the protons and ensures a tight separation between bunches in the beam.

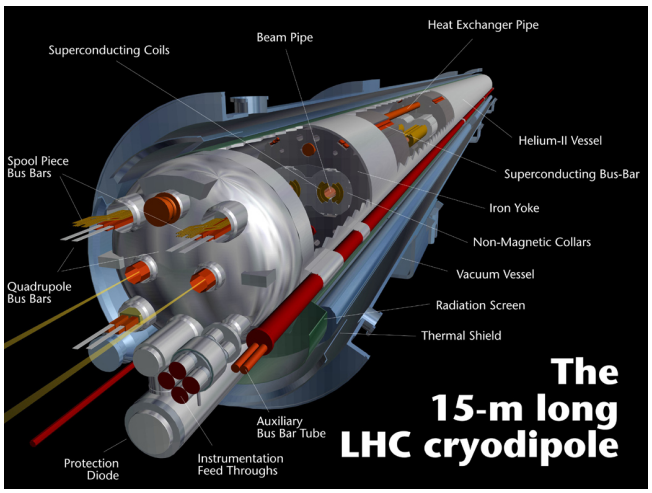


Figure 2.4.: Modelisation of the LHC dipole segment [77].

Table 2.1 presents the various parameters of the LHC for each data taking period up to early 2017. The amount of data during collisions is expressed in terms of the *instantaneous luminosity* $\mathcal{L}$ defined by the accelerator properties as follows:

$$\mathcal{L} = \frac{N_p^2 k_b f_{\text{rev}}}{4\pi\sigma_x\sigma_y} F \tag{2.1}$$

where N_p is the number of protons per bunch, k_b the number of bunches per beam, f_{rev} the revolution frequency, $\sigma_x\sigma_y$ the bunch dispersion in the transverse plane assuming a Gaussian distribution of particle density around the beam axis, and F is a geometric correction factor to account for the crossing angle of the two beams at the interaction point. The total amount of data recorded over a certain period is called the *integrated luminosity* $L = \int \mathcal{L}\, dt$. The luminosity is related to the obtained number of events of a certain process via $N = \mathcal{L}\,\sigma\,\varepsilon$ with σ the process cross-section and ε the event selection efficiency (trigger, reconstruction and selection).

	Run 1		Run 2	
parameter	2010—2011	2012—2013	2015	2016
Beam energy [TeV]	3.5	4	6.5	6.5
Bunch spacing [ns]	50	50	50-25	25
Max number of bunches	1380	1380	2244	2200
Protons per bunch [10^{11}]	1.45	1.6	1.15	1.15
Peak luminosity [$10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$]	3.7	7.7	5.0	13.6
Integrated luminosity [cm^{-2}]	5.46	22.8	4.2	38.5
Mean pile-up	9.1	20.7	13.7	24.2

Table 2.1.: Operating parameters of the LHC for each data delivering period [78, 79].

2.1.3. Physics goals at the LHC

Four of the eight straight sections are used to collide protons and are equipped with different detectors:

- The ATLAS [80] (A Toroidal LHC ApparatuS) detector: ATLAS is a general purpose detector. It is designed to identify most of pp collision products in a large range of energy. It takes full advantage of the large luminosity offered by the LHC to cover a large range of the LHC physics program. This detector is described in section 2.2.

- The CMS [81] (Compact Muon Solenoid) detector: CMS targets the same physics as the ATLAS detector but using a different technology. Its design is based on a superconducting magnet generating a 4 T magnetic field and a strong tracking system to precisely identify tracks and measure their momentum, especially in the case of muons. As two separate experiments with different detectors the ATLAS and CMS experiments are independent and complementary. Each of them can provide a confirmation of particle discovery by the other experiment and datasets can be combined for enhanced precision.

- The LHCb [82] (LHC beauty) detector: LHCb is dedicated to heavy flavour physics and the search for BSM effects via precise measurement of beauty and charm flavoured hadrons. It is designed as a single arm spectrometer focusing on high energy $b\bar{b}$ production for which the two b-quarks are mostly in the forward or backward region. It has a forward angular coverage of ± 15 mrad to 300 (250) mrad in the bending (non-bending) plane and an η acceptance of $1.9 < \eta < 4.9$. The large amount of $b\bar{b}$ production allows LHCb to work at lower luminosity compared to the ATLAS and CMS detectors ($\mathcal{L} \sim 10^{34}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$) while still recording a large amount of data.

- The ALICE [83] (A Large Ion Collider Experiment) detector: ALICE focuses on QCD measurements for strongly interacting matter and quark-gluon plasma description at large energy densities and high temperature in ion collisions.

- The TOTEM [84] (TOTal Elastic and diffractive cross section Measurement) detector: TOTEM is a low luminosity $\sim 2 \times 10^{29}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$ independent experiment but integrated in the CMS detector area. It aims at measuring the total pp cross-section and at the understanding of the proton structure via elastic scattering.

- The LHCf [85] (LHC forward) detector: LHCf is a small detector placed on both sides of the ATLAS detector 140 m away from the interaction point for neutral particle detection in the for-

ward regions. Its goal is to constrain interaction models used for the description of atmospheric showers induced by very high energy cosmic rays hitting the atmosphere.

2.2. The ATLAS experiment

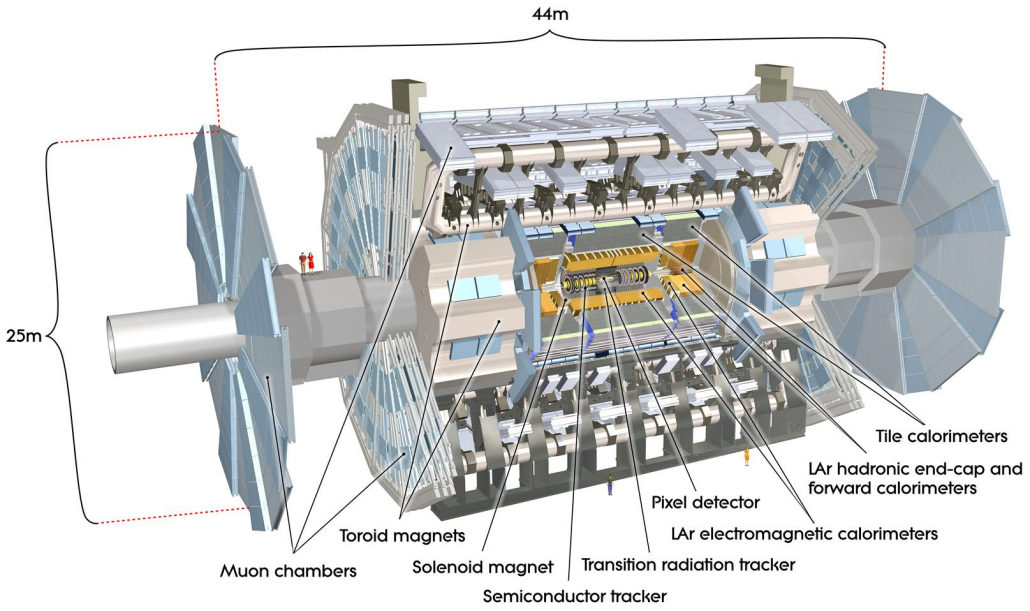


Figure 2.5.: The ATLAS detector overview [86].

The ATLAS detector is the outcome of the collaboration of over 3000 scientists from over 35 countries. This 44 m long and 25 m high cylinder design reflects the goals of a general purpose experiment at the LHC and is shown in figure 2.5. To access a large spectrum of the sought physics, ATLAS combines three main blocks, each of them targeting specific measurements:

- The **Inner Detector (tracker)** described in section 2.2.1 spots the path of charged particles while they cross each layer of this detector and measure their momentum.
- The **electromagnetic (EM) and hadronic calorimeters** described in section 2.2.2 provoke the showering of incoming particles and measure their energy.
- The **muon chambers** embedded in the **toroid magnet** described in section 2.2.3 reveal muons escaping inner parts of the detector and measure their momentum.

The coordinate system used in ATLAS is described thereafter and is used in this report. Its origin is set at the nominal interaction point in the center of the detector and the beam pipe is taken as the z-axis. The transverse plane is then defined by the x- and y-axis which points to the center of the LHC ring and towards the sky, respectively. The ATLAS geometry around the LHC pipe leads to a natural cylindrical coordinate system (r, ϕ, θ). The azimuthal angle ϕ measures the angular distance to the x-axis in the transverse plane and θ is the polar angle with respect to the $z-$axis. The *rapidity* y, or for ultra-relativistic particles ($E \gg mc^2$) the *pseudo-rapidity* η are usually used instead of the polar angle.

For a particle of energy E and momentum $\vec{p}$ these quantities are defined by:

$$y = \frac{1}{2}\log\left(\frac{E+p_z}{E-p_z}\right) \quad \xrightarrow{E \gg mc^2} \quad \eta = -\frac{1}{2}\log\left(\tan\frac{\theta}{2}\right) \tag{2.2}$$

The angular distance in the $\phi - \eta$ plane is defined as $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$.

The wide range of physics targeted by the ATLAS experiment requires a high resolution of both the energy and the momentum in the transverse plane p_{T} of incoming objects. Moreover to capture the maximum of information from the detector a full azimuthal coverage and a large η acceptance are required. It is worth mentioning that these requirements are effective in a large range of energy and transverse momentum, from of few GeV to measure the Higgs boson mass to a few TeV for searches of heavy BSM particle decays. The main requirements for the design of ATLAS are shown in table 2.2.

Detector part	Resolution	η coverage	
		Measurement	Trigger
Inner tracking	$\sigma_{p_T}/p_T = 0.05\%\ p_T \oplus 1\%$	± 2.5	
EM calorimetry	$\sigma_E/E = 10\%/\sqrt{E} \oplus 0.7\%$	± 3.2	± 2.5
Hadronic calorimetry			
barrel and end-cap	$\sigma_E/E = 50\%/\sqrt{E} \oplus 3\%$	± 3.2	± 3.2
forward	$\sigma_E/E = 100\%/\sqrt{E} \oplus 10\%$	$3.1 < \lvert\eta\rvert < 4.9$	$3.1 < \lvert\eta\rvert < 4.9$
Muon spectrometer	$\sigma_{p_T}/p_T = 10\%$ at $p_T = 1$ TeV	± 2.7	± 2.4

Table 2.2.: Design performance of the ATLAS sub-detectors [86].

2.2.1. The inner detector

The innermost part of the detector is devoted to the reconstruction of the path of charged particles (*tracks*) with their momentum as well as the reconstruction of interaction vertices and the identification of electrons. These requirements are fulfilled by the precise measurement of tracks offered by the combination of the *pixel* and *Semi-Conductor Tracker (SCT)* near the interaction point with a *Transition Radiation Tracker (TRT)* at larger radii. The inner detector is surrounded by the central solenoid generating a 2 T magnetic field. The bending of tracks under this magnetic field is used to extract the momentum of charged particles.
The layout of the Inner Detector (ID) and the design architecture are shown in figure 2.6. The three ID sub-detectors are divided in two regions. The barrel, at low η, is composed of concentric cylinders around the beam axis while in the high η region the end-caps are arranged in disks perpendicular to the beam axis. This configuration offers a maximum η coverage, up to $|\eta| < 2.5$ for the pixels and the SCT and up to $|\eta| < 2.0$ for the TRT and covers the full ϕ range. Beside the coverage, the barrel and end-cap designs minimize the material volume. It is indeed crucial that particles escape the inner detector and avoid multiple scattering that would reduce the precision of the measurements of the position and momentum in the ID and energy resolution in the calorimeters. The *radiation length*[a] of the inner detector is shown in figure 2.7. Beside two peaks around $|\eta| \simeq 1.5$ and $|\eta| \simeq 3$ due to the

[a] The radiation length measure the absorption power of a material through electromagnetic processes. It is defined as the distance at which an electron with $E > 10$ MeV keeps only $1/e \sim 37\%$ of its energy the rest being lost through bremsstrahlung radiation. For photons the radiation length correspond to 7/9 of the mean free path for e^+e^- pair production.

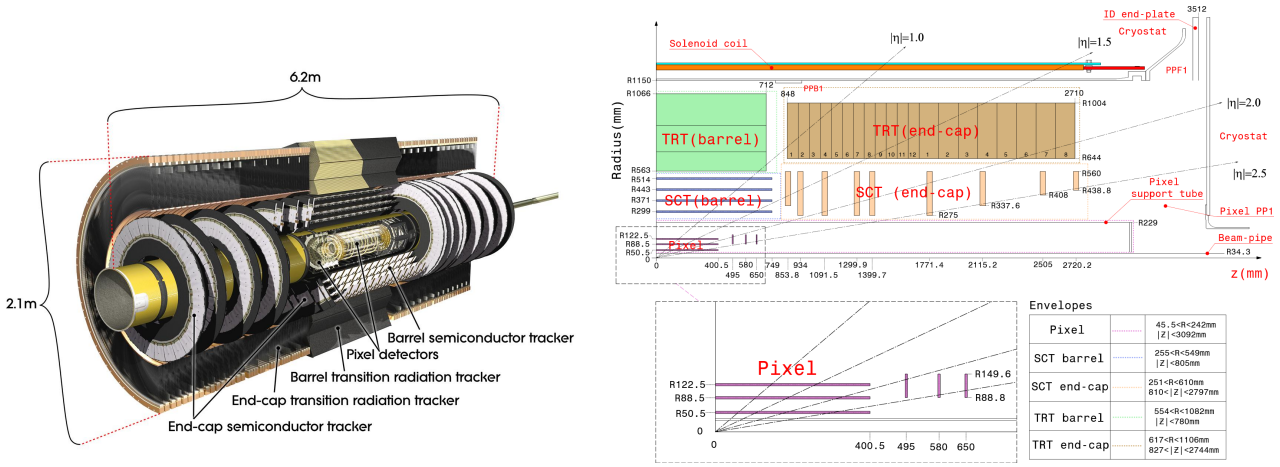


Figure 2.6.: (left) Layout of the ATLAS inner detector. (right) Quarter section of the inner detector (r, z)-plane with the detector element positions. Taken from [86].

barrel and end-cap end-plates the radiation length is dominated by the TRT in the central region of the inner detector and by the pixel support tube at larger η. The contribution from the new Insertable-B-Layer (see the pixel detector paragraph 2.2.1) is not included but is found negligible in the region where sensors are placed (in terms of X_0 the IBL is twice as light as the first pixel layer).

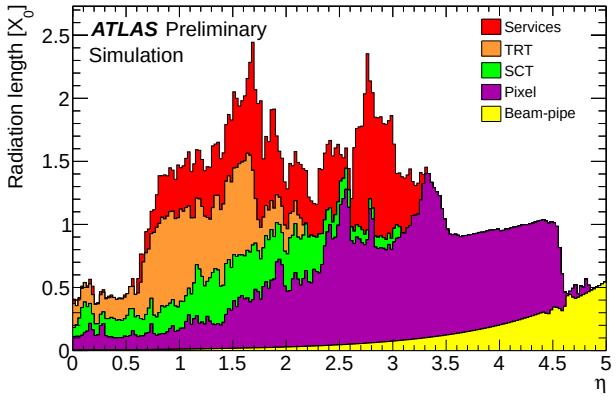
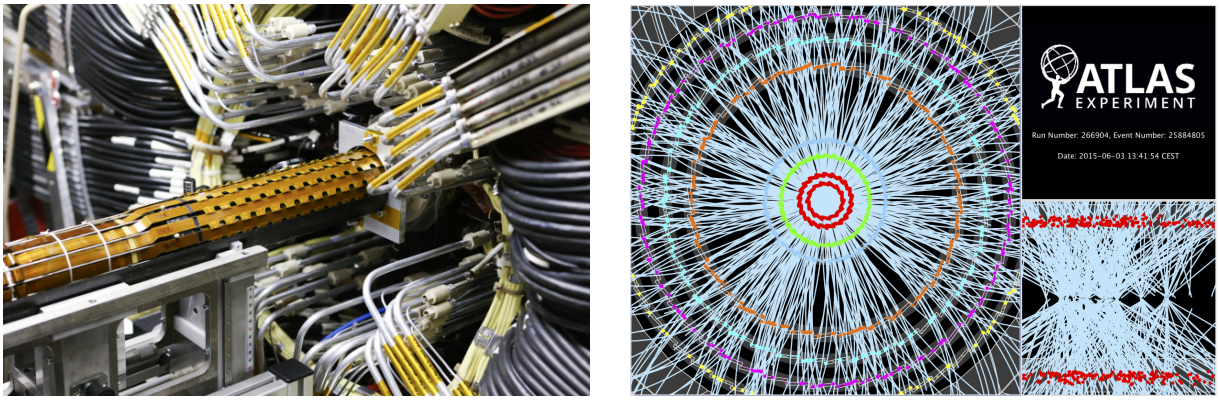


Figure 2.7.: Cumulative material thickness of the ATLAS inner detector components in terms of radiation length X_0 as a function of $|\eta|$ and averaged over ϕ [87].

The pixel detector

The pixel detector is the closest device to the interaction point (IP). Its high spatial resolution provides high precision measurement of the tracks trajectory at the vicinity of the IP. It thus offers the best precision for the recognition of displaced vertices which is essential for algorithms based on the identification of long-lived particles such as b-tagging (see chapter 3). The pixel detector original design includes three layers and was upgraded during the long shutdown after Run 1 with a fourth layer inserted at a closer point to the beam axis (the *Insertable-B-Layer (IBL)* shown in figure 2.8(a)). Figure 2.8(b) highlights the challenges for the inner detector with an event display showing the extreme concentration of particles around the IP.



(a) (b)

Figure 2.8.: (left) Picture of the Insertable-B-Layer insertion in the ATLAS inner detector [88]. (right) Event display zoomed on the inner detector and showing the particle hits in the pixel and SCT detector as well as the reconstructed tracks for a simulated event with Run 2 conditions [89].

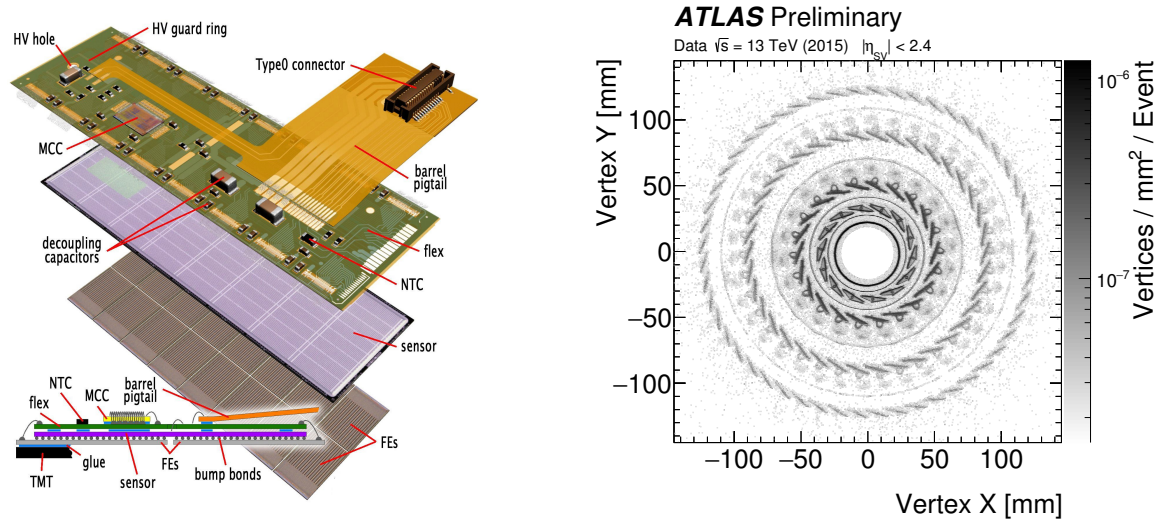
The pixel detector is based on the silicon semi-conductor technology. Each pixel collects the ionization charge deposit of charged particles crossing in a p-n junction. The fast transition of the n-bulk to a p-bulk is compensated by oxygen n-doping on the back-side and n^+ implants on the read-out for enhanced charge collection. 47232 of these pixels form the sensor which is bump-bounded to 16 front-end chips with 2880 read-out channels each. This assembly is called a *module* and is shown in figure 2.9(a). 90% of the pixels have their size constrained by the front-end chip pitch. Others are longer to allow some free space between adjacent chips. The *staves* are composed of 13 modules aligned along the beam axis. Each layer is then composed of 14, 22, 38, 52 staves from the innermost to the outermost with a tilt in ϕ to obtain the cylindrical shape and ensure a full azimuthal coverage. Figure 2.9(b) shows a cross-section in the transverse plane of the pixel layers illustrating its geometrical design and the staves assembly. This illustration uses reconstructed tracks to identify vertices compatible with material interactions. All vertices are then projected on the transverse plane.

The pixel detector parameters are summarized in table 2.3. The inner detector fulfills the $|\eta|$ coverage requirements and provides a very good resolution of the hit position, in particular the IBL which is closer to the IP and has smaller pixels.

Barrel layer:	r (mm)	N(modules)	Pixel size (μm^2)	Intrinsic resolution (μm^2)
Insertable-B-Layer	33.2	280	50×250	$8(r \cdot \phi)$ $40(z)$
B-Layer	50.5	280	50×400	$10(r \cdot \phi)$ $115(z)$
Layer-1 and -2	88.5, 122.5	280	50×400	$10(r \cdot \phi)$ $115(z)$
End-cap:	z (mm)	N(modules)	Pixel size (μm^2)	Intrinsic resolution (μm^2)
Disk $\times$ 3	495, 580, 650	48	50×400	$10(r \cdot \phi)$ $115(R)$

Table 2.3.: Parameters and intrinsic resolution of the inner detector components in the barrel and end-cap regions.

To maintain a very high performance over time and reduce radiation damage, the old layers of the pixel detector are cooled down to $-10\,^{\circ}$C with C_3F_8 gas. This cooling system is chosen for its resistance to radiations. To avoid the accumulation of humidity the pixels are embedded in N_2 gas.



(a) (b)

Figure 2.9.: The ATLAS pixels module layout on the left [86]. The right plot illustrates the pixel detector using reconstructed vertices corresponding to material interaction in $\sqrt{s} = 13$ TeV data collected with the ATLAS detector [90]. It highlights both the geometry of the pixels and the layout of the staves.

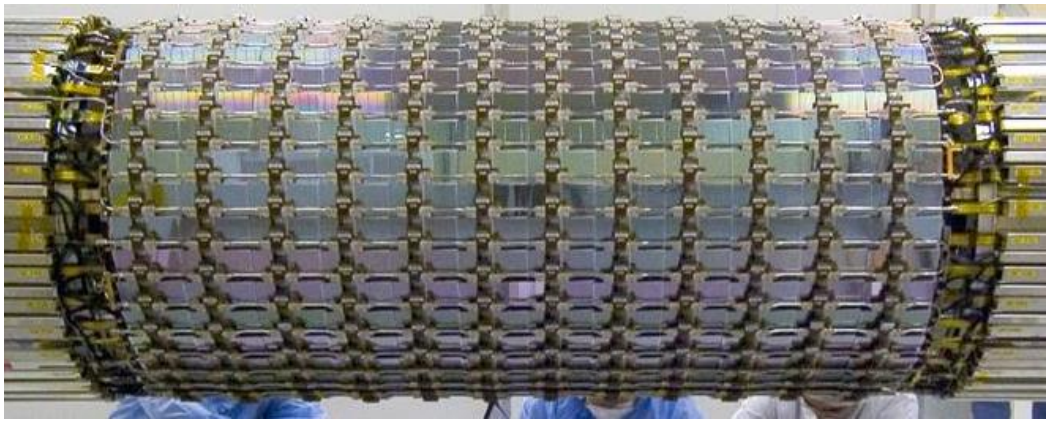
The IBL benefits from a cooling pipe with CO_2 gas which gives the same performance as for the other layers but reduces the radiation length of this layer.

The Semi-Conductor Tracker

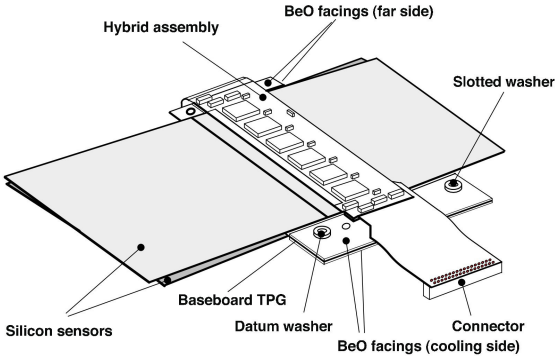
The Semi-Conductor Tracker adds four barrel layers and nine end-cap disks in the transition volume between the pixel detector and the TRT. It contributes to the high precision measurement of track impact parameters combined with the pixel detector and to the high curvature resolution with the TRT.

The SCT detection system is similar to the pixel detector. 768 p-n junctions strips with a 80 μm pitch coupled with their readout each form rectangular barrel sensor. Sensors are paired in a daisy-chain of $\sim$ 12 cm and two pairs are used to build a module, one on the top and one on the bottom. A tilt of ± 20 mrad around the geometrical center of one of the sensor pairs is introduced to allow a 2D measurement of particle hits. The layout of the SCT barrel detector and its 2112 modules is shown in figure 2.10 together with the module assembly. The same strategy is used to build the end-cap sensors but favoring a trapezoidal geometry with strip pitch of 56.9 μm to 90.4 μm. A total of 1976 modules compose the nine disks of each of the end-caps. With this design the SCT detector achieves an intrinsic accuracy of 17 μm in $r\phi$ and 580 μm in z for the barrel and 17×580 μm^2 in $r\phi \times r$ for the disks.

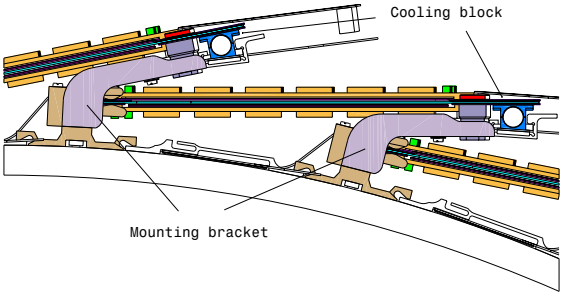
The SCT is cooled down to $-7\,^{\circ}$C using C_3F_8 circulating in cooling pipes attached to each module.



(a)



(b)



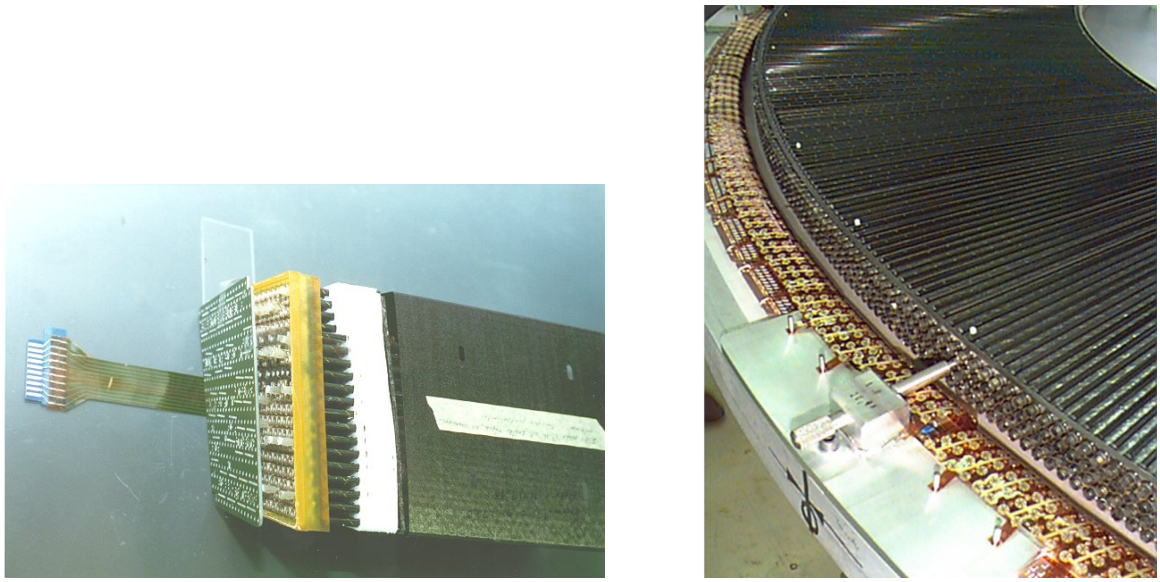
(c)

Figure 2.10.: (top) Picture of the ATLAS SCT. (bottom left) Drawing of the SCT module showing its component. (bottom right) Mounting brackets on the SCT cylinders. Taken from [86].

The Transition Radiation Tracker

The TRT is designed for precise track curvature measurement rather than high hit position resolution. It adopts a different technology than the previous layers. Its fundamental elements are 4 mm diameter wide *drift tubes* commonly called *straws*. The straw surface acts as the cathode and the anode is a 31 μm diameter tungsten wire plated with 0.5 to 0.7 μm of gold. Straws are filled with a gas mixture of 70% Xe, 27% CO_2 and 3% O_2. A charged particle entering the gas will ionize it and the charge will drift to the closest extremity of the straw where it will reach the electronics.

In the barrel straws are 144 cm long. Each module is made of up to 73 layers of straws interleaved with polypropylene fibers (fig 2.11(a)). These modules are then arranged to form a cylinder with the straw parallel to the beam axis. In the end-caps, straws are 37 cm long and 160 straws are interleaved in a plane forming the end-cap disks (fig 2.11(b)).

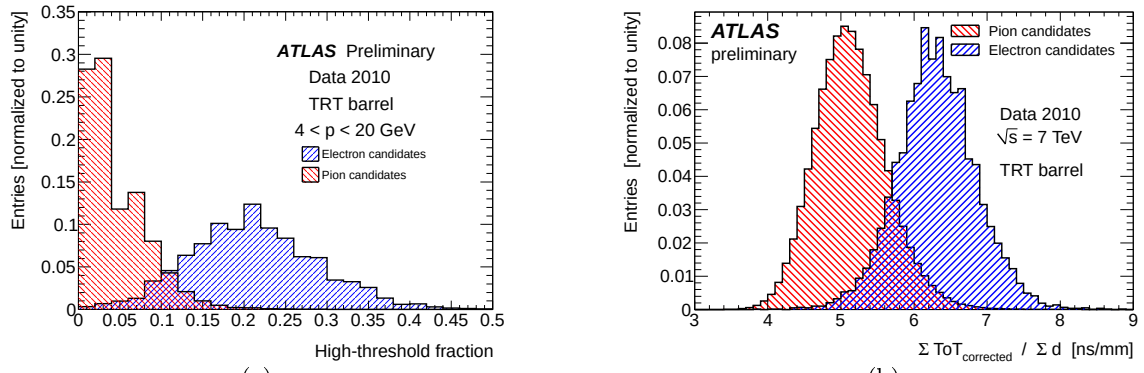


(a) (b)

Figure 2.11.: Photography of the ATLAS TRT barrel module (left) and end-cap disk (right) showing the straw layout in these two regions [86].

The TRT does not offer any measurement in z. However it offers typically 36 (22) consecutive measurements of the charged particle path in the barrel (end-cap) with an intrinsic accuracy in $r\phi$ of $\sim$ 130 μm (driven by the drift time). These consecutive and numerous hits strongly enhance the p_{T} resolution of the tracks.

The TRT straw signal can also be used directly for particle identification, in particular electrons. In facts, high thresholds on the signal energy provide discrimination between pions and electrons as shown in figure 2.12(a). The *Time-over-Threshold (ToT)* of the straw response can also be used to identify electrons as shown in figure 2.12(b). These results and more can be found in ref [91].



(a) (b)

Figure 2.12.: Particle identification based on the TRT as a standalone detector. The left plot shows the electron identification based on the energy of the straw signal output. The right plot shows the discrimination of pions and electrons using the time-over-threshold on the TRT response divided by the transverse track path length inside the straw. Taken from [91]

The inner detector alignment

The tracking performance strongly depends on the quality of the inner detector alignment. In order to correct the mis-alignment, each component of the ID can be moved in its local frame[b] by translations and rotations around each axis. Several levels of alignment are performed:

- level 1: Alignment of the IBL, pixel detector, SCT and TRT as four standalone blocks.
- level 2: Layers and disks of each detector are aligned separately. In the case of the TRT, layers are made of 32 modules each.
- level 2.7: Alignment of each stave individually.
- level 3: Allow further alignment of each module of the IBL, pixel, SCT detectors and of each TRT straw. This last step represents in total 701 696 degrees of freedom.

The inner detector alignment is performed using muons from cosmic rays and collision data collected in 2015 [92]. The TRT is used as a reference and kept fixed. The detector components are aligned using successively each level in increasing order of precision. Once a satisfying alignment is obtained the procedure is stopped for this layer. Figure 2.13 compares the obtained alignment with 2015 data to the optimal configuration from simulation. A great improvement is obtained after adding more data and the latest alignment approaches the optimal configuration.

[b] The local frame of a device is a right-handed reference frame (x, y, z). For the global device the local frame coincides with the ATLAS frame. For the TRT modules, the y-axis defines the wire and the x-axis is orthogonal to the y-axis and the radial direction. z is then orthogonal to the (x, y) plane. For modules and staves of the pixel and SCT, the z direction is defined as the orthogonal direction to the sensor. The (x, y) plane defines the sensor plane with the x-axis pointing towards the most sensitive direction of the device (shorter pitch for the pixel, perpendicular to the strip orientation for the SCT).

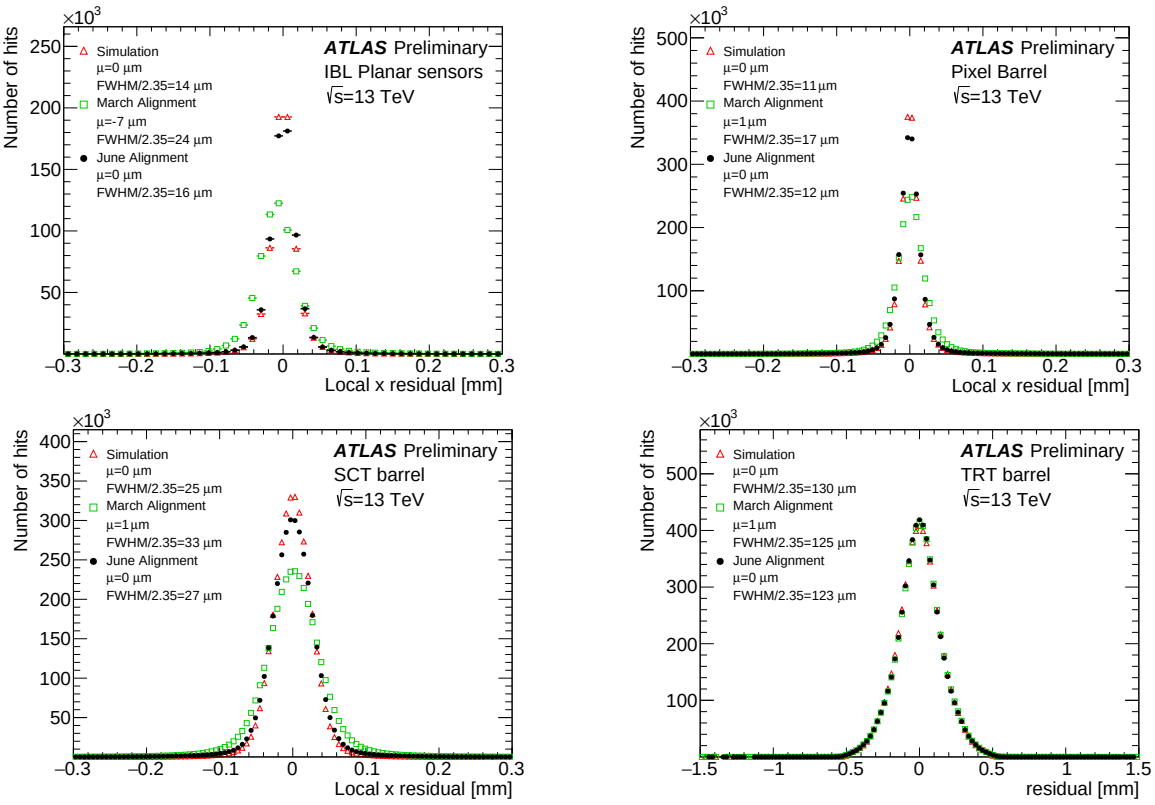


Figure 2.13.: Distance from the extrapolated track position in a given detector element to the hit recorded in the same element in the local-x-axis of the Insertable-B-Layer (top-left), pixel (top-right), SCT (bottom-left) and TRT (bottom right) detectors. The distribution obtained from simulated data with perfect alignment is compared to distribution from $\sqrt{s} = 13$ TeV data using the alignments performed in March and June 2015 [92].

2.2.2. Calorimeters

The calorimeters are made to stop incident particles, other than muons and neutrinos, and precisely measure their energy. As mentioned before the calorimeters are designed to give a high energy resolution for particles with transverse momentum from a few GeV up to several TeV. The search of the Higgs boson through WW and ZZ fusion involves forward jets which require to extend the η coverage compared to the tracker. Moreover the presence of stable supersymmetric particles which interact very weakly with the detector require a precise measurement of the *missing transverse energy* $E_{\mathrm{T}}^{\mathrm{miss}}$ (also called MET, details are found in section 2.4.6) for which the hermeticity of the detector is vital.

The ATLAS calorimetry is divided in three main parts: the electromagnetic calorimeter, the hadronic calorimeter and the forward calorimeter which extends the coverage to $|\eta| < 4.9$. The overall structure of the calorimeters is shown in figure 2.14.

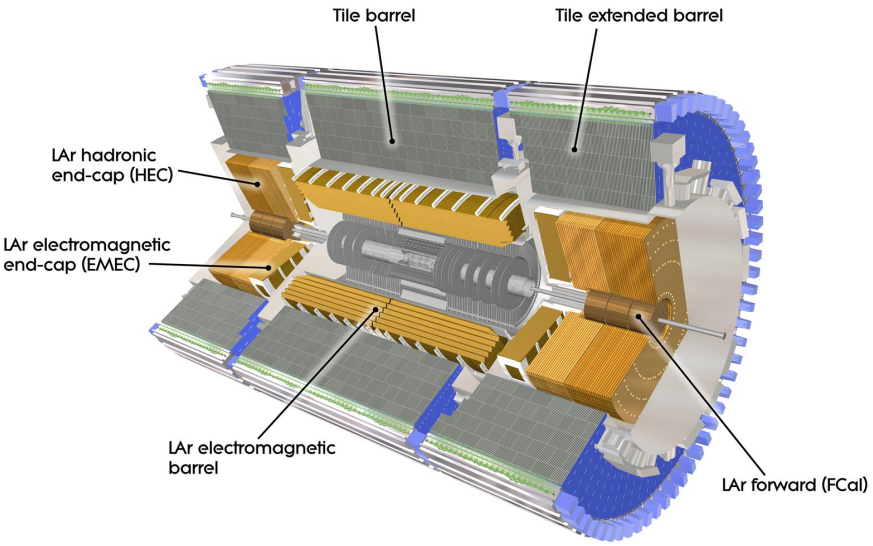
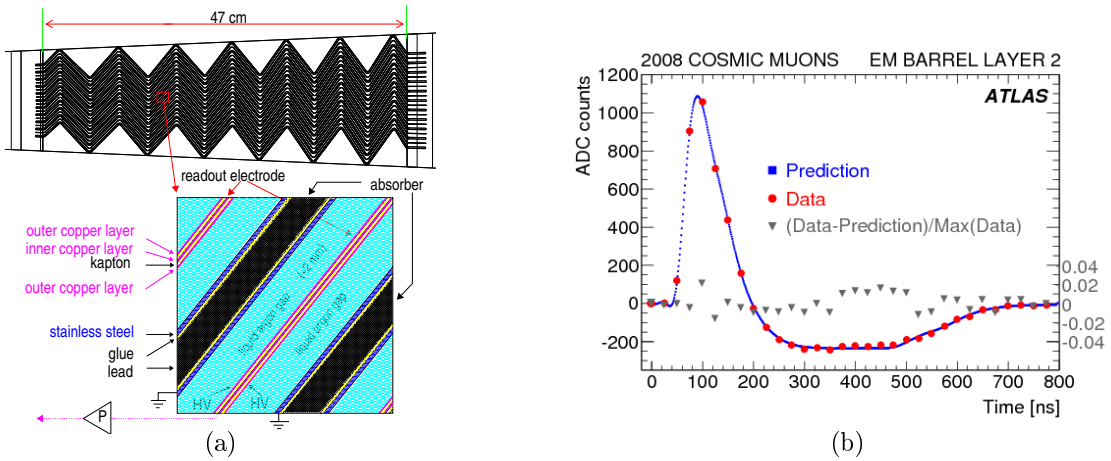


Figure 2.14.: Layout of the ATLAS calorimeters [86].

Electromagnetic calorimeter

The EM calorimeter is composed of a barrel covering the region up to $|\eta| < 2.5$ and two end-caps in the region $2.5 < |\eta| < 3.2$. An accordion geometry allows a full ϕ coverage. Figure 2.15(a) presents the accordion layers and their components. A *lead and Liquid-Argon (LAr)* detector with kapton electrodes is chosen for the EM calorimeter. Lead plate absorbers are 1.53 mm (1.13 mm) thick at $|\eta| < 0.8$ ($|\eta| > 0.8$) and trigger the electromagnetic showering of particles. Charges are deposited in liquid argon and drifted to a kapton electrode at 2.1 mm from the absorber. The kapton electrode is made of three plates separated by a polyamide sheet. The two outer layers are connected to a high voltage potential and induce the charge drift in the LAr while the inner one reads the signal and sends it to the electronics. While this technology allows to absorb incoming particles and measure the energy deposit, it suffers from a relatively long drift time. The EM calorimeter outputs triangular signals stretched over nearly 600 ns. The output signal is shaped to give the ~ 250 ns long signal shown in figure 2.15(b). The EM signal is sampled at the bunch crossing frequency (each 25 ns) and kept in a pipeline for ~ 3.6 μs.

Three layers with different granularity form the EM calorimeter as shown in figure 2.16. The first



(a) (b)

Figure 2.15.: (left) The electromagnetic calorimeter accordion geometry and the Lead Liquid-Argon technology [93]. (right) Shaped output signal of the electromagnetic calorimeter as a function of time [94].

layer is nearly 4.3 long in radiation length and offers a very high η granularity $\Delta\eta = 0.0031$. The second layer offers a coarser η granularity of 0.025 but a higher granularity in ϕ with $\Delta\phi = 0.0245$. This layer is the longest in the barrel and measure $\sim$ 16 radiation lengths. For shower tails of high energy particles a third layer is installed with a two times coarser η granularity.

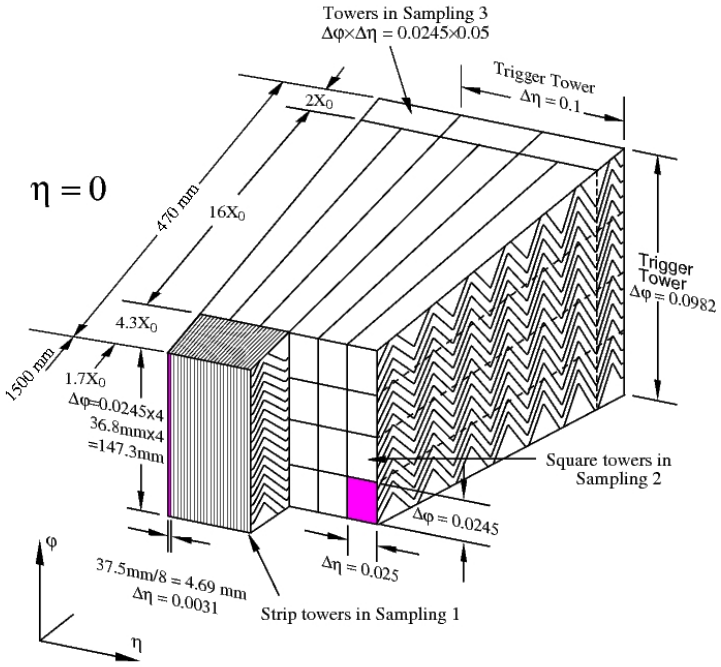
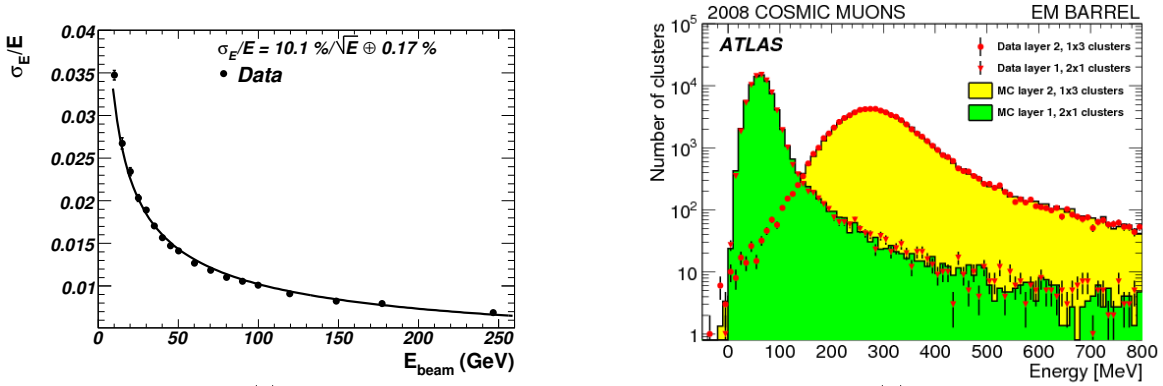


Figure 2.16.: Sketch of the barrel module divided in layers and cells [86]. The dimension of each object is also shown.

The design of the calorimeter offers a very good energy resolution as shown in figure 2.17(a). The energy dependent component is found at the design value (see section 2.2) and the constant term is $\sim$ 4 times better than the design requirements. Figure 2.17(b) shows the excellent agreement between

in data and simulation of the energy deposit of cosmic muon rays in adjacent cell clusters.

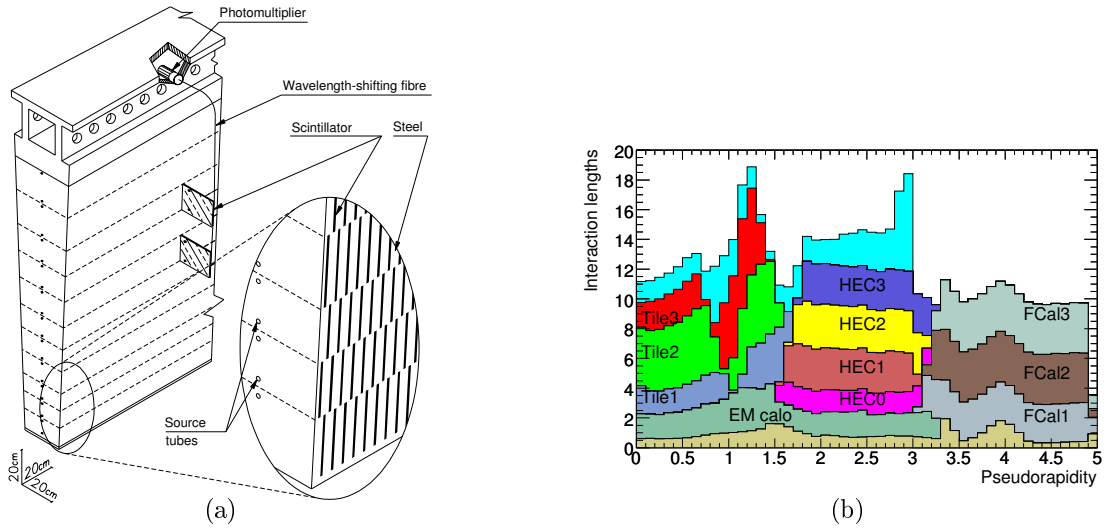


(a) (b)

Figure 2.17.: (left) The ATLAS EM calorimeter measured energy resolution σ_E/E as a function of the beam energy in simulated data together with the best fit value of σ_E/E [86]. (right) Energy of 1×3 and 2×1 clusters in units of $N^{\eta}_{\text{cells}} \times N^{\phi}_{\text{cells}}$ for simulated and observed cosmic muons [94].

Hadronic calorimeter

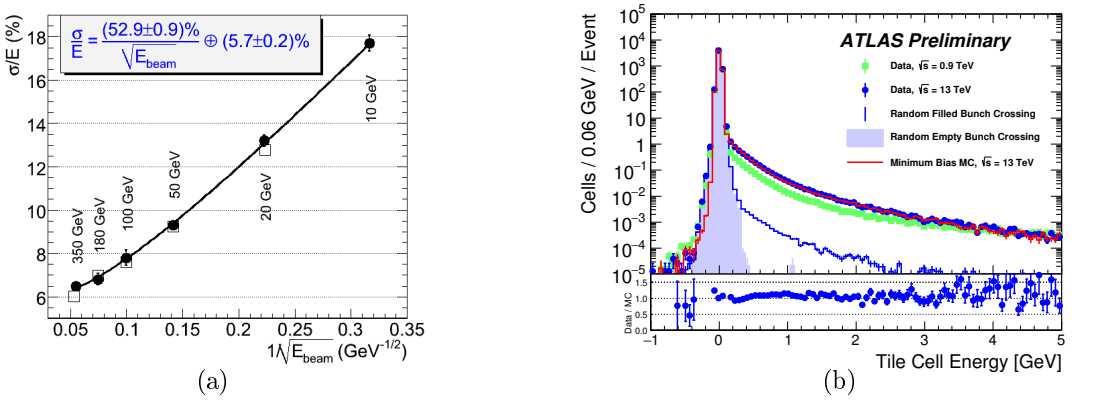
The hadronic calorimeter is designed for hadronic shower energy measurement. It is composed of three items, the *tile barrel* and *tile extended barrel* for the region $|\eta| < 1.7$ and two endcaps covering the region $1.5 < |\eta| < 3.2$. The end-cap uses the LAr technology described in section 2.2.2 and offers a granularity in $\phi \times \eta$ of 0.1×0.1 for $|\eta|$ below 2.5 and 0.2×0.2 above. The tile components are made of successive scintillating tiles 3 mm thick and separated by 15 mm thick steel absorber plates. 64 of the modules as shown in figure 2.18(a) are deployed along the azimuthal direction offering a granularity of $\phi \times \eta = 0.1 \times 0.1$. The interaction length of the hadronic calorimeter is shown in 2.18(b).



(a) (b)

Figure 2.18.: (left) Drawing of the ATLAS hadronic calorimeter tile module. (right) Cumulative material thickness of the ATLAS calorimeter components in terms of interaction length as a function of $|\eta|$ and averaged over ϕ. Taken from [86].

The hadronic calorimeter energy resolution is close to the design requirements (see figure 2.19(a)) and the agreement between expected and observed cell energy deposition is very good for both $\sqrt{s} = 0.9$ TeV and $\sqrt{s} = 13$ TeV data (see figure 2.19(b)).



(a) (b)

Figure 2.19.: (left) Fraction of energy for pions σ_E/E as a function of the beam energy. Simulated events (open squares) are compared to test beam data (full circles) [95]. (right) Distribution of the energy deposit in tile cells for $\sqrt{s} = 0.9$, 13 TeV data overlaid with the random filled or empty bunch crossing and the minimum bias simulation [96].

The forward calorimeters

The coverage of the calorimeters is extended to $3.1 < |\eta| < 4.9$ by the forward calorimeter (FCAL). This detector allows the measurement of forward particle production, it also reduces the background radiations on the muon spectrometer.

The FCAL is divided in three modules based on LAr technologies:

- The **EM FCAL** is the first component reached by incoming particles. It uses copper absorber for optimal resolution and heat removal.
- Two **hadronic FCAL** come after. They both use tungsten absorbers to avoid lateral spread of hadronic shower.

The three FCALs are composed of tubes parallel to the beam pipe. The electrodes are implemented in the tube as small diameter rods, allowing thin LAr gaps. This design is motivated by the high η covered by the FCAL at ~ 4.7 m from the interaction point which exposes the FCAL to high radiation fluxes.

2.2.3. The muon spectrometer

The *muon spectrometer (MS)* is made to spot the crossing of muons and measure their bending in the magnetic field generated by the toroid magnet. Since the toroid magnet provides a bending power in the (R, z) plane, the MS is designed to provide precise measurements along the η parameter.

The overview of the MS is presented in section 2.20. Both the barrel and the end-caps are composed of three precision muon chambers layers at a radius of 5, 7.5 and 10 m for the barrel chambers and $z = 7.4$, 10.8 and 21.5 m for the end-cap. Up to $|\eta| = 2.7$ (2.0 for the inner-most end-cap layer) the chambers are made of *Monitored Drift Tubes (MDT)*, themselves composed of three to eight layers of drift tubes. A gas mixture of Ar (97%) and CO_2 (3%) goes through the tubes. Deposited charges in

the gas mixture drift to a 50 μm diameter wide wire formed of tungsten and rhenium. MDT's achieve a mean resolution of 80 μm in the η direction.

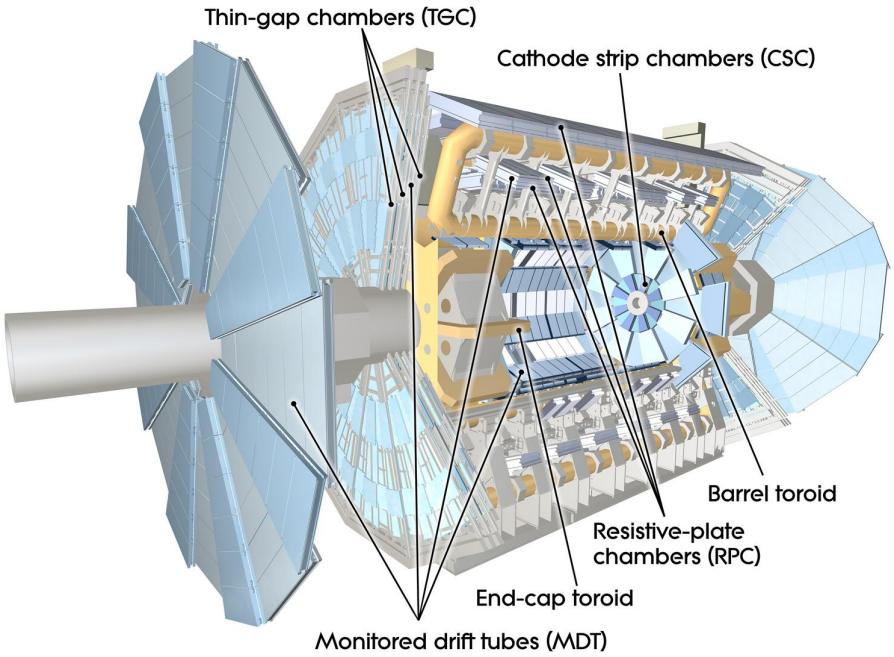
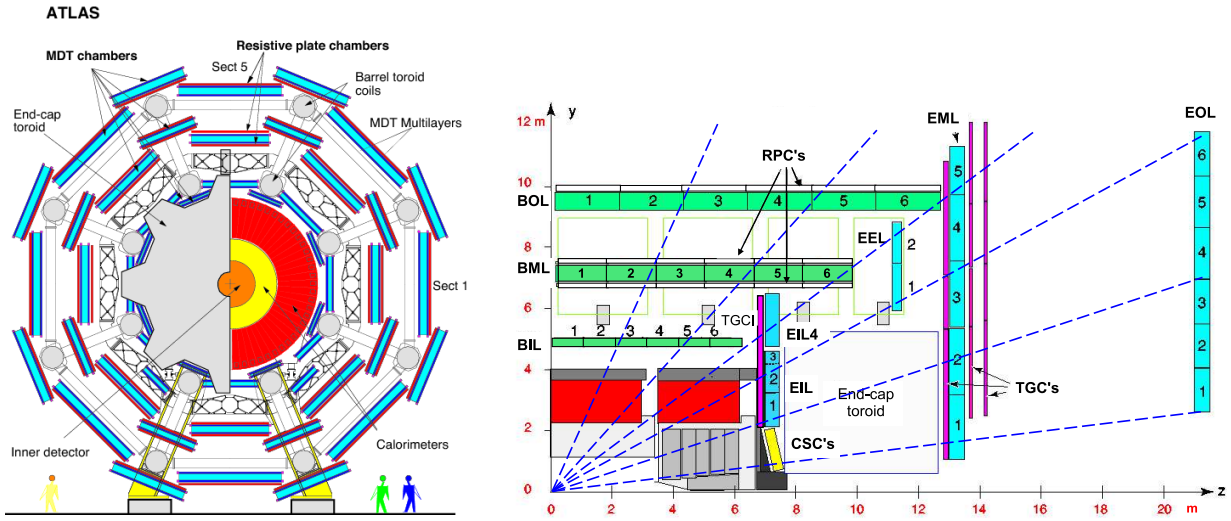


Figure 2.20.: The ATLAS muon chambers and toroidal magnets layout [86].

The inner-most layer of the end-cap ($2.0 < |\eta| < 2.7$) is made of *Cathode-Strip Chambers (CSC)*. The CSCs are multi-wire proportional chambers with cathode planes segmented into strips in orthogonal directions. They offer a higher resolution in the bending plane (~ 40 μm) than the MDTs and can work at higher rates to resolve the higher radiation at the inner-most point of the MS.

The MS is also required to provide fast trigger information of muons tracks. Therefore ATLAS is equipped with *Resistive Plate Chambers (RPC)* in the barrel and *Thin Gap Chambers (TGC)* in the end-cap. These systems measure both the η and ϕ hit coordinates and deliver a signal in 15 to 25 ns, i.e. faster than the bunch-crossing frequency.

The detailed layout of the MS is shown in figure 2.21. A high precision alignment of the detector components is required to achieve the required resolution of the muon tracks. MDTs are thus complemented with an optical alignment system which is monitored via track-based algorithms and continuously corrects the position or deformations in the MDT chambers.



(a) (b)

Figure 2.21.: (left) The x-y projection of the muon spectrometer [97]. (right) Cross-section of the muon system in a plane containing the beam axis (bending plane). Infinite-momentum muons would propagate along straight trajectories which are illustrated by the dashed lines and typically traverse three muon chambers [86].

2.2.4. Triggering data

The very high frequency of bunch-crossing (40 MHz) does not allow to store all data even with the most recent recording devices. To reduce the rate the ATLAS detector is equipped with a *trigger* system that selects events of interest for physics, detector commissioning and performance studies. The trigger strategy has been very successful for Run 1 data taking. However the trigger architecture is re-visited to compensate for the higher energy and collision frequency in Run 2.

In Run 1 the trigger system consisted of a hardware based *level 1* (L1) followed by software based *level 2* trigger, used to refine level 1 decisions, and *event filter*, combining the incoming information with offline reconstruction algorithms to keep or reject the event. In Run 2 the level 2 and event filter are combined in the *high level trigger* (HLT). This architecture allows to decrease the data input frequency to $\sim$ 100 kHz after level 1 and to 1.5 kHz after the HLT. Each level can access various detector information:

- The level 1 triggers can access the information of the built-in triggers of the calorimeters (L1Calo) and the MS (L1Muon) to decide on the event quality. It can also compile the information of the L1Calo and/or the L1Muon in the topological trigger module (L1Topo). This last module can build composite objects and extract topological information at the event level. All the required information is then sent to the Central Trigger Processing which takes the L1 decision of keeping or rejecting the events.

- The HLT can access all the detector read-out systems and uses complex reconstruction algorithms such as multivariate discrimination of electrons and photons versus hadrons.

The read-out of the detector components is directly followed by the level 1 triggers which decide to keep or reject an event in at most 2.5 μs. If a level-1 trigger finds a region of interest, the detector read-out is sent to a data collection network. The HLT combines the data collection network information with the level 1 trigger information. It provides a decision in 0.2s on average. If an HLT decides to keep an event, the corresponding data are sent to the event builder and recorded.

2.3. Production of Monte Carlo samples

The complexity of the SM and the high level of precision targeted require a complex simulation procedure. Collision events are generated using the Monte Carlo method (MC). The detector response to the generated particles is simulated either including all the detector elements (*fullsim*) or using only a reduced fast chain flow (*fastsim*). pp collision data and simulated data are then stored in the same format to allow a direct comparison. In this document simulated data are referred to as *predictions* or *MC simulations*.

2.3.1. Event generation

The first step of the simulation is the *event generation*. It describes the event parameters from the incoming protons up to the *stable particles* that fly in the detector. This is done in four main steps: the extraction of the partons from the protons, the *hard scatter*, the *parton shower* and the *hadronisation* which includes the subsequent decay of hadrons.

- The extraction of the partons from the protons is a non-perturbative process and thus cannot be computed explicitly with the SM. *Probability Density Functions (PDF)* extracted from actual measurements are used to compute the fraction of the proton energy taken by the partons and the flavour of the parton extracted from the proton.
- The hard scatter usually called matrix element (ME) is the computation of the Feynman diagrams of interest. It uses a perturbative approach and different order of precision can be used, mainly depending on the complexity of the final state. Some algorithms also generate one or several additional partons to the main process. Algorithms providing the output up to this step are referred to as *MC generator*.
- The showering brings corrections to the final state. Partons are allowed to emit gluons or photons via QCD and QED. The event kinematics is then recomputed including these new partons and the procedure is repeated. When coupled to a generator with additional partons to the main process, the overlap is removed using various methods such as the *ME+PS method* [98].
- Finally the hadronisation is also a non-perturbative process which relies on empirical models tuned to data. In this step partons are combined in hadrons and unstable hadrons are further decayed to stable particles.

The hard scatter is the only step which purely relies on the theoretical predictions and the ability to compute numerically a process with several orders in the expansion of the matrix element (see section 1.1.1). Several MC generators are used in this thesis: Sherpa [99], Madgraph5_aMC@NLO [100], Powheg-Box [101–103]. The showering relies on theoretical predictions tuned to data. MC generators are interfaced with either PYTHIA [104, 105] or Herwig [106, 107] for the showering and hadronisation steps. The Sherpa generator has its own showering and hadronisation model and does not need interfacing. All the information from the event generation is stored in the so called *MC history* and the particles it contains are referred to as *true-*(or *truth-*) *particles*.

2.3.2. Detector simulation

The detector response to stable particles from the event generation is then simulated via fastsim or fullsim. The fullsim is based on the GEANT 4 package. It simulates the interaction of stable particles with the detector components. This step is usually the one that requires the highest CPU time and can

last for several minutes per event for typically several millions of events per sample. The detector response is then simulated by the *digitization* step. Finally the *reconstruction* runs the ATLAS algorithms to produce the physical objects: e.g jets and tracks. This last step is done both on MC simulations and data.

The large amount of time needed for the simulation motivates the use of a fast simulation at the cost of reduced precision. Each of the three sub-detectors simulation times can be reduced. However the main contributor to the simulation time is the shower of particles in the calorimeters. The fastsim thus uses pre-simulated electromagnetic showers of low energy particles to skip the simulation of their interaction with the detector.

In general the fullsim provides a higher precision and is preferred for the main samples of the analyses while the fastsim package allows to produce multiple alternative samples that are compared to choose the nominal sample, or assess theoretical systematic uncertainties.

2.4. Object reconstruction and physics quantities

This section describes the algorithms used in the ATLAS experiment to reconstruct physics objects. Figure 2.22 shows the basic concepts for physics objects identification in the detector.

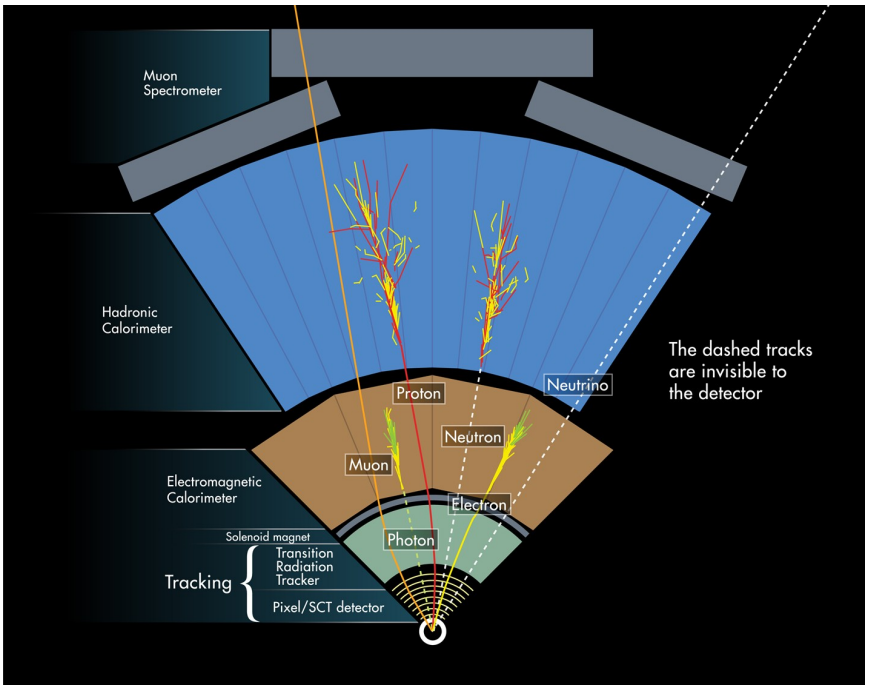


Figure 2.22.: An overview of particle identification in the ATLAS detector [88]. The solid and dashed curves show the tracks of charged and neutral particles. Arising from the interaction region (beam axis), the muon goes through the whole detector while being tracked by the Inner Detector and the Muon Spectrometer. The electron and the photon are caught by mainly the EM calorimeter and can be differentiated from the presence or absence, respectively, of a track pointing to the energy deposit in the calorimeter. The proton and the neutron are trapped by mainly the hadronic calorimeter with and without leaving a track in the ID respectively. The neutrino passes through the entire detector without leaving any signature.

2.4.1. Tracks and primary vertex

The *tracking* [108, 109] and *vertexing* [110] algorithms are both based on the inner detector information (see section 2.2.1). A charged particle path in this sub-detector generates hits in the different layers which are combined to obtain a track. Tracks extrapolation towards the beam axis are then used to reconstruct the primary vertices. The estimated distance of closest approach of tracks to their associated *primary vertex (PV)*, the *Impact Parameter (IP)*, is one of the track parameters. This induces a strong interplay between the two algorithms. This section briefly introduces the tracking and vertexing algorithms which are of key importance for data analyses, in particular the b-tagging described in chapter 3.

Tracking algorithms

In the ATLAS coordinate system, the helices produced by tracks in the magnetic field are characterized by 5 parameters to exploit the full geometry and kinematics of the incoming particles. The parameters are defined in what follows and illustrated in figure 2.23. Most of them involve the point of closest approach to the PV (*perigee*). When the PV is not yet defined, the coordinate origin $\mathcal{O}$ is used to define the perigee:

- **Inverse transverse momentum Q/p_{T}**: is the electric charge divided by the track transverse momentum. This ratio is determined by the curvature radius R_{curv} in the magnetic field B by $Q/p_{\mathrm{T}} = (0.3BR_{\mathrm{curv}})^{-1}$ and the electric charge sign is extracted using the curvature direction.
- **Azimuthal angle ϕ**: is the azimuthal angle of the track $\overrightarrow{p}$ at the perigee. A second azimuthal angle ϕ_0 is determined taking the angle between the x-axis and the vector pointing to the perigee in the transverse plane.
- **Polar angle θ**: is the polar angle of the track $\overrightarrow{p_{\mathrm{T}}}$ at the perigee.
- **Transverse IP d_0**: is the track's distance of closest approach to the PV or $\mathcal{O}$ in the transverse plane. It is defined positive if $\phi_0 - \phi = \frac{\pi}{2}$ and negative if $\phi_0 - \phi = -\frac{\pi}{2}$.
- **Longitudinal IP z_0**: is the track's perigee z coordinate.

The main tracking algorithm adopts an *inside-out* strategy. Hits in the silicon detector are first translated into *space-points*. In the pixel detector the hit position simply corresponds to the pixel position while in the SCT space-points are defined using the hits in the two superimposed sensors of each module. Seeds are then formed of helix trajectories connecting 3 space-points in either the SCT or pixel detector or two space-points from the pixels and one from the SCT. For a higher accuracy, the seed formation can be interfaced with a fast primary vertexing. Space-point pairs are formed instead of three space-points and an inclusive estimation of the primary vertices position, called the beamspot, is computed. The beamspot is then used to constrain the addition of other space-points to the seeds. Seeds are then extrapolated towards the full silicon detector using a recursive combinatorial Kalman filter.

At this point a very high number of track candidates are build with a significant fraction of fake tracks. This is partially resolved using a score-based ranking scheme. The track ranking is based on several inputs. A re-fitting of tracks' parameters is performed with a detailed map of the ID material and the resulting χ^2 measuring the fit goodness is included in the track score. In addition, a bonus is given for each hit associated to the track and a penalty is given if the track contains *holes*[c]. A hit can

[c] Holes are defined as intersections between the fitted track trajectory and active modules of the detector in which no hit is observed.

also be associated to several tracks, such hits are called *shared-hits* and are taken into account when scoring the track. Shared hits are induced either by the presence of a fake track or when the detector granularity is insufficient to resolve close-by particles. A Neural Network (NN) [111] is trained to differentiate between these two cases. Hits satisfying a quality cut on the NN output are identified as coming from multi-particle hits, labelled *split-hits*, while the others are labelled *merged-hits*. Tracks are required to have at most one merged-hit to enter the ranking and proceed to the next steps.

Tracks passing the quality requirements are then extrapolated towards the TRT and associated to corresponding drift-circles built from TRT measurements.

If this procedure allows a reconstruction efficiency of tracks between 70% and 90%, it is not suited for tracks originating from secondary vertices of long-lived particles (Λ or K_S), photon conversions, and material interactions which can be found inside the ID. To account for such topologies an *outside-in* algorithm is implemented after the PV reconstruction. It extends non used TRT drift circles towards the silicon detector to build the tracks.

Finally, track parameters are re-fitted once the primary vertices are reconstructed for enhanced precision of impact parameters.

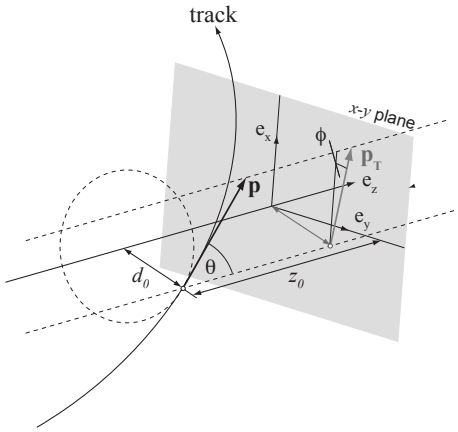


Figure 2.23.: Illustration of the track helix parameters [112].

Figure 2.24 shows the reconstructed tracks IP resolutions $\sigma(d_0)$ and $\sigma(z_0)$ for the 2012 and 2015 datasets. In the transverse plane a resolution of ~ 150 μm at low p_{T} to ~ 20 μm at high p_{T} is obtained and a resolution of ~ 220 μm at low p_{T} to ~ 80 μm at high p_{T} is seen in the longitudinal axis. The gain between the Run 1 and Run 2 configurations is mostly coming from the addition of the IBL. Indeed this last is expected to significantly improve the resolution at low p_{T} thanks to its closer distance to the PV compared to the other layers and smaller pitch of the pixels. At high p_{T}, the resolution is mostly due to the pitch of the pixels with a small contribution from the IBL radius. This explains the large gain in the longitudinal direction where the IBL pitch is reduced compared to the B-Layer and the marginal gain in $\sigma(d_0)$ since the IBL and B-layer have the same pitch in ϕ. Notice that these plots do not include the re-fitting after the primary vertex reconstruction which also benefits from the IBL inclusion. Thus a larger improvement and higher resolutions can be expected.

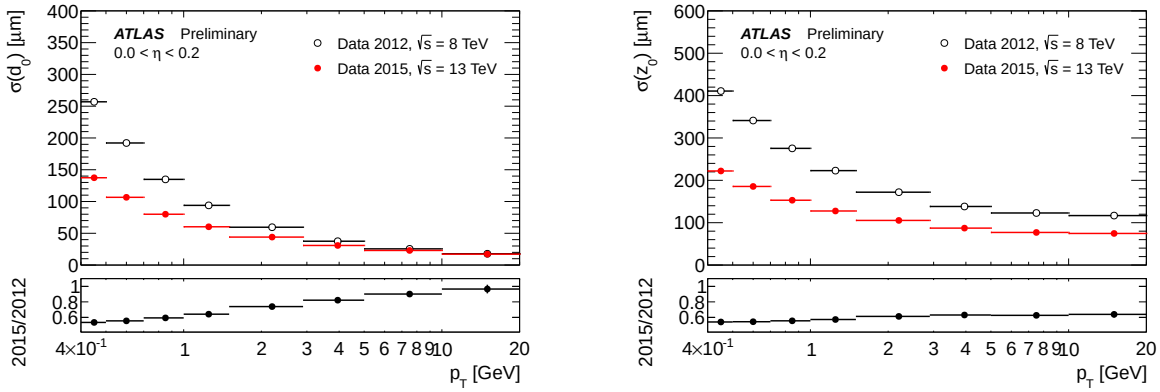


Figure 2.24.: Measured transverse (left) and longitudinal (right) impact parameter resolutions as a function of the transverse momentum comparing the Run 1 (black open circles) and Run 2 (red points) configurations [113]. The main difference between the two configurations is the insertion of the IBL for Run 2. The ratio of Run 1 data to Run 2 data is shown in the lower panel.

Primary vertexing algorithms

The primary vertices are iteratively reconstructed using a selected set of tracks and the beamspot information. In the first step a vertex seed is formed. The x- and y- coordinates of the seed are placed at the center of the beamspot and the z-coordinate is calculated from the mode of the input tracks points of closest approach to the seed. Then the optimal vertex position is obtained from an iterative fit with the seed and reconstructed tracks as input. At each iteration tracks are weighted according to their compatibility with the new seed and the fit is re-launched. Once the vertex is found the weight of each track is recomputed and tracks that are incompatible with the vertex at more than 7σ are removed from this vertex. This two step procedure is then repeated using tracks incompatible with existing vertices until no tracks are left or no vertex can be formed. The resolution of the obtained vertices depends on the number of tracks used. The vertexing algorithms typically achieves a resolution of 30 μm in the (x, y)-plane and of 50 μm in the z-axis.

2.4.2. Muons

The muon reconstruction [114, 115] relies on both the MS (see section 2.2.3) and the ID (see section 2.2.1). Two independent measurements are done in the sub-detectors and then combined giving a very high efficiency of muon track identification and an excellent momentum resolution up to a few TeV.

In the MS, MDT hits are translated into straight line segments per muon chambers and combinatorial search is performed to associate segments. Track candidates are then built fitting together all hits from associated track segments. Some quality cuts (fit quality, number of segments and shared segments) are applied to track candidates to obtain the final list of MS tracks.

The MS tracks are then combined with ID tracks reconstructed by the standard ID tracking algorithms 2.4.1. MS and ID tracks are associated using an outside-in pattern recognition coupled to a complementary inside-out algorithm. A χ^2 test or a full track fit with ID and MS hit information is then performed to obtain the full muon track. Tracks are classified depending on the association outcome:

1. **Combined muons**: muon tracks fall in this category when the association algorithm successfully finds a MS and an ID track. Tracks falling in this category are the most accurate.

2. **Segment-tagged muons**: muon tracks that are not combined muons and are made of ID tracks with at least one additional segment from the MS. This category is mostly populated with low p_{T} muons crossing only the first MS layers or muons falling in MS regions of reduced acceptance.

3. **Calorimeter-tagged muons**: due to a hole in the MS at $|\eta| < 0.1$ ID tracks that are not used in category 1 or 2 and fall in this region are compared with calorimeter energy deposit compatible with minimum ionizing particles.

4. **Standalone or extrapolated muons**: MS tracks that are not associated to any ID information. The track trajectories are extrapolated towards the beam axis taking into account the energy loss in the calorimeters. Most of the standalone muons are in the $2.5 < |\eta| < 2.7$ region which is not covered by the ID.

To discriminate muon tracks coming from prompt muons against muon tracks induced by particles escaping the inner parts of the ATLAS detector an *identification* procedure is applied. Four identification working points are provided to the analyses using basic muon quality cuts:

- **Medium muons**: only combined and standalone muons can pass the medium requirements. Combined muons are further required to have at least 3 hits (1 hit and at most 1 hole) in at least two muon chambers for $|\eta| > 0.1$ ($|\eta| < 0.1$) and standalone muons are only used in the $2.5 < |\eta| < 2.7$ region. A further requirement is added on the ID and MS p_{T} compatibility. Medium muons are used in the analysis presented in this thesis.

- **Loose muons**: in addition to medium muons, segment-tagged and calorimeter-tagged muons in the $|\eta| < 0.1$ region are included in loose muons. This category mainly aims at large muon acceptance for searches of particle decaying in multiple leptons like $H \to 4l$.

- **Tight muons**: this category maximizes the muon purity while losing in efficiency. Tight muons are required to be medium combined muons with enhanced track quality cuts.

- **High p_{T} muons**: targets muons at high p_{T} requiring medium combined muons only in specific MS regions with a high p_{T} resolution.

Figure 2.25 illustrates the outstanding performance of the muon reconstruction and identification. It shows the high muon identification efficiency as well as the invariant mass of opposite-sign muon candidate pairs in events selected by a single muon trigger at > 15 GeV.

To reject muons coming from heavy-flavoured decays, seven *isolation* criteria (or working points) are provided. They depend on a track isolation parameter using the momenta of the muon track and of surrounding tracks, as well as a calorimeter isolation looking at energy deposits in a cone around the muon tracks.

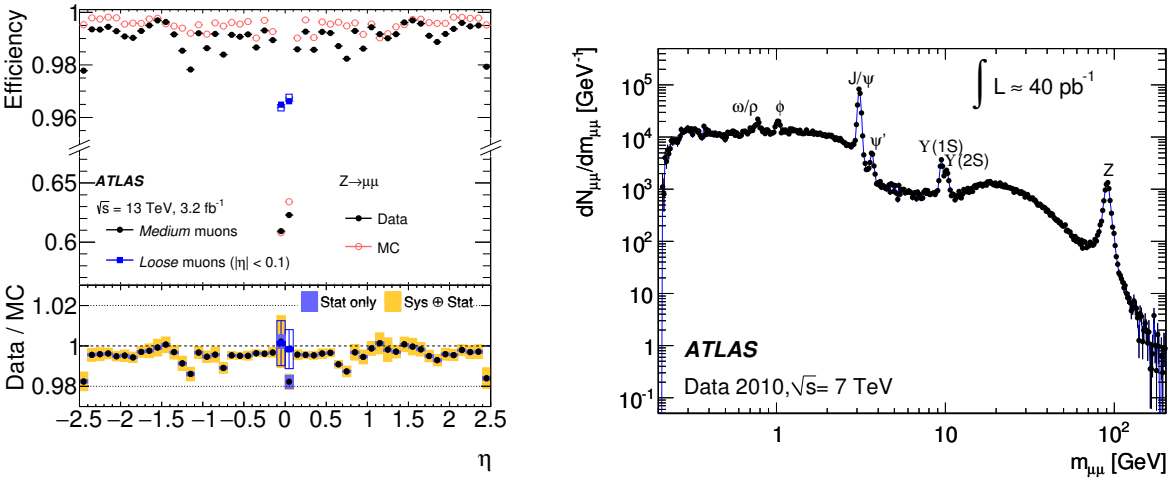


Figure 2.25.: (left) Reconstruction efficiency of muons at the loose and medium identification working points as a function of η comparing Run 2 data and simulations in $Z \to \mu\mu$ events [114]. (right) Distribution of the reconstructed di-muon invariant mass in $\sqrt{s} = 7$ TeV data with the re-observation of known particles [115].

2.4.3. Electrons and photons

Electrons and photons reconstruction algorithms [116] are deeply connected due to the similar signature they have in the EM calorimeter. One of the key ingredient to the electrons and photons reconstruction is the clustering of EM showers [117]. The first step is the search of cluster seeds with a size corresponding to $\eta \times \phi = 3 \times 5$ cells in the middle of the EM calorimeter (25×25 μm^2 squares). A sliding window algorithm scans the full EM calorimeter acceptance and saves cluster seeds if their energy is higher than the detector noise background ($E_{\mathrm{T}} > 2.5$ GeV). The contributions to the seed from all layers are integrated over the r-coordinate to form a tower out of each square.

In parallel the standard ATLAS tracking is extended to account for the larger bremsstrahlung radiation of electrons. Obtained tracks are then matched to EM cluster towers using the η and ϕ distance of the track to the tower barycenter in the EM calorimeter. EM cluster towers with no ID tracks are set as unconverted photon candidates. Cluster towers associated to ID tracks are used as converted photon candidates if the track is compatible with a photon conversion secondary vertex, and set as electron candidates otherwise. Electron and photon candidates are re-formed using towers with enlarged size as shown in table 2.4.

	Cluster size in $\eta \times \phi$ [N(towers)]	
Cluster type	Barrel	End-cap
Electron	3×7	5×5
Un-converted photon	3×5	5×5
Converted photon	3×7	5×5

Table 2.4.: Cluster size given in $N_{\eta}^{\mathrm{towers}} \times N_{\phi}^{\mathrm{towers}}$ for each particle type in the EM calorimeter barrel and end-caps.

Electron identification

The electron identification [118] is based on a multivariate likelihood method discriminating prompt electrons from background-like objects, mainly hadronic jets and photon conversions which are not removed in the cluster-track association step. The likelihood includes shower shape information since isolated electron showers in the EM calorimeter are more collimated than hadronic showers or electrons from photon conversions [119]. It also includes information from the ID such as track quality, variables sensitive to the bremsstrahlung effect and composite variables accounting for the compatibility between the clusters and their associated tracks. Three working points are defined cutting on the likelihood score, the loose, medium and tight working points. The corresponding signal and background efficiencies are shown in figure 2.26.

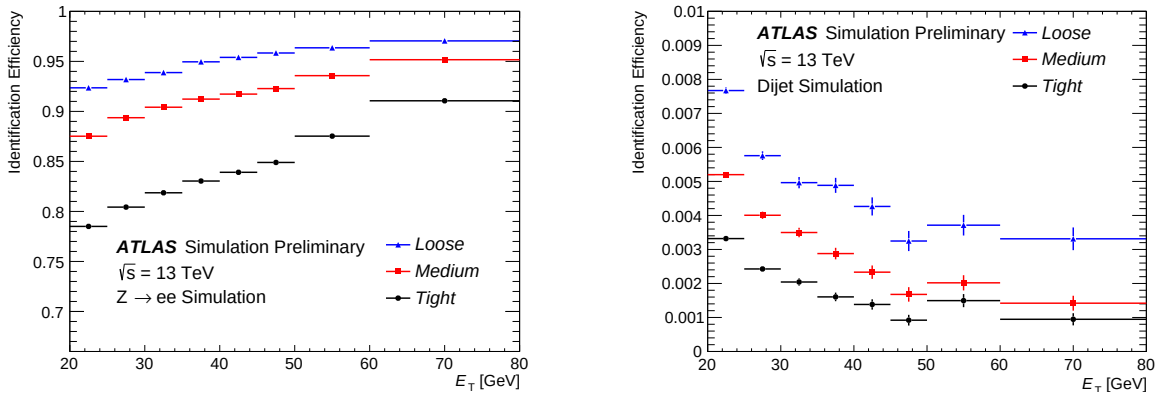


Figure 2.26.: (left) Efficiency to identify prompt electrons in $Z \rightarrow ee$ simulated events. (right) Efficiency to identify fake electrons in multi-jet simulated events. [118].

Furthermore electron isolation criteria are provided to reject electrons coming from heavy-flavour decays similarly to the muon isolation. The isolation is based on both the energy deposit in the EM and hadronic calorimeters in a cone around the cluster and the tracks p_{T} density around the electron track. Several working points are provided using either fixed isolation cuts (LooseTrackOnly, Loose and Tight) or E_{T} dependent cuts (Gradiant and Gradient-Loose) [118].

Photons identification

The photon identification [120] uses the same ingredients as for electron but uses a set of cuts rather than a likelihood. Two working points are provided: loose and tight. The tight working point is separately tuned to differentiate converted and unconverted photons.

2.4.4. Jets

The QCD confinement forces quarks and gluons to hadronize almost instantaneously. *Jets* are collimated sprays of energetic hadrons reconstructed with a dedicated clustering of energy deposit in the calorimeters.

In the ATLAS experiment the standard jet reconstruction [121] is based on the anti-k_t algorithm, which is precisely described in ref [122] and summarized here. k_t algorithms are sequential recombination algorithms, where the acronym k_t refers to the usual label of transverse momenta. Distances

d_{ij} and d_{iB} between particle i and particle j or the beam B are introduced as follows:

$$d_{ij} = \min\left((p_T)_i^{2p}, (p_T)_j^{2p}\right)\frac{\Delta R(i,j)^2}{R^2} \quad , \quad \text{and} \quad d_{iB} = (p_T)_i^{2p} \tag{2.3}$$

where p is an arbitrary parameter and R is a cut-off radius parameter defining an approximate cone size of jets. A negative value of p implies that high energy particles aggregate soft particles. The anti-k_t algorithm illustrated in figure 2.27 is obtained when setting $p = -1$ and iteratively associates the two particles minimizing d_{ij}. The standard cone size used for ATLAS analyses is set to 0.4 but alternative jet collections are built with other sizes, mainly for the study of *boosted jets* originating from composite objects such as top quarks (1 b-quark and 2 other quarks from the W).

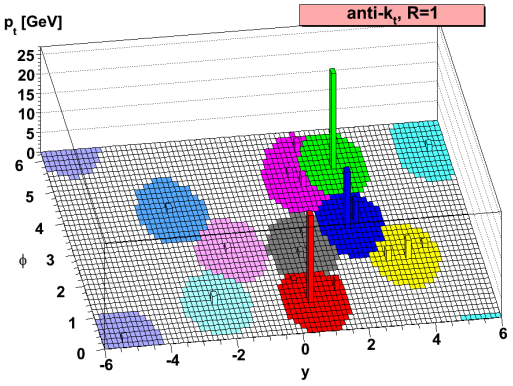


Figure 2.27.: A sample parton-level event, together with many random soft momentum particles (called ghost), clustered with the anti-k_t algorithm [122].

Two main jet collections are used in ATLAS. The *truth-jet* collection is formed from stable truth-particles generated in MC samples. *AntiKt4EMTopo*-jets, called reco-jets in this thesis, are formed from the measurements in the calorimeters and are the most commonly used jets for ATLAS measurements. Connected calorimeter cells measuring a significant signal over the detector noise are merged in topological clusters in both the EM and hadronic calorimeter. Topological clusters are then used as inputs to the anti-k_t algorithms.

The energy and direction of reco-jets are corrected by a EM+JES calibration scheme [121, 123] of the jet p_{T} and η. The first step is an offset pile-up correction derived from in-situ measurements to account for pile-up contribution to the topoclusters. The second step corrects the jet 4-vectors to move the jet origin from the ATLAS detector center to the primary vertex coordinates. Finally the energy and direction of reco-jets are corrected by constants derived from the comparison of the reco-jet kinematic to the one of truth-jet.

Pile-up jets

The *pile-up jets* (jets coming from pile-up interactions) suppression described in ref [124] is essential for proper measurement of the hard scatter process. Since the hadronic calorimeter does not give any information about the origin of clusters the pile-up jet suppression relies on the properties of tracks associated to jets. Pile-up jets are rejected using the jet-vertex-tagger (JVT) discriminant. It is constructed using a 2D likelihood from the two following variables:

- **The corrected-jet-vertex-fraction, corrJVF**: corrJVF accounts for the p_{T} fraction of tracks associated to the jet that come from the hard scatter primary vertex (HS-PV):

$$p_{\mathrm{T}}\left(trk \in \{\text{jet} \cap \text{HS-PV}\}\right) = \sum_{trk \in \text{jet} \cap \text{HS-PV}} p_{\mathrm{T}}(\mathrm{trk}) \tag{2.4}$$

  It is corrected by the number of pile-up tracks in the event $n_{\mathrm{trk}}^{\mathrm{PU}}$ to reduce JVT dependence on pile-up. The corrJVF variable is defined as follows:

$$\mathrm{corrJVF} = \frac{p_{\mathrm{T}}\left(trk \in \{\text{jet} \cap \text{HS-PV}\}\right)}{p_{\mathrm{T}}\left(trk \in \{\text{jet} \cap \text{HS-PV}\}\right) + \frac{p_{\mathrm{T}}(trk \in \{\text{jet} \cap \text{HS-PV}\})}{0.01 n_{\mathrm{trk}}^{\mathrm{PU}}}} \tag{2.5}$$

- $\boldsymbol{R_{p_{\mathrm{T}}}}$: is the ratio of the sum of the p_{T} of tracks associated to the jet and originating from the HS-PV over the fully calibrated jet p_{T}.

Figure 2.28 shows the efficiency versus fake rate curves for several variables that discriminate pile-up jets from hard-scatter jets.

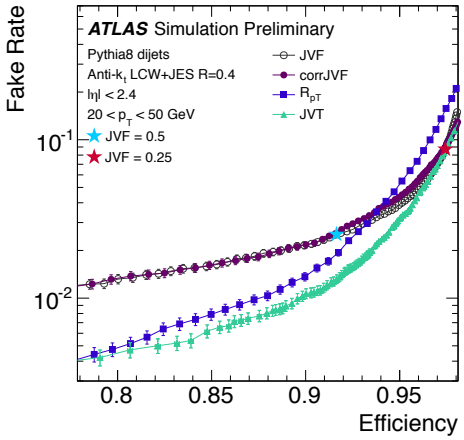


Figure 2.28.: Fake rate from pileup jets versus hard-scatter jet efficiency curves for JVF, corrJVF, $R_{p_{\mathrm{T}}}$, and JVT. The figure and definitions of all variables are found in [124].

2.4.5. Taus

Taus are the heaviest leptons. With a mass of 1.777 GeV and a lifetime of 2.9×10^{-13} s, a τ-lepton at $p_{\mathrm{T}} \sim 50$ GeV decays after traveling only ~ 2 mm in the transverse plane, i.e. before any detector layer. In $\sim 35\%$ of the cases the τ-lepton will decay in a lepton and two neutrinos and in $\sim 65\%$ of the cases the τ-lepton decays hadronically with an accompanying neutrino.

Ideally leptonically decaying taus would be identified as an electron or a muon associated to a track not pointing towards the primary vertex and with missing energy. However due to their short traveled distance taus decaying to electrons or muons are very difficult to differentiate from prompt-leptons and are not reconstructed.

Only hadronically decaying τ-leptons are identified [125] using jets and their associated tracks[d].

[d]For the τ-lepton identification a track is associated to a jet if it is found in a cone of size $\Delta R < 0.2$ around the jet.

Since τ-leptons decay via weak interaction they are expected to give narrower jets and low track multiplicities compared to gluons or quarks. This feature together with the kinematic information from tracks and jets are combined in two multi-variate analyses, a Boosted Decision Tree and a projective likelihood. The output distributions of these techniques are used to discriminate the τ-hypothesis from the QCD-jets and electrons hypotheses.

2.4.6. Missing Transverse Energy

Missing energy [126] is generated by particles escaping the detector. The precise extraction of this quantity is thus vital for BSM searches and processes involving decays to neutrinos. The initial partons energy being unknown the full missing energy computation is impossible. However since pp collisions are produced along the z-axis, initial partons can be assumed to have a negligible momentum in the transverse plane. The Missing Transverse Energy, noted MET or $E_{\mathrm{T}}^{\mathrm{miss}}$, is then accessible requiring momentum conservation.

The MET measurement is based on objects reconstructed in the EM and hadronic calorimeters as well as muons. For all physics objects a MET term is computed as the vectorial negative sum of all transverse momenta. $E_{x(y)}^{\mathrm{miss,term}} = \left(\sum_{\mathrm{obj} \in \mathrm{term}} (-1)\, \overrightarrow{p_{\mathrm{T}}} \right)_{x(y)}$. An additional term is added to account for soft emission using ID tracks matching the HS-PV and not associated to any physics object. The obtained missing energy terms are then added along the x and y axes. The final MET is given by the vectorial sum of the x and y components. Figure 2.29 shows the distribution of the reconstructed MET in $Z \to \mu\mu$ selected events. A good agreement is observed between the prediction and ATLAS data.

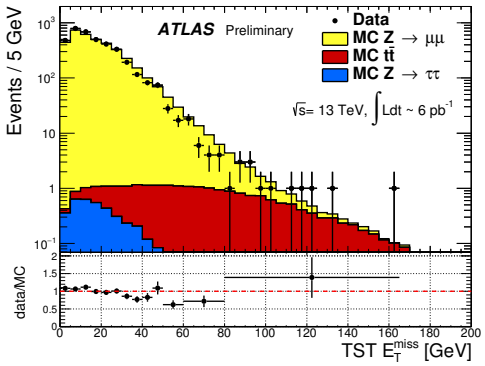


Figure 2.29.: Distribution of the reconstructed missing energy in $Z \to \mu\mu$ selected events. Run 2 data (black points) is compared to the cumulative distribution of predictions from processes passing the selections [126].

3. Identification of b-flavoured-jets and bb-flavoured-jets

The identification of jets containing b-hadrons, called *b-tagging*, is a key ingredient for many analyses ranging from top-quark measurements to new physics searches, along with Higgs boson studies. An overview of the b-tagging procedure in the ATLAS experiment is presented in section 3.1. A precise description of b-jets requires a deep understanding of the b-jet definition in MC predictions. However there exists no recipe to disentangle the various ambiguities in the b-jet definition. Section 3.2 reviews my studies of the b-quark fragmentation in jets and the detailed studies of the b-jets particle and track content that I have performed to choose the most suitable definition for b-jets. This definition is used as the default definition by the ATLAS collaboration in LHC Run 2 for b-tagging studies. An overview of the b-tagging algorithms is presented in section 3.3. The increased center of mass energy and the high statistics provided by LHC Run 2 allow to explore rare topologies such as boosted objects with merged b-jets. Moreover the identification of $g \to b\bar{b}$ initiated jets can provide valuable information on the description of the gluon splitting to heavy flavour quarks which is badly described by current MC simulations. Section 3.4 describes my studies to develop, understand and improve a tagging algorithm aiming to identify $g \to b\bar{b}$ initiated jets (bb-jets) in LHC Run 2 conditions.

3.1. b-tagging in ATLAS

For b-tagging studies in the ATLAS experiment, jets are classified into 4 categories:

- b-jets: jets containing b-flavoured hadrons.
- c-jets: non-b-jets containing c-flavoured hadrons.
- τ-jets: jets that are neither b- nor c- jets and contain a τ-lepton.
- $light$-jets: jets that do not belong to any of the above categories, i.e $udsg$-jets

b-tagging algorithms benefit from unique properties of b-hadrons to separate b-jets from c-jets and $light$-jets. This section shortly reports which (section 3.1.1) and how (section 3.1.2) b-hadron properties are used. Further details will be presented in section 3.3.

3.1.1. Exploiting b-hadron properties

The long b-hadron lifetime, of the order of 1.5 ps, is the main ingredient of the b-jet identification. A b-hadron of transverse momentum $p_{\mathrm{T}} = 50$ GeV and a mass around 5 GeV, will have a flight path length in the transverse plane $L_{xy} = \beta\gamma c\tau$ of around 4 mm before decaying. This translates into the presence of a *secondary vertex*, corresponding to the decay vertex of the b-hadron, disjoint from the primary vertex. Charged decay products of the b-hadrons, originating from this secondary vertex, lead to observed tracks with *large impact parameters*. Figure 3.1 shows an illustration of a b-jet with

displaced tracks and a secondary vertex. These two signatures of the long b-hadron lifetime can be resolved by the ATLAS detector.

The second b-hadron property that can be exploited is their high mass, around 5 GeV, which is at least two times higher than the mass of c- and $light$- hadrons. Thanks to their high masses, b-hadrons produce a large number of charged particles, resulting in higher track multiplicities in b-jet than in c- and $light$- jets.

b-(c-)hadrons can decay to electrons and muons. This signature can be exploited to identify b-jets and c-jets. Taggers based on these *semi-leptonic* decays of the b-(c-)hadrons are not discussed in this thesis.

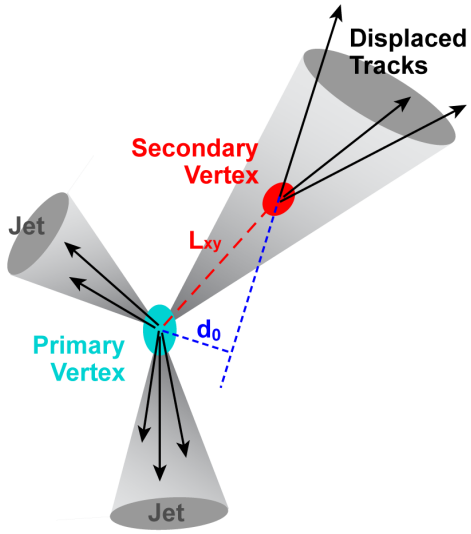


Figure 3.1.: Famous artist view of a b-jet and two $light$-jets with their track content. The large flight path length L_{xy} of b-hadrons allows to resolve the secondary vertex and to observe tracks with a large transverse impact parameter d_0.

3.1.2. Basic principles

Tracks are the most vital inputs for b-tagging. The track-to-jet association is based on the angular distance $\Delta R(\text{track}, \text{jet})$. Since the decay products of high p_{T} particle are more collimated than the ones of low p_{T} particles, tracks are associated to a jet if they fall in a cone around the jet axis which size, $\Delta R_{\text{trk}-\text{jet}}$, depends on the jet p_{T}:

$$\Delta R_{\text{trk}-\text{jet}}\,(p_{\text{T}}) = a_0 + e^{a_1 + a_2 p_{\text{T}}} \tag{3.1}$$

with $a_0 = 0.239$, $a_1 = -1.22$ and $a_2 = -1.64 \cdot 10^{-5}$ MeV^{-1}, resulting in a narrower cone for high p_{T} jets. The a_i parameters are chosen such that on average 95% of the tracks corresponding to b-hadron decay products are associated to the corresponding jet while minimizing the background track contamination (such as pile-up tracks) [127].

b-tagging algorithms are built in two stages resulting in a single discriminating variable for b-jets against c-jets and $light$-jets. The first stage is composed of two classes of algorithms exploiting different properties of the b-hadrons:

- *Impact parameter (IP) based algorithms*: These algorithms use likelihood discriminants for the b-jets against c-jets and $light$-jets hypotheses based on the signed IP significance of tracks in jets $s(d_0) = d_0/\sigma(d_0)$ and $s(z_0) = z_0/\sigma(z_0)$, with d_0 and z_0 the IP of tracks in the transverse plane and along the longitudinal axis, respectively. The errors on d_0 and z_0, $\sigma(d_0)$ and $\sigma(z_0)$ respectively, are introduced to allow a larger contribution to the likelihood from high quality tracks. The signed IP significance is defined positive if the point of closest approach to the primary vertex is in the same direction as the jet momentum.

- *Secondary vertex (SV) based algorithms*: Two vertexing algorithms use tracks to reconstruct the b-hadron decay vertices. The first algorithm is the *Single Secondary Vertex Finder (SSVF)* and aims at the reconstruction of a single effective secondary vertex from all b-hadron decays in the jet. The second, *JetFitter*, aims at the reconstruction of the two vertices in the b-hadron to c-hadron decay chain. The properties of the reconstructed vertices (mass, number of associated tracks, ...) are then either directly propagated to stage 2 or combined in multi-variate techniques (log likelihood ratio for SSVF and neural network for JetFitter).

The information from the reconstructed SVs and the IP based algorithms LLRs are combined using *Multi-Variate Analyses (MVA)* to create the MV1 tagger in Run 1 and the MV2 tagger (with three variations: MV2c00, MV2c10, MV2c20) in Run 2 as described in section 3.3.2. The MV2c10 algorithm output is shown in figure 3.2.

The b-tagging output distribution is used to define selection cuts which are called *working points*. Once a working point is chosen, any jet with a b-tagging output higher than the threshold is b-tagged. b-tagging working points are usually chosen to correspond to a given b-jet *global efficiency*, $\epsilon_b = N_b^{\text{tagged}}/N_b^{\text{true}}$, in $t\bar{t}$ simulations, the $\epsilon_b = 70\%$ working point being the most commonly used in ATLAS analyses. b-tagging performance in ATLAS is expressed in terms of background rejections $R_{\text{light}} = 1/\epsilon_{\text{light}}$ and $R_{\text{c}} = 1/\epsilon_{\text{c}}$ for a given b-jet efficiency ($R_{\text{light}} = 381$ and $R_{\text{c}} = 12$ for $\epsilon_b = 70\%$ with MV2c10).

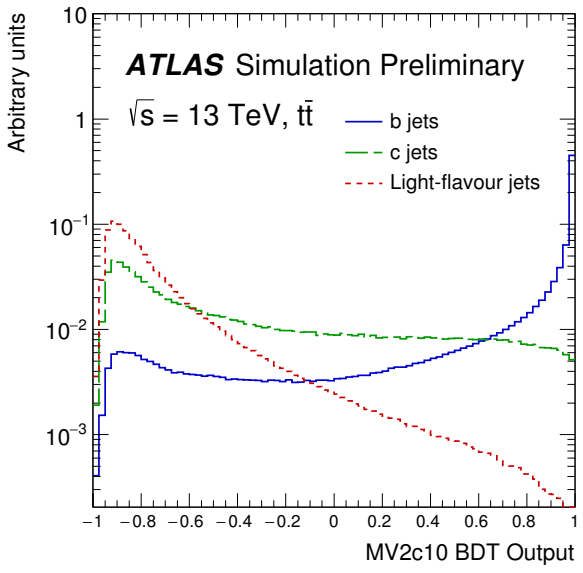


Figure 3.2.: The MV2c10 BDT output discriminant in $t\bar{t}$ simulated events as presented in [128]. The b-jets (blue) are very well separated from the c-jets (green) and the $light$-jets (red).

3.2. The b-jet definition

The study of the definition of b-jets, c-jets and $light$-jets in ATLAS simulation is a complex problem. On the one hand, a definition which is suitable to all analyses is desirable to avoid large corrections from one analysis to another. On the other hand MC efficiencies are corrected to match the observed efficiencies in data and systematic uncertainties are assigned on these corrections. Thus the definition of b-jets in MC predictions has to be close to what can be identified as b-jets in data to avoid large systematic uncertainties arising from the MC extrapolation in non-covered regions of the phase space.

However the definition of b-jets in simulation is ambiguous in several cases. For example b-hadron products can split into two jets (see figure 3.3 left). One has to decide which jet to associate to the b-hadron, or whether both jets are defined as b-jets, in which case one would need a specific algorithm to identify split jets that should be merged. Several b-hadrons can also fall in the same jet (see figure 3.3 right) and a separate category can be considered for these jets. A b-hadron can also represent a small fraction of the jet content in case of late shower or merged jets (the latter being even more important in boosted topologies). Such jets typically have a low b-tagging efficiency and separate categories for such jets can also be considered for specific analyses.

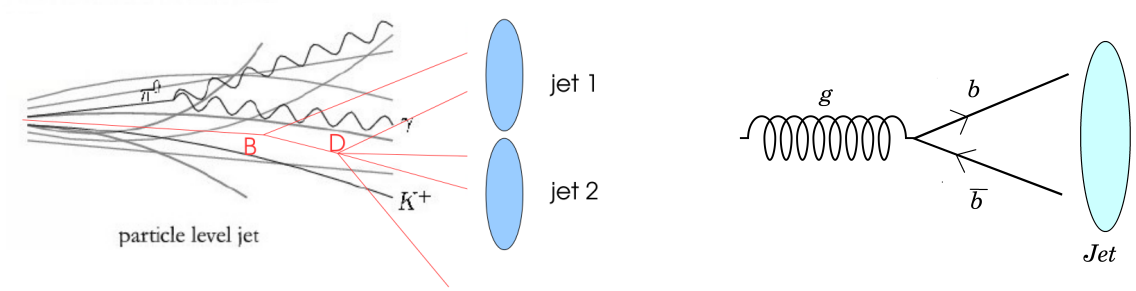


Figure 3.3.: Schematic view of a b-hadron which decay products are split in two jets (left). Schematic view of two b-hadrons merging in a single jet (right).

The jet heavy flavour truth labelling, referred to as *labelling* from now on, is the procedure which defines b-jets, c-jets and $light$-jets in the ATLAS simulation. In particular, the particle to jet association is a critical point in the labelling procedure. Two main particle to jet association algorithms are studied and compared in this section. The properties of the obtained b-jets are also presented.

3.2.1. MC samples and jet definition

The b-jet definition studies are mostly performed in a 13 TeV $t\bar{t}$ simulation. The $t\bar{t}$ samples offer a large amount of b-jets, together with a large number of c-jets and $light$-jets from the W-boson hadronic decay and the particle radiation.

Monte Carlo events are generated using the Powheg-Box method showered with PYTHIA6 [104]. The EvtGen [129] interface is used to correct the decay rates and lifetimes of b-hadrons to match the latest measurements. Finally the standard ATLAS fullsim procedure (see section 2.3) is applied.

Two jet definitions are used:

- *Reco-jets* are reconstructed from calorimeter cluster seeds as explained in section 2.4.4.
- *Truth-jets* are reconstructed from stable truth-particles of the MC history. Leptons and neutrinos which are not coming from Z and W decays are also included in the jet reconstruction to improve the resolution of the b-hadron energy in the jet.

For the two jet definitions the clustering is done by the anti-k_t algorithm with $R = 0.4$ (see section 2.4.4). Jets with $p_{\mathrm{T}} < 20$ GeV are not considered. Jets are also required to have $|\eta| < 2.5$ in order to be within the inner detector coverage.

In this section jets are tagged based on their MV1 output as the MV2 algorithm was not yet fully available when this study was done.

3.2.2. Particles to jets association

Reco-jets and truth-jets are labelled based on the flavour of their associated particles. In Run 1 a cone-based labelling was used where jets are associated to quarks found within $\Delta R(\text{quark,jet}) < 0.3$. In this definition only *final-state quarks* (i.e after all parton shower radiations) with $p_{\mathrm{T}} > 5$ GeV are used. Any jet associated to a b-quark is labelled b-jet. Similarly if a c-quark is found but no b-quark the jet is labelled c-jet. Jets not falling in the previous categories but containing a τ-lepton are labelled τ-jets. The remaining jets compose the $light$-jet category.

Past studies have shown that quark-based and hadron-based, which uses the same algorithm but associating jets to hadrons rather than quarks, labelling gives similar results. However the hadron-based labelling is better defined than the quark-based labelling. Indeed, the definition of quarks in MC simulations depends on the choice of generator and parton shower. A particle level labelling reduces the dependence of b-tagging on the choice of MC and thus reduces the MC to MC extrapolations and their uncertainties.

For Run 2, two main jet to hadron association schemes are proposed and studied in this document.

- The $\Delta R < 0.3$ exclusive algorithm, referred to as "ΔR " from now on, matches each hadron h with $p_{\mathrm{T}} \geq 5$ GeV to its closest jet j within a cone of size 0.3: $\Delta R(h, j) = \min_{\text{jet}\in\text{jets}} \Delta R(h, \text{jet}) \leq 0.3$. This algorithm is called *exclusive* in the sense that a hadron can only be matched to one jet while one jet can matched to several hadrons. Several ΔR cut are studied. A cone size of 0.3 gives a good compromise between high matching efficiency of b-hadrons and the removal of jets with low fraction of constituents coming from the b-hadrons.

- The *ghost association (GA)* [130, 131] is a hadron to jet matching algorithm based on the jet clustering algorithm used to build the jets, in our case the anti-k_t algorithm with $R = 0.4$. b-hadrons, c-hadrons and τ are added to the list of inputs to the jet clustering algorithm with a p_{T} close to 0. The algorithm is rerun and particles are associated to a given jet if they are part of this jet constituents. A p_{T} close to 0 prevents these new inputs from modifying the original kinematics of the jet. Indeed, it was shown in section 2.4.4 that the aggregation power of a particle is proportional to its p_{T} squared. Thus particles with null p_{T} do not aggregate any other particles. A p_{T} of exactly 0 can not be used either since the anti-k_t algorithm involves the inverse of the particle p_{T}.

In this labelling section b-jets are further separated in b-jets and bb-jets if they carry exactly one b-hadron or at least two b-hadrons respectively. This additional split with respect to the default procedure aims at quantifying the $g \to bb$ initiated jet candidates. The identification of $g \to bb$ jets described in section 3.4 requires two b-hadrons inside a jet.

3.2.3. Comparison of the ΔR and ghost association algorithms

Table 3.1 summarizes the truth-jet flavour composition and the agreement between both algorithms. The ghost association and ΔR matching schemes are very consistent with 99.1% of the truth-jets labelled the same way by both algorithms. Such jets have their heavy-flavour content invariant under

a change of the hadron to jet matching and are thus taken as a reference. They will be referred to as pure-b-truth-jets, pure-c-truth-jets or pure-$light$-truth-jets. A very low fraction of bb-truth-jets is present in $t\bar{t}$ events and such topologies are left for further discussion in section 3.4.

GA labelled / ΔR labelled	b-truth-jet	bb-truth-jet	c-truth-jet	$light$-truth-jet
b-truth-jet	40.3%	0.2%	<0.1%	<0.1%
bb-truth-jet	<0.1%	0.4%	<0.1%	<0.1%
c-truth-jet	0.1%	<0.1%	9.4%	<0.1%
$light$-truth-jet	0.4%	<0.1%	0.3%	49.0%

Table 3.1.: Fraction of truth-jets per labelling category in $t\bar{t}$ events. Rows represent the obtained label from the ΔR hadron to jet matching scheme. Columns show the obtained label from the Ghost Association hadron to jet matching scheme.

Truth-jets labelled "b" with the ΔR algorithm are called ΔR-b-truth-jets. The ones labelled "b" with the ghost association are called GA-b-truth-jets. Similar nomenclature is used for the other flavors, e.g. ΔR-$light$-truth-jets and GA-$light$-truth-jets. The $\epsilon_b^{\Delta R} = 70\%$ working point corresponds to the cut on the MV1 output corresponding to a 70% efficiency to select ΔR-b-truth-jets. The rejection of ΔR-$light$-truth-jets obtained at this working point is referred to as $R_{light}^{\Delta R}$. Similarly, the $\epsilon_b^{\text{GA}} = 70\%$ working point selects 70% of the GA-b-truth-jets, corresponding to a GA-$light$-truth-jets rejection of R_{light}^{GA}.

Even though the two matching algorithms are very consistent, the impact of the labelling choice on expected b-tagging performance is large. In particular, the $light$-jets rejection changes by 10% when moving from the ΔR algorithm to the GA matching scheme.

Table 3.2 shows the efficiencies of each labelling category for the $\epsilon_b^{\text{GA}} = 70\%$ working point. ΔR-b-truth-jets show a higher probability to be tagged than GA-b-truth-jets. It is especially interesting to look at the migration between the b-truth-jet and $light$-truth-jet samples when changing the labelling scheme. Indeed, the very high $light$-jet rejection of b-tagging algorithms makes the $light$-jet sample very sensitive to a contamination with very few jets with high b-tagging efficiencies. Two cases are particularly interesting:

- GA-$light$-ΔR-b-truth-jets: truth-jets labelled as $light$-truth-jets by the ghost association algorithm but b-truth-jets by the ΔR algorithm. Their efficiency to be b-tagged at the $\epsilon_b^{\text{GA}} = 70\%$ working point is 111 times higher than the one of pure-$light$-truth-jets. When going from the ΔR matching scheme to the ghost association matching scheme, they will contaminate the $light$-truth-jet sample with relatively high efficiency b-tagged truth-jets. Even if the contribution of GA-$light$-ΔR-b-truth-jets to the $light$-truth-jets sample is below 0.1%, they represent 6.4% of the GA-$light$-truth-jets passing the b-tagging cuts.

- GA-b-ΔR-$light$-truth-jets: truth-jets labelled as b-truth-jets by the ghost association algorithm but $light$-truth-jets by the ΔR algorithm. Similarly to the previous category, these are potential b-jets which would contaminate the $light$-jet sample when going from the ghost association matching scheme to the ΔR matching scheme. Their efficiency at the $\epsilon_b^{\text{GA}} = 70\%$ working point is smaller than the one of GA-$light$-ΔR-b-truth-jets, only 3 times higher than $light$-truth-jets. However the GA-b-ΔR-$light$-truth-jets category is 12 times larger than the GA-$light$-ΔR-b-truth-jets category. If the ΔR association scheme is chosen, they represent 2.4% of the $light$-truth-jets passing the b-tagging cuts.

These two categories are thus very important to discriminate between the two association schemes and understand the definition of b-jets. They are further described in section 3.2.6.

GA labelled / ΔR labelled	b-truth-jet	bb-truth-jet	c-truth-jet	*light*-truth-jet
b-truth-jet	70.8%	66.1%	45.9%	33.3%
bb-truth-jet	60.0%	76.4%	62.5%	50.0%
c-truth-jet	8.7%	12.0%	20.3%	6.7%
light-truth-jet	1.0%	0.0%	0.4%	0.3%

Table 3.2.: Efficiency of identifying truth-jets as b-truth-jets with the MV1 algorithm per labelling category. Rows represent the obtained label from the ΔR hadron to jet matching scheme. Columns show the obtained label from the Ghost Association hadron to jet matching scheme.

3.2.4. Fragmentation of the b-hadron energy inside jets

The fragmentation function of b-quarks [48] shows that b-hadrons receive in general $\sim 80\%$ of the energy of the originating b-quark. The b-hadron energy is thus expected to contributes to $\sim 80\%$ of the b-jet energy and focus is given to b-hadron decay products rather than all particles coming from the hadronisation process. The transverse momentum of the b-hadron decay products associated to a b-truth-jet is defined by the vectorial sum of the 4-vectors of the jet constituents originating from the b-hadron:

$$p4\,(\text{jet-}b\text{-hadron-constituents}) = \sum_{\text{obj}\in\{b\text{-hadron products}\}\cap\{\text{jet constituents}\}} p4(\text{obj}) \tag{3.2}$$

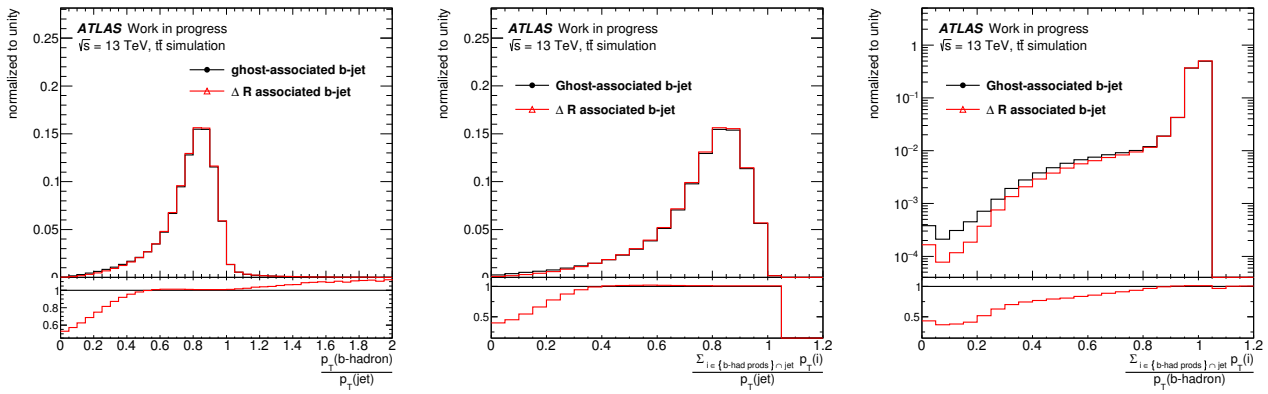
Three quantities are used in order to evaluate the flow of energy between the b-hadrons and the b-truth-jets:

- b-p_{T}-ratio $= \frac{p_{\mathrm{T}}(b\text{-hadron})}{p_{\mathrm{T}}(\text{jet})}$: ratio of the b-hadron p_{T} to the jet p_{T}.
- jet-p_{T}-fraction-from-b $= \frac{p_{\mathrm{T}}(\text{jet-}b\text{-hadron-constituents})}{p_{\mathrm{T}}(\text{jet})}$: ratio of the transverse momentum of the b-hadron decay products found inside the corresponding b-truth-jet to the jet p_{T}. This quantity provides a measure of the fraction of the jet energy which comes from the b-hadron.
- b-p_{T}-fraction-in-jet$= \frac{p_{\mathrm{T}}(\text{jet-}b\text{-hadron-constituents})}{p_{\mathrm{T}}(b\text{-hadron})}$: ratio of the transverse momentum of the b-hadron decay products found inside the corresponding b-truth-jet to the b-hadron p_{T}. This quantity provides a measure of the fraction of the b-hadron energy which ends up in the corresponding jet.

Figure 3.4 shows how the energy propagates from the b-hadron to the truth-jet, and the truth-jet energy composition for the ΔR and ghost association labelling schemes. The b-p_{T}-fraction-in-jet distribution (right) shows that over 85% of the b-hadrons have over 95% of their energy going into the truth-jet that they are associated to by both matching algorithms. Furthermore, the distribution of the fraction of the jet energy coming from the b-hadron (jet-p_{T}-fraction-from-b) reflects the fragmentation function of b-quarks. In particular, over 60% of the b-truth-jets have at least 75% of their energy coming from the b-hadron.

The two matching schemes agree for the bulk of the distribution. However the higher acceptance

of the ghost association scheme introduces more b-truth-jets with low transverse momentum fraction from the b-hadron.



(a) b-p_{T}-ratio (b) jet-p_{T}-fraction-from-b (c) b-p_{T}-fraction-in-jet

Figure 3.4.: Fragmentation of the b-hadron energy inside b-jets for the ΔR matching and ghost association of heavy-flavoured-hadrons to jets. (a) p_{T}-ratio of b-hadrons to their associated jet. (b) ratio of the fraction of b-hadron p_{T} in the jet to the jet p_{T}. (c) ratio of the fraction of b-hadron p_{T} in the jet to the b-hadron p_{T} (see text for more details).

3.2.5. Association of tracks originating from b-hadron decay products to b-jets

For b-tagging the association of tracks coming from the b-hadron to b-jets is crucial. The track content of jets is constructed by analogy with the previous section (3.2.4), using tracks associated to the jets instead of jet constituents. The track transverse momentum originating from the b-hadron and found inside the corresponding jet is defined as the vectorial sum of the 4-momenta of tracks associated to the jet and to the b-hadron decay products:

$$p4\,(\text{jet-}b\text{-hadron-tracks}) = \sum_{\text{trk}\in\{b\text{-hadron products}\}\cap\{\text{jet}\}} p4(\text{trk}) \tag{3.3}$$

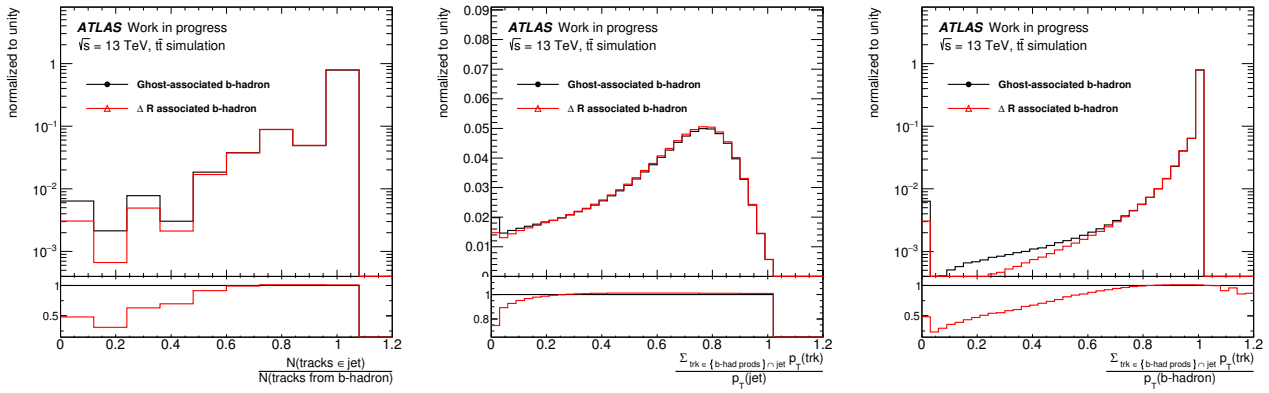
Three quantities are built:

- $N(\text{track})$-ratio $= \frac{N(\text{tracks from the } b\text{-hadron inside this jet})}{N(\text{tracks from } b\text{-hadron})}$: ratio of the number of tracks originating from the b-hadron and associated to the jet to the total number of tracks originating from the b-hadron.
- track-jet-p_{T}-fraction-from-b $= \frac{p_{\mathrm{T}}(\text{jet-}b\text{-hadron-tracks})}{p_{\mathrm{T}}(\text{jet})}$: ratio of the track transverse momentum originating from the b-hadron found inside the corresponding b-truth-jet to the b-truth-jet p_{T}. This quantity gives a measure of the fraction of the jet energy coming from tracks associated to the jet and originating from the b-hadron.
- track-b-p_{T}-fraction-in-jet$= \frac{p_{\mathrm{T}}(\text{jet-}b\text{-hadron-tracks})}{p_{\mathrm{T}}(b\text{-hadron})}$: ratio of the track transverse momentum originating from the b-hadron found inside the corresponding b-truth-jet to the b-hadron p_{T}. This quantity gives a measure of the fraction of the b-hadron track energy which is associated to the jet.

The track to jet association has been optimized for a large acceptance of the tracks coming from b-hadrons as explained in section 3.1.2. Figure 3.5 shows the $N(\text{track})$-ratio (left), track-jet-p_{T}-fraction-from-b (middle) and track-b-p_{T}-fraction-in-jet (right). Most of the tracks originating from

a b-hadron are found in the corresponding b-truth-jet. The fraction of the jet momentum due to associated b-hadron tracks shows the same peak at 0.8 as the b-truth-jet energy composition in b-hadron constituents. However the large tail towards low fractions of momentum originating from b-hadron tracks indicates a significant contamination from background tracks (other hadron decays and pile-up).

Similarly to the truth-jet energy composition, the ΔR and ghost association schemes are in very good agreement in the bulk of the distribution. However the ghost association has a larger fraction of truth-jets associated to the b-hadron with a smaller fraction of the corresponding associated tracks coming from the b-hadron.



(a) $N(\text{track})$-ratio (b) track-jet-p_{T}-fraction-from-b (c) track-b-p_{T}-fraction-in-jet

Figure 3.5.: Association of the b-hadron tracks to b-jets for the exclusive $\Delta R < 0.3$ matching and ghost association of heavy-flavoured-hadrons to jets. (a) number of tracks originating from the b-hadrons associated to the jet divided by the number of b-hadron tracks. (b) ratio of the fraction of b-hadron track p_{T} in the jet to the jet p_{T}. (c) ratio of the fraction of b-hadron track p_{T} in the jet to the b-hadron p_{T} (see text for more details).

3.2.6. Properties of jets with different labelling between ΔR and ghost association

As shown above, the ΔR and ghost association labelling schemes differ only in the tails of the b-truth-jet track and energy composition. Among the differences in labelling outcome the GA-$light$-ΔR-b-truth-jet and GA-b-ΔR-$light$-truth-jet categories are important for b-tagging.
A difference in labelling for b-jets can only occur in two cases:

- Case 1, *distant isolated jets*: in this case the jet is only associated to the b-hadron by the ghost association algorithm which is looser than the ΔR algorithm.
- Case 2, *close-by-jets*: the second possibility to obtain differences in labelling is to have two close-by jets around a b-hadron which is associated to one jet with the ΔR matching and to the other jet with the ghost association.

3.2.6.1. Distant isolated jet topologies

Most of the truth-jets labelled "b" by the GA algorithm but "$light$" by the ΔR algorithm (GA-b-ΔR-$light$-truth-jet) fall in this category. Over 99% of the GA-b-ΔR-$light$-truth-jet are composed of less than 50% of

b-hadron energy and have low b-p_{T}-ratio. They are thus truth-jets built around high p_{T} particles close to the b-hadron with high clustering power in the jet algorithm. The shift in the truth-jet axis towards the surrounding particles induces a partial loss of the b-hadron energy which falls out of the truth-jet cone, or for low p_{T} b-hadrons out of the detector acceptance. Moreover more than 90% of these truth-jets are composed of less than 25% of b-hadron tracks and up to 30% of the associated b-hadrons have all their tracks outside the truth-jet. These truth-jets merge a fraction of the b-hadron energy with the surrounding hadronic activity. They are unlikely to be tagged as b-truth-jets.

3.2.6.2. Close-by jet topologies

Truth-jets labelled "*light*" by the GA algorithm but "b" by the ΔR algorithm (GA-*light*-ΔR-b-truth-jet) fall in this category. Indeed, figure 3.6 shows the ΔR between the b-truth-jet and its closest neighboring jet for all b-truth-jets[a] and for GA-*light*-ΔR-b-truth-jet. In 91% of the cases GA-*light*-ΔR-b-truth-jet have a nearby-jet found within $\Delta R < 0.7$ which is close to the maximal distance allowing the second truth-jet to be matched to the b-hadron via ghost association. In 89% of the cases, this nearby-truth-jet is associated to the same b-hadron by the ghost association algorithm.

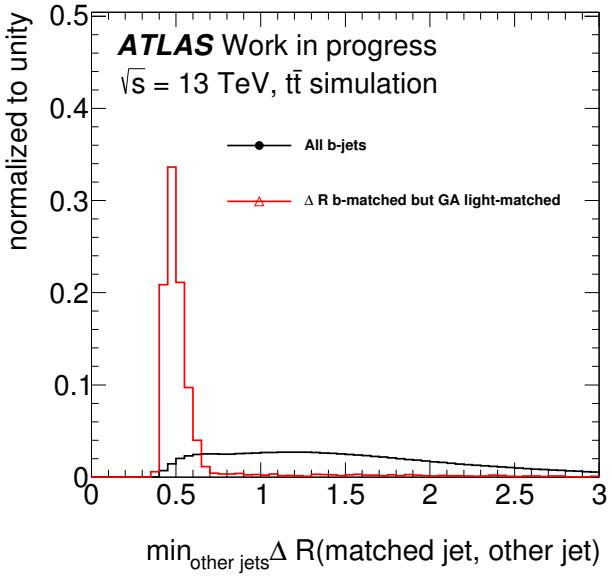


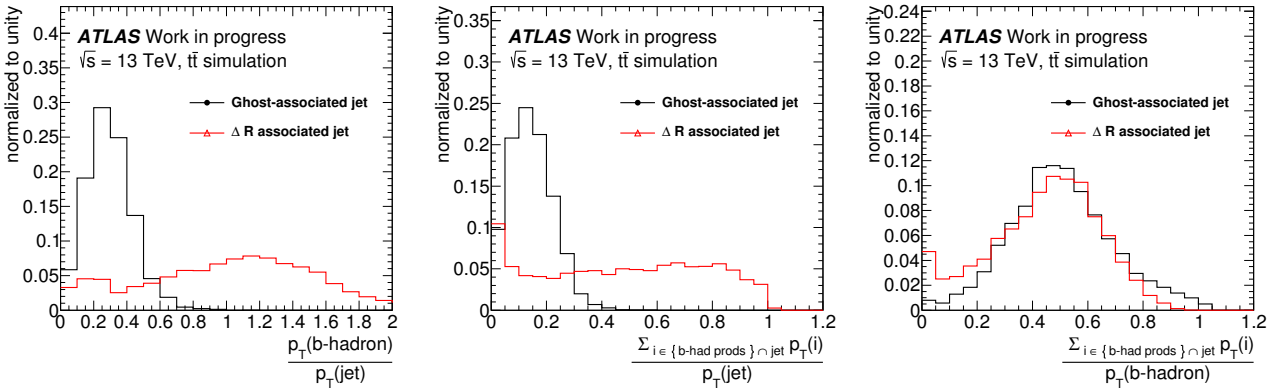
Figure 3.6.: Normalized distribution of ΔR distance between b-truth-jets and their closest jet. The black curve shows the ΔR distance computed for all jets matched to a b-hadron by the ghost or the exclusive $\Delta R < 0.3$ association. The red curve shows the ΔR distance computed for jets labelled "b" using the ΔR matching but labelled "*light*" using the ghost association.

The subset of close-by-truth-jets with one GA-*light*-ΔR-b-truth-jet gives a natural set of truth-jets to study the labelling procedure. Indeed one truth-jet is associated to a b-hadron with the ΔR procedure, the *ΔR-associated-truth-jet*, and the other truth-jet is associated to the same b-hadron by ghost association, the *ghost-associated-truth-jet*.

Figure 3.7 shows the b-p_{T}-ratio (left), jet-p_{T}-fraction-from-b (middle) and b-p_{T}-fraction-in-jet (right) of these two truth-jets. The b-p_{T}-fraction-in-jet distribution clearly shows that the b-hadron energy

[a] Any truth-jet labelled "b" by either the ΔR algorithm or the ghost association enters the "all truth-b-jets" category.

is split between two truth-jets. The ΔR association scheme associates the b-hadron to the closest jet while the p_{T} dependence of the anti-k_t algorithm jet-clustering algorithm enforces the b-hadron to be associated to the highest p_{T} jet when using the ghost association. This results in low b-p_{T}-ratio and low jet-p_{T}-fraction-from-b of the ghost-associated-truth-jet.

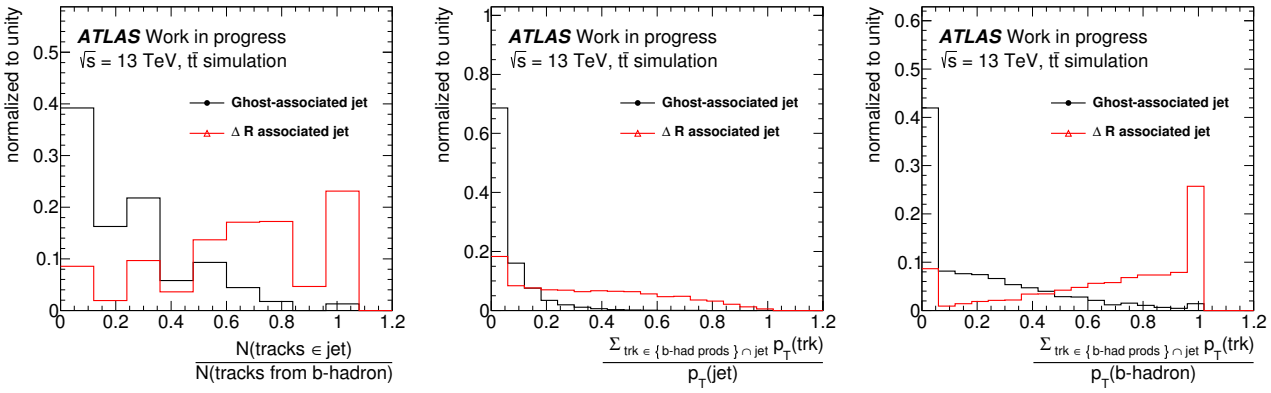


(a) b-p_{T}-ratio (b) jet-p_{T}-fraction-from-b (c) b-p_{T}-fraction-in-jet

Figure 3.7.: Fragmentation of the b-hadron energy inside b-jets for the ΔR-associated-truth-jet and the ghost-associated-truth-jet in case of close-by-jets. In this case the same b-hadron is associated to two jets depending on the association scheme. (a) p_{T}-ratio of b-hadrons to their associated jet. (b) ratio of the b-hadron p_{T} in the jet to the jet p_{T}. (c) ratio of the fraction of b-hadron p_{T} in the jet to the b-hadron p_{T}.

Even though the decay products of the b-hadron are equally split between the two jets, the tracks originating from the b-hadron are not which explains the large difference in b-tagging efficiencies of the two jets. Figure 3.8 shows the N(track)-ratio (left), track-jet-p_{T}-fraction-from-b (middle) and track-b-p_{T}-fraction-in-jet (right) for the ΔR-associated-truth-jet and the ghost-associated-truth-jet. Even though both truth-jets are heavily polluted by background tracks and thus have low track-jet-p_{T}-fraction-from-b, the ΔR-associated-truth-jet gather most of the b-hadron tracks and have large track-b-p_{T}-fraction-in-jet.

The track to jet association is a cone-based algorithm with tracks associated to the closest jet. The choice of the ΔR algorithm follows the track to jet association and thus the ΔR algorithm is expected to pair better with the b-jet identification algorithms than the ghost association in the context of close-by-jets.



(a) (b) (c)

Figure 3.8.: Association of the b-hadron tracks to b-jets for ΔR-associated-truth-jet and the ghost-associated-truth-jet in case of close-by-jets. In this case the same b-hadron is associated to two jets depending on the association scheme. (a) number of tracks originating from the b-hadrons associated to the jet divided by the number of b-hadron tracks. (b) ratio of the fraction of b-hadron track p_{T} in the jet to the jet p_{T}. (c) ratio of the fraction of b-hadron track p_{T} in the jet to the b-hadron p_{T}.

3.2.7. Summary of the jet labelling study

The definition of b-jets is studied looking at the two main association schemes between hadrons and jets, the ΔR matching and the ghost association. These two algorithms are consistent with 99.1% of the jets being invariant under labelling choice in $t\bar{t}$ events. However the difference between the two association schemes can go up to 10% for the $light$-jet tagging efficiency. It is also important to study the differences between the algorithms in special topologies that are not dominant for $t\bar{t}$ production but important in analyses containing close-by-jets and boosted environment.
Both algorithms associate the b-hadron to the truth-jet capturing most of the b-hadron energy. Truth-b-jets are also mostly composed of b-hadron decay products reflecting the high fragmentation function of b-quarks. It is shown that the differences between ΔR- and ghost association-based labelling arise from either distant isolated b-truth-jets or close-by-jets around a b-hadron. In the case of distant isolated b-truth-jets, the b-hadron is only matched to the jet by the ghost association algorithm. These jets have a low fraction of the b-hadron energy and tracks and give low b-tagging efficiencies. In close-by-jet topologies one of the jets can be associated to the b-hadron by the ΔR matching while the other jet is associated to the same b-hadron using the ghost association. It is shown that the b-hadron energy is equally split between the two jets. However tracks are ΔR matched to the jet and thus mostly associated to the ΔR-associated-jet.

For all these reasons the ΔR matching pairs better with the b-jet identification algorithms than the ghost association, and is chosen as default labelling scheme for the b-tagging group in ATLAS for Run 2.

3.3. b-tagging in ATLAS

The basic principles of b-tagging have been explained in section 3.1. This section provides further details on the b-tagging algorithms and their respective performance.

b-tagging performance and the properties of reconstructed objects in b-tagging algorithms are esti-

mated in simulated $t\bar{t}$ events from 13 TeV pp-collisions. Only reco-jets with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.5$ are considered. The jet flavour is assessed via the ΔR labelling procedure. Tracks are associated to jets with the standard procedure (see section 3.1.2) and each b-tagging algorithm uses its own selections on track p_{T}, impact parameter and number of hits in the inner detector. SV-based algorithms typically use loose track requirements in order to maximize the vertex reconstruction efficiency and take advantage of the vertex reconstruction procedure to enhance the purity. On the other hand, impact parameter based algorithms use tight track selection to remove undesired tracks.

3.3.1. Basic algorithms

b-tagging in ATLAS relies on low level algorithms which extract b-hadron properties from the tracks associated to the jets. The IP-based and secondary vertex-based algorithms are described here.

3.3.1.1. Impact parameter based algorithms

Impact parameter based algorithms evaluate the (in)compatibility of the b-track candidate with the primary vertex. The signed IP significances of selected tracks in the jet, $s(d_0)$ and $s(z_0)$ (see section 3.1.2) are used to define *Probability Density Functions (PDFs)* of single tracks to fulfill the b-jet (P_b), c-jet (P_c) and $light$-jet (P_u) hypotheses. Jet-flavour discriminants are then defined by *Log Likelihood Ratios (LLRs)*; $\mathrm{LLR}_{bu} = \sum_{\text{selected tracks}} \log(P_b/P_u)$ is used for the b-jet versus $light$-jet separation. Similarly b-jets against c-jets and c-jets against $light$-jets discriminant variables are built.

Two algorithms are defined that way, the IP2D algorithm is based on LLRs built only with the transverse IP. The IP3D algorithm combines the transverse and longitudinal IPs in two dimensional PDFs, taking into account their correlations. This additional information makes the IP3D algorithm more performant than the IP2D algorithm. However with a longitudinal component the IP3D tagger is also less robust against pile-up than the IP2D tagger.

To maximize performance, tracks are divided into 14 categories depending on their quality which is assessed using the hit information in the detector layers: missing hits, shared hits and split hits as defined in section 2.4.1. Reference templates for the $s(d_0)$ and $s(z_0)$ distributions are built separately for each track category.

Figure 3.9 shows the transverse signed IP significance for the best quality track category (tracks without any defect), called *good tracks*, and the final LLR for the b-jet against $light$-jet hypotheses in the IP3D algorithm.

3.3.1.2. Secondary vertex based algorithms

SV based algorithms are a very important input to b-tagging as they give access to an estimation of the b-hadron mass, decay path length and number of tracks from charged decays of the b-hadron. Two such algorithms, SSVF and JetFitter (see section 3.1.2), are optimized to give both a high reconstruction efficiency and a good rejection of fake vertices.

Single Secondary Vertex Finder: SSVF

The SSVF algorithm reconstructs explicitly one secondary vertex per jet. As we have seen in section 3.2.4 b-jets are contaminated with background tracks coming from pile-up interaction or surrounding hadronic activity. Moreover SV algorithms adopt a loose track selection to maximize the reconstruction efficiencies. The first part of the SSVF algorithm is thus dedicated to the removal of background tracks. Two track vertices are formed based on the geometrical compatibility between

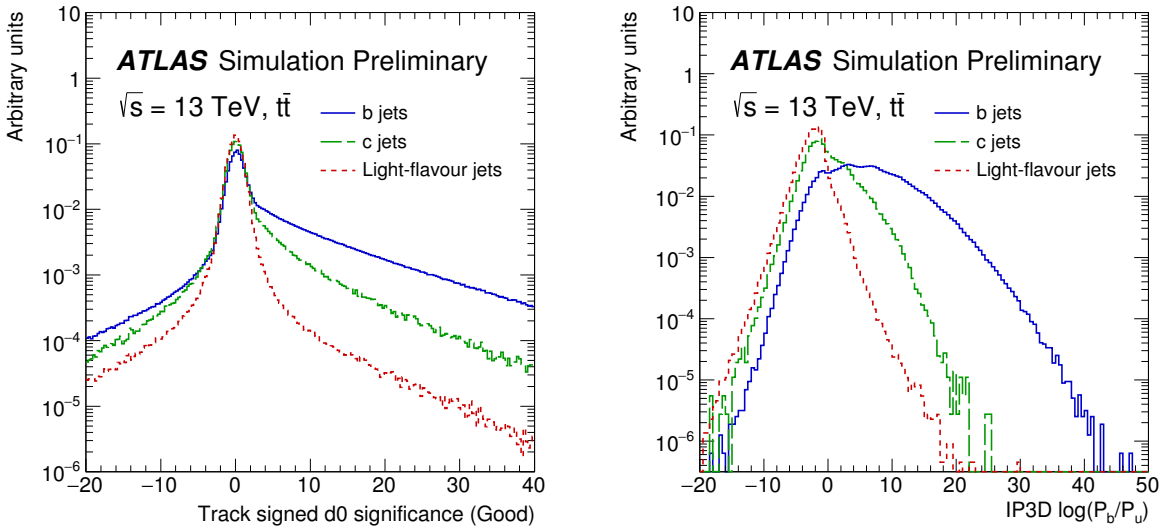


Figure 3.9.: Signed transverse impact parameter significance of *good tracks* for b-jets, c-jets, and *light*-jets in $t\bar{t}$ simulated events at $\sqrt{s} = 13$ TeV (left). Log likelihood ratio of the b-jet against the *light*-jet hypotheses for the IP3D b-tagging algorithm for b-jets, c-jets, and *light*-jets in $t\bar{t}$ simulated events at $\sqrt{s} = 13$ TeV (right) [128].

each pair of tracks. Tracks are then rejected if they belong to a vertex compatible with one of the three following backgrounds:

- *Fake vertices*: correspond to the crossing of two tracks originating from different vertices. These background vertices become important in high pile-up events with large number of tracks or in boosted topologies where tracks are more collimated. Fake vertices are highly reduced requiring that each track should not have a hit in the inner detector before the vertex.
- *Hadronic material interaction*: these vertices are removed comparing the vertex position to a simplified map of the innermost detector layers and of the beam pipe.
- *Photon conversion and decays of long-lived particles* (e.g K_s and Λ): such vertices are rejected comparing the mass of the expected particles to the mass of the reconstructed vertex, evaluated assuming that the track pair is produced by a e^+e^-, $p\pi$ or $\pi^+\pi^-$ pair. In order to keep long-lived particle from b-hadron decays, the incoming particle direction is evaluated. If this particle is found to come from the PV the tracks are removed.

A single SV is then iteratively fitted using the remaining tracks. At each iteration if the χ^2 is too large or the vertex invariant mass is ≥ 6 GeV the track with the largest contribution to the χ^2 is removed.

The SSVF algorithm achieves an average reconstruction efficiency of 80% in b-jets in $t\bar{t}$ events. The p_{T} dependence of the SSVF reconstruction efficiency is shown in figure 3.10 (left). The vertex reconstruction efficiency is maximum for jets with p_{T} between 100 GeV and 150 GeV and drops by around 15% for very low and high jet p_{T}. On the other hand the fake rate in *light*-jets constantly increases with the jet p_{T} making the separation between *light*-jets and b-jets harder at high jet p_{T}. Figure 3.10 (right) shows the mass distribution of the reconstructed vertex. The b-jets at high masses are very well separated from *light*-jets and c-jets dominating at low masses.

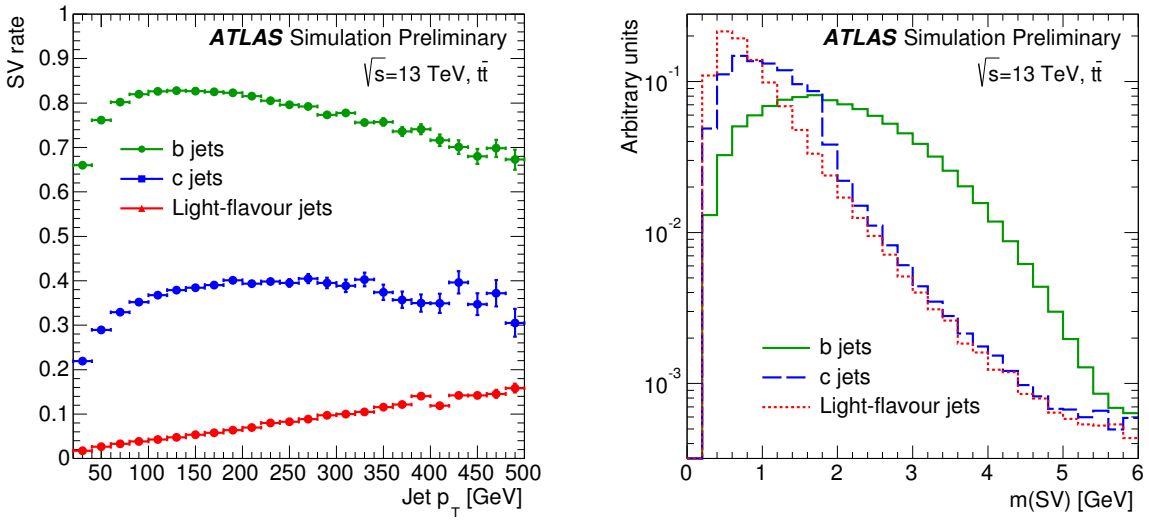


Figure 3.10.: Single secondary vertex reconstruction efficiency for b-jets, c-jets, and *light*-jets in $t\bar{t}$ simulated events at $\sqrt{s} = 13$ TeV (left). Mass of the reconstructed single secondary vertex for b-jets, c-jets, and *light*-jets in $t\bar{t}$ simulated events at $\sqrt{s} = 13$ TeV (right) [128].

b-hadron decay chain reconstruction: JetFitter

The JetFitter algorithm aims at the reconstruction of a secondary and a tertiary vertex to exploit the full decay chain PV$\rightarrow b$-hadron $\rightarrow c$-hadron which is not resolved by the SSVF algorithm. The JetFitter algorithm assumes that the c-hadron vertex lies on the extrapolation of the b-hadron flight path. This common line connecting the PV, the b-hadron vertex and c-hadron vertex is estimated from a Kalman filter. In this approach, all track candidates are used to build single track vertices along the first approximation of the flight axis, the jet direction. Then an iterative clustering algorithm merges at each iteration the two vertices with the highest probability to originate from the same vertex and the complete fit is re-performed. The decay chain is obtained once no two-vertex clusters above a certain probability are found.

This algorithm allows high reconstruction efficiency of the b-hadron decay chain, even in incomplete topologies. It can indeed build vertices from single tracks which are found compatible with the b-hadron flight path. The vertex reconstruction efficiency is shown in figure 3.11 (left). The addition of one-track vertices increases the vertex reconstruction efficiency in b-jets. However it also increases significantly the efficiency to reconstruct vertices in *light*-jets. The number of reconstructed two-track vertices is shown in figure 3.11 (right).

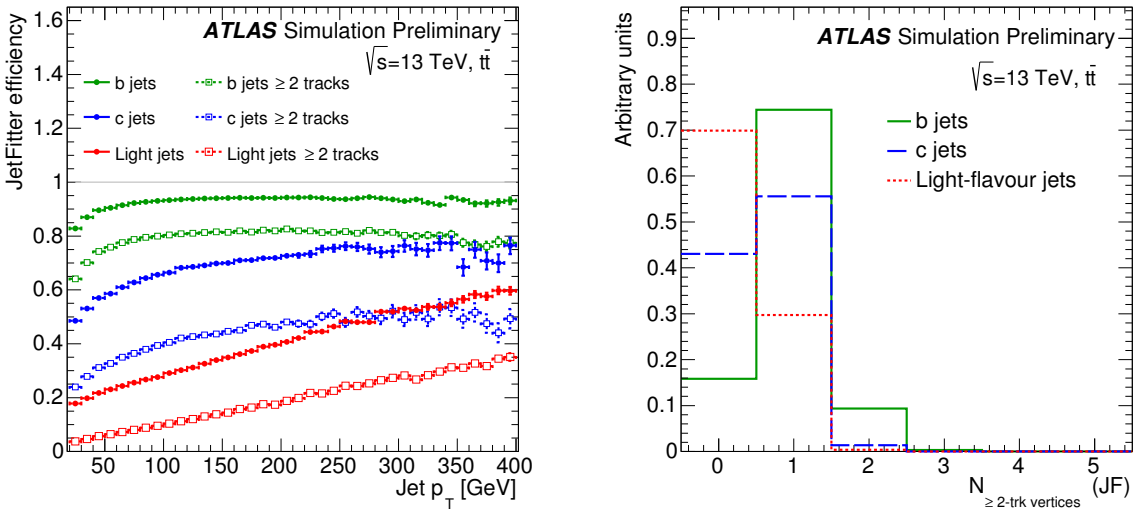


Figure 3.11.: JetFitter secondary vertex reconstruction efficiency for b-jets, c-jets, and *light*-jets in $t\bar{t}$ simulated events at $\sqrt{s} = 13$ TeV (left). Number of two-track vertices reconstructed by the JetFitter algorithm for b-jets, c-jets, and *light*-jets in $t\bar{t}$ simulated events at $\sqrt{s} = 13$ TeV (right) [128].

3.3.2. MVX algorithms

The properties of the reconstructed SV and the IPxD LLRs are combined in MVAs providing the final b-tagging discriminant. This final MVA maximizes the separation between b-jets and other jet flavours.

In Run 1 the three intermediate MVA taggers are used:

- The IP3D output: Log-Likelihood-Ratios for the different jet-flavour hypotheses.
- The SV1 output: likelihood discriminant based on the properties of the reconstructed single secondary vertex.
- the JetFitterCombNN output: output of a neural network based on the properties of the reconstructed vertices by JetFitter.

They are combined in a neural network (NN) whose output gives the final jet discriminant. A first NN trained exclusively against the *light*-jet background results in the MV1 tagger. The MV1 tagger is the most commonly used b-tagging discriminant for physics analyses in Run 1. An alternative (MV1c) tagger is also developed including c-jets in the training. This tagger has mainly been used for $H \to b\bar{b}$ searches due to its increased c-jet rejection.

For Run 2 the b-tagging algorithm chain is simplified. The intermediate MVAs of the secondary vertex-based algorithms are removed and the IPxD LLRs are directly combined to the properties of the SSVF and JetFitter SVs in a BDT resulting in the MV2 tagger. Three variations of the MV2 taggers are available MV2c00, MV2c10, MV2c20 based on the fraction of c-jets included in the training (respectively 0%, 7% and 15% and the rest are *light*-jets).

The expected performance in $t\bar{t}$ events for the b-jet versus *light*-jet discrimination of the MV2c20 tagger is compared to the MV1c tagger performance in figure 3.12 (left). A large improvement of a factor 4 in *light*-jet rejection at the typical $\epsilon_b = 70\%$ working point is observed which significantly reduces the backgrounds in analyses with b-jets in the final state. For the same background rejection an improvement of 10% in b-jet efficiency is achieved. For an analysis with four b-quarks in the final

state such as $t\bar{t}H(H \to b\bar{b})$ this gain in efficiency represents a 40% to 50% increase in signal acceptance. Most of this gain comes from the inclusion of the IBL in the inner detector and proper use of this new information in the tracking and b-tagging algorithms. In the $t\bar{t}H(H \to b\bar{b})$ analysis a further optimization of the MV2 algorithm provided by the b-tagging group has been used. Figure 3.12 (right) shows a comparison on the c-jet rejection for the three MV2 trainings in the optimized setup and the MV2c20 tagger in the previous optimization. Four different working points are shown in table 3.3 for the 2016 MV2c10 tagger (see figure 3.2). The efficiencies of each jet flavour for these working points are corrected to match data and can be used for physics analyses.

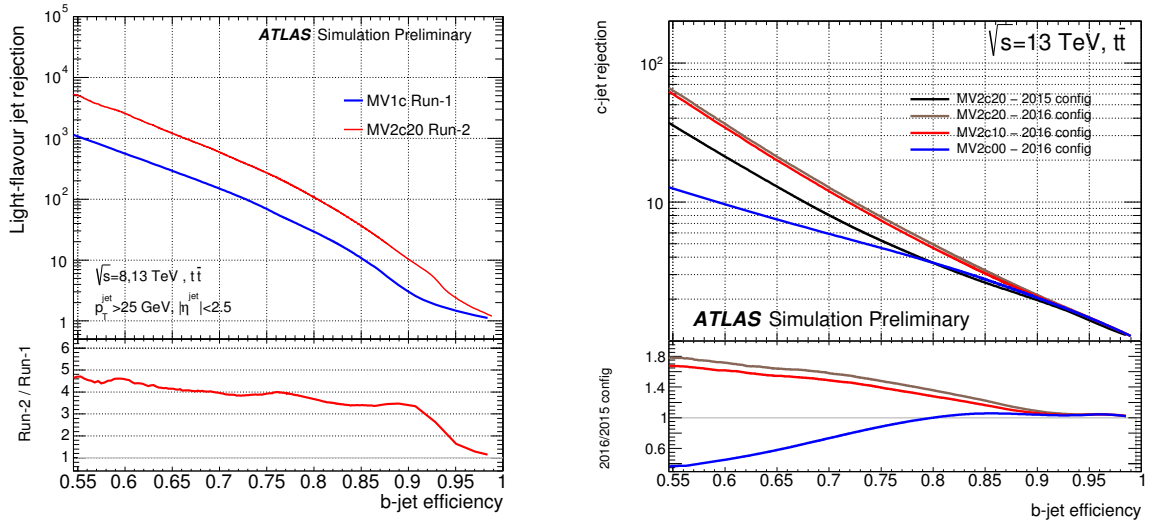


Figure 3.12.: *light*-jet rejection against b-jet efficiency for the MV2c10 (Run 2 default algorithm for 2015 data) and MV1c algorithms (Run 1 algorithm with enhanced c-jet rejection). Performance is evaluated in $t\bar{t}$ simulated events at $\sqrt{s} = 13$ TeV with the Run 2 detector geometry and $\sqrt{s} = 8$ TeV under Run 1 conditions for the MV2c20 and MV1c algorithm, respectively (left) [128]. c-jet rejection against b-jet efficiency for the three trainings of the MV2 algorithms for 2016 data overlaid with the MV2c20 algorithm in 2015 data conditions. Performance is evaluated in $t\bar{t}$ simulated events at $\sqrt{s} = 13$ TeV under Run 2 conditions(right) [132].

MV2c10 cut value	b-jet efficiency	c-jet rejection	*light*-jet rejection
0.9349	60%	34	1538
0.8244	70%	12	381
0.6459	77%	6	134
0.1758	85%	3.1	33

Table 3.3.: Cut values and performance of the four working points provided by the b-tagging group to analyses. Performance is evaluated in $t\bar{t}$ simulated events at $\sqrt{s} = 13$ TeV under Run 2 conditions [132].

3.4. The $g \to b\bar{b}$ identification

Standard b-tagging algorithms provide a very good separtion of b-jets against other jet flavours. However they do not provide any information on the number of b-hadrons inside a b-tagged jet. In particular MVX algorithms are not tuned to separate jets containing a single b-hadrons from jets arising from a gluon splitting into a $b\bar{b}$ pair at small opening angle.

In QCD the large differences between the NLO and LO predictions in the heavy flavour production rate imply large theoretical uncertainties on the inclusive b-jet spectrum. This difference mainly arises from the presence of new heavy flavour production diagrams in the NLO calculation, namely the flavour excitation and gluon splitting channels. It is shown in [133] that the identification of $g \to b\bar{b}$ initiated jets can significantly improve the constraints on the gluon splitting production mode and reduce the theoretical uncertainties on the b-jet spectrum.

An enhanced precision of gluon splitting into a $b\bar{b}$ pair would also be benificial to $H \to b\bar{b}$ searches. In particular the $t\bar{t}H(H \to b\bar{b})$ and $VH(H \to b\bar{b})$ channels suffer from large backgrounds coming from $t\bar{t} + b\bar{b}$ and W+b-jets, respectively, where the additional b's can be produced by gluon splitting. Moreover, the increased center-of-mass energy for Run 2 and the high collected luminosity that is foreseen could allow to access boosted $H \to b\bar{b}$ topologies. In such topologies a significant contribution to the background from $g \to b\bar{b}$ at small opening angle is expected and a $g \to b\bar{b}$ initiated jet identifier could significantly increase the signal purity.

This section reviews the $g \to b\bar{b}$ intiated jet identification based on multi-secondary-vertices reconstruction in the ATLAS experiment that was initially developped in [134].

3.4.1. Samples and physics objects

A mix of W+jets and multi-jets events is used in order to have a representative sample of bb-jets, b-jets, cc-jets, c-jets and $light$-jets with a wide range of transverse momenta.

The W+jets sample is based on $W^{\pm} \to \mu^{\pm}\nu$ events with 0, 1 or 2 additional jets at NLO and up to 4 jets at LO generated with Sherpa [99] under the CT10 parton distribution function set [135]. In order to improve the effective statistics for each jet flavours in a wide range of p_{T}, the W+jets sample is sliced depending on the additional partons flavour and on the p_{T} of the W boson.

The multi-jet sample is generated with PYTHIA8 [105] with the A14 tune [136], the NNPDF2.3LO parton distribution function [137], and interfaced with EvtGen [129]. Similarly to the W+jets samples, a high effective statistics for various ranges of p_{T} is obtained by generating several samples for different slices of leading jet p_{T} and merging them back afterwards.

In this section standard reco-jets with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.5$ are considered. Pile-up jets are removed by rejecting jets with $p_{\mathrm{T}} < 50$ GeV and $|\eta| < 2.4$ if their JVT output (see section 2.4.4) are below 0.64. The standard ΔR labelling procedure is used to define the b-jets, c-jets and $light$-jets samples. The b-jets are further separated into bb-jets if they have two b-hadron within $\Delta R = 0.4$ and into single-b-jets otherwise. The ΔR cut is looser compared to the default labelling scheme in order to improve the matching efficiency of $g \to bb$ initiated jets. The same split is also applied on c-jets to separate cc-jets from single-c-jets.

In order to increase the multi-vertex reconstruction efficiency in bb-jets the cone size of the track to jet association (see section 3.1.2) is increased to: $\Delta R_{\mathrm{trk-jet}}(p_{\mathrm{T}}) = 0.315 + e^{-0.367-1.56\cdot10^{-5}\,p_{\mathrm{T}}}$. For a typical jet of $p_{\mathrm{T}} = 50$ GeV the track to jet association cone size is thus raised from $\Delta R_{\mathrm{trk-jet}} \leq 0.37$ to $\Delta R_{\mathrm{trk-jet}} \leq 0.63$. This capture 97% of all tracks coming from both b-hadron decay in bb-jets.

3.4.2. The Multi-Secondary-Vertex Finder algorithm: MSVF

The MSVF algorithm aims at the reconstruction of all possible vertices in a jet, taking as input the tracks asssociated to the jet. It is implemented as an alternative of the single vertex reconstruction in the SSVF algorithm. It uses the same two-track fake vertices rejection as the SSVF algorithm (see section 3.3.1.2). The MSVF strategy is sketched in figure 3.13 and reads as follows. The selected tracks are used to form all possible two-track vertices from a simple geometrical matching where each track can be used to build several vertices. Two track vertices are interpreted as a graph whose nodes are tracks and links are added between tracks forming vertices. The full graph is divided in its sub-components where all nodes are connected to each other via an algorithm included in the BOOST GRAPH library [138]. The sub-components are the vertex candidates and can be formed of an arbitrary number of tracks. At this stage tracks can belong to several vertex candidates (cliques). The final vertices are obtained after an iterative cleaning based on three procedures:

- The first procedure maximizes the probability of each vertex to be real. If a vertex shows a high χ^2, the track with the largest contribution to the χ^2 is removed from this vertex. The other track from the vertex which shows the highest compatibility with the removed track (minimizing the χ^2 of a two-track vertex) is used to form a vertex with the removed track. The additional vertex is added to the list of vertices.

- In the second procedure the ambiguity of shared tracks between distant vertices is resolved. In this case the track is removed from the vertex with the highest χ^2 and the vertex position is refitted.

- Finally, close-by vertices are merged.

In these procedures one track vertices are allowed. They correspond to single tracks which are found incompatible with all other vertices. Even though half of these rejected tracks are origininating from the b-hadrons their contribution to the properties of the multi-secondary vertices is small and they are neglected.

This algorithm provides the set of all possible vertices in a jet. However b-jets have a non-negligible fraction of tracks from surrounding hadronic decays and pile-up interactions (see section 3.2.4) leading to fake vertices. Due to the finite resolution of the detector the b-hadron and c-hadron vertices are very hard to resolve. The imperfect reconstruction translates into split vertices containing a fraction of the tracks originating from the decay vertex, or merged vertices whose tracks have different origins. Properties of the reconstructed vertices are studied in the following section (3.4.2.1).

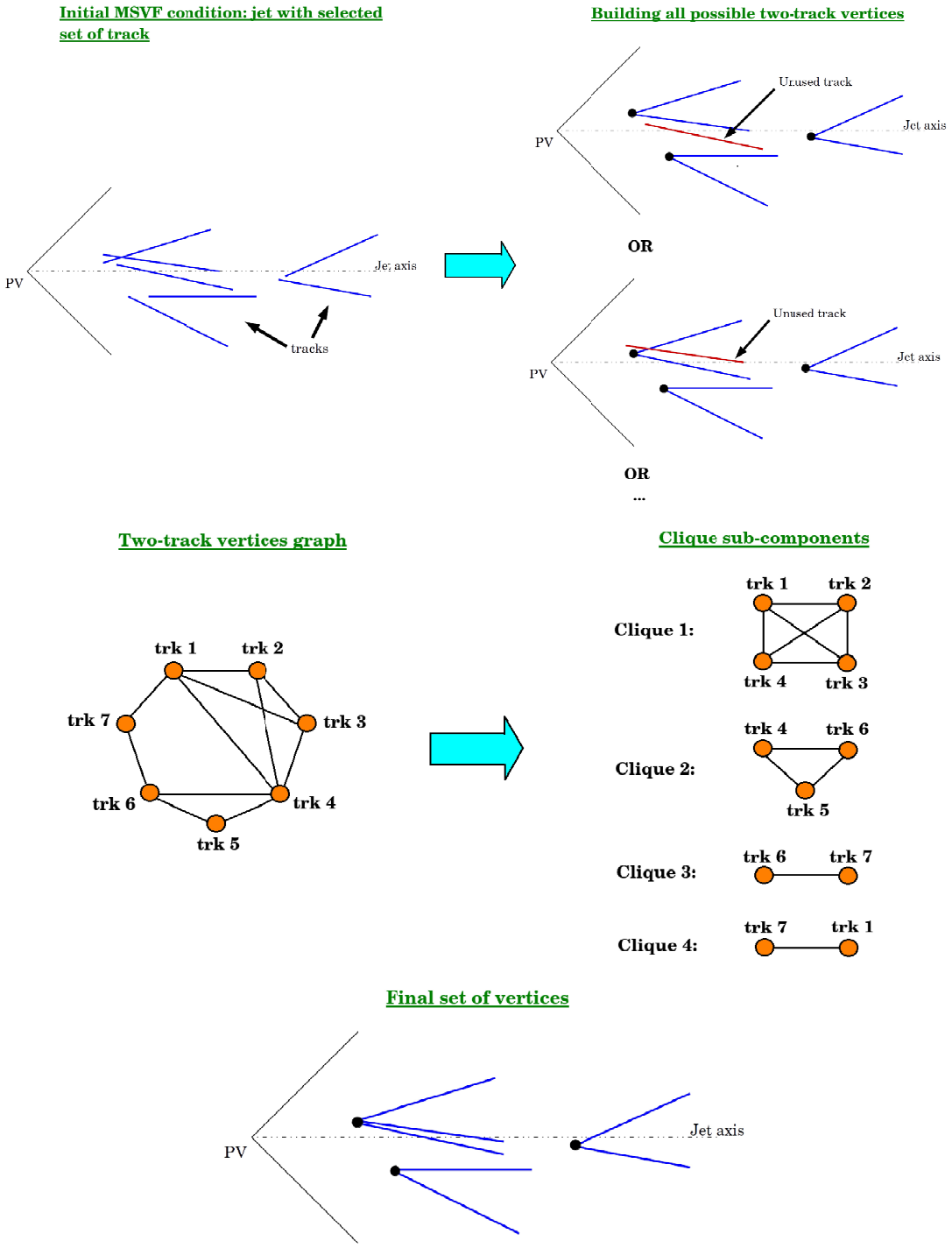


Figure 3.13.: Schematic view of the MSVF algorithm. Input tracks are selected using the same two track vertices rejection as the SSVF algorithm.

3.4.2.1. Properties of the reconstructed Multi-Secondary-Vertices (MSV)

Figure 3.14 shows the number of vertices with at least two tracks. The MSVF algorithm shows high performance with an inclusive efficiency of reconstructing at least one vertex in single-b-jets of 77% and of 88% in bb-jets. 22% of the single-b-jets have at least two vertices while 55% of the single-b-jets have exactly one vertex. This means that at least 55% of the single-b-jets with at least one reconstructed vertex have the b-hadron and c-hadron decay chain merged in a single vertex. In bb-jets up to four vertices are expected and at least two are required for the MultiSVbb algorithms (see section 3.4.3). 51% of the bb-jets satisfy the requirements of having at least two two-track vertices.

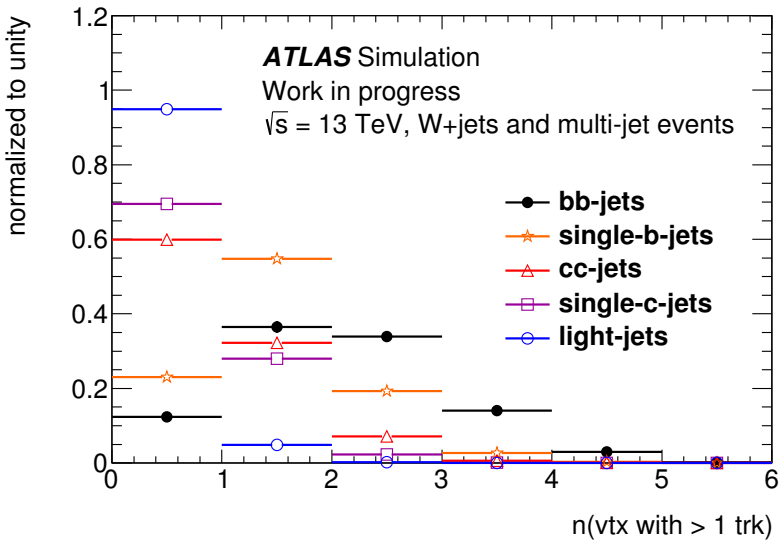


Figure 3.14.: Number of vertices for different jet flavours in a mixture of di-jet and W+jets events.

Quality of the reconstructed vertices in single-b-jets

Single-b-jets are used to estimate the quality of the reconstruction of the b-hadron to c-hadron vertex decay chain. In bb-jets the presence of two such decay chains makes the study of the resolution of the b-hadron and c-hadron decay vertices more difficult.

The purity in b-hadron tracks of the reconstructed vertices is estimated using the set of vertices with at least two tracks ($\geq$ 2-track-vertices). As said previously most of the vertices merge the tracks from the b-hadron and c-hadron decay (62%). Only 11% of the vertices are purely made of tracks coming from the b-hadron direct decay vertex (excluding tracks from the c-hadron decay vertex). A significant fraction (23%) of the vertices is contaminated by background tracks. Finally the fraction of remaining fake vertices, i.e., vertices made of no tracks originating from the b-hadron and its c-hadron child, is only 4% thanks to the SSVF vertex cleaning.

Single-b-jets are also used to estimate the quality of the reconstruction of the b-hadron properties. Because of merged and split vertices this quality is difficult to estimate. Moreover neutral decays of the b-hadron are not considered by MSVF since vertices are built from the tracks of charged particles in the inner detector. In order to isolate the reconstruction quality of MSV, the b-hadron four-momentum is evaluated from three sets of objects:

- *Charged decay particles*: The b-hadron 4-momentum is evaluated from the sum of the 4-momenta of the charged decay particles of the b-hadron (including c-hadron charged decays). This allows to evaluate the loss of information due to neutral decays of the b-hadron. It gives the maximum quality that can be achieved using tracks.

- *Tracks from b-hadron*: The b-hadron 4-momentum is evaluated from the sum of the 4-momenta of the reconstructed tracks originating from the b-hadron (including tracks of c-hadron charged decays). The comparison with the b-hadron evaluated from charged decay particles gives the loss of information from the tracking reconstruction and track to jet association efficiencies. It gives the maximun reconstruction quality that can be achived by MSV.

- *≥ 2-track vertices*: The b-hadron 4-momentum is evaluated from the sum of the 4-momenta of all reconstructed secondary vertices which contain at least one track originating from the b-hadron (including tracks of c-hadron charged decays). The MSV reconstruction quality is then estimated by comparing with the previous categories.

Figure 3.15 (left) shows the obtained p_{T}-ratio of the b-hadron evaluated using the three methods described above over the one of the original b-hadron in single-b-jets. A mean loss of 43% of the truth b-hadron p_{T} is observed with the track based evaluation. This loss is mostly due to the absence of neutral decays and is the maximum that the MSVF algorithm can achieve. With the current setup a further mean loss of 11% of the truth b-hadron p_{T} is observed when using vertices.

The ΔR distance of the truth b-hadron and the corresponding evaluation of its momentum in single-b-jets is shown in figure 3.15 (right). The mean ΔR distance using the vertex based evaluation is 0.045 which is 9 times smaller than the jet size threshold. However significant shifts in the ΔR distance are observed when going from the track based to the MSV based evaluation.

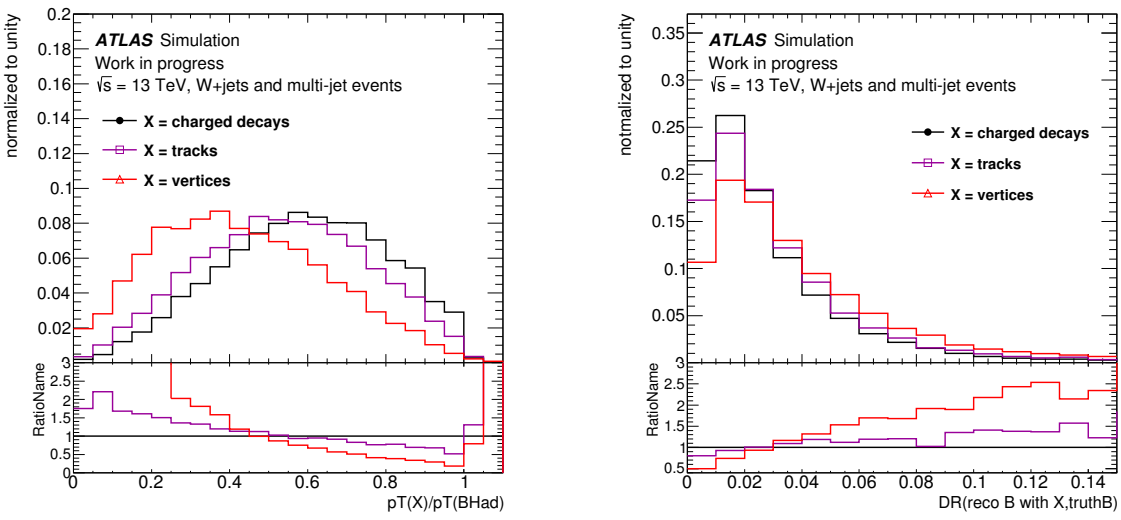


Figure 3.15.: p_{T}-ratio of the b-hadron evaluated using three methods (see text) over the one of the original b-hadron (left) and their ΔR distance (right) in single-b-jets. The b-hadron momentum is evaluated using either its charged decays (black), the tracks originating from this b-hadron (magenta) or the vertices which include at least one track originating from the b-hadron (red).

Quality of the reconstructed vertices in bb-jets

Up to four vertices are expected in bb-jets. In practice 51% of bb-jets have at least two vertices with at least two tracks which can be used as candidates for the two b-hadrons. In what follows only vertices with at least two tracks are considered. For Run 1 the bb-jet identification is based only on the two highest mass vertices. In addition 17% of the bb-jets have at least three such vertices. The usage of only two vertices in these jets correponds to a potential loss of information from the b-hadron and more complex b-hadron reconstructions are investigated.

Figure 3.16 shows the fraction of b-hadron tracks that are kept in vertices and the purity of these vertices in terms of b-hadron tracks for two configurations: the two highest mass vertices, and the set of all vertices. On average 56% of the tracks coming from both b-hadrons inside the jets are already recovered with the two maximum mass vertices. However the distribution is very broad and the fraction of recovered b-hadron tracks can vary by $\pm 17\%$ within a 1σ probability. 58% of the two maximum mass vertices are good vertices, i.e., with all their tracks originating from a single-b-hadron and the subsequent c-hadron. The fraction of vertices sharing tracks originating from different particles is 39% in bb-jets which is 15% higher than in single-b-jets. These shared vertices are separated in two categories:

- Shared vertices with tracks coming from the two b-hadrons only: 15% of the vertices. This category strongly reduces the bb-jet indentification performance as the vertex position and energy are averaged over the two b-hadrons.
- Shared vertices with background tracks: 24% of the vertices (similar to single-b-jets), the major contribution to vertices contaminated by background tracks.

The fraction of fake vertices, i.e. vertices with no track originating from a b-hadron, is only 4%. However, $\sim 17\%$ of the bb-jets have no vertices with tracks originating from one of the b-hadrons. These bb-jets typically have a low bb-jet identification efficiency.

On the other hand the set of all vertices increases the fraction of b-hadron recovered tracks compared to the two highest mass vertices. A limited cost of a 5% higher contribution from shared vertices with background tracks is observed. Furthermore, the fraction of fake vertices is reduced by $\sim 34\%$ in the set of all vertices compared to the two highest mass vertices.

In order to estimate the potential benefit from using all vertices in the jet, the b-hadron is reconstructed using the two sets:

- The two highest mass vertices are matched to their closest b-hadrons in (x, y, z).
- In the second set, all vertices are matched to their closest b-hadrons in (x, y, z). Two clusters are formed from the vertices matched to each b-hadrons. The 4-momentum and position of each cluster is estimated from the sum of the 4-momenta and the barycenter of the corresponding vertices, respectively. The two clusters are ordered by mass.

Figure 3.17 compares the distance in (x, y, z) of the reconstructed b-hadrons to the truth b-hadrons using the two sets described above. Jets where no vertex contains tracks from one of the b-hadrons are not considered. The leading and subleading mass vertices give a mean resolution of 1.1 mm and 1.7 mm, respectively. Even though the additional b-hadron tracks are recovered using all vertices, the two sets of reconstruction give the same distance between the reconstructed and the truth b-hadron within statistics.

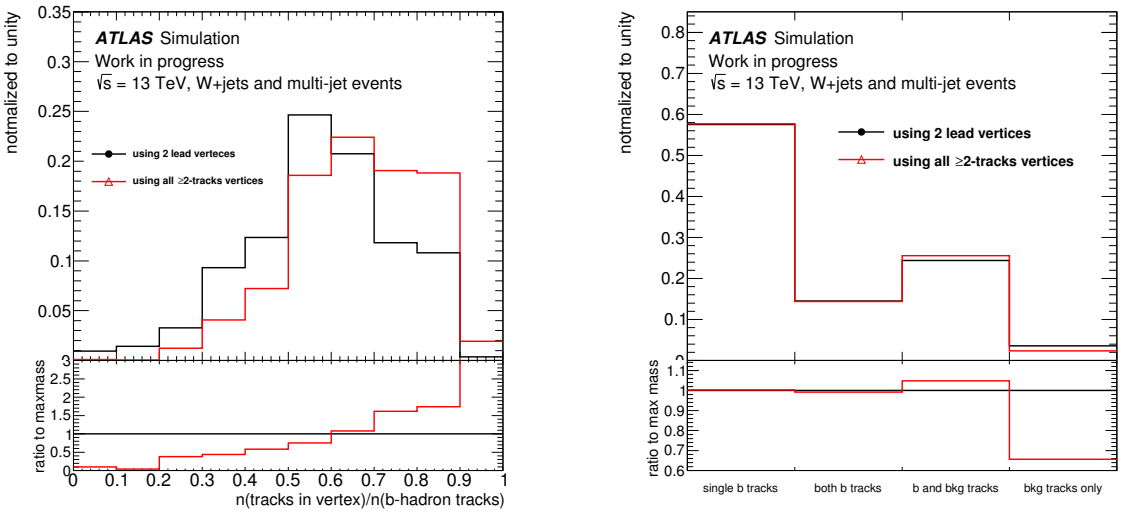


Figure 3.16.: Fraction of tracks coming from b-hadrons found in selected vertices (left) and purity of selected vertices (right). The vertex purity is evaluated separating vertices in four categories: vertices made of tracks originating from a single b-hadron decay chain (first bin), vertices made of tracks originating from both b-hadrons decay chains (second bin), vertices made of b-hadron decay chain tracks and background tracks (third bin), vertices made of only background tracks (fourth bin).

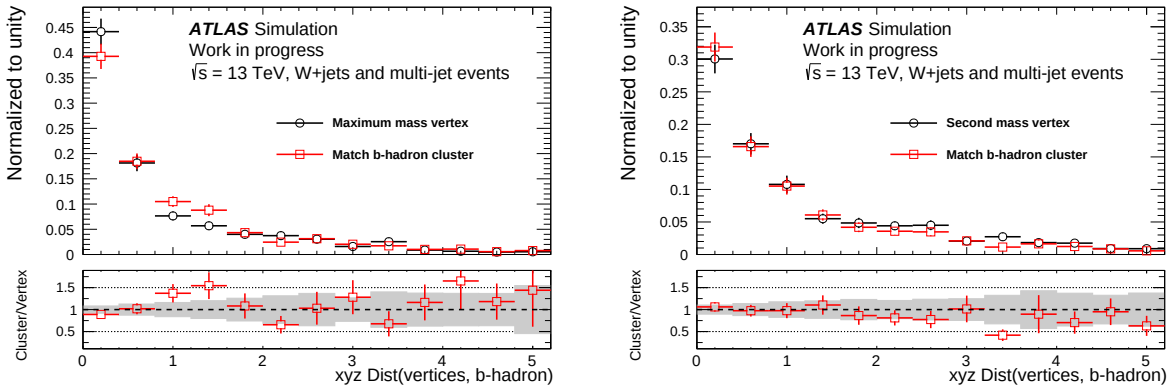


Figure 3.17.: Distance of the maximum mass (left) and second highest mass (right) clusters/vertices to the truth-b-hadron in the (x, y, z) volume. The black distribution is obtained using only the two highest mass vertices while red points are obtained from vertex clusters.

The set of all vertices provides a larger fraction of b-hadron tracks than the two highest mass vertices. However the reconstruction quality of the b-hadron after truth matching of the vertices does not show a significant difference between these two configurations. Thus the default $g \to b\bar{b}$ initiated jets taggers are based on the two highest mass vertices.

A vertex clustering technique using a MVA is investigated in appendix B. As expected from the truth matching study, no improvement is observed on the bb-jet identification.

3.4.3. The MultiSVbb algorithms

The MultiSVbb algorithms are $g \to b\bar{b}$ intiated jets identification algorithms. The main goal is to separate bb-jets from single-b-jets taking advantage of the reconstructed vertices to differentiate the b-hadron to c-hadron decay topology from $g \to b\bar{b}$ topologies. Boosted Decision Trees (see appendix A) are trained to separate bb-jets from single-b-jets, cc-jets, single-c-jets and $light$-jets using the parameters listed in table 3.4.

Parameter	Value
Maximun depth	4
Minimum node size	4%
Boosting	Gradient boost
Number of trees	250

Table 3.4.: Settings of the Boosted Decision Tree for the MultiSVbb algorithms. The definitions of the parameters are given in appendix A.

3.4.3.1. The MultiSVbb taggers inputs

We have seen in section 3.4.2.1 that the two highest mass vertices reconstructed by the MSVF algorithm give a good aproximation for the two b-hadrons in the jet. The MultiSVbb taggers are BDTs trained for the bb-jets identification against all other jet flavours, especially single-b-jets. The kinematic and topological information of the reconstructed vertices provides the first set of variables to be used in the BDTs. Several variables combining global information in the jet also provide a good separation between bb-jets and the other flavours such as the number of vertices (shown in figure 3.14). Two trainings are performed using different sets of variables corresponding to the MultiSVbb1 and MultiSVbb2 taggers. The MultiSVbb2 tagger includes more topological variables providing a higher performance than the MultiSVbb1 algorithm. With more topological variables MultiSVbb2 is however more sensitive to bias from the MC modelling. The full list of MultiSVbb variables is shown in table 3.5.

To prevent the BDT from focusing on a specific range of jet p_{T}, the jet p_{T} distribution is re-weighted to be flat for each jet flavor. The jet p_{T} variable is however kept in the training to benefit from correlations with the other variables.

The MultiSVbb taggers are trained with bb-jets as signal and a background composed of a mixture of single-b-jets, single-c-jets, cc-jets and $light$-jets. The fraction of each background component is optimized to maximise the single-b-jet rejection while keeping other jet flavour rejections at a good rate. These fractions are optimized for 13 TeV collisions and a background composition of 48% single-b-jets, 22% single-c-jets, 20% cc-jets and 10% $light$-jets is found to give the best performance.

Variable	Description	MultiSVbb1	MultiSVbb2
Jet properties:			
Jet p_{T}	Jet transverse momentum	✓	✓
$N(\mathrm{trk})$	Total number of tracks in the jet	✓	✓
$N(\mathrm{vtx})$	Total number of vertices in the jet	✓	✓
Total mass	Scalar sum of the vertices mass	✓	✓
SSVF diff $N(\mathrm{trk})$	$N(\mathrm{trk} \in \mathrm{jet}) - N(\mathrm{trk\ selected\ for\ SSVF})$	✓	✓
Mean $s(\mathrm{flight})$	Decay length significance averaged over vertices	✓	✓
Maximun E-frac	$\max_{\mathrm{vtx}_i \in \mathrm{jet}} [E(\mathrm{vtx}_i)/E(\mathrm{jet})]$, E = energy	–	✓
Leading and sub-leading mass vertices kinematic properties:			
$M(\mathrm{vtx}_1)$	Highest vertex mass	✓	–
$M(\mathrm{vtx}_2)$	Second highest vertex mass	✓	–
E-frac(vtx_1)	$E(\mathrm{vtx}_1)/E(\mathrm{jet})$	✓	–
E-frac(vtx_2)	$E(\mathrm{vtx}_2)/E(\mathrm{jet})$	✓	✓
vtx_1 $s(\mathrm{flight})$	Leading mass vertex decay length significance	✓	✓
vtx_2 $s(\mathrm{flight})$	Sub-leading mass vertex decay length significance	✓	–
Leading and sub-leading mass vertices topological properties:			
$\mathrm{Dist}_{xy}(\mathrm{vtx}_1, \mathrm{vtx}_2)$	(x, y) distance between the two leading mass vertices	–	✓
$\Delta R(\mathrm{vtx}_1, \mathrm{vtx}_2)$	ΔR between the two leading mass vertices	–	✓
$\Delta R(\mathrm{vtx}_1, \mathrm{jet})$	ΔR between the leading mass vertex and the jet	–	✓
$\Delta R(\mathrm{vtx}_2, \mathrm{jet})$	ΔR between the subleading mass vertex and the jet	–	✓
$\mathrm{Angle}(\mathrm{vtx}_1, \mathrm{vtx}_2)$	Angle between the (PV,vtx_1) and (PV,vtx_2) axes	–	✓

Table 3.5.: Input variable list of the MultiSVbb taggers.

3.4.3.2. The MultiSVbb taggers performance

Figure 3.18 shows the output of the MultiSVbb1 and MultiSVbb2 BDTs for each jet label. The rejection of the different backgrounds against the bb-jet efficiency for the MultiSVbb1 and MultiSVbb2 taggers is shown in figure 3.19. $light$-jets and c-jets are the easiest backgrounds to reject and are very well separated already by standard b-tagging techniques. However the cc-jets and single-b-jets backgrounds are very challenging. Indeed single-b-jets have one b-hadron in the jet and thus have at least one heavy vertex. On the other end cc-jets have two c-hadrons in the jet which are lighter and fly less than the vertices in bb-jets but have a similar topology.

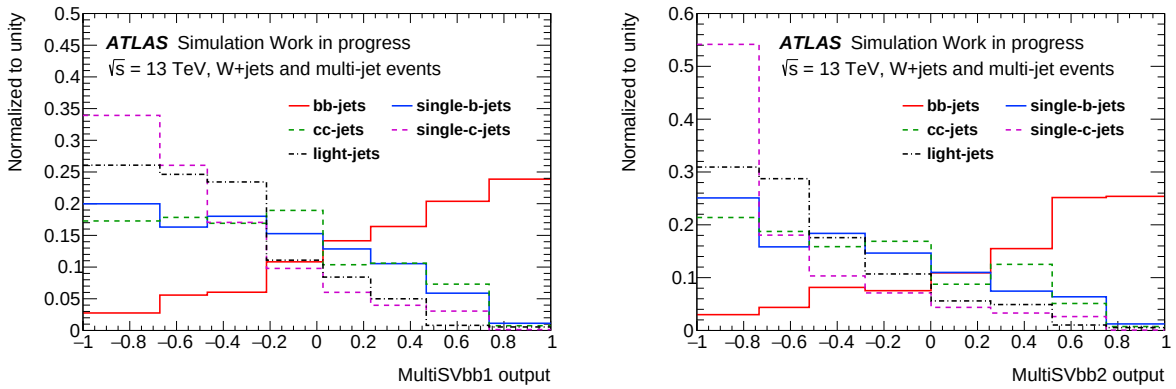


Figure 3.18.: BDT output of the MultiSVbb1 (left) and MultiSVbb2 (right) algorithms split according to the jet flavour.

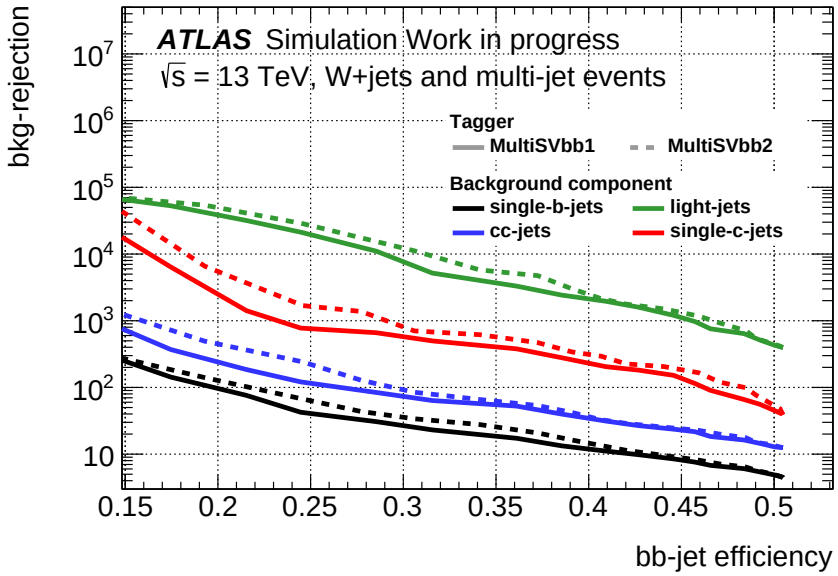


Figure 3.19.: Expected MultiSVbb1 and MultiSVbb2 bb-jet efficiency versus the various background rejections in a mixture of di-jet and W+jets events in $\sqrt{s} = 13$ TeV simulations.

As mentioned in section 3.4.2.1 the maximum efficiency of the MultiSVbb algorithms (51%) is limited by the efficiency to reconstruct at least two vertices. Thus a typical working point for bb-jet taggers is $\epsilon_{bb} = 35\%$. The rejection values of the different backgrounds for the Run 1 configuration tested with 8 TeV multi-jet and W+jet events at the $\epsilon_{bb} = 35\%$ working point [134] is compared to the background rejections obtained for the same bb-jet efficiency with the Run 2 configuration at 13 TeV in table 3.6. Similar performance is observed for the single-b-jet rejection while all the other backgrounds show an improved separation with bb-jets for Run 2. MultiSVbb algorithms typically achieve a rejection of 20 for a bb-jet efficiency of $\epsilon_{bb} = 35\%$ where the MV2c20 b-tagging algorithm only gives a single-b-jet rejection of 3 at the same working point. The low operating efficiency of MultiSVbb taggers makes these algorithms better suited to measurements, which can benefit from high single-b-jet rejection, than searches, which also need high signal efficiencies.

Flavour	single-b-jet	cc-jet	single-c-jet	*light*-jet
Run 1				
MultiSVbb1	18	35	200	2400
MultiSVbb2	23	38	250	3200
Run 2				
MultiSVbb1	19	55	390	3600
MultiSVbb2	24	63	600	5500

Table 3.6.: Comparison of the Run 1 and Run 2 background rejections of MultiSVbb algorithms at 35% bb-jet efficiency in a mixture of di-jet and W+jets events.

Similarly to standard b-tagging, the bb-tagging performance depends on the jet kinematics, especially the jet p_{T}. The single-b-jet rejection for a global $\epsilon_{bb} = 35\%$ (upper row) and a flat $\epsilon_{bb} = 35\%$[b] (lower row) as a function of the jet p_{T} in the same conditions as above is shown in figure 3.20 for the MultiSVbb1 (left column) and MultiSVbb2 (right column) algorithms. The Run 1 and Run 2 configurations have different shapes. While the Run 1 algorithms reject more single-b-jets at low p_{T}, the Run 2 algorithms give better performance for medium p_{T} jets. A large improvement is brought by the 13 TeV optimization. In particular, when requiring a flat 35% bb-jet efficiency, the single-b-jet rejection is improved in all jet p_{T} bins.

[b] A flat efficiency $\epsilon_{bb} = X$ is obtained requiring that each bin in the distribution has $\epsilon_{bb} = X$ independently of the other bins.

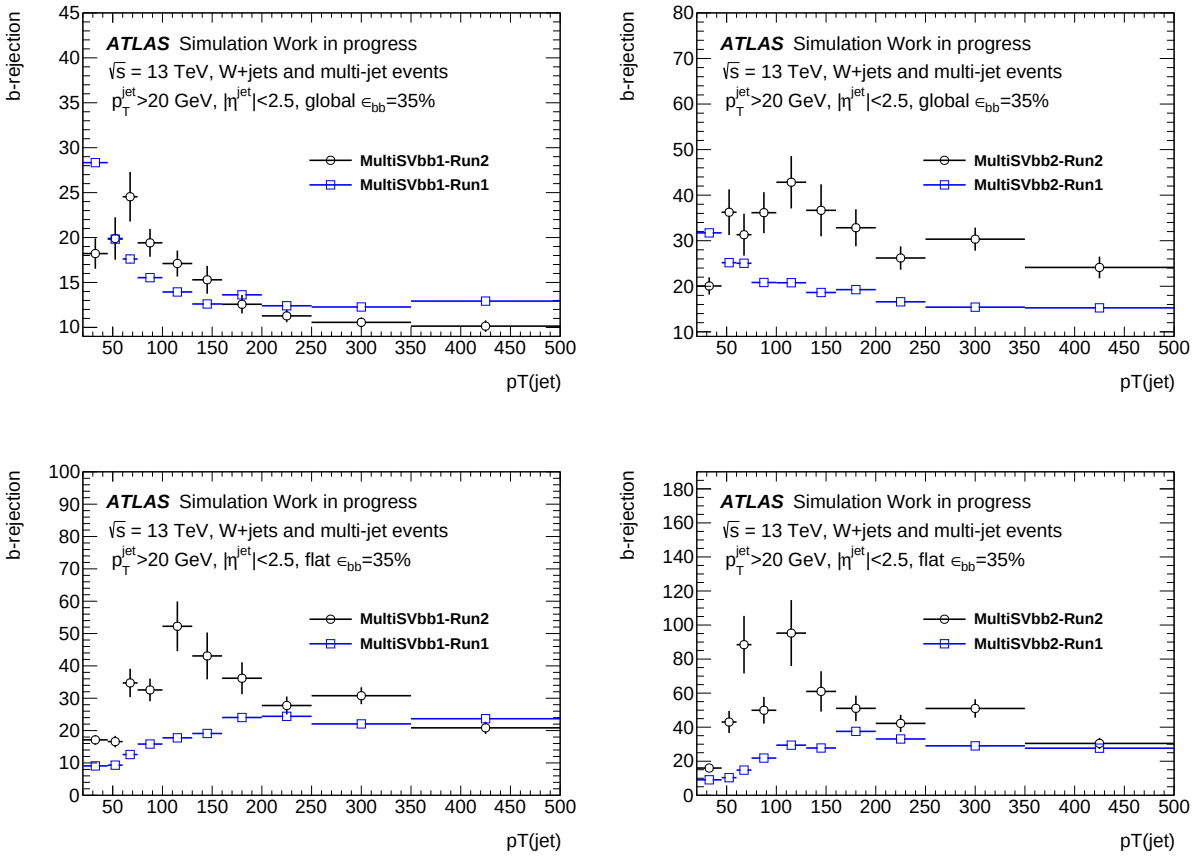


Figure 3.20.: b-jet rejection as a function of the jet p_T for a global (fixed) 35% efficiency of bb-jets for the top (bottom) row using the MultiSVbb1 (MultiSVbb2) algorithm in the left (right) column.

3.5. Summary

b-tagging is a crucial ingredient to many physics analyses, especially the $t\bar{t}H(H \to b\bar{b})$ search. In order to fully benefit from the improvements in b-tagging performance for Run 2, especially thanks to the new pixel layer (IBL), a detailed understanding of the b-jet definition in MC simulations is required.

The heavy flavour jet truth labelling mostly relies on the particle to jet association. For Run 2 a particle level definition using hadrons is preferred over parton level definition using quarks to avoid ambiguities in MC definitions of partons. A p_{T} cut of 5 GeV is applied to heavy flavour hadrons. This cut removes low p_{T} b-hadrons coming from the low tail of the fragmentation function, which are hard to model. I have studied two hadrons to jets association schemes: a cone-based exclusive matching, the ΔR algorithm, and the ghost association which is based on the anti-k_t jet clustering algorithm. I found that both algorithms are constitent with an overall 99% agreement in $t\bar{t}$ events.

However the small differences in jet labelling have a non-negligible impact on the $light$-jet rejection assesment. I found that this effect mostly comes from b-hadrons splitting in two components resulting in two nearby-jets carrying half of the b-hadron energy or only one jet with half of the b-hadron energy, the rest being lost by acceptance cuts. In nearby-jet topologies, the ΔR scheme associates the b-hadron to the closest jet while the ghost association scheme matches the b-hadron to the highest p_{T} jet. If the b-hadron energy splits in the two jets, tracks are matched to the jet using a cone-based association and thus are mostly associated to the ΔR-associated b-jet. Thus the sample of $light$-jets obtained from the ghost association procedure contains a fraction of jets with high efficiencies. These jets are not in the $light$-jet sample obtained with the ΔR algorithm. This translates in a reduced expected rejection of $light$-jets when using the ghost association with respect to the ΔR algorithm.

In order to avoid large MC corrections with large uncertainties when measuring the b-tagging efficiencies in data, the b-jet definition has to pair with the b-tagging algorithms. Thus the ΔR matching scheme is chosen for the default Run 2 labelling in b-tagging.

b-tagging algorithms do not separate b-jets containing a single b-hadron from $g \to b\bar{b}$ initiated jets. However a bb-jet identifier would be beneficial to constrain the gluon splitting to heavy flavour quarks in QCD. Moreover in Run 2 $H \to b\bar{b}$ searches in boosted topologies become accessible. In such topologies the $g \to b\bar{b}$ at small opening angles becomes an important background.

The MultiSVbb taggers are $g \to b\bar{b}$ initiated jets identifiers based on multi-secondary-vertices. I studied the input vertices of the MultiSVbb algorithms, comparing the set composed of the two highest mass vertices to the set of all vertices in the jet. I have shown that the set of all vertices allows to recover more b-hadron tracks than the set of the two highest mass vertices. However these additional vertices do not provide an improved precision on the b-hadron reconstruction and thus do not improve the MultiSVbb performance. Therefore, the set of the two highest mass vertices is kept as input for the MultiSVbb algorithm.

I have also optimized MultiSVbb taggers for Run 2, revisiting the background composition, the BDT training parameters and the input variables corresponding to properties of the reconstructed vertices. I obtained rejection of typically 20 for single-b-jets at the $\epsilon_{bb} = 35\%$ working point which is seven times higher than the rejection obtained with MV2c20 at the same working point. MultiSVbb taggers performance depends on the jet p_{T} with a peak at $p_{\mathrm{T}}(\mathrm{jet}) \sim 100$ GeV. It is shown that the Run 2 setup for the MultiSVbb taggers outperforms the Run 1 setup for most of the jet p_{T} bins.

4. Search for the Higgs boson in the $t\bar{t}H(H \to b\bar{b})$ channel

The production of the Higgs boson in association with top quarks, in particular the $t\bar{t}H$ channel (see section 1.3.1), provides a unique access to the Yukawa coupling of the Higgs boson to the top quark. This coupling is of substantial importance to asses the SM behavior of the observed Higgs boson. A comparison of the direct measurement of this coupling to its indirect measurement in ggH (see 1.3.1) allows to characterize the content of the loop in ggH and reveal potential BSM contributions.

The $t\bar{t}H$ production mode is split into three main analyses depending on the Higgs boson decay mode: $H \to b\bar{b}$, $H \to$multi-leptons and $H \to \gamma\gamma$. This thesis focuses on the $t\bar{t}H(H \to b\bar{b})$ channel. Even being challenging the $t\bar{t}H(H \to b\bar{b})$ channel is expected to give high sensitivity to the Higgs boson coupling to quarks as it involves (at leading order) only the couplings of the Higgs boson to top or bottom quarks. The Higgs boson coupling to the top quark can then be constrained in the combination with the other decay modes.

This chapter reviews the Run 1 $t\bar{t}H(H \to b\bar{b})$ searches in section 4.1 followed by an overview of the Run 2 analysis in section 4.2. I have contributed to various studies that lead to the initial object and event selections as well as the choice of MC generators which are described in sections 4.3 and 4.4 respectively. In particular, I have studied in details the MC modelling of data for the $t\bar{t}+$ jets process; part of these studies are shown in section 4.5. I also contributed to the improvement of the event classification strategy with respect to the Run 1 analysis; the adopted event classification strategy is described in 4.6. The $t\bar{t}H(H \to b\bar{b})$ analysis uses multi-variate techniques in order to enhance the signal sensitivity. The separation of the $t\bar{t}H(H \to b\bar{b})$ signal from the $t\bar{t}+$ jets background is highly improved by using inputs coming from the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ final state. Section 4.7 describe the available reconstruction techniques in the $t\bar{t}H(H \to b\bar{b})$ analysis and a promising novel technique that I proposed and developed for future iterations of the analysis.

4.1. The $t\bar{t}H(H \to b\bar{b})$ Run 1 legacy

The $t\bar{t}H(H \to b\bar{b})$ events are first separated in three analyses channels according to the W-bosons decay mode:

- The all-hadronic channel is obtained when both W-bosons decay hadronically. This channel presents the highest branching ratio of 46%. However the absence of any lepton (lepton stands for electron or muon for the rest of this thesis) and the presence of many jets makes this channel very difficult to separate from the multi-jet background, especially at trigger level. This channel is the subject of a stand-alone analysis and will not be presented in this thesis.
- The di-lepton channel is obtained when both W-bosons decay leptonically. It provides the cleanest topology with a very high separation from multi-jet background. However this channel suffers from a low branching ratio of 10%.

- The single lepton channel provides a compromise between high branching ratio and relatively clean topology with one W-boson decaying leptonically and the other hadronically. It offers a branching ratio close to the all-hadronic channel 44% with also one lepton allowing to extract the signal from the multi-jet background.

The $t\bar{t}H(H \to b\bar{b})$ single lepton and di-lepton channels follow the same strategy but are analyzed separately and combined in the final fit.

The ATLAS [139] and CMS [140] searches for $t\bar{t}H(H \to b\bar{b})$ production in Run 1 are performed using roughly 20 fb^{-1}of $\sqrt{s} = 8$ TeV data. In both experiments no evidence of the $t\bar{t}H$ production mode or of the $H \to b\bar{b}$ decay mode was found with an observed signal significance of 1.3σ in the ATLAS analysis.

For both analysis the $t\bar{t}H(H \to b\bar{b})$ signal production is parametrised by the ratio of the observed rate over the SM prediction: $\mu = \sigma/\sigma_{\text{SM}}$. The μ parameter is usually referred to as signal strength. The best fit values for the signal strength in both experiments are shown in figure 4.1. The observed signal strength is $\mu = 1.5 \pm 1.1$ for the ATLAS search and $\mu = 1.2^{+1.6}_{-1.5}$ for the CMS analysis. They both are compatible with the SM prediction within 1σ.

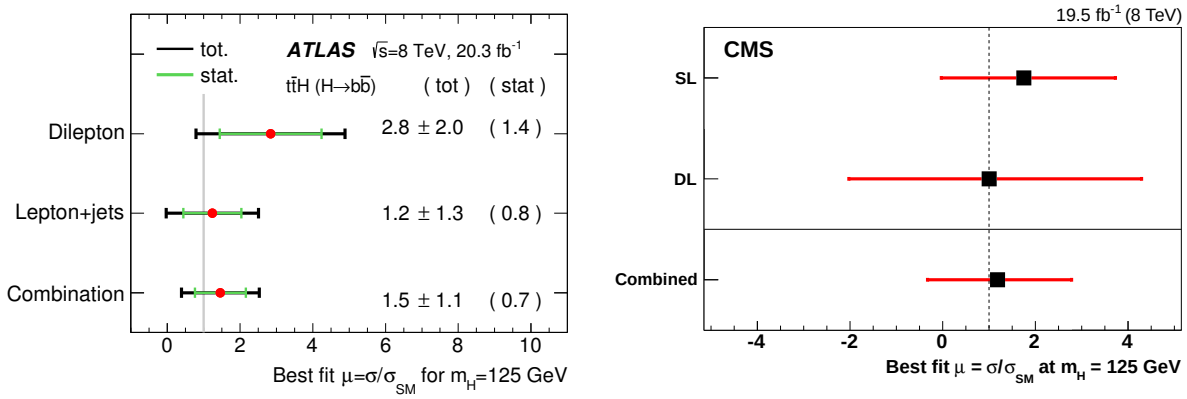


Figure 4.1.: Observed values of the $t\bar{t}H(H \to b\bar{b})$ signal strength $\mu = \sigma/\sigma_{\text{SM}}$ obtained from the best fit to 20 fb^{-1}of $\sqrt{s} = 8$ TeV data within the ATLAS (left) [139] and CMS (right) [140] experiments.

The $t\bar{t}+jets$ background is one of the main challenges of the $t\bar{t}H(H \to b\bar{b})$ analysis. In the phase space of the $t\bar{t}H(H \to b\bar{b})$ signal, with high number of jets and b-tagged-jets, this background has large theoretical uncertainties and is poorly constrained by existing data measurements. The final sensitivity of the analysis is driven by the modelling of the $t\bar{t}+$ jets background and the most important sources of uncertainties on the signal strength are systematic uncertainties on the $t\bar{t}+$ jets (mainly $t\bar{t}$ plus additional b-jets, $t\bar{t}+\geq 1b$) process as can be seen in figure 4.2.

Due to the presence of 4 b-quarks in the $t\bar{t}H(H \to b\bar{b})$ final state, b-tagging plays a key role in the analysis. Figure 4.2 shows that systematic uncertainties on b-tagging efficiencies also have a significant impact on the $t\bar{t}H(H \to b\bar{b})$ sensitivity.

In Run 2 a *preliminary result* is presented at the ICHEP conference in 2016 using the 3.2 fb^{-1}of data available from the 2015 run, and 10.0 fb^{-1}of data from early 2016. The analysis follows the Run 1 strategy. The main improvement in the analysis chain is the addition of a new MVA technique for the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ final state. The background and fit models are also revisited to improve the analysis sensitivity, in particular by constraining the $t\bar{t}+$ jets background. The sensitivity is similar to the Run 1 analysis with the available luminosity which is two third compared to the Run 1

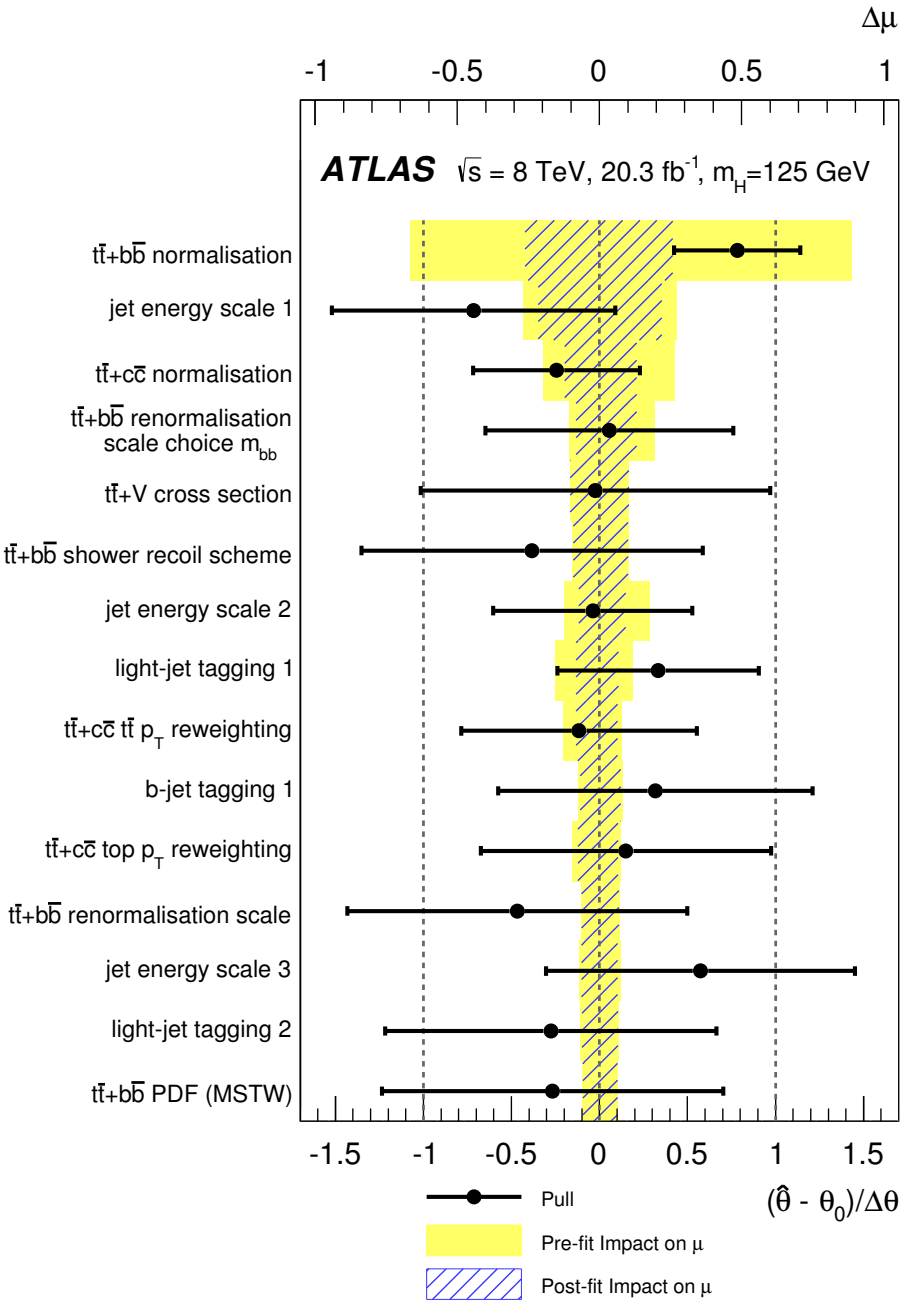


Figure 4.2.: Ranking of the nuisance parameters used in the fit according to their impact on the sensitivity of the $t\bar{t}H(H \to b\bar{b})$ analysis in the ATLAS experiment [139]. Only the 15 leading ones are shown. The black points are plotted according to the bottom axis and there displacement with respect to 0 show the deviation of the measured value of the nuisance parameter in units of pre-fit 1σ variation. The black error bars show the post-fit uncertainty after applying the constraint from data. The blue hashed (yellow filled) areas correspond to the post(pre)-fit effect of the systematic uncertainty on the signal strength. The post(pre)-fit impact on sensitivity is computed performing the fit fixing the nuisance parameter at the post(pre)-fit $\pm 1\sigma$ variation and taking the difference in the fitted μ with the default fit.

luminosity. A signal strength of 1.6 ± 1.1 is observed in single lepton channel.

An *updated analysis* is ongoing with the full $2015 + 2016$ datasets for an overall collected luminosity of 36.1 fb^{-1} and is expected to be published soon [141]. A great fraction of the tools and methods presented in this thesis are developed to understand and produce the preliminary result and improved for the updated analysis. This document presents the studies done using the full 2015 and 2016 data.

4.2. Overview of the $t\bar{t}H(H \to b\bar{b})$ analysis

The $t\bar{t}H(H \to b\bar{b})$ analysis strategy for Run 2 is similar to the Run 1 strategy. Standard leptons and jets are first reconstructed (see section 4.3) and the analysis is split in the single lepton and di-lepton channels if exactly one or two isolated leptons are found, respectively. A global selection on the number of jets and b-jets is then applied focusing on $t\bar{t}+$ jets event topology. Events are further categorized according to the number of jets and b-tagged-jets in order to increase the signal purity (see section 4.6.1). However, some categories are designed to be enriched in one or more backgrounds to help constraining the systematic uncertainties on these backgrounds.

Even after the event categorization, the purest signal regions at high number of jets and b-tagged-jets only reach a signal over background ratio of 5%. Thus multi-variate techniques are used to enhance the sensitivity. First a BDT aiming at the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ system is used. Topological and kinematic discriminating variables for the $t\bar{t}H(H \to b\bar{b})$ and $t\bar{t}+$ jets hypotheses are defined based on the reconstructed Higgs boson and top quarks candidates. They are combined with other MVA techniques and event kinematic variables in a classification BDT trained for the separation of the $t\bar{t}H(H \to b\bar{b})$ signal from the $t\bar{t}+$ jets background.

Finally a simultaneous fit of all background-enriched and signal-enriched categories is performed in the single lepton and di-lepton channels individually. In signal-enriched regions the classification BDT is fitted in order to maximize the sensitivity to the $t\bar{t}H(H \to b\bar{b})$ signal. Background-enriched regions are used to constrain the large uncertainties in the $t\bar{t}+$ jets modelling. One bin or the $H_{\mathrm{T}}^{\mathrm{had}} = \sum_{\mathrm{jets}} p_{\mathrm{T}}(\mathrm{jet})$ distributions are fitted in these categories. The final result is then extracted from a combined fit of both channels to data. The single lepton fit and combined fit of the $t\bar{t}H(H \to b\bar{b})$ analysis are presented in great details in chapter 5.

This thesis focuses on the single lepton channel and the combined fit with the di-lepton channel. Therefore the di-lepton channel is not detailed in the following sections.

4.3. Object and event selections

The analysis is performed using data events from pp-collisions at $\sqrt{s} = 13$ TeV recorded with the ATLAS detector in years 2015 and 2016. Events are selected only in periods where all sub-components of the detector are operational giving an integrated luminosity of 3.2 ± 0.1 fb^{-1} in 2015 and 32.9 ± 0.7 fb^{-1} in 2016. Data events typically contain 10 to 30 vertices from multiple pp-collisions in each bunch crossing. Only events containing at least one primary vertex associated with two or more tracks with $p_{\mathrm{T}} > 0.4$ GeV are kept. The primary vertex of interest, or hard scatter vertex, is separated from pile-up vertices by selecting the vertex with the largest squared sum of p_{T} of its associated tracks.

Events are selected using unprescaled single-lepton triggers with requirements depending of the lepton p_{T} as described in table 4.1.

Electrons and muons are reconstructed using standard ATLAS algorithms described in section 2.4.2 and 2.4.3. A common set of selections between all $t\bar{t}H$ decay channels in the various final states is first applied to avoid the overlap in the combination. These requirements are typically looser than the final requirements used in each of the channels. In the single lepton channel of the $t\bar{t}H(H \to b\bar{b})$

Reconstructed object	p_{T} threshold [GeV]	Identification menu	Isolation menu
2015 (2016) datasets			
Electrons:	$\geq 24(26)$	Medium (Tight)	Gradient
	≥ 60	Medium	None
	≥ 120	Loose	None
Muons:	$\geq 20(26)$	Loose	Loose (Gradient)
	≥ 50	None	None

Table 4.1.: Lepton triggers used for the $t\bar{t}H(H \to b\bar{b})$ analysis.

analysis, events with exactly one such lepton are kept. Further requirements are then applied for individual channels. In the single lepton channel, leptons are required to have $p_{\mathrm{T}} > 27$ GeV and electrons are further required to pass the tight likelihood identification criterion [118]. The contribution of leptons originating from hadronic decays (non-prompt leptons) is reduced by applying the gradient isolation [118]. Finally cuts on the longitudinal impact parameter, $|z_0 \sin\theta| < 0.5$ mm, and on the transverse impact parameter significance, $s(d_0) < 5(3)$ for electrons (muons), ensure the compatibility of lepton tracks with the hard scatter vertex.

Lepton efficiencies in simulations are corrected using *scale-factors (SF)* to match the ones measured in data. They are obtained from tag and probe methods in $Z \to ll$ and $J/\Psi \to ll$ events for both the muons [118] and electrons [142] efficiency measurements.

Standard *reco-jets* reconstructed with the anti-k_t $R = 0.4$ algorithm, using topological clusters in the calorimeters as inputs, and corrected by the JES calibration (see section 2.4.4) are used in the $t\bar{t}H(H \to b\bar{b})$ analysis. Calibrated jets are required to have $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$. To reduce the contribution from pile-up jets, any jet with $p_{\mathrm{T}} < 60$ GeV and $|\eta| < 2.4$ is rejected if its JVT output (see 2.4.4) is below 0.59. An overlap procedure between jets and leptons avoids the usage of single EM calorimeter detector responses twice in an event. Events are required to have at least 5 selected jets in the single lepton channel.

The identification of b-jets is done using the MV2c10 algorithm (see section 3.3.2). As mentioned in chapter 3 the b-tagging efficiencies for each working point and each jet flavour in MC simulations are corrected by scale-factors to match data. The $light$-jets efficiencies are corrected to match measured $light$-jet rates in 13 TeV di-jet events, using reversed b-tagging algorithms which are meant to identify $light$-jets. b-jets (c-jets) efficiencies are measured in 13 TeV $t\bar{t}$ di-lepton data events with two (two or three) selected jets. The b-tagging scale factors (and the corresponding uncertainties) are measured independently for each b-tagging working point. These are then combined in the so called *pseudo-continuous* calibration which allows the simultaneous usage of several working points. Selected events are required to have at least two b-tagged-jets at the $\epsilon_b = 60\%$ working point or at least three b-tagged-jets at the $\epsilon_b = 77\%$ working point.

Boosted objects are also considered in the single lepton channel. The so called *large-R-jets* are reconstructed with a re-clustering [143] of reco-jets using the anti-k_t algorithm with $R = 1.0$. Constituent jets of the large-R-jet which p_{T} account for 5% or less of the large-R-jet p_{T} are likely to originate from pile-up or soft radiation and are removed [144].

To be selected as boosted, an event is required to contain at least one standard jet and at least two large-R-jets with $p_{\mathrm{T}} > 200$ GeV, $M > 50$ GeV and $|\eta| < 2.0$. Additionally one of the large-R-jets must be identified as a Higgs boson candidate and one of the remaining large-R-jets as a hadronically-decaying

top (hadronic-top) candidate. Boosted Higgs boson candidates are identified as large-R-jets with two constituent jets being b-tagged at $\epsilon_b = 85\%$. Among the candidates the one with the highest sum of jet b-tagging weights is selected. The hadronic-top is found as the large-R-jet, which is not selected as a Higgs candidate, of highest mass with $p_{\mathrm{T}} > 250$ GeV. At least one constituent b-tagged-jet and one constituent jet not b-tagged at $\epsilon_b = 85\%$ working point are required in the hadronic-top candidate.

The missing transverse energy is reconstructed as described in section 2.4.6. It is not used in the event selection but enters the event reconstruction to describe the neutrino from the leptonic decay of the W-boson.

4.4. Signal and background predictions

All Monte Carlo (MC) simulations of the $t\bar{t}H(H \to b\bar{b})$ analysis are produced with the standard ATLAS simulation software described in section 2.3. Nominal MC predictions for all processes are simulated with the fullsim procedure to achieve the best precision of data modelling. The fastsim procedure is used for alternative samples entering the definition of systematic uncertainties (see also section 5.2). The same selections as for data are applied to MC events. The averaged number of pp collisions in simulated events is based on the foreseen pile-up profile of the LHC runs. It is then corrected to match the observed distribution in data. EvtGen [129] is used to simulate heavy-flavour hadron decays for all samples but the ones simulated with Sherpa.

4.4.1. $t\bar{t}H$ MC simulation

The Madgraph5_aMC@NLO package provides the $t\bar{t}H$ signal model at NLO accuracy for the matrix element. The Higgs boson mass is fixed to 125 GeV. All Higgs boson decays are included using the latest branching ratio calculation reported in [59]. The NNPDF3.0NLO parton distribution function set [145] is used for the matrix element. The parton shower is modelled with PYTHIA8 with the A14 tune [136] and NNPDF2.3NLO parton distribution function set [137]. The $t\bar{t}H$ cross section is corrected to match the latest calculations also reported in [59].

4.4.2. $t\bar{t}$ background MC simulation

The nominal sample of $t\bar{t}+$ jets events is generated using the Powheg-Box NLO generator and the NNPDF3.0NLO parton distribution function set. The p_{T} of the first additional emission to the $t\bar{t}$ system is controlled by the $hdamp$ parameter. This parameter is optimized and a value of $1.5 \cdot m_t$ is found to give the best description of $\sqrt{s} = 8$ and 13 TeV data [146]. The showering is done in PYTHIA8 [105] using the A14 tune and NNPDF2.3NLO parton distribution function set. The $t\bar{t}$ cross section is set to the latest NNLO QCD calculation with NNLL resummation of soft gluons terms of 832^{+46}_{-51} pb [147]. The POWHEG+PYTHIA8 sample is referred to as *PP8* from now on.

The $t\bar{t}$ production with additional heavy flavoured jets (especially b-jets) represent the most important background in the $t\bar{t}H(H \to b\bar{b})$ analysis. Therefore, the $t\bar{t}+$ jets sample is divided into three orthogonal components. This division is performed before reconstruction using "truth" objects from the MC simulation. It is based on the presence of truth-jets in the event which do not originate from the $t\bar{t}$ decay chain (additional jets). These truth-jets are required to have $p_{\mathrm{T}} > 15$ GeV and $|\eta| < 2.5$. The three $t\bar{t}+$ jets components are defined as follows:

- The $\boldsymbol{t\bar{t}+\geq 1b}$ component is obtained when at least one of the additional jets is truth-matched to at least one b-hadron within $\Delta R(\text{jet}, b\text{-hadron}) < 0.4$. At least one b-hadron in the jet is required to have $p_{\text{T}} > 5$ GeV.
- Non $t\bar{t}+\geq 1b$ events are labelled $\boldsymbol{t\bar{t}+\geq 1c}$ if one the additional jets is truth-matched to at least one c-hadron with the same requirements.
- Other events are labelled $\boldsymbol{t\bar{t}}$ **+ light**.

The $t\bar{t}+$ heavy flavours sample refers to the $t\bar{t}+\geq 1c$ and $t\bar{t}+\geq 1b$ contributions together.

A finer decomposition of the $t\bar{t}+$ heavy flavours contributions in sub-components is used to apply corrections on the relevant sub-components or define systematics. In this classification truth-jets matched to several b-hadrons are referred to as B-jets and single-b-jets include only jets matched to exactly one b-hadron. The $t\bar{t}+\geq 1b$ sample is then split in four sub-categories. Events with exactly one single-b-jet are labelled $t\bar{t}+b$, those with exactly two single-b-jets are labelled $t\bar{t}+b\bar{b}$, and with exactly one B-jet are labelled $t\bar{t}+B$, the rest of the $t\bar{t}+\geq 1b$ background enters the $t\bar{t}+\geq 3b$ sub-component. Events that contain b-jets from Multi-Particle Interactions (MPI) or from the showering of $t\bar{t}$ decay products (FSR) are considered in separate categories. These represent a small fraction of events and are not present in all MC generators. The $t\bar{t}+\geq 1c$ background is divided analogously.

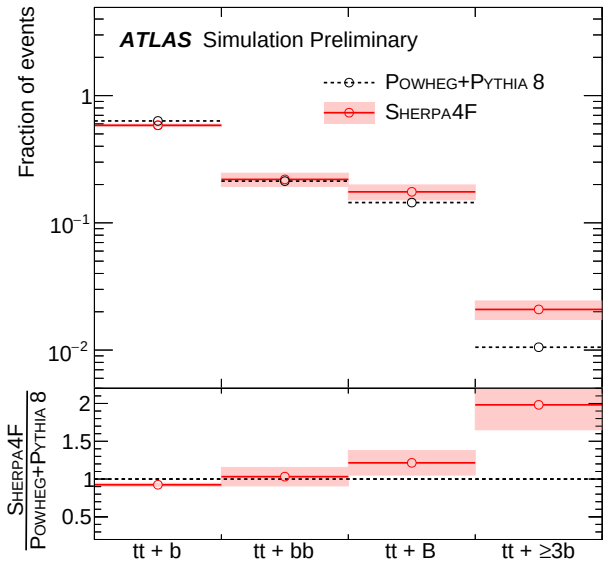


Figure 4.3.: Predicted fractions of the $t\bar{t}+\geq 1b$ sub-components in the inclusive POWHEG+PYTHIA8 $t\bar{t}$ sample and in the four-flavour SHERPA+OPENLOOPS NLO $t\bar{t}+b\bar{b}$ sample. The shaded area represent the SHERPA+OPENLOOPS systematic uncertainties as explained in section 5.2.2.

In signal regions the $t\bar{t}+\geq 1b$ contribution is overwhelming the other backgrounds and the signal. In order to achieve the best possible precision on the $t\bar{t}+\geq 1b$ background a SHERPA+OPENLOOPS (referred to as Sherpa+OL) $t\bar{t}+b\bar{b}$ sample at NLO is produced. It provides a matrix element prediction for the $t\bar{t}$ plus two b-jets while the PP8 simulation can only produce additional b-jets from the parton shower. It is produced with SHERPA 2.1 and the CT10 four-flavour scheme (4FS) PDF set which allows to use massive b-quarks.

There exist no clear theoretical prescription on how to mix the $t\bar{t}+b\bar{b}$ sample with the inclusive $t\bar{t}$+jets samples and remove the overlap with the $t\bar{t}$ production with additional b-jets. This complicates the usage of the Sherpa+OL sample as the nominal prediction for the $t\bar{t}+\geq 1b$ components. Instead, the relative contribution of the $t\bar{t}+\geq 1b$ sub-categories ($t\bar{t}+b$, $t\bar{t}+b\bar{b}$, $t\bar{t}+B$, $t\bar{t}+\geq 3b$) in the PP8 sample are corrected to match those of the Sherpa+OL sample. The differences between the two predictions in the fractions of each sub-component are shown in figure 4.3.

Although Sherpa+OL provides state-of-the-art accuracy (at NLO) for the $t\bar{t}+b\bar{b}$ process, no evidence, as of summer 2017, is found to indicate that the Sherpa+OL prediction better describes the $t\bar{t}+\geq 1b$ background in data compared to the PP8 prediction. Thus a second model based on the PP8 prediction, without any corrections of the $t\bar{t}+\geq 1b$ sub-components, is proposed in this thesis and is described in section 5.2.2. The PP8 $t\bar{t}$ sample with corrected fractions of $t\bar{t}+\geq 1b$ sub-components is referred to as the default $t\bar{t}$ sample. The one without these correction is referred to as the PP8-based sample.

4.4.3. Other backgrounds

Several other backgrounds are considered in the $t\bar{t}H(H\rightarrow b\bar{b})$ analysis in total accounting for 8% of the total amount of background. The merged contribution of all these backgrounds is referred to as *non-*$t\bar{t}$ in the rest of this document.

The first category of additional backgrounds comes from W+jets, Z+jets and di-boson plus jets processes which represent together 3% of the total amount of background. They are generated with the Sherpa 2.2.1 generator [99] and its integrated parton shower model. The Comix and OpenLoops matrix element generators are used to account for up to two additional partons at NLO and up to four additional partons at LO. They are then merged with the parton shower using the ME+PS method (see section 2.3.1). In addition the Z+heavy-flavour-jets contribution is increased by a factor 1.3 in order to match the data in Z+jets control regions.

$t\bar{t}Z$ and $t\bar{t}W$ processes which represent 0.4% of the background are generated with the same setup as the $t\bar{t}H$ signal.

4% of the total amount of background is coming from other top production modes. Single-top processes are all generated with Powheg-Box using the CT10 parton distribution function set [148] and interfaced with the PYTHIA6 parton shower with the Perugia 2012 tune. The electroweak t-channel uses a four-flavour-scheme calculation accounting for massive b-quarks. The overlap between the Wt and $t\bar{t}$ production modes is removed from the Wt sample using the "diagram removal" scheme [149]. Additional sources of tops from the $t\bar{t}WW$, WtZ, tZ and 4-top production modes are considered. They are all generated with Madgraph5_aMC@NLO. The tZ sample is obtained using the PYTHIA6 parton shower with the Perugia 2012 tune and considering massive b-quarks. The others three components use the PYTHIA8 parton shower with the A14 tune and NNPDF2.3LO parton distribution function set.

Finally single-top plus Higgs boson production modes (see section 1.3) are a negligible contribution to the main $t\bar{t}H(H\rightarrow b\bar{b})$ analysis. They are however a valuable input to the Higgs boson to top quark coupling measurement in the $t\bar{t}H$ combination as their cross section is asymmetric with respect to the sign of the Higgs to top coupling. Thus they are included as additional backgrounds and are enabled to contribute to the coupling of the Higgs boson to the top-quark in the $t\bar{t}H$ combination. The WtH sample is generated with Madgraph5_aMC@NLO interfaced to HERWIG++ with the CTEQ6L1 parton distribution function set. The $tHb+$jets sample is produced with MADGRAPH 5 interfaced to PYTHIA8 with the CT10 parton distribution function set.

In the single lepton channel, 1% of the background includes fake or non-prompt leptons (referred

to as fakes in what follows) with an important contribution from multi-jet heavy-flavour production. Such processes are hard to model theoretically and are thus extracted from data using the Matrix method [146].

This method is based on the measurement of the efficiencies of fakes and real leptons to pass a loose identification requirement and the tight criteria of the $t\bar{t}H(H \to b\bar{b})$ analysis. Fake efficiencies are measured in multi-jet enriched regions. Real efficiencies are extracted using a $Z \to ll$ tag-and-probe method.

The fake background contribution is neglected in some of the analysis categories where the corresponding estimate is statistically compatible with zero.

4.5. $t\bar{t}$ sample modelling in data

The selection described in section 4.3 leads to a sample dominated by $t\bar{t}+$jets events. This selection is referred to as *inclusive selection*. Basic quantities are used to assess the $t\bar{t}$ MC description of data. The default $t\bar{t}$ model is used unless stated otherwise. The fraction re-weighting of the $t\bar{t}+\geq 1b$ sub-components affects mainly the number of b-jets per event. Other quantities like the number of jets and the kinematics of jets and leptons are left almost unchanged. The MC predictions are shown without applying the corrections from the fit procedure (pre-fit) described in chapter 5. The uncertainties on the prediction (described in section 5.2) are also shown.

Figure 4.6 displays the number of selected jets per event. A very good modelling of this quantity is observed up to relatively high jet multiplicities (up to 9 jets).

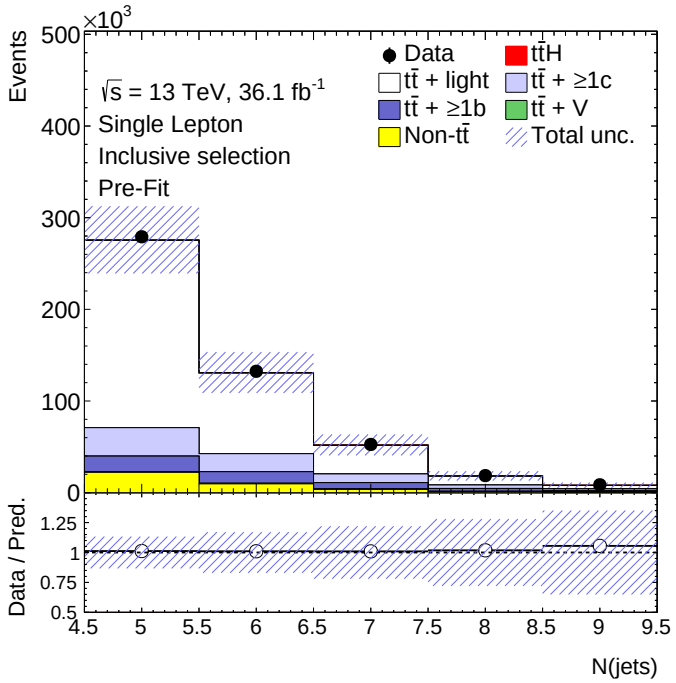


Figure 4.4.: Comparison of the predicted number of jets to the one observed in data using the inclusive single lepton channel selections. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations which are free parameters of the fit.

The number of b-jets tagged at the different working points are shown in figure 4.5 for the PP8-based

$t\bar{t}$ sample. Clear slopes are observed with data overshooting, by up to 15%, the PP8-based $t\bar{t}$ sample prediction at high b-jet multiplicities. However, the discrepancies between data and predictions are covered by the uncertainties on the prediction. Figure 4.6 shows the same distributions for the default $t\bar{t}$ model. In this model the data overshoot at large number of b-jets is reduced to at most 10% thanks to increased fractions of $t\bar{t}+ \geq 2b$ events with multiple b-jets, and reduced fraction of $t\bar{t}+b$ events (see figure 4.3). In the default model the uncertainties on the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations are not included as they are free parameters of the fit. Thus the total uncertainty appears to be reduced compared to the PP8-based model where the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations are systematic uncertainties with priors (see section 5.2.2).

Figure 4.7 displays the transverse momentum of the six leading jets and figure 4.8 shows the electron and muon p_{T}. Some mismodelling due to the $t\bar{t}$ MC modelling (across the whole p_{T} range) and to the non-$t\bar{t}$ backgrounds (mainly at low p_{T}) is observed. However, the uncertainties on the predictions cover the observed mismodelling.

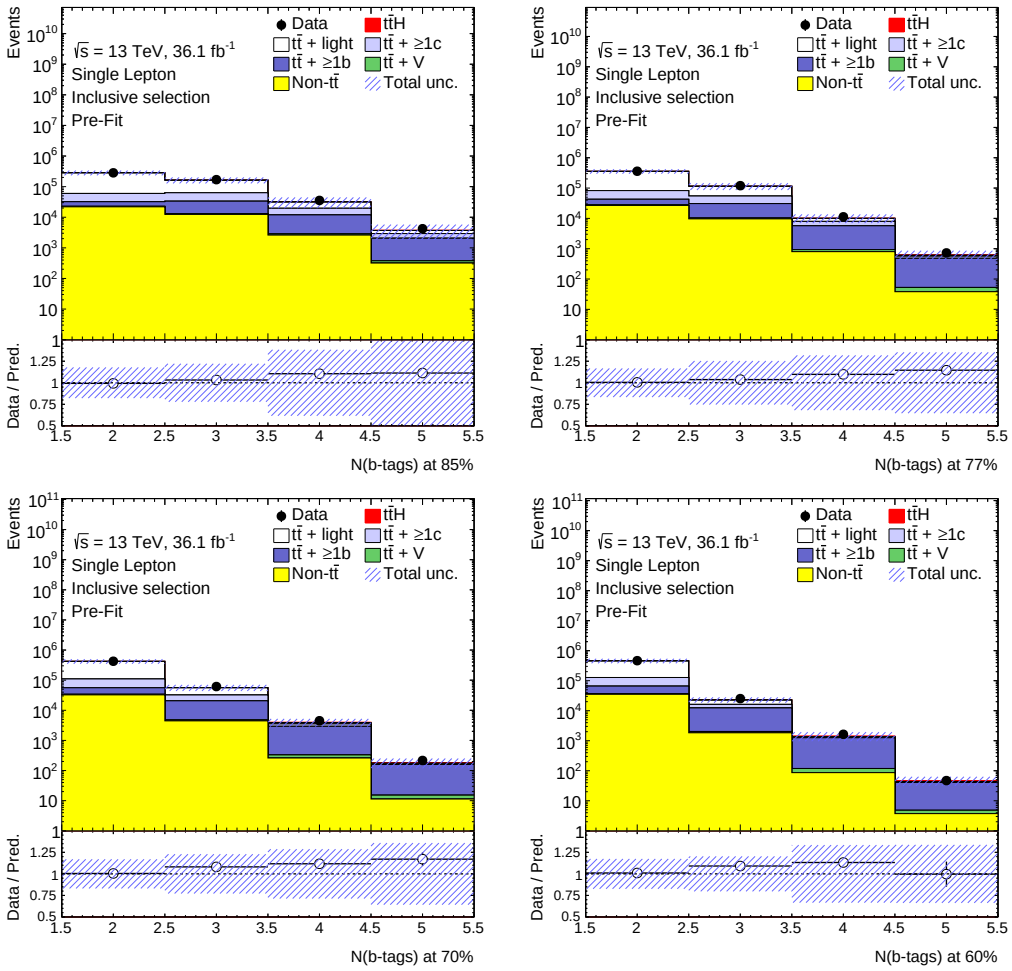


Figure 4.5.: Comparison of the predicted number of b-jets at the various working points using the PP8-based $t\bar{t}$ sample to the one observed in data using the inclusive single lepton channel selections. The hashed area represent the statistical and systematic uncertainties.

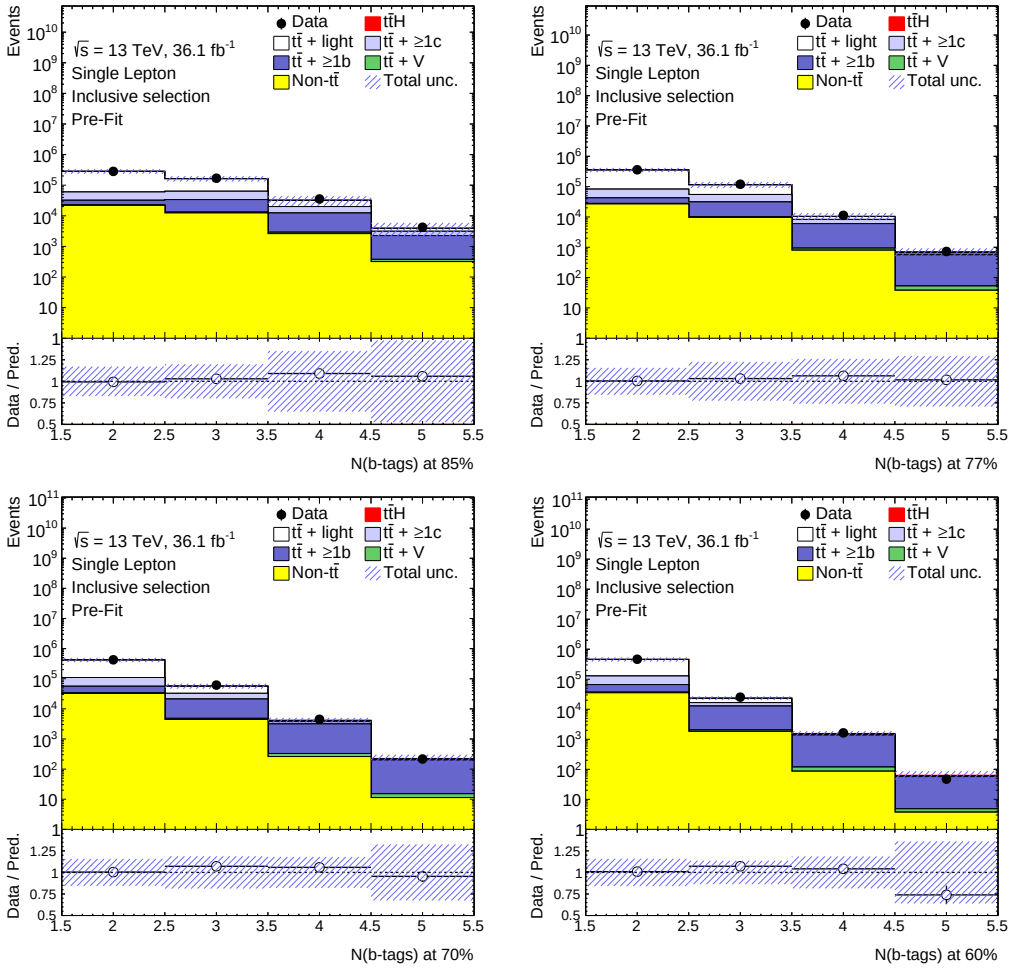


Figure 4.6.: Comparison of the predicted number of b-jets at the various working points to the one observed in data using the inclusive single lepton channel selections. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+≥1b and $t\bar{t}$+≥1c normalisations which are free parameters of the fit.

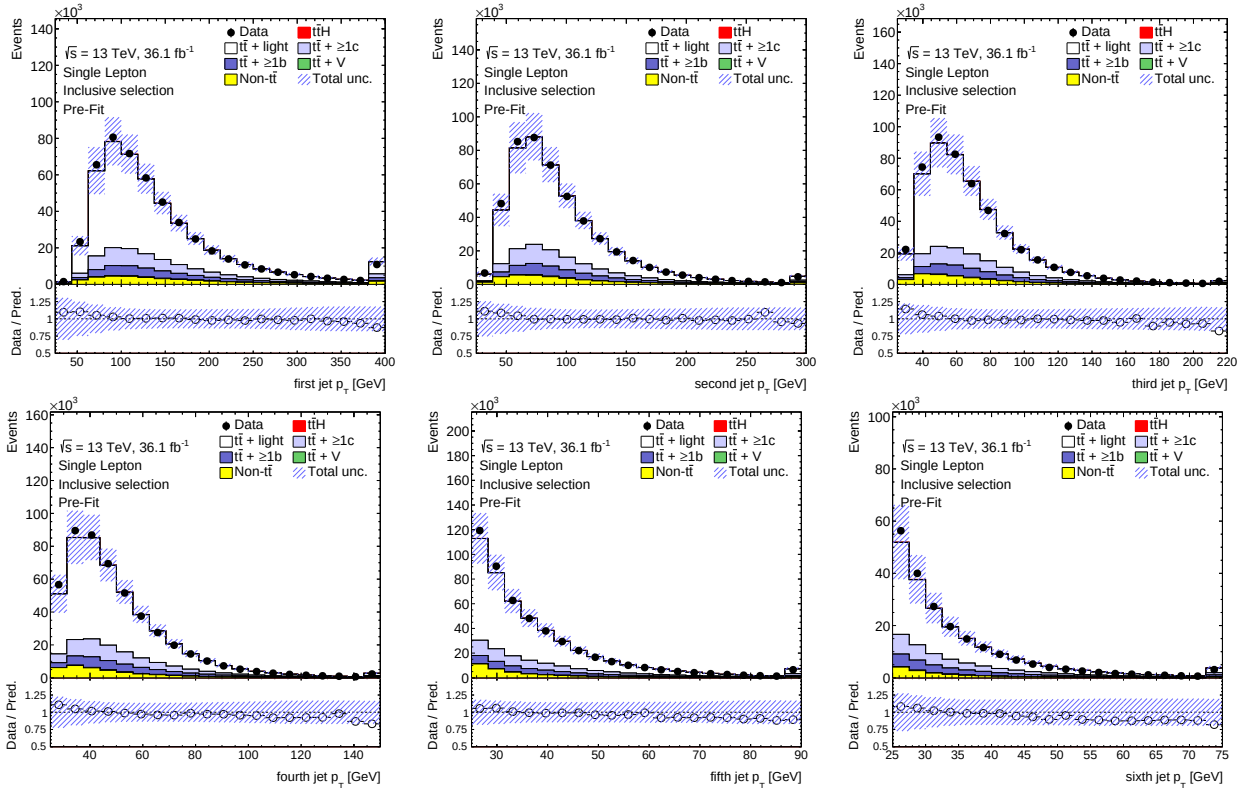


Figure 4.7.: Comparison of the predicted p_{T} distribution of the six leading jets to the one observed in data using the inclusive single lepton channel selections. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit.

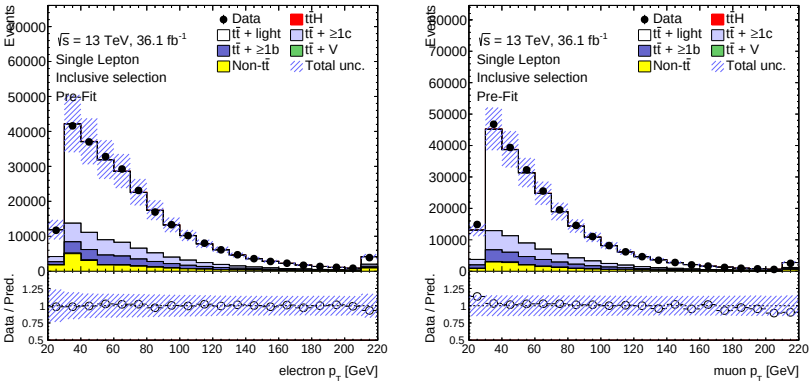


Figure 4.8.: Comparison of the predicted p_{T} distribution of the electron (left) and muon (right) to the one observed in data using the inclusive single lepton channel selections. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit.

4.6. Analysis strategy

Selected events are distributed in several categories according to the number of jets and b-tagged-jets. MVA techniques are used to further improve the $t\bar{t}H(H \to b\bar{b})$ sensitivity in the purest categories.

4.6.1. Event categorization

For the Run 2 result, an individual boosted category is considered. Events satisfying the boosted selections explained in section 4.3 are categorized as boosted events and the remaining ones are called *resolved events*.

Selected resolved events are first divided into two categories: one with exactly 5-jets and one with 6 or more jets. In the 5-jet and $\geq$ 6-jet resolved categories events are further categorized based on the number of b-tagged-jets at the four available working points: $\epsilon_b = 85, 77, 70, 60\%$ (see section 3.3.2). The four working points correspond to five MV2c10 ranges of efficiencies: $[100\%, 85\%]$, $[85\%, 77\%]$, $[77\%, 70\%]$, $[70\%, 60\%]$, $[60\%, 0\%]$. Events are grouped together if they have the same number of jets in the various ranges. Only the four jets with the highest MV2c10 weights are considered for this grouping. The obtained groups are then merged depending on the background composition with a sequential merging starting from high $t\bar{t}+\geq 1b$ purity bins towards high $t\bar{t}$ + light purity bins. The merging sequence is detailed in table 4.2.

Category	Label	Merging condition
5 jet categories:		
$\rightarrow t\bar{t}H(H \to b\bar{b})$ enriched	5j SR1	$\geq 60\%$ of $t\bar{t}+ \geq 2b$
$\rightarrow t\bar{t} + 1b$ enriched	5j BR($t\bar{t} + b$)	$\geq 20\%$ of $t\bar{t} + 1b$
$\rightarrow t\bar{t}+ \geq 2b$ enriched	5j SR2	$\geq 20\%$ of $t\bar{t}+ \geq 2b$
$\rightarrow t\bar{t}+ \geq 1c$ enriched	5j BR($t\bar{t}+\geq 1c$)	$\geq 20\%$ of $t\bar{t}+ \geq 1c$
$\rightarrow t\bar{t}+$light enriched	5j BR($t\bar{t}$ + light)	Rest
$\geq$ 6 jet categories:		
$\rightarrow t\bar{t}H(H \to b\bar{b})$ enriched	$\geq$ 6j SR1	$\geq 60\%$ of $t\bar{t}+ \geq 2b$
$\rightarrow$ Highly $t\bar{t}+ \geq 2b$ enriched	$\geq$ 6j SR2	$\geq 45\%$ of $t\bar{t}+ \geq 2b$*
$\rightarrow t\bar{t}+ \geq 2b$ enriched	$\geq$ 6j SR3	$\geq 30\%$ of $t\bar{t}+ \geq 2b$*
$\rightarrow t\bar{t} + 1b$ enriched	$\geq$ 6j BR($t\bar{t} + b$)	$\geq 60\%$ of $t\bar{t} + 1b$
$\rightarrow t\bar{t}+ \geq 1c$ enriched	$\geq$ 6j BR($t\bar{t}+\geq 1c$)	$\geq 60\%$ of $t\bar{t}+ \geq 1c$
$\rightarrow t\bar{t}+$light enriched	$\geq$ 6j BR($t\bar{t}$ + light)	Rest

Table 4.2.: Sequential merging of events groups corresponding to the different number of b-tagged-jets at different b-tagging working points (see text). The merging in 5-jet and $\geq$ 6-jet regions is done starting from the signal enriched category and going to the next line at each step. Signal enriched categories are referred to as SR and background enriched categories are referred to as BR.

The event categorization is largely improved with respect to Run 1, especially thanks to the availability of the pseudo-continuous b-tagging calibration which allows to use multiple working points. In total the $t\bar{t}H(H \to b\bar{b})$ single lepton channel includes 12 categories. Their signal purity, in terms of S/B and $S/\sqrt{B}$ ratios, is shown in figure 4.9 and the background composition of all categories is shown in figure 4.10. Categories with $S/B > 1.5\%$ are considered as signal-enriched categories and MVA

techniques are used to further discriminate the signal and the background. Other categories are considered as background-enriched and are used to control the various background components in the fit. Signal enriched categories are mainly dominated by the $t\bar{t}+\geq 1b$ background. In the most sensitive category, which is made of events with six or more jets and at least four b-jets at the $\epsilon_b = 60\%$ working point, a signal over background ratio of 5.3% is achieved. Background-enriched categories are labelled according to the background component they are designed to control. The $t\bar{t}+$ light-enriched categories are mainly made of the $t\bar{t}+$ light events with a non-negligible contribution from $t\bar{t}+\geq 1c$ events. $t\bar{t}+\geq 1c$-enriched categories have a relatively large contribution from $t\bar{t}+\geq 1c$ events. However they are still dominated by $t\bar{t}+$ light events and have a large contribution from the $t\bar{t}+\geq 1b$ background. It is hard to select categories pure in $t\bar{t}+\geq 1c$ events and this background is hard to control as discussed in chapter 5. $t\bar{t}+\geq 1b$-enriched categories increase the fraction of $t\bar{t}+\geq 1b$ events but the $t\bar{t}+$ light and $t\bar{t}+\geq 1c$ contributions are still large. Thus, these categories also help constraining the $t\bar{t}+\geq 1c$ background. The $t\bar{t}+\geq 1b$ background is mainly constrained in the signal-enriched categories where it dominates.

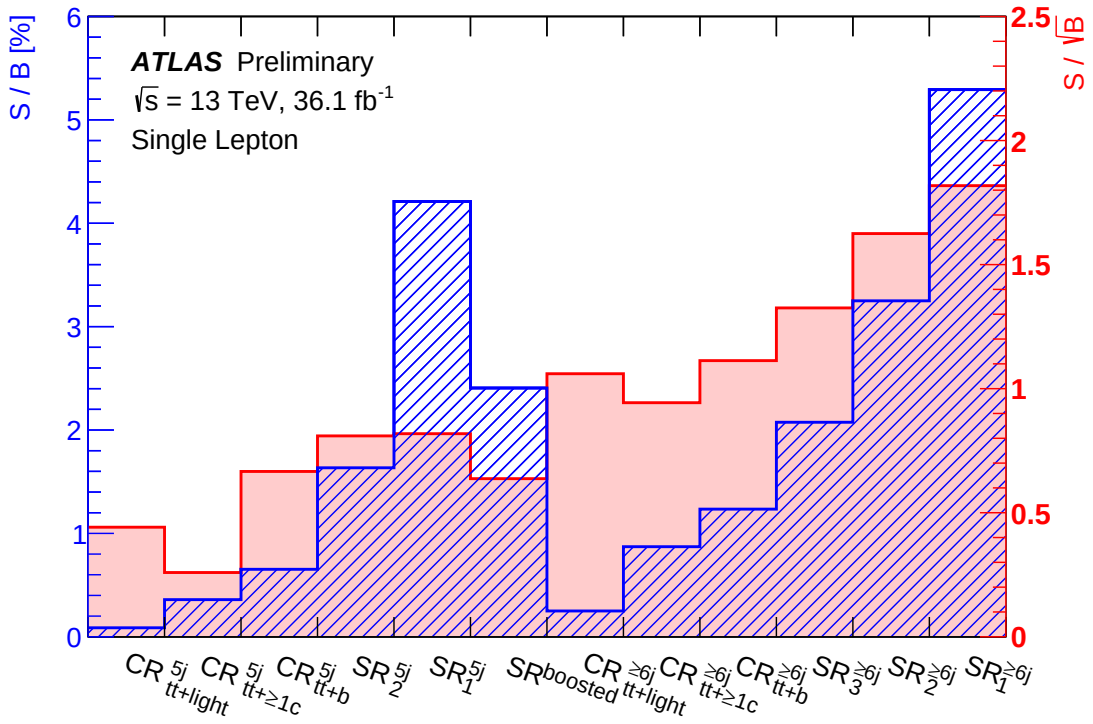


Figure 4.9.: $t\bar{t}H(H \to b\bar{b})$ event categories in the single lepton channel. Events are split in 11 categories based on pseudo-continuous b-tagging plus a boosted category. The blue hashed histogram displays the S/B ratio and the red histograms shows the $S/\sqrt{B}$ ratio in all categories.

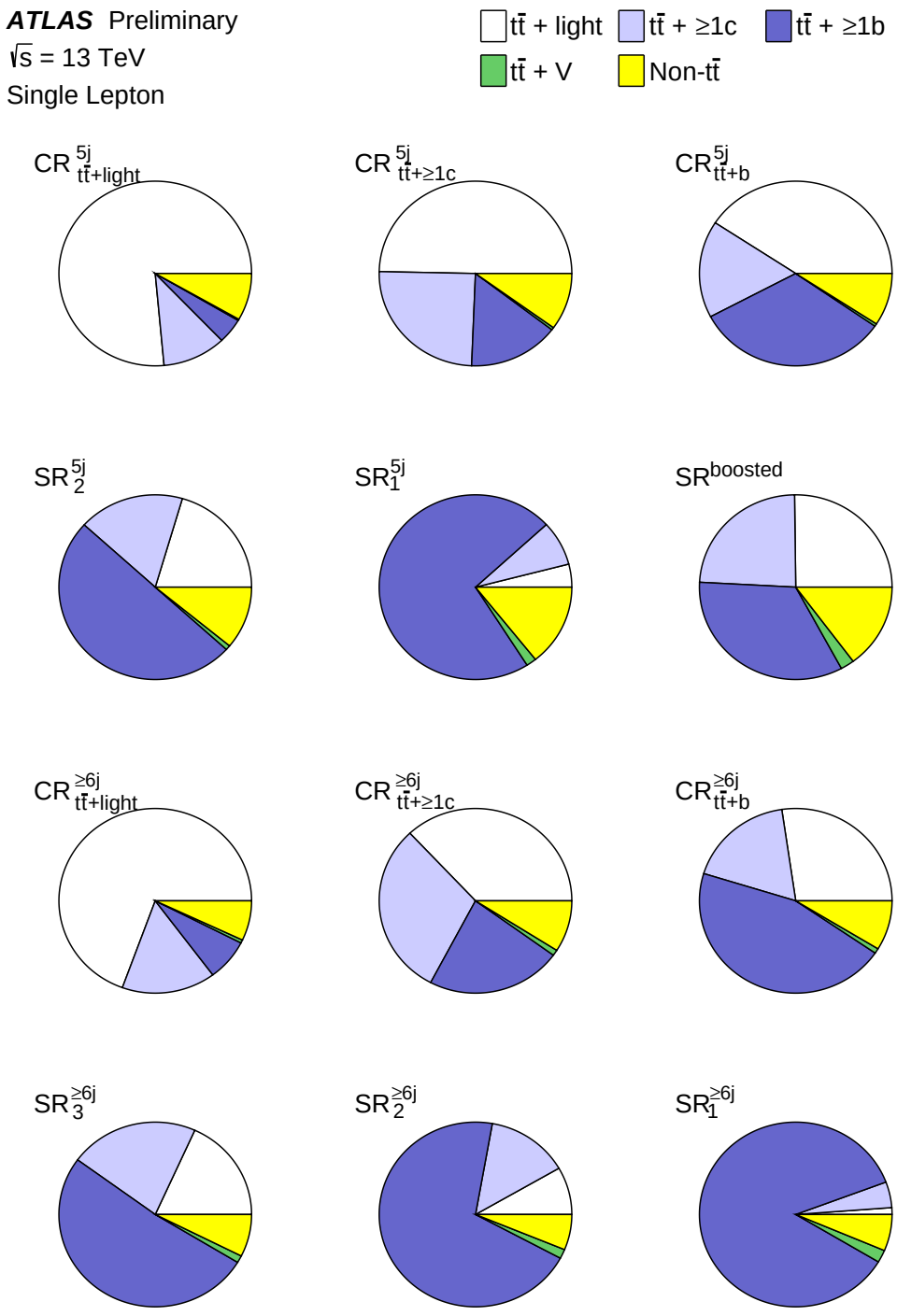


Figure 4.10.: $t\bar{t}H(H \to b\bar{b})$ event categories in the single lepton channel. Events are split in 11 categories based on pseudo-continuous b-tagging plus a boosted category. The fractional amount of all backgrounds to the total amount of background in each category is shown as pie-charts.

4.6.2. Multi-Variate techniques

High signal purity bins are obtained using multivariate techniques to separate $t\bar{t}H(H \to b\bar{b})$ from $t\bar{t}+$ jets events. The $t\bar{t}H(H \to b\bar{b})$ analysis uses two layers of MVAs.

In the first layer, three methods use the topology and the kinematic of the objects to build discriminating variables for the $t\bar{t}H(H \to b\bar{b})$ or the $t\bar{t}+\geq 1b$ hypotheses:

- *The reconstruction BDTs*: aim at the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ system by finding the best correspondence between the observed jets and quarks originating from the $t\bar{t}H(H \to b\bar{b})$ decay products. This method is further described in section 4.7. The properties of the reconstructed Higgs boson and top quark candidates are used to define discriminating variables for $t\bar{t}H(H \to b\bar{b})$ against $t\bar{t}+$ jets.

- *The Matrix Element Method (MEM)*: gives a single likelihood discriminant for each event to satisfy the $t\bar{t}H(H \to b\bar{b})$ or $t\bar{t}+\geq 1b$ hypotheses in the ≥ 6 $t\bar{t}H(H \to b\bar{b})$-enriched category only. It is based on the integration of the matrix element for each event assuming the signal or background Feynman diagrams at leading order. Details of this technique can be found in [139].

- *The Likelihood Discriminant (LD)*: also provides a single likelihood discriminant for each event to satisfy the $t\bar{t}H(H \to b\bar{b})$ or $t\bar{t}+$ jets hypotheses. For each event the probability to be signal and the probability to be background are computed using reference distributions for the kinematics of the final state objects. The signal and background probabilities are then combined within a likelihood discriminant. Details of this technique are not in the scope of this thesis but can be found in [141].

The outputs of the first layer MVAs are combined with general event kinematic variables, and variables based on the b-tagging weights of jets, in the classification BDT. The $t\bar{t}H(H \to b\bar{b})$ process is used as signal and the classification BDT is trained to separate it from the $t\bar{t}+$ jets background. For each of the 5-jet and $\geq$ 6-jet categories, the inclusive training on events with at least four b-tagged-jets at the $\epsilon_b = 85\%$ working point, which cover all our signal-enriched categories, performs as well as a dedicated training in each of the signal-enriched categories and is taken as default method. In addition a dedicated training in the ≥ 6 $t\bar{t}H(H \to b\bar{b})$-enriched category is done to include the MEM discriminant.

4.7. $t\bar{t}H(H \to b\bar{b})$ system reconstruction techniques

The search for $t\bar{t}H(H \to b\bar{b})$ events requires advanced MVA techniques to separate the signal from the $t\bar{t}+$ jets background. Valuable inputs to the classification BDT are coming from the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ system. This section describes the available reconstruction techniques, including novel techniques, and their performance for the Run 2 $t\bar{t}H(H \to b\bar{b})$ analysis. Comparisons of various techniques are performed in events with at least six jets and at least four b-tagged-jets at the 70% working point, called SR($\geq 6j, \geq 4b$). This allows for a higher statistic than the $\geq$ 6j SR1 category while keeping a high purity.

For reconstruction studies the parton from which each jet originates must be identified at the *truth level*. This identification is done by matching jets to quarks from the hard scatter process if ΔR(jet,quark). The two b-quarks from the Higgs boson are found in around 90% of the events which is the maximum efficiency that can be achieved by the Higgs boson reconstruction. In the remaining 10% of the events at least one jet from the Higgs boson decay is not selected by acceptance cuts. In only 40% of the events, 6-jets are found to match the 6-quarks from the $t\bar{t}H(H \to b\bar{b})$ decay in the

single lepton channel. This is limited by the low efficiency to find the sub-leading quark from the hadronic decay of the W-boson which is produced at relatively low p_{T}. The reconstruction techniques are built to find the best correspondence, defined by the truth matching, between the observed jets and final state particles in $t\bar{t}H(H \to b\bar{b})$ events.

4.7.1. Reconstruction BDTs

The reconstruction BDTs are the baseline reconstruction technique for the $t\bar{t}H(H \to b\bar{b})$ analysis. BDTs are trained to identify the correct matching of quarks to jets and reject any other combination using the topology and kinematic of the $t\bar{t}H(H \to b\bar{b})$ decay products. Discriminating variables between signal and background are built from the reconstructed objects:

- The **leptonic W** is built from the sum of the lepton and neutrino 4-momenta. The neutrino transverse momentum is obtained from the missing transverse energy. However, the neutrino longitudinal momentum $p_z(\nu)$ is not measured. This component is obtained constraining the invariant mass of the lepton-neutrino system to the one of the W boson, leading to a quadratic equation with $p_z(\nu)$ the only unknown. Solving this equation allows to access the $p_z(\nu)$ as described in ref [150].

- The set of **hadronic W** candidates is composed of all combinations of two jets that are not b-tagged. In the 5-jet categories the sub-leading quark from the W boson is not matched in most of the events. Thus the hadronic W-boson is not reconstructed for events with exactly 5-jets.

- The **top quark** candidates are reconstructed from the association of a reconstructed W-boson with a b-tagged jet. In the case of the 5-jet categories, a *partial top* is reconstructed from a b-tagged-jet and a non-b-tagged-jet.

- The **Higgs boson** candidates are finally reconstructed from all possible pairs of b-jets.

The use of variables based on the Higgs candidate allows a high reconstruction efficiency of the Higgs boson. However they potentially bias the mass and ΔR spectra of $b\bar{b}$ pairs associated to a Higgs boson candidate in background events. Thus two reconstruction BDTs are trained:

- The reconstruction BDT with Higgs boson variables, referred to as *reco BDT with Higgs*, allows a high reconstruction efficiency of the Higgs boson.

- the reconstruction BDT without Higgs boson information, referred to as *reco BDT without Higgs*, do not bias the mass and ΔR spectra of $b\bar{b}$ pairs associated to a Higgs boson candidate in background events.

The transverse momentum of the Higgs and top candidates are shown in figure 4.11 for events in the signal-enriched categories. The Higgs boson mass, the ΔR between the two b-quarks from the Higgs candidate and ΔR between the two top candidates are shown in figure 4.12 for events in the signal-enriched categories. All these variables are well modelled by the default $t\bar{t}$ sample within the uncertainties on the predictions.

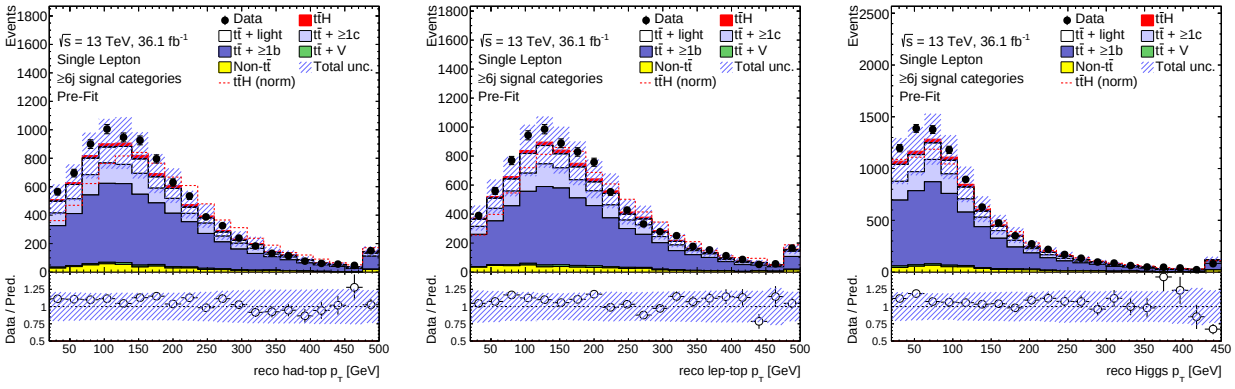


Figure 4.11.: Comparison of the predicted p_{T} distribution of the hadronic-top candidate (left), leptonic-top candidate (middle), and Higgs boson candidate (right) to the one observed in data for events in the signal-enriched categories. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+≥1b and $t\bar{t}$+≥1c normalisations which are free parameters of the fit.

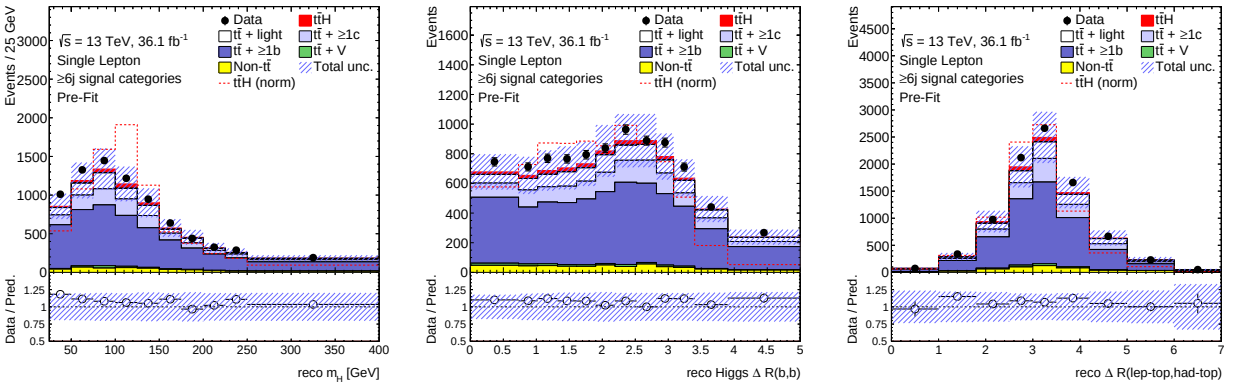


Figure 4.12.: Comparison of the predicted distributions of the Higgs candidate mass (left), of the ΔR between the two b-jets from the Higgs candidate (middle), and the ΔR between the hadronic-top and leptonic-top candidates (right) to the one observed in data for events in the signal-enriched categories. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+≥1b and $t\bar{t}$+≥1c normalisations which are free parameters of the fit.

4.7.2. The Network based reconstruction

The Network based reconstruction is a novel technique based on the interpretation of all jets and the lepton as vertices of graph $\mathcal{G}$, called the $t\bar{t}H(H \to b\bar{b})$ *network*, where a link is implemented between all pairs of vertices. A weight is then attached to each link representing the probability of the pair to originate from the same particle decay. This probability is built from a BDT output, *the pairing BDT*, which is trained to identify pairs of objects originating from the same particle.

Two methods are applied on the obtained network. The *clustering* method merges objects until a set of terminating conditions are satisfied. The *network solving* seeks for the *true pattern* of the $t\bar{t}H(H \to b\bar{b})$ system on the graph, i.e. two clusters composed of two objects (corresponding to the leptonic-top and Higgs boson) and a third with three objects (corresponding to the hadronic-top). All permutations of jets satisfying this pattern are formed and the combination maximizing the probability to be the correct combination is kept. Figure 4.13 provides a schematic view of the Network based reconstruction.

The network based reconstruction is meant to be a complementary technique to the reconstruction BDTs. In fact, the former is based on object pair properties and the event properties are extracted from the network while the reconstruction BDTs directly access the properties of the reconstructed candidates. Thus the network based reconstruction potentially provides additional information.

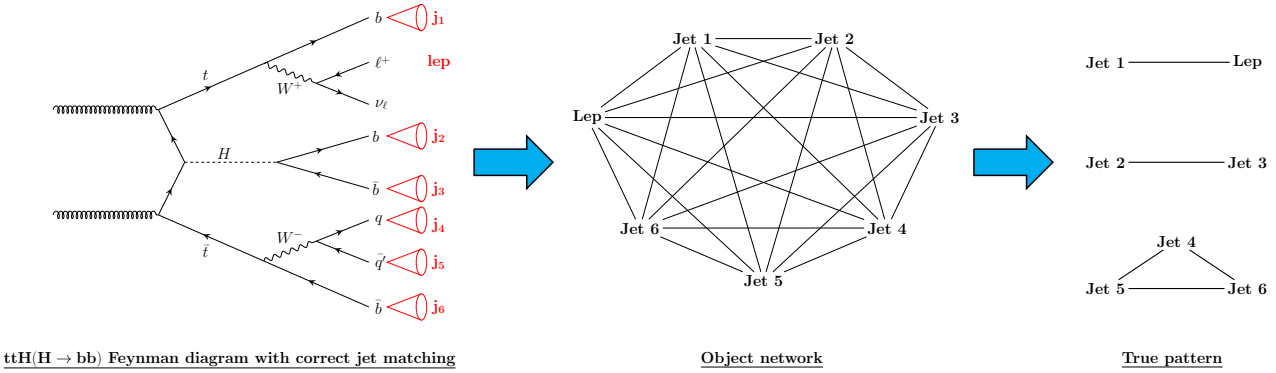


Figure 4.13.: Schematic view of the network based reconstruction. Starting from the Feynman diagram with jet associated to partons (left) a graph is formed with all objects linked together (middle). The clustering and network algorithms then seek for correct jet pattern (right).

4.7.2.1. The link probability and the pairing BDT

The paring BDT is build to discriminate pairs of objects originating from the same particle (signal) from all other pairs formed of two objects (background).

In the single lepton channel, the lepton origin does not need to be identified as it necessarily originates from the leptonic top quark. The lepton is removed from the pairing BDT training allowing the BDT to focus on the jets. The efficiency to match the correct jet to the lepton is recovered using specific conditions for the lepton association in the clustering and network solving algorithms.

A significant fraction of jets are not matched to any final state particle. There is no proper way to treat these jets. In the default pairing BDT setup pairs formed of at least one un-match jet are added to the background sample in order to maximize the performance of the network reconstruction methods to identify the Higgs boson.

Five input variables are used in the default pairing BDT setup:

- Pair ΔR: ΔR between the two objects forming the pair.
- Pair $\Delta\phi$: absolute value of the $\Delta\phi$ between the two objects forming the pair.
- $M(\text{pair})$: invariant mass of the two objects forming the pair.
- p_{T} (pair): p_{T} of the pair system.
- $\frac{M(\text{pair})}{\sum_{x\in\text{pair}} p_{\text{T}}(x)}$: invariant mass of the pair divided by the scalar sum of the p_{T} of objects in the pair.

Two additional variables sets are tested. The first set is based on additional variables from the topology and kinematics of the pair system (*pair vars*). In the second set variables including information of the full event shape are also considered (*all vars*). The background rejection against signal efficiency curves of the pairing BDTs for the different sets of variables are shown in figure 4.14. The additional variables of the second set provides higher performance. However, these curves give an estimation of the overall performance of all signal and background pairs. In order to estimate the performance on the reconstruction efficiency of the clustering and network solving algorithms are run on the three sets. Even though the differences after reconstruction are small, the baseline set of variables provides the highest Higgs boson reconstruction efficiency by a few percent.

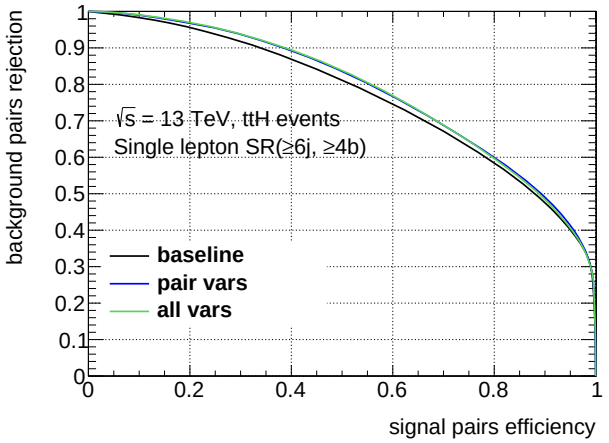


Figure 4.14.: Object pairing BDT performance for several sets of variables. The efficiency to identify pairs of objects originating from different particles is shown as a function of the efficiency to identify pairs of objects originating from the same particle.

In order to solve the network and reconstruct the event one needs to create a link probability between the jets and the lepton which are considered as vertices v_i of the network.

The links weight $w(v_i \to v_j)$ from the vertex v_i to the vertex v_j represents the "probability" that v_i originates from the same particle as v_j. The pairing BDT output gives a natural candidate for this weight.

In order to have a finer description of the network, other variables are considered. In particular, a "participation ratio like variable" defined as the ratio of the link weight over the vertex strength [151] is used. It is referred to as *participation ratio* in what follows for simplicity. The participation ratio gives a measure of the importance of the oriented link $l(v_i \to v_j)$ for the vertex v_i compared to all links starting from v_i. In our context the participation ratio is defined as the pairing BDT output of the

vertices v_i and v_j divided by the sum of the BDT outputs of all pairs starting from v_i:

$$pr(v_i \to v_j) = \frac{\text{BDT output}(v_i,\, v_j)}{\sum_{v_k \in \mathcal{G}} \text{BDT output}(v_i,\, v_k)}. \tag{4.1}$$

The participation ratio provides a probability for the vertex v_i to be attached to the vertex v_j.

4.7.2.2. The network based clustering

The clustering algorithms target the merging of objects into clusters. Beside the definition of the algorithm itself, many parameters can be tuned to constrain the clustering towards a certain pattern and reject certain topologies: content of the cluster, number of clusters, number of occurrence, etc. However constraints on the clustering can result in a bias towards certain topology and reduce the performance.

A jet like clustering algorithm, called *jet-like* in the future, is used as a reference clustering algorithm due to its simplicity. Starting with the full set of pairs, the pair of vertices with the highest participation ratio is merged in a cluster. Then the procedure is repeated with the new cluster included in the list of objects to be merged until one of the terminating conditions is fulfilled. In order to increase the efficiency to reconstruct the correct pattern, three conditions based on the $t\bar{t}H(H \to b\bar{b})$ final state are imposed on the cluster candidates to be considered:

- The mass of the cluster can not be less than 50 GeV nor more than 300 GeV. This condition is rarely un-satisfied but avoids combinations which can not represent a top quark, a W boson or a Higgs boson.
- The cluster containing the lepton can only have one b-jet. The permutation of b-jets is the largest source of wrong jet to particle assignment in the reconstruction BDTs. Requiring that the lepton can only be merged with one b-jet improves the separation of the four b-jets in the final clusters.
- For the same reason as above, clusters with more than two b-jets are not allowed.

The algorithm is stopped either when all initial objects are merged into three final objects (clusters or single objects), or when all new cluster candidates are forbidden by the above three rules.

The *figure clustering* algorithm is based on the reduction of the network in a subset of maximum clustering power. This is obtained by requiring at each vertex to keep only the oriented link of maximum weight. On this new graph $g \subset \mathcal{G}$ a *figure* $f = (v_i, ..., v_j)$ is found if following the links starting at a vertex v_i one ends up at the same vertex v_i without passing through all the vertices. All figures are considered as cluster candidates. Cluster candidates are kept if they satisfy the clustering conditions of the jet like algorithm. A new graph is build taking all remaining objects and the obtained clusters and the procedure is repeated until the same terminating requirements as the jet-like algorithm are fulfilled. The procedure is sketched in figure 4.15.

An important property of the figure based clustering is its ability to identify patterns with more than two vertices. However it can be shown that a symmetric weight $w(v_i \to v_j) = w(v_j \to v_i)$, such as the pairing BDT output, does not allow any figure with more than two vertices unless two different BDT outputs are numerically the same. The participation ratio breaks the weight symmetry. However the ordering of weights around a vertex obtained with the pairing BDT output is not changed when switching to the participation ratio, i.e $\forall (v_j, v_k)$ if BDT output$(v_i, v_j) >$ BDT output(v_i, v_k) then $pr(v_i \to v_j) > pr(v_i \to v_k)$. Thus it only allows figures formed of pairs of vertices.

In order to fully exploit this algorithm, more complex weights need to be considered or the pairing

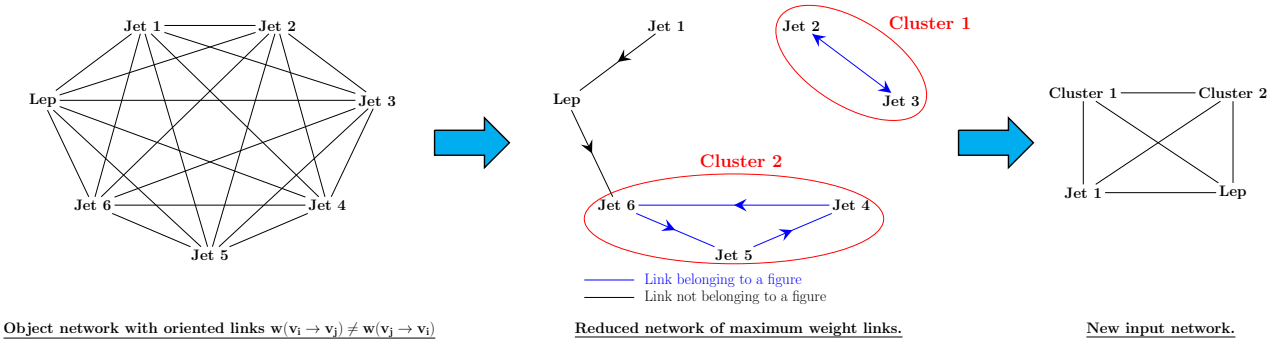


Figure 4.15.: Schematic view of the figure clustering.

BDT strategy should be revisited. This is one possible future improvement for this method.

Figure 4.16 shows the basic properties of the clustering algorithms: the number of formed clusters, the number of objects in each cluster and the number of objects which are not clustered. The jet-like algorithm finds slightly more clusters than the figure clustering. However these clusters have less objects compared to the figure clustering.

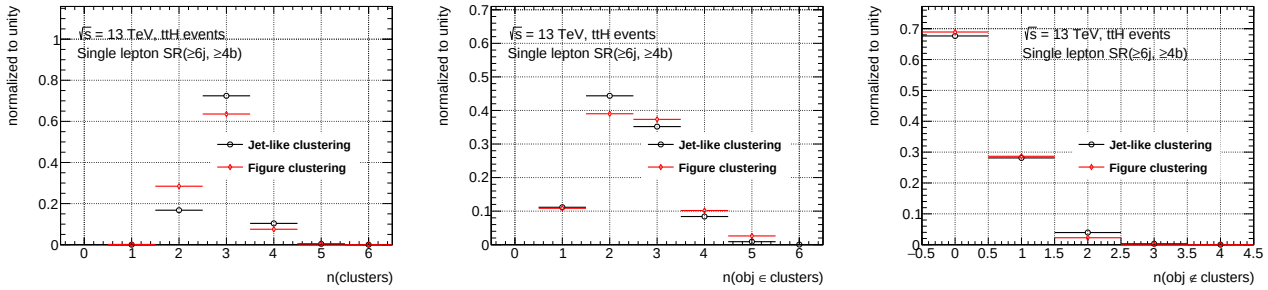


Figure 4.16.: Basic properties of the jet-like (black) and figure (red) clustering algorithm: number of reconstructed clusters per event (left), number of objects contained in clusters per event (middle), number of objects that are not clustered per event (right).

After forming the clusters, the leptonic top candidate is defined as the cluster containing the lepton. In 12(7)% of the events the lepton is not clustered with any jet with the jet-like (figure) clustering. The Higgs boson cluster is identified as the highest mass cluster not containing the lepton. The hadronic-top candidate is taken as the next cluster in mass order not containing the lepton. Indeed, the hadronic top is usually partially reconstructed leading to a lower mass than the one of the Higgs boson cluster. Further improvements are expected from a more precise classification, for example using the number of b-tagged-jets in clusters.

4.7.2.3. The network solving

The network solving aims at finding the best combination of jets satisfying the true pattern. Three clusters are thus built:

- The **leptonic-top cluster** is formed of the lepton and any b-tagged-jet.
- The **Higgs boson cluster** is formed of two b-tagged-jets exclusively.
- The **hadronic-top cluster** is formed of three jets out of which at most two can be b-tagged.

In addition, each cluster is required to satisfy the merging condition of clustering algorithms. In order to avoid loosing events, if no combination is found the requirements are loosened until a valid combination is found. Such events are however very rare.

Several probabilities per combination are defined to choose the correct combination. The best performance are obtained using the participation sum of the clusters:

$$pr_{\text{cl}} \text{ sum} = \frac{\sum_{\text{cl}\in\text{clusters}} \sum_{(v_i,v_j)\in\text{cl}} pr(v_i \to v_j)}{\sum_{\text{cl}\in\text{clusters}} \sum_{v_i\in\text{cl}} \sum_{v_k\notin\text{cl}} pr(v_i \to v_k)}. \tag{4.2}$$

The combination of jets maximizing the pr_{cl} sum is kept as the best matching.

4.7.3. The $t\bar{t}H(H \to b\bar{b})$ reconstruction performance

The performance of the reconstruction is estimated looking at the fraction of events with the correct matching of all objects. Figures 4.17, 4.18 and 4.19 show the comparison of the matching fractions of objects obtained with the reco BDT without Higgs to the matching fractions of objects obtained with each of the network based reconstruction algorithms.

The network solving algorithm shows the highest correct reconstruction efficiencies among network based algorithm and similar to the reco BDT techniques. In particular the efficiency to reconstruct the correct Higgs boson is 41% which is in between the efficiency obtained with the reco BDT without Higgs, 31%, and the one from the reco BDT with Higgs, 48%.

In the case of the clustering algorithms, individual objects are not identified. The efficiency to correctly identify a particle of the $t\bar{t}H(H \to b\bar{b})$ Feynman diagram is defined as the fraction of events where the particle belongs to the correct cluster. For example, the efficiency to identify the leading b-jet from the Higgs boson is defined as the fraction of events where this b-jet is found in the Higgs boson cluster. The clustering algorithms correctly merge the two b-jets from the Higgs boson in the same cluster in $\sim 40\%$ of the events. However the Higgs boson reconstruction efficiency drops by 10% when identifying the clusters by their mass order. Moreover the clustering algorithms allow the Higgs cluster to be made of more than two objects. This is especially true for the jet-like clustering which is less exigent than the figure based algorithm and thus gives higher efficiencies but spoil the kinematic properties of the cluster by adding additional jets. The matching fractions of the other objects is lower for the clustering methods than the network solving and reconstruction BDTs.

The overlap between the reco BDT without Higgs and the network based reconstruction techniques is limited. In particular, 42% of the correctly reconstructed Higgs with the reco BDT without Higgs and 56% of the correct Higgs boson clusters from the network solving algorithm are coming from non-overlapping events. Thus the two methods are complementary and their combination can improve the $t\bar{t}H(H \to b\bar{b})$ separation from $t\bar{t}+$ jets.

The Higgs candidate mass and the ΔR of its associated $b\bar{b}$ pair are shown in figure 4.20 for the reco BDT without Higgs and the network solving algorithms in $t\bar{t}H(H \to b\bar{b})$ and $t\bar{t}+$ jets events. If the network solving provides high Higgs boson matching fraction it also significantly biases the mass distribution and shifts $\Delta R_{b\bar{b}}$ distribution. Indeed, both the mass and ΔR of objects in a pair are included in the pairing BDT.

The separation power of the Higgs boson mass is reduced when using network based algorithms rather than the reco BDT without Higgs. However the two algorithms provide complementary information and the combination of the two distributions can still be beneficial to the classification BDT. The $\Delta R_{b\bar{b}}$ distributions in both $t\bar{t}+$jets and $t\bar{t}H(H \to b\bar{b})$ events are pushed to the left. The separation power of the $\Delta R_{b\bar{b}}$ variable is similar for the two techniques.

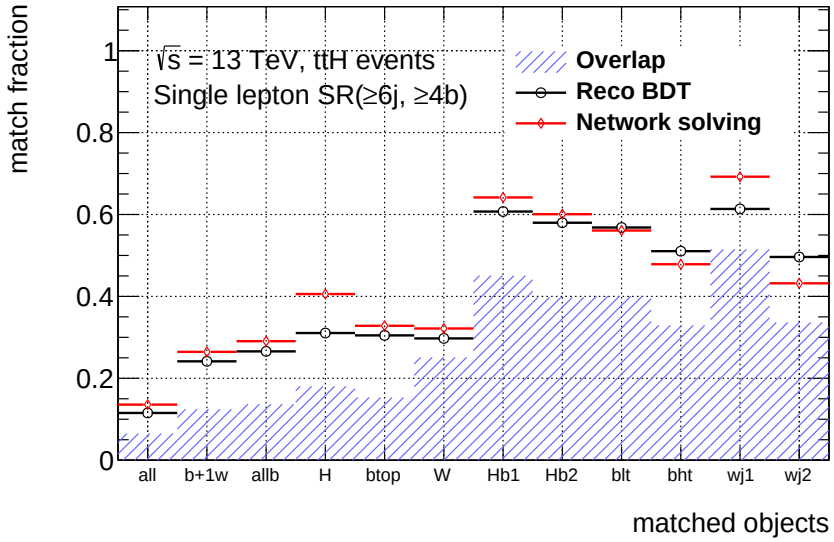


Figure 4.17.: Fraction of events where all jets (all), all b-jets and one of the *light*-jets from the W-boson decays (b+1w), all b-jets (allb), the two b-jets from the Higgs boson (H), the b-jets from the top-quarks (btop), the *light*-jets from the W-boson (W), the leading and sub-leading b-jets from the Higgs boson (Hb1 and Hb2 respectively), the b-jet from the leptonic and hadronic tops (blt and bht respectively), the leading and sub-leading *light*-jets from the W-boson (wj1 and wj2 respectively) are correctly assigned by the reconstruction BDT without Higgs boson variables (black) and the network solving algorithm (red). The hashed band represent the overlap between the two methods, i.e. the fraction of events where the two methods find the same correct candidate.

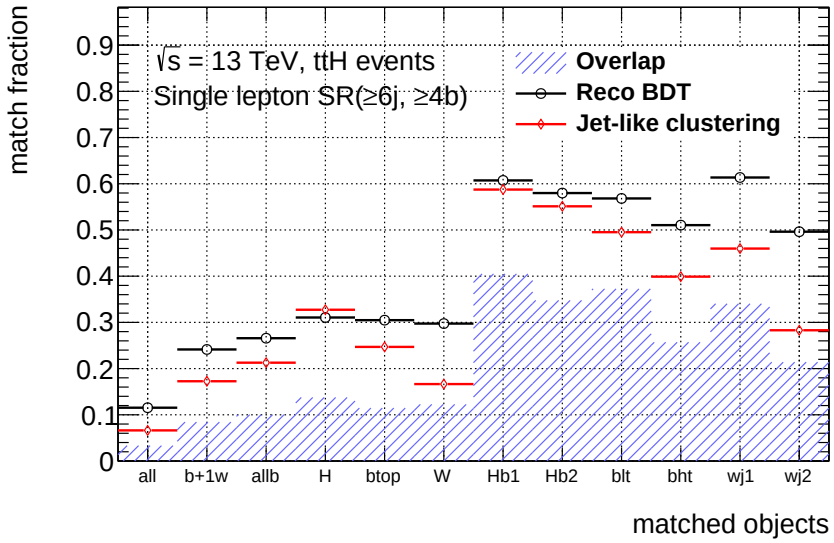


Figure 4.18.: Fraction of events with correctly assigned objects (see figure 4.17) by the reconstruction BDT without Higgs boson variables (black) and the jet-like clustering (red) algorithms. The hashed band represent the overlap between the two methods.

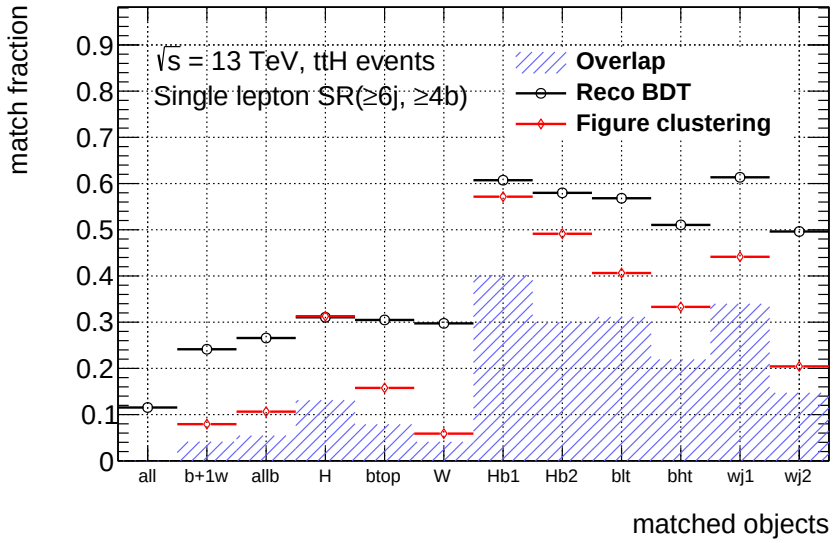


Figure 4.19.: Fraction of events with correctly assigned objects (see figure 4.17) by the reconstruction BDT without Higgs boson variables (black) and the figure based clustering (red) algorithms. The hashed band represent the overlap between the two methods.

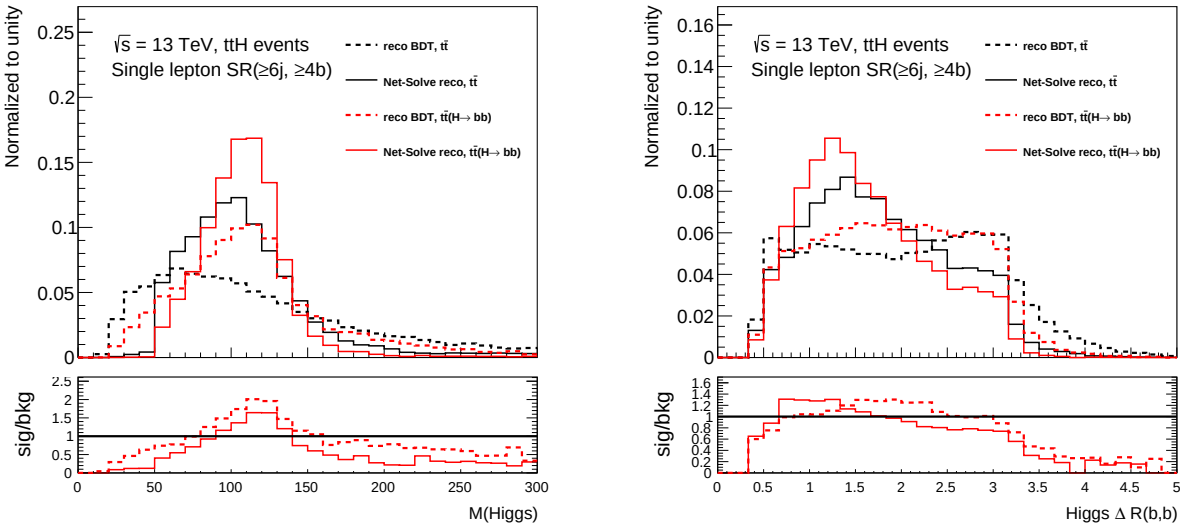


Figure 4.20.: Higgs candidate mass (left) and Higgs candidate $\Delta R(b,b)$ (right) variables from the reconstruction BDT (dashed) or the network solving (full).

4.7.4. Network based reconstruction for the discrimination between $t\bar{t}H$ and $t\bar{t}$

The reconstruction of the $t\bar{t}H(H \to b\bar{b})$ is a major component of the separation of the $t\bar{t}H(H \to b\bar{b})$ signal with the $t\bar{t}+$ jets background. The network based reconstructions provide additional events with the correct Higgs boson matching and thus potential new information for the classification BDT.

In addition to the properties of the reconstructed object, new variables can be extracted from the structure of the network. In particular two variables separate the signal from the background:

- *The Higgs confinement* is the sum of the participation ratios of all oriented links in the Higgs cluster $\sum_{v_i \in \text{Higgs}} \sum_{v_j \in \text{Higgs}} pr(v_i \to v_j)$. This variable gives a measure of the strength of the connection between the two b-jets in the Higgs cluster. In the $t\bar{t}+\geq 1b$ background the Higgs cluster is supposed to be a $g \to b\bar{b}$ cluster. A $g \to b\bar{b}$ pair have higher $\Delta R_{b\bar{b}}$ and lower $m_{b\bar{b}}$ compared to a $H \to b\bar{b}$ pair, and thus lower participation ratios. The Higgs confinement is shown in figure 4.21 (left).

- *The Higgs attraction* is the sum of the participation ratios of all links going out of the Higgs cluster $\sum_{v_i \in \text{Higgs}} \sum_{v_k \notin \text{Higgs}} pr(v_i \to v_k)$. It gives a measure of the attraction power of the top decays on the Higgs cluster vertices. The Higgs attraction is shown in figure 4.21 (left).

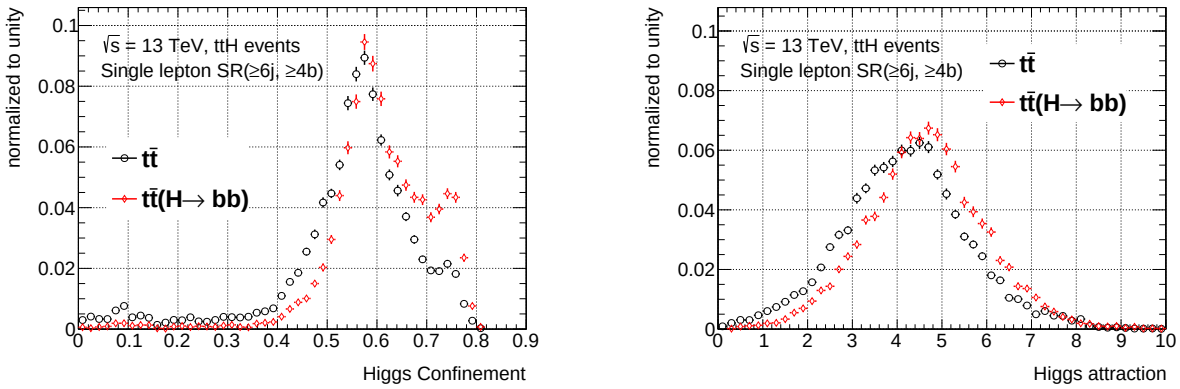


Figure 4.21.: Higgs cluster confinement (left) and Higgs cluster attraction (right) in $t\bar{t}H$ and $t\bar{t}$ events.

The classification BDT is retrained including the reconstruction BDTs based variables and the variables from one of the network based reconstruction. The performance are shown in figure 4.22. Unfortunately, the new techniques do not improve the final separation between the $t\bar{t}H(H \to b\bar{b})$ and $t\bar{t}+$jets processes.

The classification BDT includes also several variables from the event shape. All variables coming from the network based reconstruction have $\sim 20\%$ to 50% correlations with several other variables from both the reconstruction BDTs and the event variables. Thus the first addition of the network based reconstruction to the classification BDT does not improve the $t\bar{t}H(H \to b\bar{b})$ separation from the $t\bar{t}+\geq 1b$ background even though it adds information with respect to the reconstruction. However these techniques are not yet used to their full potential and many parameters can be improved for future analyses.

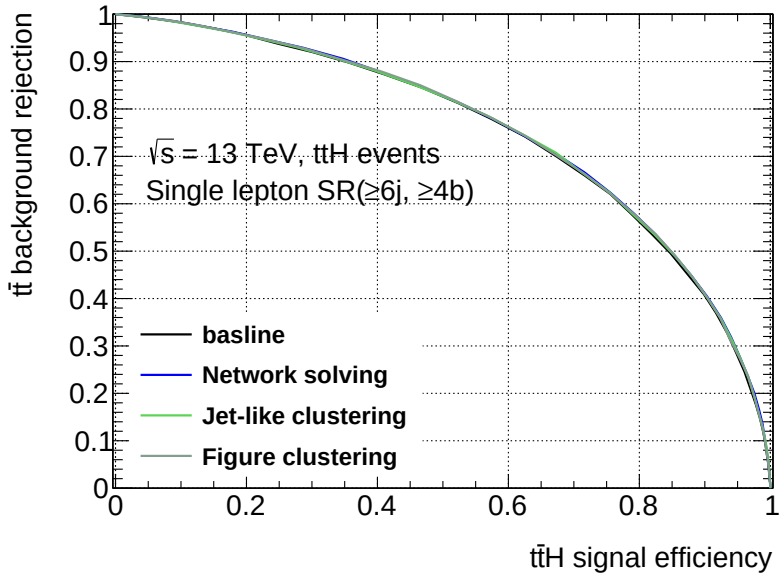


Figure 4.22.: $t\bar{t}$ rejection as a function of the $t\bar{t}H$ efficiency for the baseline classification BDT, or the classification BDTs with additional information from the network based reconstruction.

4.8. Summary

The $t\bar{t}H$ analyses are challenging but rewarding searches as they provide a unique access to the Yukawa coupling of the Higgs boson with the top-quark. The search for $t\bar{t}H(H \to b\bar{b})$ events in Run 1 data was limited by the the $t\bar{t}+$jets model. The $t\bar{t}+$jets background is poorly constrained by data in the phase space of $t\bar{t}H(H \to b\bar{b})$. Thus, a high separation of the signal from the $t\bar{t}+$jets background is required.

The $t\bar{t}H(H \to b\bar{b})$ analysis adopts a divide and conquer strategy. $t\bar{t}+$jets like events are selected demanding one(two) isolated leptons for the single(di-) lepton channel and several jets. In the single lepton channel events are categorized based on the number of selected jets and the number of b-tagged-jets at the four different b-tagging working points. This classification allows to build categories with signal over background ratios up to 5.3%.

In signal enriched categories, a BDT is used to further discriminate $t\bar{t}H(H \to b\bar{b})$ events from $t\bar{t}+$jets events. In the Run 2 analyses a reconstruction BDT is implemented before the classification BDT. The former aims at finding the best matching of jets to final state particles of the $t\bar{t}H(H \to b\bar{b})$ process. The performance of the classification BDT is improved by the addition of variables based on the kinematic of the reconstructed objects.

In this thesis I propose a novel technique for the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ signal. This new reconstruction technique is based on the network interpretation of the $t\bar{t}H(H \to b\bar{b})$ final state with all objects being vertices of a graph all connected by weighted links. The link weights represent the probability to originate from the same particle and are obtained from a pairing BDT which output is transformed in a participation ratio probability. Several techniques to solve the network are investigated.

The network reconstruction algorithms give similar performance to the reconstruction BDT. However, these two techniques provide complementary information since about half of the correctly reconstructed Higgs candidates found by one method are not found by the other method. A direct

application to the classification BDT does not improve the separation between $t\bar{t}H(H \to b\bar{b})$ events and $t\bar{t}$ + jets events. However this new network based reconstruction is a promising technique for future optimization. Indeed, they are based on single object reconstruction and the information on the event shape is recovered during the clustering or solving algorithms. The network can thus be used to extract new information at different levels which are not accessible in standard reconstruction techniques.

5. The $t\bar{t}H(H \to b\bar{b})$ statistical analysis of data

The result of the $t\bar{t}H(H \to b\bar{b})$ analysis is obtained from a simultaneous fit of the predicted distribution of all processes in all categories to observed data which I am responsible of for the paper to come. Section 5.1 describes some of the preprocessing procedures that I helped developing to improve the stability of the *profile likelihood fit*. The $t\bar{t}H(H \to b\bar{b})$ analysis relies on a complex fit model, especially for the description of the $t\bar{t}+\geq 1b$ systematic uncertainties. I was a key person in developing the main input model described in section 5.2; in addition I proposed and developed a second model for the $t\bar{t}+\geq 1b$ background that has been used to validate the first model.

In order to avoid biasing the signal towards a certain result, the $t\bar{t}H(H \to b\bar{b})$ analysis is first done without looking at the signal in data. This includes not looking at the distributions in high $t\bar{t}H(H \to b\bar{b})$ purity bins, but also not doing fits in bins sensitive to signal. This is referred to as the blinding procedure. However, the analysis sensitivity is systematically limited. Thus the performance of the analysis and the completeness of the systematic model needs to be evaluated after the fit but before looking at the final result in data. As part of my responsibilities as the coordinator of the fitting group in the single lepton channel, I developed specific techniques to validate the fit model and to improve the fit procedure stability. The performance of the $t\bar{t}H(H \to b\bar{b})$ analysis in terms of expected sensitivity and background constraints are evaluated from a fit to the *Asimov data-set* [152] and are described in section 5.3. This section also includes my optimization studies for the binning choice in the final fit and my investigations of the various constraints induced by the fit on different systematic uncertainties. My studies dedicated to the single leading systematic uncertainty in terms of the impact on signal sensitivity are presented in section 5.4. The quality and completeness of the systematic model is evaluated using a fit to pseudo-data built from an alternative $t\bar{t}$ model. The results I have obtained from this fit to pseudo-data are described and discussed in section 5.5. Section 5.6 examine the fit to data and includes dedicated studies I have performed to validate its results. Finally, the results that I have obtained from the fit to data and the combination with the di-lepton channel are presented in section 5.7.

5.1. Statistical analysis

The statistical matching of expected distributions to data is done in a template profile likelihood fit. For a given discriminant variable, template distributions for the signal and each of the backgrounds are confronted to data. Nuisance parameters, and normalisation factors are assigned to each template. They provide the degrees of freedom that the fit can use to correct the predicted templates and match the data.

5.1.1. The profile likelihood fit

The fit procedure compares the number of data events $N_{c,i}^{\text{data}}$ in each bin i of each category c to the expected bin content:

$$N_{c,i}^{\text{exp}}(\mu, k_1, ..., k_m, \theta_1, ..., \theta_n) = \mu \cdot N_{c,i,\text{sig}}^{\text{exp}}(\theta_1, ..., \theta_{n_{\text{sig}}}) + \sum_{b \in \text{bkg}} k_b \cdot N_{c,i,b}^{\text{exp}}(\theta_1, ..., \theta_{n_b}) \, . \tag{5.1}$$

where n is the total number of nuisance parameters, $(\theta_1, ..., \theta_{n_i})$ is the set of n_i nuisance parameters affecting the sample i, k_b is the normalisation factor on background b (referred to as k-factor), m the number of backgrounds and $\mu = \sigma_{t\bar{t}H}/\sigma_{t\bar{t}H}^{\text{SM}}$ the signal strength. In what follows, $\boldsymbol{k}$ is used for the set of all normalisation factors and $\boldsymbol{\theta}$ for the set of all nuisance parameters. In this approach each nuisance parameter θ_i modify the shape and normalisation of the templates according to the systematic uncertainty it parametrizes. The normalisation factors and the signal strength modify only the normalisation of the template distributions.

In each bin, data is expected to follow a Poisson probability. The primary likelihood function is obtained as a product of the Poisson probability for each bin:

$$L_{\text{main}}(\mu, \boldsymbol{k}, \boldsymbol{\theta}) = \prod_{c \in \text{cats}} \prod_{i \in \text{bins}} \frac{\left(N_{c,i}^{\text{exp}}(\mu, \boldsymbol{k}, \boldsymbol{\theta})\right)^{N_{c,i}^{\text{data}}}}{N_{c,i}^{\text{data}}\,!} \cdot e^{N_{c,i}^{\text{exp}}(\mu, \boldsymbol{k}, \boldsymbol{\theta})} \, . \tag{5.2}$$

Systematic uncertainties are defined by a central value $\theta = 0$ corresponding to the best knowledge of a specific parameter, and a $\pm 1\sigma$ variation which corresponds to the 1σ uncertainty. The continuous nuisance parameters are usually referred to as $\alpha(\theta)$ but no distinction will be made between α and θ in what follows. They are defined by the extrapolation ($|\theta| > 1$) and interpolation ($|\theta| < 1$) functions with the constraints that $\theta = 0$ corresponds to no corrections and $\theta = \pm 1$ shifts the distribution by $\pm 1\sigma$ systematic uncertainty. The prescription [153] is to use a linear and exponential extrapolations for the shape and normalisation components of systematic uncertainties, respectively. The exponential extrapolation of normalisation nuisance parameters is especially important to avoid generating negative yields. The interpolation is done using two polynomial functions, one for the shape and one for the normalisation components of systematic uncertainties. The deviation of a nuisance parameter (and also of normalisation factors and μ) is called a *pull*. In the likelihood, nuisance parameters are implemented with Gaussian constraints reflecting *prior* knowledge of the systematic uncertainty:

$$L(\mu, \boldsymbol{k}, \boldsymbol{\theta}) = L_{\text{main}}(\mu, \boldsymbol{k}, \boldsymbol{\theta}) \cdot \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta_t{}^2}{2}} \, . \tag{5.3}$$

On the other hand normalisation factors are applied without any prior and are referred to as free floating parameters. The best estimate for the parameter set $(\mu, \boldsymbol{k}, \boldsymbol{\theta})$ is obtained maximizing the likelihood function. The minimization of the negative log likelihood $-\log L$ is an alternative providing the same result while being numerically more stable and is generally used. The minimization is done using the minuit2 package [154]. In the $t\bar{t}H(H \to b\bar{b})$ analysis the uncertainties on signal strength (μ) and the $t\bar{t}+\geq 1b$, $t\bar{t}+\geq 1c$ normalisation factors are computed using minos which provides an improved precision on the uncertainty calculation and support asymmetric errors.

5.1.2. Averaging and pruning

In order to avoid statistical fluctuations in the systematic model and the fit, two procedures are applied: the *averaging* and *pruning* procedures. The first step in the systematic averaging is the sys-

tematic *symmetrization*. Systematics are separated in two categories with their own symmetrization algorithms:

- *One-sided* systematics: These are systematics for which the 1σ variation is available only in one direction, by convention the up variation. This is mainly the case of systematic uncertainties arising from comparing two MC samples. For one-sided uncertainties the symmetrization provides the down variation as the symmetric of the up variation around the nominal prediction.
- *Two-sided* systematics: These are uncertainties with both the up and down variations provided. In this case the symmetrization takes the mean difference between the up and down variations and uses it to re-define the up variation. The one-sided symmetrization is then applied to define the down variation.

The second step in the averaging procedure is the *smoothing* which averages systematic uncertainties across bins in each category. It is meant to remove fluctuations in the systematic model, in particular for uncertainties derived from the comparisons of several MC simulations with limited amount of generated events. In the $t\bar{t}H(H \to b\bar{b})$ analysis two procedures are implemented:

- The *root-smoothing* is directly based on the smooth function of histograms in one dimension TH1::Smooth. It averages bin contents based on neighboring bin information and the histogram integral. This procedure is only applied to the $t\bar{t}$+$\geq$1b NLO generator uncertainty (see section 5.4).
- The *main-smoothing* is applied to all other systematic uncertainties. It is based on two parameters. The first is the number of variations, i.e. the number of changes in the sign of the derivative of the distribution. The second is a statistical threshold which controls the minimal statistical uncertainties of bins used to define the systematic shape. The initial statistical threshold is set to 8% of the number of events in the bin. Bins are merged in groups until the relative statistical uncertainty is smaller than 8%. The number of variations in the distribution of merged bins is computed and if four or less changes in the derivative sign are found the histograms is kept. If not the procedure is repeated dividing the statistical threshold by two at each iteration until a configuration with four or less variations is found. The obtained histogram then defines the systematic shape. The normalisation effect is kept fixed to the integral of the original distribution and the smoothing only affects the shape of the systematic uncertainty.

In the case of small backgrounds with large statistical uncertainties the smoothing procedure is also applied to the nominal distribution. In the $t\bar{t}H(H \to b\bar{b})$ analysis the smoothing procedure is applied to the fake and non-prompt process as well as the templates of non-$b\bar{b}$ decays of the Higgs boson in $t\bar{t}H$ samples.

The pruning consists of the removal of small systematic components which do not affect the result to speed up the fit procedure and make it more robust. In the $t\bar{t}H(H \to b\bar{b})$ analysis shape and normalisation components are pruned separately if there effect is found to be smaller than 1%[a] after the averaging procedure.

5.2. Fit model

The *fit model* compiles all the inputs to the likelihood function. It is fully characterized by the templates of all processes, i.e. the variables used in each category and their binning, the systematic

[a] The shape component of an uncertainty is kept if at least one bin has an effect larger than 1%.

uncertainties and their correlation scheme for the various templates, and the normalisation factors on specific processes.

5.2.1. Fitted distributions

As explained in section 4.6.1, the $t\bar{t}H(H \to b\bar{b})$ single lepton channel is composed of 12 categories meant to be fitted simultaneously. Figure 5.1 shows the predicted number of events in each category before the fit compared to the amount of observed data events. Data overshoot the prediction in several categories with large fractions of $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ backgrounds. However the difference is covered by the systematic uncertainties and the free floating normalisations of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ components.

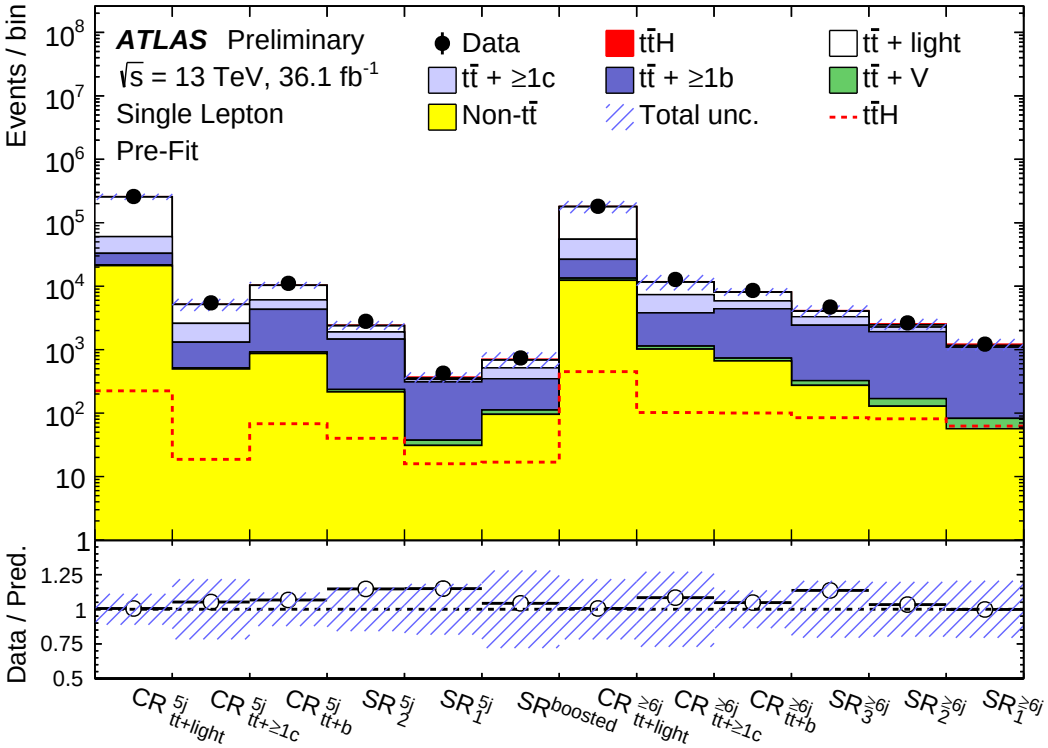


Figure 5.1.: Comparison of the predicted and observed yields in all categories of the single lepton channel before applying corrections from the fit to data. The signal contribution is shown both as a filled red area stacked on top of the backgrounds and as a separate dashed red line. The hashed band represent the statistical and systematic uncertainties. Uncertainties on the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ background normalisations are not included as those are free floating parameters of the fit.

Three distributions are used to define templates in each category:

- In the $5j$ and $\geq 6j$ $t\bar{t}+$light enriched and $t\bar{t}+b$ enriched categories only one bin is used. Indeed, the $H_{\mathrm{T}}^{\mathrm{had}}$ variable shape is not well modelled in these categories as shown in figure 5.2. In the fit to data, the correction of these shapes require several pulls of various uncertainties. It is not always clear if these corrections should be extrapolated to signal regions or not. In the case of

these four categories the shape mismodelling is not fully covered by the systematic uncertainties and the extrapolation of the corrections to the signal categories is not trustable. Moreover the $t\bar{t}$ + light background contributes significantly to these categories. The systematic model of this component is less flexible than the one of the $t\bar{t}$+$\geq$1b background and several pulls are observed in the non-$t\bar{t}$ systematic uncertainties to correct the shape in the $H_{\mathrm{T}}^{\mathrm{had}}$ distribution of the $t\bar{t}$+light background.

- In the $t\bar{t}+c$ enriched categories the $H_{\mathrm{T}}^{\mathrm{had}}$ distribution is kept. Here the $H_{\mathrm{T}}^{\mathrm{had}}$ shape mismodelling is smaller than in the other background-enriched categories and is covered by the systematic uncertainties. Moreover, the $H_{\mathrm{T}}^{\mathrm{had}}$ shape allows to disentangle the contributions from the $t\bar{t}$ + light, $t\bar{t}$+$\geq$1c and $t\bar{t}$+$\geq$1b background components and thus improves significantly the signal sensitivity.
- In the signal enriched categories the classification BDT output is used for a better separation of the $t\bar{t}H(H \to b\bar{b})$ signal from the $t\bar{t}$+$\geq$1b background.

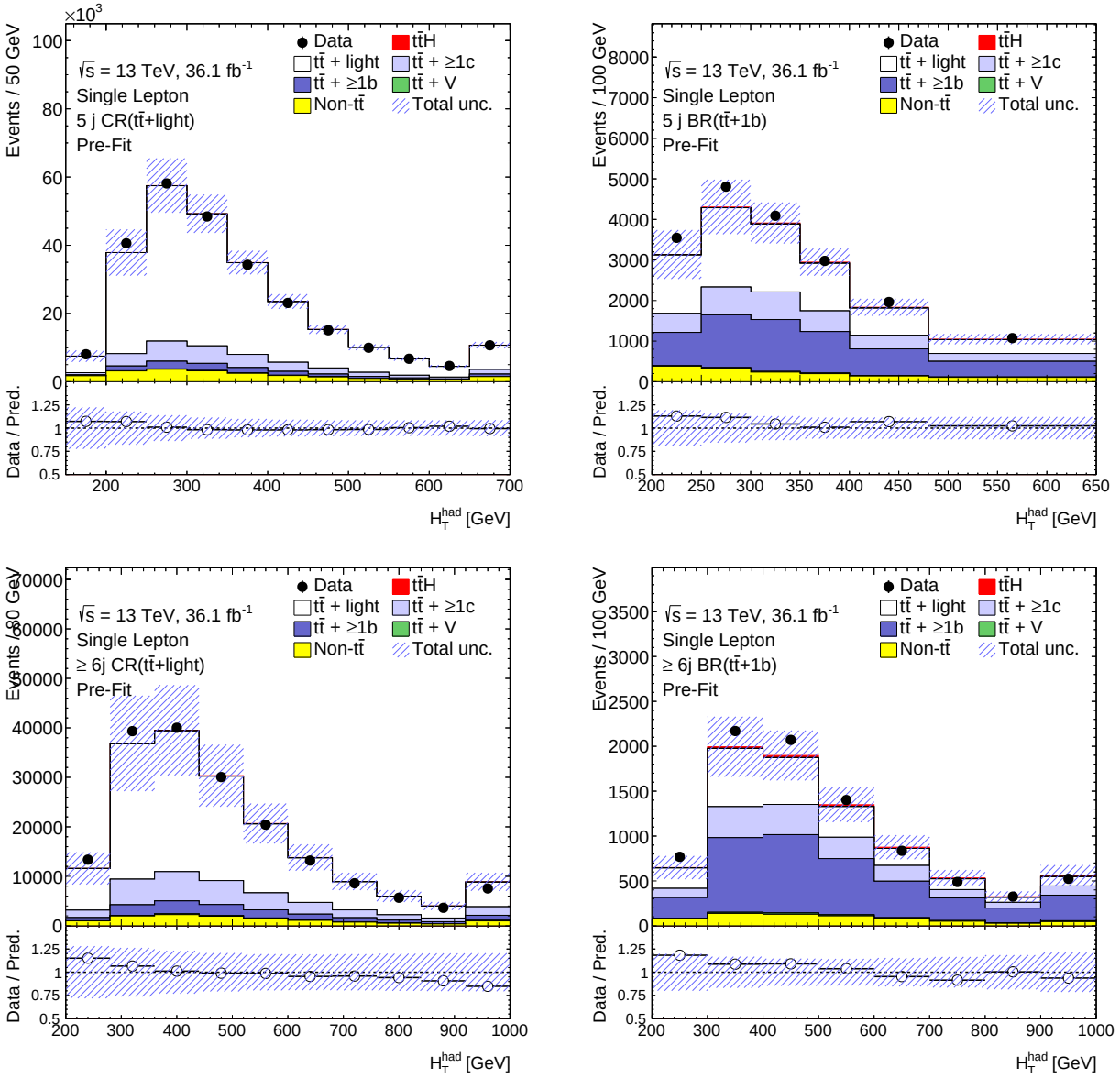


Figure 5.2.: Comparison of the predicted $H_{\mathrm{T}}^{\mathrm{had}}$ distribution to the one observed in data for the 5-jet (up) and ≥ 6 (bottom) categories enriched in the $t\bar{t}+\text{light}$ (left) and $t\bar{t}+b$ (right) backgrounds. The predicted distribution is shown before the corrections from the fit to data. The hashed area represent the statistical and systematic uncertainties. The uncertainties do not include the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit.

5.2.2. $t\bar{t}$ + jets models

The $t\bar{t}$ + jets model is a critical part of the $t\bar{t}H(H \to b\bar{b})$ analysis. As described in section 4.4.2 the $t\bar{t}$ + jets inclusive background modelling is based on the POWHEG+PYTHIA8 sample (PP8) which is found to describe data better than other available MC generators [146].

The $t\bar{t}$+jets background has a complex systematic model with large theoretical uncertainties coupled to a relatively weak constraint from data measurements. An uncertainty of $\pm 6\%$ is set on the inclusive $t\bar{t}$ NNLO+NNLL cross-section [147]. In the Run 1 result the $t\bar{t}$+$\geq 1b$ normalisation is found to be under-estimated by 35 to 40% in MC simulations. Moreover preliminary studies show that the $t\bar{t}$ + heavy flavours normalisation are potentially higher in data than in simulations. Thus the individual normalisations of both the $t\bar{t}$+$\geq 1b$ and the $t\bar{t}$+$\geq 1c$ components are left free to float in the fit.

$t\bar{t}$ + jets shape uncertainties are based on the differences between the PP8 predictions and other MC simulations. Four alternative $t\bar{t}$ + jets samples are produced to define three systematic uncertainties:

- The *NLO Generator* uncertainty is defined as the relative difference between the PP8 and the Sherpa prediction. The Sherpa[b] sample is an inclusive $t\bar{t}$ sample generated using SHERPA 2.2.1 [99] with the NNPDF3.0NNLO parton distribution function set. The renormalisation and factorisation scales are set to $\sqrt{m^2_{\mathrm{T},t} + m_{\mathrm{T},\bar{t}}}$.

- The *Parton Shower (PS) and hadronisation* uncertainty is derived from the comparison of the PP8 to the POWHEG+HERWIG7 simulations. The later being produced with the same generator settings as the nominal $t\bar{t}$ sample but interfaced with the HERWIG7 version 7.0.1 [107] showering with the H7-UE-MMHT tune.

- The *radiation* uncertainty accounts for the modelling of the initial and final state radiations. Two alternative PP8 samples are generated and compared to the nominal PP8 sample to define this uncertainty. The up variation is generated with the nominal PP8 $t\bar{t}$ setup but with the renormalisation and factorisation scales decreased by a factor two, the $hdamp$ parameter increased by a factor 2 and the Var3cUp variation of the A14 tune in PYTHIA. The down variation is generated with the nominal PP8 $t\bar{t}$ settings but using the Var2cDown variation of the A14 tune and increasing the scales in POWHEG by a factor two.

These three uncertainties are applied uncorrelated between the $t\bar{t}$ + light, $t\bar{t}$+$\geq 1c$ and $t\bar{t}$+$\geq 1b$ components. In order to avoid large fluctuations in the systematic model, especially in the $t\bar{t}$+$\geq 1b$ component (see 5.4) additional samples are generated with a $t\bar{t}$+$\geq 1b$ filter before the simulation step. Given that the overall normalisation of the $t\bar{t}$+$\geq 1b$ and $t\bar{t}$+$\geq 1c$ backgrounds are free to float, the relative fractions of $t\bar{t}$+$\geq 1b$, $t\bar{t}$+$\geq 1c$ and $t\bar{t}$ + light in each alternative sample are corrected to match the PP8 predictions. This procedure avoids double counting ot the uncertainties on the normalisation of the $t\bar{t}$ + jets components in the fit.

The $t\bar{t}$+$\geq 1b$ background modelling is one of the most challenging component of the $t\bar{t}H(H \to b\bar{b})$ analysis. The large uncertainties on the $t\bar{t}$+$\geq 1b$ prediction have a large impact on the signal sensitivity (see section 5.3.2) but there exist little guidance on which MC generator describes data the best. Thus, several $t\bar{t}$+$\geq 1b$ models are confronted. In particular the Run 1-like model, which uses the truth re-weighting of the PP8 sample to the SHERPA+OPENLOOPS (Sherpa+OL) prediction both for the fractions and shapes of the $t\bar{t}$+$\geq 1b$ sub-component, is modified to take into account new observations from studies done in Run 2. In this thesis two models are presented. The model chosen as the baseline

[b] Sherpa refers to the $t\bar{t}$ inclusive sample generated with five flavour scheme parton distribution functions. It should not be confused with the NLO $t\bar{t} + b\bar{b}$ Sherpa+OL sample generated with four flavour parton distribution functions.

for the $t\bar{t}H(H \to b\bar{b})$ search, called the default model and published in [141], uses the Sherpa+OL prediction for the nominal fractions of the $t\bar{t}+\geq 1b$ sub-components and the systematic model. However it is observed that the PP8 sample describes better the fractions of some sub-components, especially $t\bar{t}+ \geq 3b$ (see section 5.6.3), compared to the Sherpa+OL sample. Thus an alternative model, also called PP8-based model, is presented in this thesis. It is based on the nominal PP8 prediction and uses additional flexibility on the systematic model to test the fractions and shapes of each $t\bar{t}+\geq 1b$ sub-component in data.

The default model: This model is based on the PP8 prediction for the $t\bar{t}+\geq 1b$ background with the fractions of each sub-component ($t\bar{t}+b$, $t\bar{t}+b\bar{b}$, $t\bar{t}+B$ and $t\bar{t}+ \geq 3b$) re-weighted to Sherpa+OL as described in section 4.4.2. Systematic uncertainties on Sherpa+OL prediction for the fractions of the $t\bar{t}+\geq 1b$ sub-categories are derived from the alternative Sherpa+OL samples listed in table 5.1. For each variation, the differences of the fractions is taken as a systematic correlated between $t\bar{t}+\geq 1b$ sub-components. In addition a 50% prior normalisation uncertainty is applied to the $t\bar{t}+ \geq 3b$ category to cover the overshoot in the Sherpa+OL prediction (see section 5.6.3).

Systematic source	Description
SHERPA+OPENLOOPS variations:	
→ $t\bar{t}+\geq 1b$ renorm. scale	Up or down by a factor of two
→ $t\bar{t}+\geq 1b$ resumm. scale	Vary μ_{Q} from $H_{\mathrm{T}}/2$ to μ_{CMMPS}
→ $t\bar{t}+\geq 1b$ global scales	Set μ_{Q}, μ_{R}, and μ_{F} to μ_{CMMPS}
→ $t\bar{t}+\geq 1b$ shower recoil	Alternative model scheme
→ $t\bar{t}+\geq 1b$ PDF set 1	CT10 vs. NNPDF
→ $t\bar{t}+\geq 1b$ PDF set 2	CT10 vs. MSTW
→ $t\bar{t}+\geq 1b$ FSR	Radiation variation samples
→ $t\bar{t}+\geq 1b$ UE	Alternative set of tunable parameters for the underlying event
Others:	
→ $t\bar{t}+\geq 1b$ MPI	Up or down by 50%
→ $t\bar{t}+ \geq 3b$ normalisation	$t\bar{t}+ \geq 3b$ up or down by 50%

Table 5.1.: Definitions of the systematic uncertainties related to the fractions of the $t\bar{t}+\geq 1b$ sub-components in the default model. Differences between the PP8 prediction corrected to match the baseline SHERPA+OPENLOOPS sample (default) and the PP8 prediction corrected to match a SHERPA+OPENLOOPS variation defines the systematic uncertainty.

The $t\bar{t}+\geq 1b$ model inherits from the $t\bar{t}+$ jets NLO generator, PS and hadronisation, and radiation uncertainties described above. Since the $t\bar{t}+\geq 1b$ sub-categories fractions have their dedicated systematics, all alternative samples are corrected to the Sherpa+OL prediction before deriving the corresponding systematic uncertainty. The difference after the correction of the $t\bar{t}+\geq 1b$ sub-categories is referred to as *residual difference*.

Finally, an uncertainty on the *four flavour scheme versus five flavour scheme shape differences (5FS vs 4FS shape)* is derived from the comparison of the PP8 sample to the Sherpa+OL sample. In order to account only for the shape differences, the PP8 sample is corrected to match the Sherpa+OL predic-

tion of the fractions of the $t\bar{t}+\geq 1b$ sub-categories, then the two samples are normalized to the same cross section.

Figure 5.3 shows the systematic uncertainties applied to the $t\bar{t}+\geq 1b$ sample in the most signal enriched category. As mentioned above the $t\bar{t}+\geq 1b$ is hard to model and is badly constrained by data. This leads to large shape differences between the various MC samples considered as well as inter-category normalisation corrections. Among all uncertainties, the $t\bar{t}+\geq 1b$ NLO generator systematic provides the largest shape effect and its down variation mimics the shape of the signal in the BDT distribution. The PS and hadronisation has the largest impact on the normalisation of the $t\bar{t}+\geq 1b$ background in the most sensitive region ($\geq$ 6j SR1) together with a relatively small contribution to the BDT output shape. The 4FS vs 5FS shape uncertainty is flattened by the smoothing procedure and ends up rather small in $\geq$ 6j SR1. However this uncertainty have a large impact on sensitivity due to large shape components in the other categories (see section 5.3.2). The uncertainties from the Sherpa+OL variations affect the relative fractions of $t\bar{t}+b$, $t\bar{t}+b\bar{b}$, $t\bar{t}+B$ and $t\bar{t}+\geq 3b$ and are relatively small. As the $\geq$ 6j SR1 category is dominated by the $t\bar{t}+b\bar{b}$ sub-component (70% of the $t\bar{t}+\geq 1b$ background), the Sherpa+OL variation systematic uncertainties are not expected to have a large shape effect in this category. However these uncertainties allow inter-category normalisation corrections as the categories have different fractions of each $t\bar{t}+\geq 1b$ sub-components. They also provide a shape correction in categories which are not dominated by a single $t\bar{t}+\geq 1b$ sub-component. The uncertainties for all signal-enriched categories can be found in appendix C.

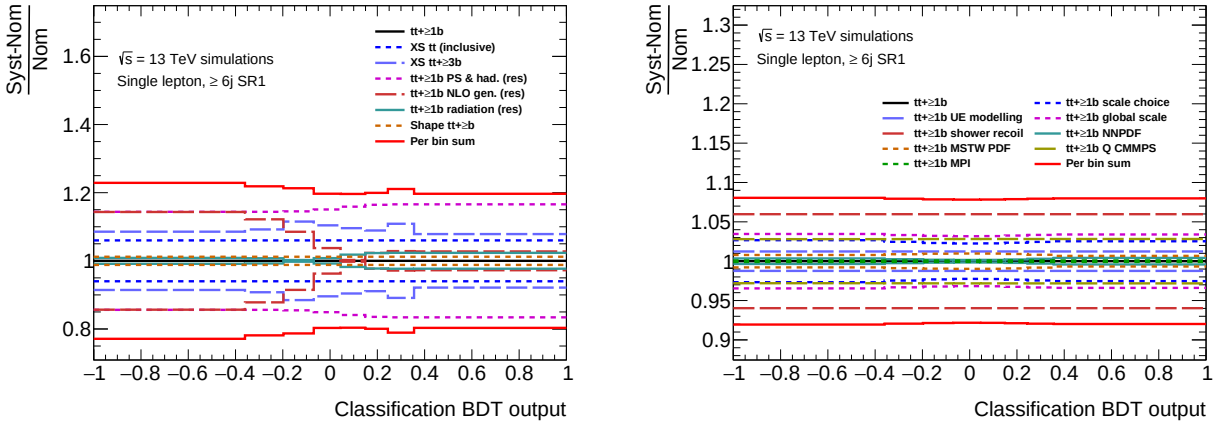


Figure 5.3.: Relative variations induced by the $t\bar{t}+\geq 1b$ systematic uncertainties on the $t\bar{t}+\geq 1b$ sample at $\pm 1\sigma$ for the classification BDT distribution in the $\geq$ 6j SR1 category. (left) common uncertainties between the two $t\bar{t}+\geq 1b$ models: $t\bar{t}$ cross section, 50% normalisation uncertainty of the $t\bar{t}+\geq 3b$ sub-component and various MC to MC comparison uncertainties. (right) systematic uncertainties from SHERPA+OPENLOOPS variations. In both plots the red lines show the quadratic sum of the up and down effects systematic uncertainties in each bin.

The PP8-based model: In this model the nominal fractions of the $t\bar{t}+\geq 1b$ sub-categories are directly taken from the PP8 sample. The flexibility of the systematic model is increased by considering the normalisation and shape of each sub-component as an independent nuisance parameter.

The $t\bar{t}+\geq 1b$ normalisation factor is replaced by 50% normalisation uncertainties decorrelated across different $t\bar{t}+\geq 1b$ sub-components. In fact, it is shown that the $k-$factor (free floating normalisation parameter) of the $t\bar{t}+\geq 1b$ background constraint is higher than 50% and its value is below 1.5 in Run 2

data (see section 5.6.3). In addition the 4FS vs 5FS shape uncertainty is derived independently for the $t\bar{t}+b$, $t\bar{t}+b\bar{b}$, $t\bar{t}+B$ and $t\bar{t}+\geq 3b$ backgrounds and applied as uncorrelated nuisance parameters for each sub-component of the $t\bar{t}+\geq 1b$ background. Similarly to the default model, only the residual component of the Generator, PS and hadronisation, and radiation systematic uncertainties are kept.

The systematic uncertainties applied to the $t\bar{t}+\geq 1b$ sample in the PP8-based model are shown in figure 5.4 for the most signal like category. The uncertainties from MC to MC comparisons are similar to the ones of the default model. In particular, the $t\bar{t}+\geq 1b$ NLO generator uncertainty is the largest source of uncertainty on the BDT output shape. Once decorrelated across $t\bar{t}+\geq 1b$ sub-components, the 4FS vs 5FS shape uncertainty has significant shape contributions for most of the categories. The uncertainties for all signal-enriched categories can be found in appendix C.

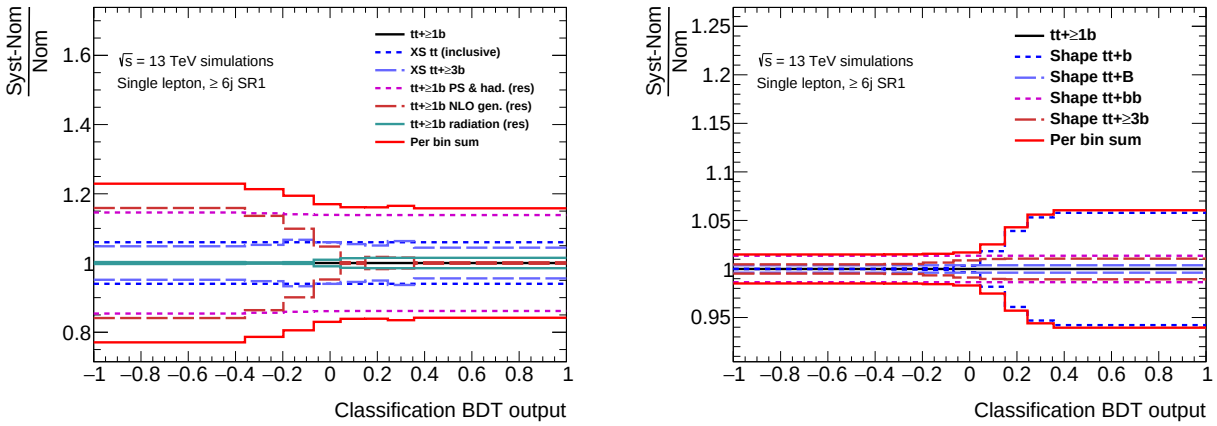


Figure 5.4.: Relative variations induced by the $t\bar{t}+\geq 1b$ systematic uncertainties on the $t\bar{t}+\geq 1b$ sample at $\pm 1\sigma$ for the classification BDT distribution in the $\geq$ 6j SR1 category. (left) common uncertainties between the two $t\bar{t}+\geq 1b$ model: $t\bar{t}$ cross section, 50% normalisation uncertainty of the $t\bar{t}+\geq 3b$ sub-component and various MC to MC comparison uncertainties. (right) 4FS to 5FS shape comparison uncertainties. In both plots the red lines show the quadratic sum of the up and down effects systematic uncertainties in each bin.

Similarly to the $t\bar{t}+\geq 1b$ background, there exists little guidance from theory or experimental measurements for the $t\bar{t}+\geq 1c$ background. However this background has a smaller impact on the $t\bar{t}H(H\to b\bar{b})$ signal than the $t\bar{t}+\geq 1b$ background. The differences between the PP8 prediction and a Madgraph5_aMC@NLO +HERWIG++ $t\bar{t}+c\bar{c}$ prediction at NLO using the three flavour scheme is applied as an additional systematic on the $t\bar{t}+\geq 1c$ sample.

The $t\bar{t}+\geq 1c$ background is hard to control precisely as there is no control region dominated by this component. When using the PP8-based model a 50% Gaussian prior is added to the normalisation of the $t\bar{t}+\geq 1c$ background while it is a free floating parameter in the default model.

5.2.3. Non-$t\bar{t}$ modelling uncertainties

The $t\bar{t}H(H\to b\bar{b})$ analysis includes several systematic uncertainties to account for the modelling of all backgrounds and of the signal. Theses uncertainties (excluding the ones related to the $t\bar{t}+$ jets background) are listed in table 5.2. Most of them are small and contribute marginally to the signal sensitivity. Only the signal uncertainties, especially the $t\bar{t}H(H\to b\bar{b})$ parton shower uncertainties con-

tribute significantly to the signal sensitivity. However, their impact is still sub-dominant with respect to the $t\bar{t}+$jets systematic uncertainties. More details on the non-$t\bar{t}$ background uncertainties can be found in ref [141].

Channel	Type	Systematic uncertainties
$t\bar{t}H$	N	PDF, QCD scale norm uncertainties
	N	Branching ratios: uncorrelated normalisations per decay mode
	SN	Parton shower: comparison to MG5_aMC@NLO+Herwig++
$t\bar{t}+V$	N	PDF, QCD scale norm uncertainties
	SN	NLO generator: comparison to Sherpa
W/Z+jets	N	40% cross section unc.
	N	30% V+2-heavy-flavoured-jets norm uncertainty
	N	30% $V+\geq$3-heavy-flavoured-jets norm uncertainty
Di-boson	N	50% cross section uncertainty
Single top: W-channel	N	5% cross section uncertainty
	SN	Parton shower and radiation: similar to $t\bar{t}+$jets
	SN	Diagram subtraction: MC-MC comp for overlap removal with $t\bar{t}$
Single top: t-channel	N	5% cross section uncertainty
	SN	Parton shower and radiation: similar to $t\bar{t}+$jets
Single top: s-channel	N	5% cross section uncertainty
tZ	N	PDF and QCD scale variations
WtZ	N	50% cross section uncertainty
4-tops	N	50% cross section uncertainty
$t\bar{t}WW$	N	PDF and QCD scale variations
Fakes & non-prompts	N	50% norm uncertainty
	N	$\rightarrow$ Decorrelated for e and μ channels
	N	$\rightarrow$ Decorrelated across N(jets) and for boosted category

Table 5.2.: Systematic uncertainties on the non-$t\bar{t}$ backgrounds and on the signal categorized per process. The second column shows the type of the uncertainty where N stands for normalisation, S for shape and SN for both.

The MC statistical errors are also included in the fit. The uncertainty on each bin is computed as the quadratic sum of the MC statistical error of all background components. These uncertainties are added as nuisance parameters, called γ's, affecting all background components and treated as uncorrelated across all bins of the analysis.

Nuisance parameters associated to MC statistical errors are large and affect all background components. Thus they can easily be used to cover mismodelling. Studies in other analyses have shown that to avoid a bias in the signal extraction the binning needs to be chosen such that none of the γ nuisance parameters is larger than 20% of the background yields in the corresponding bin. The largest MC statistical uncertainty in a given bin is 12% after the binning optimisation procedure described in section 5.3.1.

5.2.4. Experimental uncertainties

An uncertainty of 2.1% is applied to the normalisation of all processes determined by MC simulation. It accounts for the uncertainty on the estimation of the $2015 + 2016$ integrated luminosity and is derived with a similar methodology as described in [155].

Twenty systematic uncertainties are assigned to jet energy scale including the uncertainties from the jet calibration, the high p_{T} extrapolation, the jet flavour, pile-up treatment and η interpolation [123]. In addition systematic uncertainties are added to account for the jet energy resolution (JER) and the efficiency to pass the JVT cut. Even though these uncertainties are relatively small on individual jets they are inflated by the large number of jets in the final state. In particular the JER uncertainty is used to correct the $t\bar{t}+$ jets prediction in the $t\bar{t}+\geq 1c$ enriched categories of the single and di-lepton channels. In order to avoid the propagation of this pull to the signal regions the JER uncertainty is applied as two uncorrelated nuisance parameters, one in $t\bar{t}+\geq 1c$-enriched categories and one in the other categories (see section 5.6.4).

The uncertainties on the b-jet, c-jet and $light$-jet tagging efficiency are extracted from measurements on data [156]. The b-jet, c-jet and $light$-jet are calibrated in several jet p_{T} bins for each working point. In addition two η bins are used for the $light$-jet tagging efficiencies. A diagonalization procedure of the error matrix allows to keep the correlation between each bin, while providing variations that can be considered uncorrelated. In total, 30 nuisance parameters correspond to the b-jet tagging efficiency, 15 to the c-jet tagging efficiency and 80 to the $light$-jet tagging efficiency.

Uncertainties on the lepton trigger, reconstruction, identification and isolation efficiencies as well as lepton energy scale and momentum are considered. The lepton related systematic uncertainties are very small in the $t\bar{t}H(H \to b\bar{b})$ analysis and contributes marginally to the fit.

Some of the experimental uncertainties are correlated with the $t\bar{t}$ modelling, in particular b-tagging related uncertainties. These systematics are studied in details, especially the ones which have a significant impact on the signal sensitivity. Two examples of such studies are shown in section 5.6.4.

5.3. Performance of the $t\bar{t}H(H \to b\bar{b})$ analysis

As mentioned in the introduction, the $t\bar{t}H(H \to b\bar{b})$ analysis is conducted as blinded. Its performance, both on signal sensitivity and for the constraints on the background uncertainty, is estimated running the fit procedure on the so called *Asimov data-set*. The Asimov data-set is built from the predicted distribution assuming Poisson error in each bin. The fit to Asimov data is a crucial ingredient for the $t\bar{t}H(H \to b\bar{b})$ analysis. Many studies and decisions on the analysis strategy are done based on the obtained performance of this fit: choice of the event categorization, choice of MVAs, binning optimisation, studies of the systematic uncertainties and their impact on the signal sensitivity ... A selection of these studies is presented in this section: binning optimisation, constraints on a set of important nuisance parameters and the impact of various uncertainties on the signal sensitivity. The results of the fit to the Asimov data-set are also described.

5.3.1. Binning optimization

The binning of the discriminant in $t\bar{t}+\geq 1c$-enriched regions is optimized to avoid statistical fluctuations, especially in the systematic model, while keeping a high constraining power on the $t\bar{t}+\geq 1c$ and $t\bar{t}+\geq 1b$ backgrounds.

In signal regions the binning calculation is automatized to council high separation of the $t\bar{t}H(H \to b\bar{b})$

signal from the background and avoid bins with large statistical errors. The automatic binning algorithm scans the original distribution, starting from the bin with largest BDT output, and merges bins until a certain fraction of signal and background events is obtained. The merging threshold is defined by the function Z:

$$Z = z_b \frac{n_b}{N_b} + z_s \frac{n_s}{N_s} \tag{5.4}$$

where n_s (n_b) is the number of signal (background) events in the merging bin, N_s (N_b) is the total number of signal (background) events, z_s and z_b are two tunable parameters. A bin is formed when Z becomes equal to 1 or more. The z_s (z_b) parameter controls the maximum fraction of signal (background) events in each bin with the condition $z_s + z_b = N(\text{bins})$. Figure 5.5 shows the signal and background shapes assuming $N(\text{bins}) = 8$ for three cases: $z_s = 0$ leading to a flat background template, $z_s = z_b$ and $z_b = 0$ leading to a flat signal template.

Two other automatic binning functions are implemented. The best performance in terms of signal sensitivity after the fit is obtained using the function Z mentioned above with $z_s = z_b$ in most of the categories. $N(\text{bins}) = 8$ gives the best compromise between signal sensitivity and MC statistics in the fit to the Asimov data-set. The final distributions of the classification BDT in the various categories of the single lepton channel are shown in section 5.6.1.

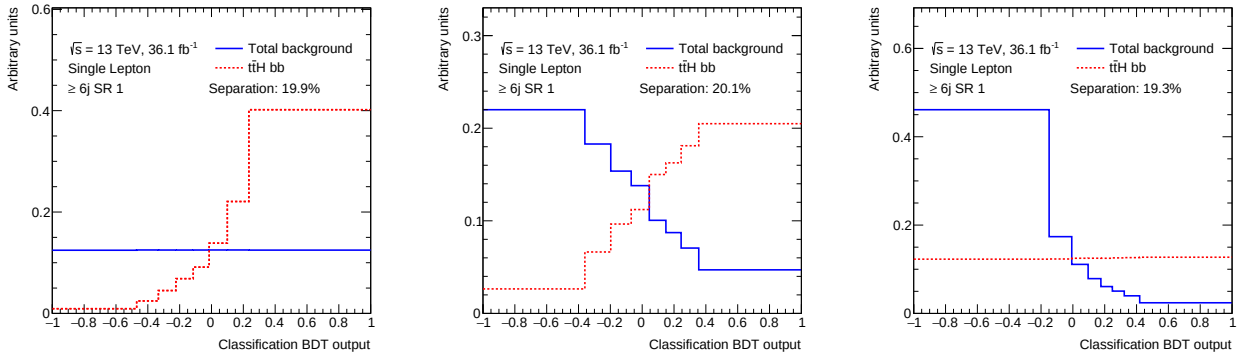


Figure 5.5.: Normalized distributions of the classification BDT output distributions for the total background (blue line) and $t\bar{t}H(H \to b\bar{b})$ signal (dashed red line) in the $\geq$ 6j SR1 category. The binning is computed with the automatic binning function Z for 8 bins with $z_s = 0$ and $z_b = 8$ (left), $z_s = z_b = 4$ (middle), $z_s = 8$ and $z_b = 0$ (right).

As explained in section 5.2.1, using only one bin in the 5j BR($t\bar{t}$ + light), $\geq$ 6j BR($t\bar{t}$ + light), 5j BR($t\bar{t} + b$) and 5j BR($t\bar{t} + b$) categories allows to reduce the pulls on the fit to data and the tensions between the different categories at the cost of a 10% loss in signal sensitivity. In the $t\bar{t}$+$\geq$1c enriched categories the shape of the $H_{\text{T}}^{\text{had}}$ variable is kept and several binnings are tested: varying the number of bins, using different automatic binning functions or no automatic binning. The usage of only one bin in these categories induces a significant loss in sensitivity due to lower constraints on the nuisance parameters mainly associated with the $t\bar{t}$+$\geq$1c modelling. For the other options, no strong difference is observed in the fit to the Asimov data-set on the nuisance parameter constraints and on the signal sensitivity. Thus the simplest option with 6 bins and 8 bins of equal size is kept for the 5j BR($t\bar{t}$+$\geq$1c) and $\geq$ 6j BR($t\bar{t}$+$\geq$1c) categories, respectively.

5.3.2. $t\bar{t}H(H \to b\bar{b})$ signal sensitivity

The signal model in the default model is compared to the one obtained with the PP8-based model in figure 5.6. The default model yields a signal strength of $1.00^{+0.68}_{-0.65}$, corresponding to a 1.5σ expected significance of the $t\bar{t}H(H \to b\bar{b})$ signal. The uncertainty on μ is dominated by the systematic uncertainties and the Asimov data statistics only induces a ± 0.32 uncertainty on μ. The statistical uncertainty also includes the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations. In fact these uncertainties are free to float in the fit and the statistical error is expected to scale with the observed values of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations. In the PP8-based model the signal strength is less correlated to the background modelling nuisance parameters and a 10% higher sensitivity is observed. In the case of the PP8-based model the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations are nuisance parameters with priors and are thus included in the systematic uncertainty. This explains the reduction of the statistical error on μ for this model.

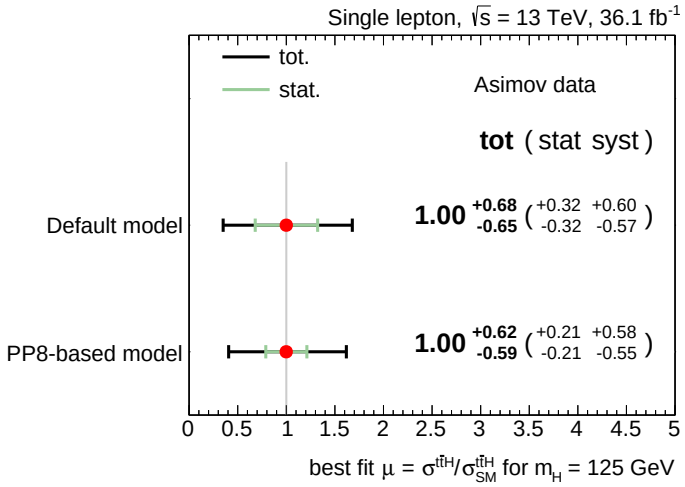


Figure 5.6.: Fitted value of the signal strength and its uncertainty from the fit with the two $t\bar{t}+\geq 1b$ models to the Asimov data-set in the single lepton channel.

Table 5.3 shows the break-down of the uncertainty on the signal strength in the various uncertainties grouped by sources. The two $t\bar{t}+\geq 1b$ models have the same behavior and thus only the break-down of the default model is shown. As for Run 1, the $t\bar{t}H(H \to b\bar{b})$ sensitivity is mostly limited by the $t\bar{t}+$ jets model, especially the $t\bar{t}+\geq 1b$ component. The uncertainty on the signal strength related to the $t\bar{t}+\geq 1b$ background nuisance parameters is $^{+0.49}_{-0.48}$ in addition to $^{+0.12}_{-0.14}$ accounting for the $t\bar{t}+\geq 1b$ normalisation.

Even with the automatic binning procedure the background MC statistics and the statistical error on the fake lepton estimation are also a limiting factor of the analysis. Indeed, their combined impact on the signal strenght uncertainty is $^{+0.29}_{-0.31}$ and is the second largest contribution after the $t\bar{t}+\geq 1b$ modelling.

The b-tagging and jet energy related systematics also have a significant impact on the signal sensitivity. This is partially due to the correlations of few of the corresponding nuisance parameters to the $t\bar{t}+$ jets modelling (see 5.6.3).

Uncertainty source	$\Delta\mu$	
$t\bar{t}+\geq 1b$ modelling	$+0.49$	-0.48
Background model statistics	$+0.29$	-0.31
$t\bar{t}H$ modelling	$+0.24$	-0.03
Jet flavour tagging	$+0.16$	-0.15
Jet energy scale and resolution	$+0.12$	-0.13
$t\bar{t}+\geq 1c$ modelling	$+0.11$	-0.12
Other background modelling	$+0.10$	-0.10
$t\bar{t}$+light modelling	$+0.06$	-0.06
Luminosity	$+0.03$	-0.03
Light lepton (e,μ) ID, isolation, trigger	$+0.03$	-0.03
Jet-vertex association, pileup modelling	$+0.01$	-0.01
Total systematic uncertainty	$+0.64$	-0.61
$t\bar{t}+\geq 1b$ normalisation	$+0.12$	-0.14
$t\bar{t}+\geq 1c$ normalisation	$+0.03$	-0.01
Statistical uncertainty	$+0.21$	-0.21
Total uncertainty	$+0.68$	-0.65

Table 5.3.: Summary of the effects on the signal strength uncertainty of the nuisance parameters grouped in categories by sources. The background model statistics refers to the statistical uncertainties from the limited number of simulated events and from the data-driven determination of the non-prompt and fake lepton background component in the single-lepton channel. The normalisation factors for both $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ are not included in the statistical component. The impact of each group is obtained running the fit without the corresponding uncertainties and subtracting the obtained error from the total uncertainty in quadrature.

5.3.3. Study of the constraints on the various nuisance parameters

The twenty most important individual systematic uncertainties, ranked by there impact on the signal strength error, is shown in figure 5.7. The first four nuisance parameters are all from the $t\bar{t}+\geq 1b$ background model and are all constrained to at least 0.5σ. The $t\bar{t}+\geq 1b$ NLO generator uncertainty is the leading source of uncertainty with a post-fit contribution to the error on μ of $^{+0.45}_{-0.43}$. Moreover this systematic uncertainty is constrained to 0.47σ. A detailed study of this systematic is shown in section 5.4.

The leading systematic uncertainty related to experimental sources is the first eigenvector in the $light$-jet efficiency uncertainty decomposition (l-tag e.v. 0). A high impact on the sensitivity from this nuisance parameter is not expected. In fact, the most signal-enriched categories have several b-tagged-jets at the tight working point which has a very high $light$-jet rejection. However, the l-tag e.v. 0 nuisance parameter is correlated to several $t\bar{t}+$jets nuisance parameters (see figure 5.32) which then have an important impact on the signal strength.

The l-tag e.v. 0 nuisance parameter is also constrained to 0.54σ. Naively, the $t\bar{t}H(H\rightarrow b\bar{b})$ analysis is not expected to constrain the l-tag e.v. 0 more than the dedicated analysis used to measure this parameter. However the uncertainties on the $light$-jet identification are large, up to 100% in several jet p_{T} and η bins. This results in large uncertainties on the predicted MC yields which are shown in

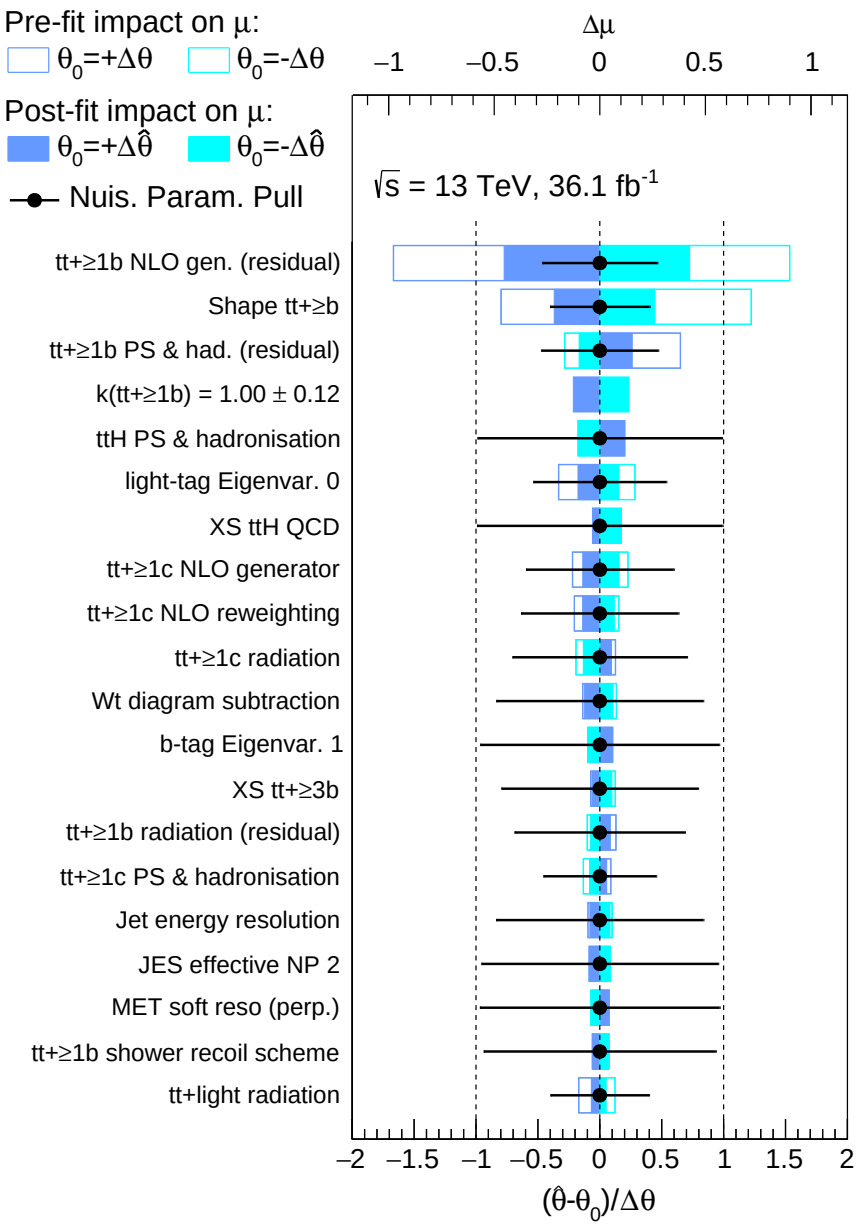


Figure 5.7.: Ranking of the nuisance parameters used in the fit according to their impact on sensitivity. Only the 20 first ones are shown. The filled (open) blue rectangles correspond to the post(pre)-fit contribution of the systematic to the uncertainty on the signal strength and is scaled with respect to the upper axis. The horizontal bar on the black points show the post-fit uncertainties after applying the constraints from the fit and scaled with respect to the bottom axis. The post(pre)-fit impact on sensitivity is computed performing the fit fixing the nuisance parameter at the post(pre)-fit $\pm 1\sigma$ variation and taking the difference in the fitted μ with the default fit.

figure 5.8. In particular, categories with several b-tagged-jets at the loose working points and with large fractions of $t\bar{t}+$ light and $t\bar{t}+bl$ ($t\bar{t}+b$ where the fourth b-tagged jet is a $light$-jet) have a large number of mistagged $light$-jets. Potential constraint on the l-tag e.v. 0 is thus possible.

The l-tag e.v. 0 affects simultaneously the shape and normalisation of all processes modelled by MC and all categories. The major contributions to the constraint are evaluated from the Asimov fit with three decorrelation schemes:

- *Region decorrelation*: the l-tag e.v. 0 uncertainty is treated uncorrelated for each of the analysis categories in the fit.
- *Sample decorrelation*: the l-tag e.v. 0 uncertainty is treated uncorrelated for the different signal and background samples used in the fit.
- *Shape/Acc decorrelation*: the shape and normalisation components of the l-tag e.v. 0 uncertainty are separated and treated uncorrelated in the fit.

Several categories contribute to constraint of the l-tag e.v. 0 nuisance parameter as can be seen in figure 5.9. The reduction of the uncertainty mostly happens in categories where the largest contamination from mistagged $light$-jets is expected. The sample decorrelation shows that the constraint originates from the $t\bar{t}+$ light and $t\bar{t}+\geq 1b$ samples as expected.

Moreover, the error on the signal strength is not affected by the choice of the correlation scheme. The difference in the signal strength uncertainty between the default correlation and the region decorrelation schemes is below 2%. However, the region decorrelation does not show a reduction of the l-tag e.v. 0 uncertainty in the two most significant categories: ≥ 6j SR1 and 5j SR1. It proves that the impact of the constraint extrapolation to the most signal like categories, when the nuisance parameter is correlated across categories, has a negligible impact on the signal sensitivity.

In consequence, the constraint on the l-tag e.v. 0 uncertainty nuisance parameter is justified. Moreover, this constraint does not impact significantly the analysis sensitivity. Thus, no further actions are taken to modify this uncertainty.

Finally, figure 5.7 shows that the signal uncertainties have a non negligible impact on the sensitivity. In particular the $t\bar{t}H$ PS and hadronisation systematic, which accounts for the differences between the PYTHIA8 and HERWIG++ showering, comes directly after the leading $t\bar{t}+\geq 1b$ uncertainties. However, their impact on signal sensitivity is still sub-dominant compared to the $t\bar{t}+\geq 1b$ modelling uncertainties and is not constrained. However future iterations of the analysis will need more accurate description of signal uncertainties, especially if a better $t\bar{t}+\geq 1b$ model reduces the associated uncertainties.

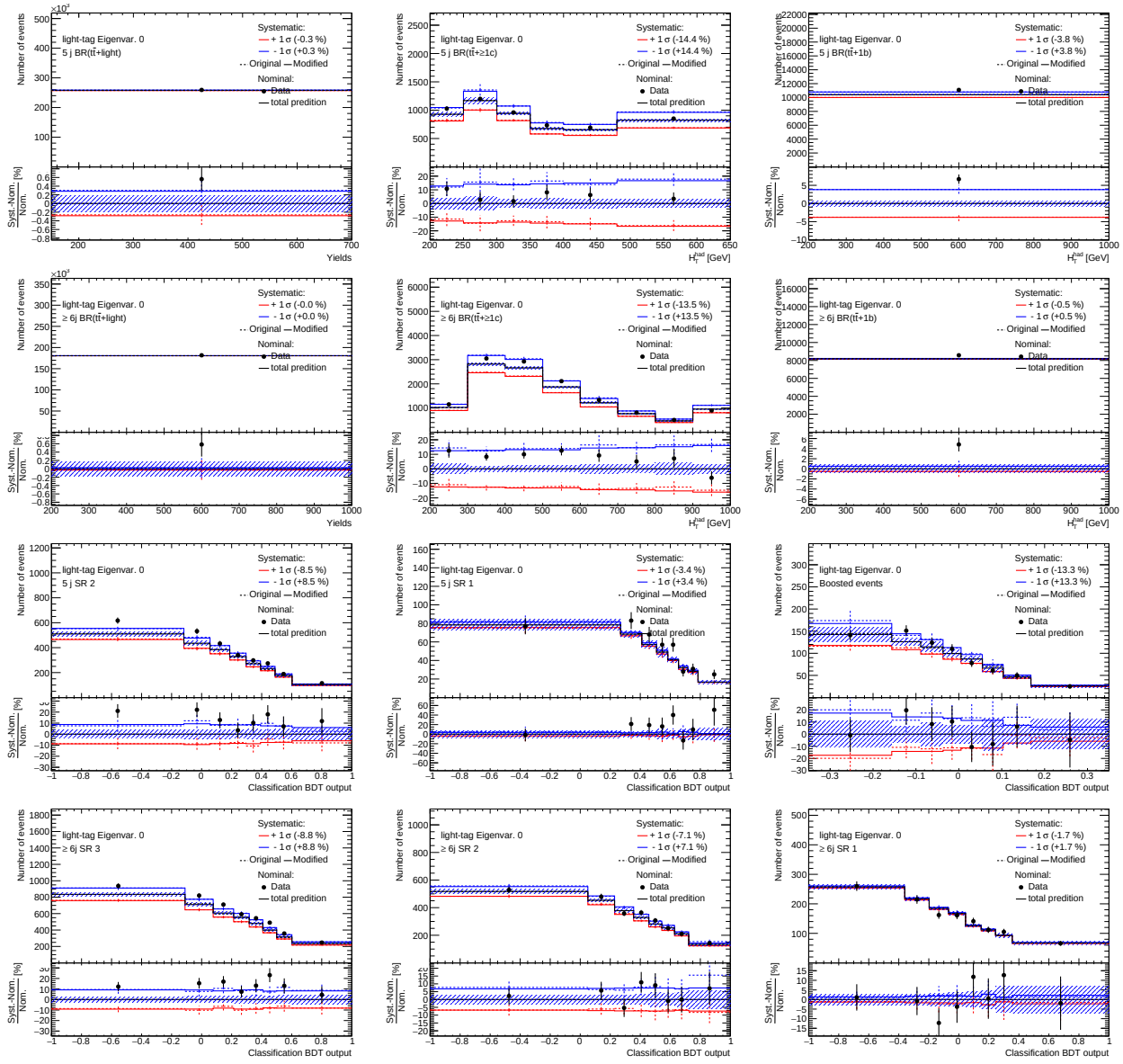


Figure 5.8.: Total prediction of all signal and background samples including the $\pm 1\sigma$ variations induced by the l-tag e.v. 0 uncertainty in all categories compared to data. Colored points (lines) displays the systematic uncertainty before(after) smoothing. The main-smoothing is used in these plots. The black points represent data and the black solid line represent the nominal prediction. The lower pad displays the relative systematic uncertainty in percent. This relative uncertainty is compared to the relative difference between the nominal prediction and data (black points).

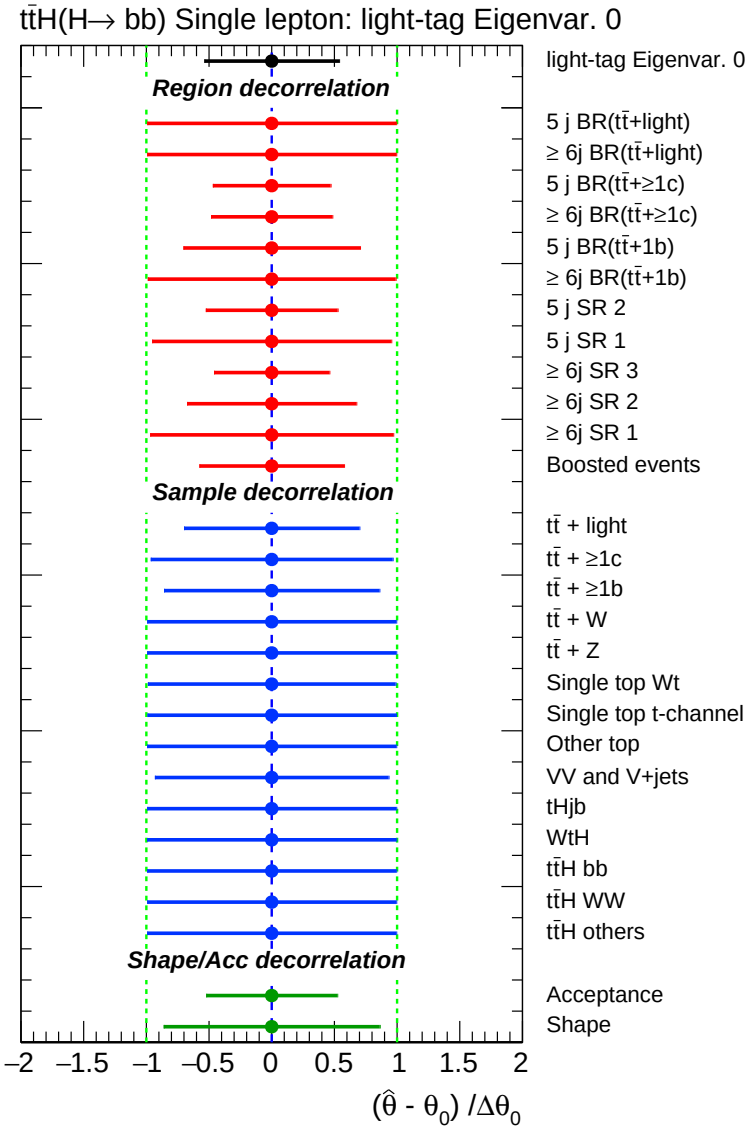


Figure 5.9.: Post-fit uncertainties on the various components of the first eigenvector in the decomposition of the uncertainty on *light*-jet efficiencies. They are obtained from four fits using different decorrelation schemes: (black) default fit with all components correlated, (red) uncorrelated per categories, (blue) uncorrelated per process, (green) uncorrelating the normalisation and shape components of the nuisance parameter.

5.4. Detailed study of the $t\bar{t}$+$\geq$1b NLO generator uncertainty

The $t\bar{t}$+$\geq$1b NLO generator systematic is the largest single source of uncertainty on the signal (see section 5.3.2). Particular attention is given to the high constraint on this systematic which can lead to an underestimation of the signal uncertainty.

Figure 5.10 shows the $t\bar{t}$+$\geq$1b NLO generator systematic uncertainty in the signal enriched categories. The lines represent the systematic uncertainty which is used in the fit, i.e. after root-smoothing (see section 5.1.2). The shape of this systematic uncertainty mimics the shape of the signal in the BDT output distribution leading to high impact on the signal sensitivity. However, the $t\bar{t}$+$\geq$1b NLO generator systematic shape is sensitive to statistical fluctuations in the two MC samples used to define the systematic. The full difference between the two MC before the smoothing procedure is shown by the points and presents several statistical fluctuations. These fluctuations are partially due to the negative weights of the Sherpa sample, which reduces the effective statistics. They are present despite the large effort to produce larger samples with $t\bar{t}$+$\geq$1b filters.

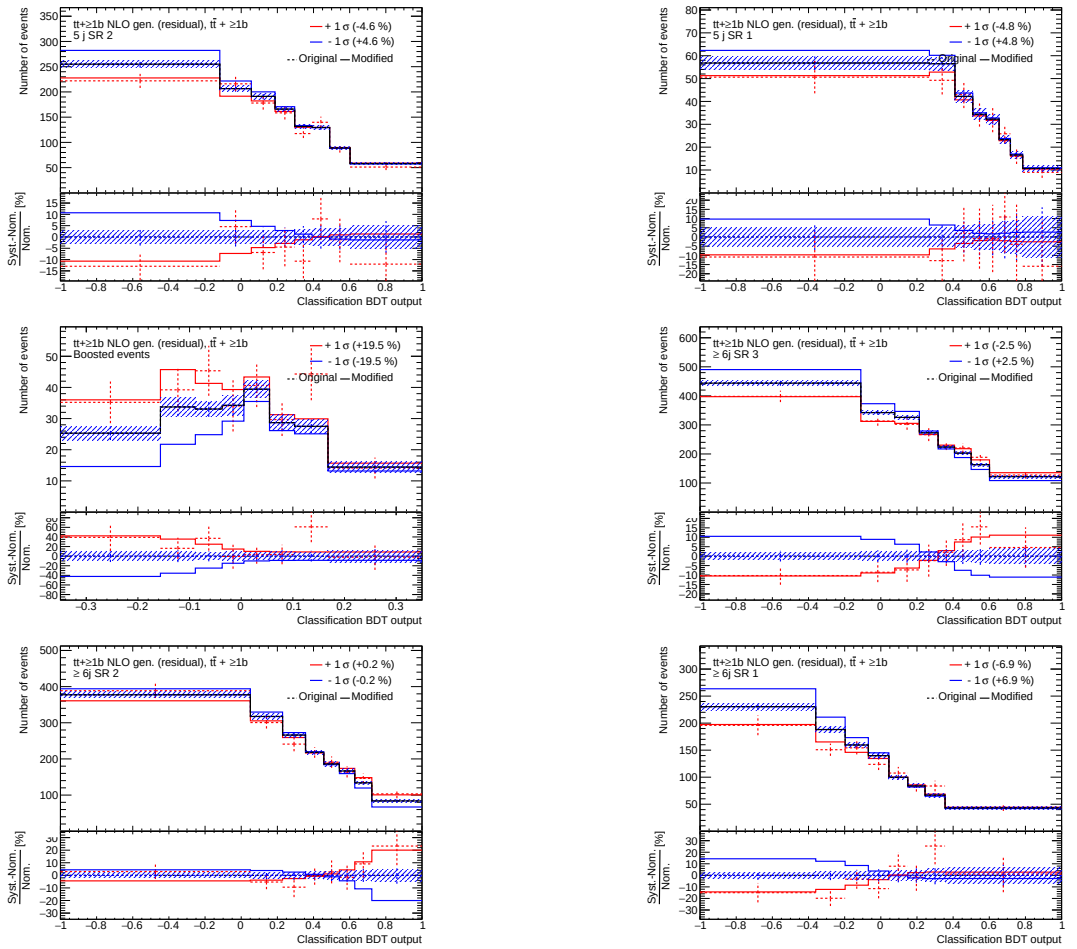


Figure 5.10.: $t\bar{t}$+$\geq$1b templates including the $\pm1\sigma$ variations induced by the $t\bar{t}$+$\geq$1b NLO generator uncertainty in the signal-enriched categories of the single lepton channel. Dashed points (solid lines) display the systematic uncertainty before(after) smoothing. The root-smoothing procedure is used in these plots. The black line represents the nominal prediction.

5.4.1. Study of the $t\bar{t}$+$\geq$1b NLO generator constraint

As for the first eigenvector of the $light$-jet efficiency uncertainty, the contribution from each category to the constraint on the $t\bar{t}$+$\geq$1b NLO generator systematic is estimated performing the fit with the $t\bar{t}$+$\geq$1b NLO generator systematic uncorrelated between all categories. The shape and acceptance effects are also separated. The constraint from the nominal fit procedure is compared to the constraints on each component of the $t\bar{t}$+$\geq$1b NLO generator uncertainty in figure 5.11.

Most of the $t\bar{t}$+$\geq$1b enriched categories constrain the differences between the POWHEG+PYTHIA8 and the Sherpa $t\bar{t}$+$\geq$1b samples. The highest reduction of the $t\bar{t}$+$\geq$1b NLO generator uncertainty comes from signal regions with at least 6-jets where the systematic uncertainty is large and constraints are expected. On the other hand both the shape and acceptance components of the $t\bar{t}$+$\geq$1b NLO generator systematic are constrained. The main contribution to the overall constraint comes from the shape of the uncertainty.

The current data-set can already differentiate the MC generators for the $t\bar{t}$+$\geq$1b sample. A dedicated $t\bar{t}$+$\geq$1b measurement would significantly improve the constraints on MC simulations, and improve the $t\bar{t}H(H \to b\bar{b})$ sensitivity.

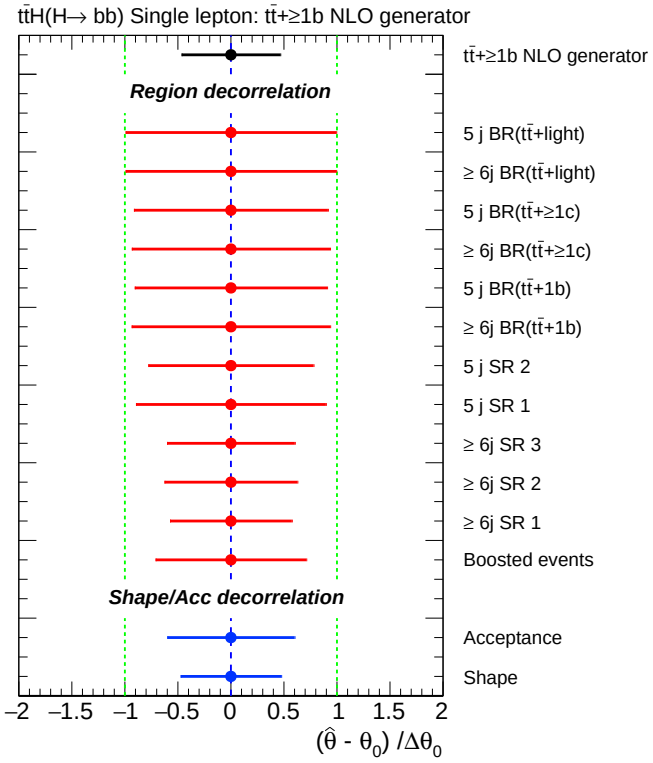


Figure 5.11.: Post-fit uncertainties on the various components of the $t\bar{t}$+$\geq$1b NLO generator systematic. They are obtained from the fits using different decorrelation schemes: (black) default fit with all components correlated, (red) uncorrelated per process, (blue) uncorrelating the normalisation and shape components of the nuisance parameter.

5.4.2. Statistical component of the $t\bar{t}$+$\geq$1b NLO generator uncertainty

The statistical component of systematics is handled by the smoothing procedure which average statistical fluctuations. However, in the case of systematics defined by the differences between two MC simulations with large statistical errors the smoothing can be sub-optimal or sensitive to statistical fluctuations. In particular, the $t\bar{t}$+$\geq$1b NLO generator systematic uncertainty has large statistical errors coming mainly from the Sherpa sample. Figure 5.12 shows a comparison of the $t\bar{t}$+$\geq$1b NLO generator systematic with the two different smoothing procedures (defined in section 5.1.2) in the $\geq$ 6j SR1 category. Both smoothing algorithms are compatible with the original prediction within statistical uncertainty. However the resulting systematic uncertainties are different, especially in this particular category where the root-smoothing predicts a large shape uncertainty while the main-smoothing predicts a flat uncertainty.

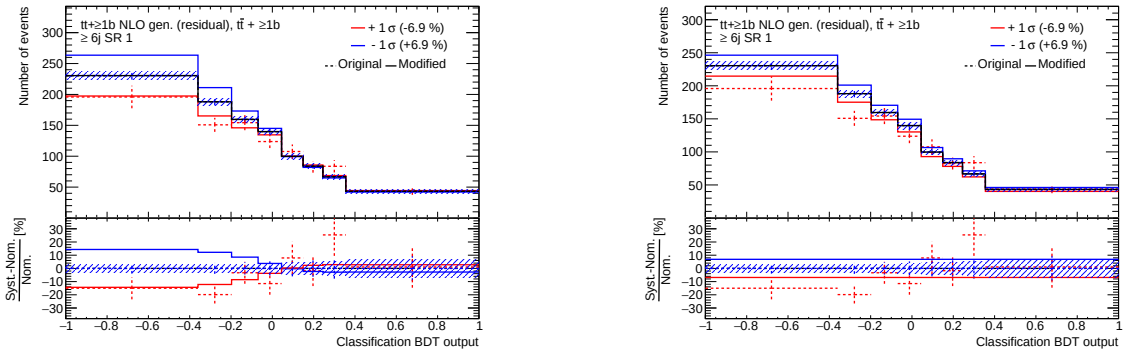


Figure 5.12.: $t\bar{t}$+$\geq$1b templates including the $\pm 1\sigma$ variations induced by the $t\bar{t}$+$\geq$1b NLO generator uncertainty in the signal-enriched categories of the single lepton channel. Colored points (solid lines) display the systematic uncertainty before(after) smoothing. The root-smoothing is used in the left figure and the main-smoothing in the right one. The black line represents the nominal prediction.

The statistical uncertainty on the systematics is not included in Maximum Likelihood fits. In order to estimate its impact toys on the Sherpa sample are used. Each toy corresponds to a fluctuation of the Sherpa distribution within its statistical error. The statistical error of the nominal prediction is lower than the one of Sherpa. Furthermore the statistical uncertainty of the nominal prediction is included in the γ nuisance parameters (see section 5.2.3). For these reasons only the Sherpa sample is considered for the toys. The Sherpa sample toys are produced with the Bootstrap method. For each event a set of 500 weights is picked with a Poisson probability around 1. The systematic uncertainty and the smoothing are re-evaluated for all toys. The distributions of the uncertainty on the signal strength (err(μ)), obtained for the 500 fits to the Asimov data-set, and for both smoothing procedures, are shown in figure 5.12.

For the main smoothing, the distribution of err(μ) is highly asymmetric and presents large tails towards higher err(μ) values. In fact, as shown in figure 5.12, the $t\bar{t}$+$\geq$1b NLO generator uncertainty shape is flattened by the main smoothing procedure. Thus the impact of this uncertainty is significantly reduced and is more likely to increase in the toys from potential addition of a shape component. This explains the asymmetric behavior in the toys and points to a potential bias from the main-smoothing leading to an underestimation of this uncertainty.

For the above reason, other smoothing procedures are considered. In particular, the root-smoothing shows a symmetric behavior in the toys. Moreover, the mean err(μ) value from the toys is 1σ lower compared to the baseline value. The real value of the $t\bar{t}$+$\geq$1b NLO generator systematic is not known

beyond the precision of the statistical uncertainty. However, the root-smoothing provides a conservative estimation of the $t\bar{t}$+$\geq$1b NLO generator systematic.

The 1σ variation of the uncertainty on err(μ) is 0.04 for the root-smoothing. This represents 6% of the total err(μ) and 10% of the contribution to err(μ) of the $t\bar{t}$+$\geq$1b NLO generator component alone.

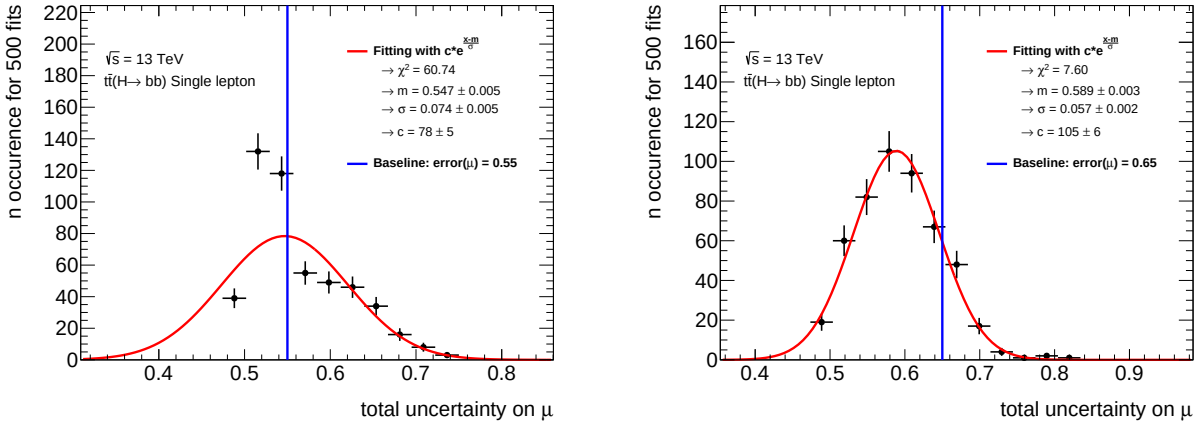


Figure 5.13.: Distribution of the signal strength uncertainties from the fits with the 500 toys of the $t\bar{t}$+$\geq$1b NLO generator systematic (see text). (left) Toys are ran using the root-smoothing. (right) Toys are ran using the main-smoothing. The distributions are fitted with a Gaussian distribution to extract the mean signal strength uncertainty and its 1σ variation from the statistical uncertainty on the Sherpa sample. The blue line displays the fitted signal strength uncertainty without applying toy weights (baseline).

The constraint on the $t\bar{t}$+$\geq$1b NLO generator uncertainty is also affected by the statistical uncertainty on the Sherpa sample as shown in figure 5.14. However in all toys the fit to the Asimov data-set is able to significantly constraint the differences between the Sherpa and the PP8 predictions.

Figure 5.15 displays the results of the 500 toy fits to data for the measured signal strength (μ). The root-smoothing is used. A large spread of $\Delta\mu = 0.21$, at one standard deviation, is observed. It represents about half the uncertainty on μ due to the impact of the $t\bar{t}$+$\geq$1b NLO generator nuisance parameter and 30% of the total uncertainty on μ. Even though a large effort is put to produce additional $t\bar{t}$+$\geq$1b samples with b-filters for both the PP8 and Sherpa predictions, the statistical uncertainty due to the Sherpa sample still impacts significantly the measurement of the signal. Further improvements, such as improved b-filters, are required for future iterations of the analysis.

The pull distribution of the $t\bar{t}$+$\geq$1b NLO generator uncertainty in the 500 toy fits is also shown in figure 5.15. Similarly to the signal strength, the pull is significantly affected by the statistical uncertainty on the Sherpa sample. A 1σ deviation of 0.19 is observed on the $t\bar{t}$+$\geq$1b NLO generator pull while the baseline fit measures 0.32 ± 0.45 for this nuisance parameter.

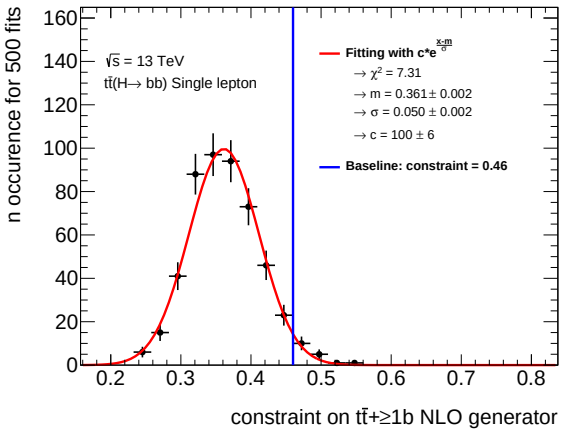


Figure 5.14.: Distribution of the post-fit uncertainty induced by $t\bar{t}$+≥1b NLO generator systematic after the fits to all toys using the root-smoothing (see text). The distribution is fitted with a Gaussian distribution to extract the mean constraint and its 1σ variation from the statistical uncertainty on the Sherpa sample. The blue line displays the fitted constraint without applying toy weights (baseline).

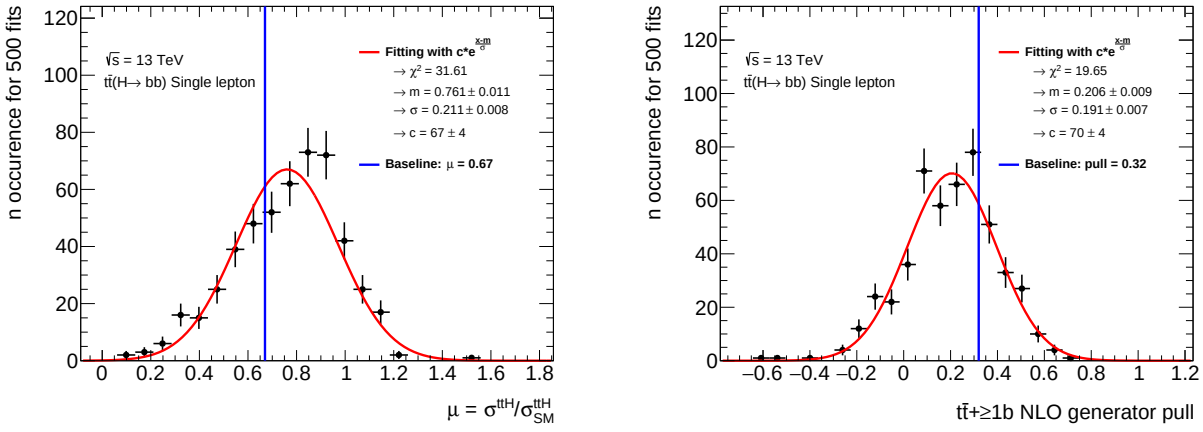


Figure 5.15.: (left) Distribution of the measured signal strength from the fits with the 500 toys of the $t\bar{t}$+≥1b NLO generator systematic (see text). (right) Distribution of the $t\bar{t}$+≥1b NLO generator uncertainty pull from the fits with the 500 toys of the $t\bar{t}$+≥1b NLO generator systematic (see text). Toys are ran using the root-smoothing. The distributions are fitted with a Gaussian distribution to extract the mean signal strength uncertainty and its 1σ variation from the statistical uncertainty on the Sherpa sample. The blue line displays the fitted signal strength uncertainty without applying toy weights (baseline).

5.5. Fits to pseudo-data

As discussed in the introduction, the analysis strategy and fit model are decided based on a blinded analysis. In order to verify the robustness of the $t\bar{t}+$jets model, fits to *pseudo-data* are performed. The pseudo-data are built from the nominal predictions of all processes but the $t\bar{t}+$ jets background for which the POWHEG+PYTHIA8 is replaced by the POWHEG+PYTHIA6 sample. A perfect fit would then use the $t\bar{t}+$ jets systematics to correct this change and leave the other nuisance parameters and the signal strength untouched. In particular, the free floating normalisations are expected to compensate for the differences in the fractions of $t\bar{t}+\geq 1c$ and $t\bar{t}+\geq 1b$ in the two samples. The expected truth k-factors are: $k(t\bar{t}+\geq 1b) = 1.03$ and $k(t\bar{t}+\geq 1c) = 0.87$.

Figure 5.17 and 5.18 shows the nuisance parameter pulls when fitting the two $t\bar{t}+\geq 1b$ models (described in section 5.2.2) to pseudo-data. The $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ fractions in pseudo-data are well identified by the fit. The default model observes $k(t\bar{t}+\geq 1b) = 0.98^{+0.13}{}_{-0.12}$ and $k(t\bar{t}+\geq 1c) = 0.70^{+0.32}{}_{-0.28}$ which are compatible with the truth k-factors. The same behavior is observed in the alternative $t\bar{t}+\geq 1b$ model where the normalisations of each of the $t\bar{t}+\geq 1b$ sub-components are not changed and the $t\bar{t}+\geq 1c$ contribution is reduced by 33%. Further tests are performed using various scalings of each of the $t\bar{t}+$heavy flavours components in pseudo-data and the fit is observed to always reproduce the scaled fractions.

The overall behavior of the nuisance parameters is close to what is expected and both models are almost identical. In particular significant pulls on the $t\bar{t}+\geq 1c$ parton shower and $t\bar{t}+$ light radiation systematics are observed. They convey the change from PYTHIA8 to PYTHIA6 in the pseudo-data as well as the increased $hdamp$ parameter in PYTHIA8. Pseudo-data fits also demonstrate that few experimental nuisance parameters (such as b-tagging and JER) can be used to correct for the $t\bar{t}+$ jets modelling, in particular b-tagging and jet related systematics. The same behavior is seen in the fits to data as explored in section 5.6.

The signal strength observed in pseudo-data is shown in figure 5.16. Both models are compatible between each other and also compatible with the truth value within the 1σ uncertainty. This enhances the confidence in the robustness of the signal extraction against the choice of the $t\bar{t}+$ jets model.

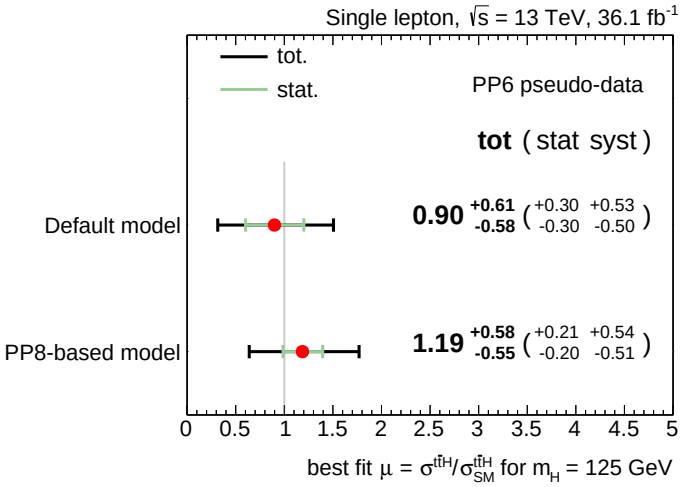


Figure 5.16.: Fitted value of the signal strength and its uncertainty from the fit with the two $t\bar{t}+\geq 1b$ models to pseudo-data in the single lepton channel.

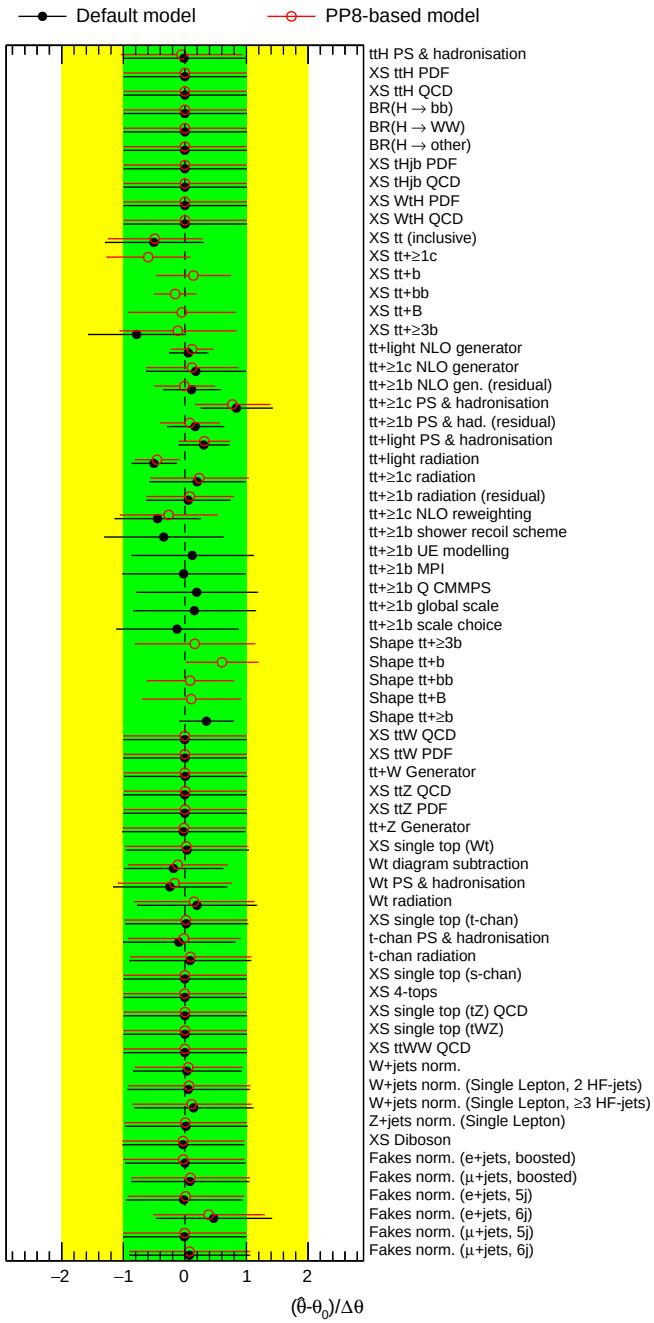


Figure 5.17.: Post-fit theoretical systematic uncertainties from the fit to pseudo-data based on the POWHEG+PYTHIA6 $t\bar{t}$ prediction. The black points display the values obtained from the default fit. The red points are obtained from the fit to data with the alternative $t\bar{t}+\geq 1b$ model. The green (yellow) area represent the $\pm 1(2)\sigma$ band on the pre-fit systematic uncertainty. The position of the black points and the size of their horizontal bars give the pulls and constraints in units of standard deviation, respectively.

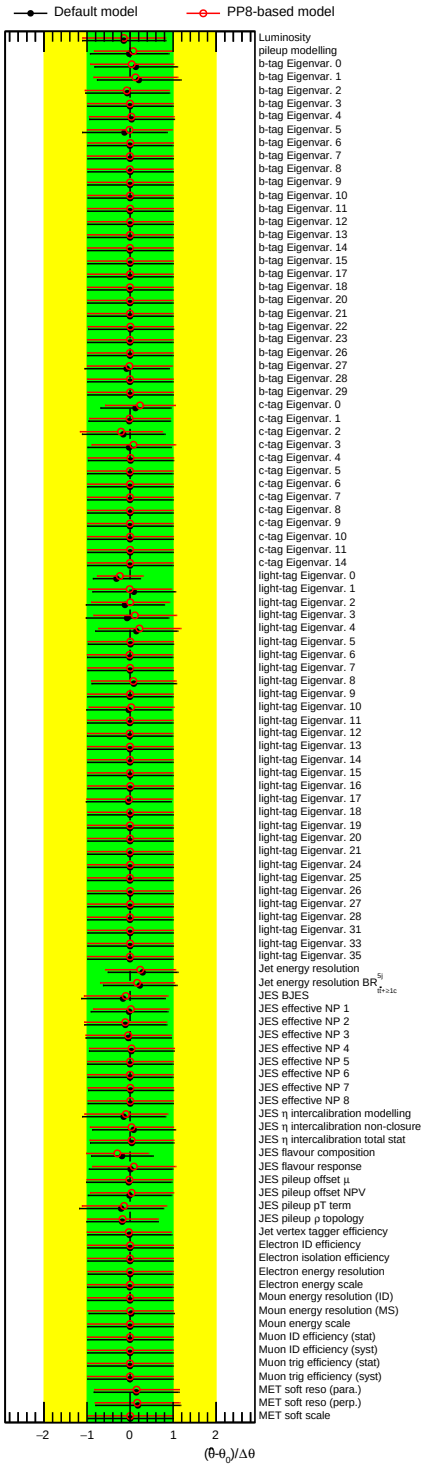


Figure 5.18.: Post-fit experimental systematic uncertainties from the fit to pseudo-data based on the POWHEG+PYTHIA6 $t\bar{t}$ prediction. The black points display the values obtained from the default fit. The red points are obtained from the fit to data with the alternative $t\bar{t}+\geq 1b$ model. The green (yellow) area represent the $\pm 1(2)\sigma$ band on the pre-fit systematic uncertainty. The position of the black points and the size of their horizontal bars give the pulls and constraints in units of standard deviation, respectively.

5.6. Data modelling

The $t\bar{t}H(H \to b\bar{b})$ fit to observed data is a complex procedure. In fact the mismodelling of the $t\bar{t}+$jets background requires many corrections from the systematic model. A great amount of attention is payed to the pulls of the data fit. In order to study the behavior of the fit to data in a blinded analysis, several fits to data are done under the background only hypothesis. In the first step, fits with the $H_{\mathrm{T}}^{\mathrm{had}}$ variable in all categories are performed in order to minimize the signal contribution and study the background modelling. In the second step, the background only fits are performed with the *blinded BDT*, i.e. using the final classification BDT discriminant but removing bins with $S/B \geq 6\%$[c], in signal-enriched categories. These studies are repeated with the final fit to confirm the blinded results. A selection of these studies is presented in this section.

5.6.1. Post-fit MC agreement with data for distribution used in the fit

Figure 5.19 shows the predicted number of events in each category after applying the corrections from the fit (post-fit) to data compared to the amount of observed data events. In all categories the data agrees with the corrected prediction. Some normalisation offset is still present in the boosted category. However the difference is well within the post-fit uncertainties.

Figure 5.20 show comparisons of the observed data and the prediction for the $H_{\mathrm{T}}^{\mathrm{had}}$ distribution in $t\bar{t}+\geq 1c$-enriched categories before applying the corrections from the fit (pre-fit) and post-fit. The fit manages to correct both the shape and the normalisation mismodelling in these categories. The uncertainty is also reduced due to the constraints on the nuisance parameters.

Similarly, figures 5.21 and 5.22 show comparisons of the observed data to the prediction of the classification BDT in signal enriched categories pre-fit and post-fit. The BDT output shape is relatively well modelled and the fit mainly corrects for the MC deficit normalisation in several of these categories.

Even though some statistical fluctuations in some bins show a 1 to 2σ deviations of data from predictions, the overall post-fit agreement is good. Indeed, no clear trend for shape mismodelling or normalisation offset are observed in post-fit distribution. Thus the simultaneous fit of all bins is able to capture and correct most of the mismodelling in the $t\bar{t}+$jets model.

[c] A 6% threshold has been chosen to remove any sensitivity to the signal while keeping enough bins to study the shape of the discriminating variables.

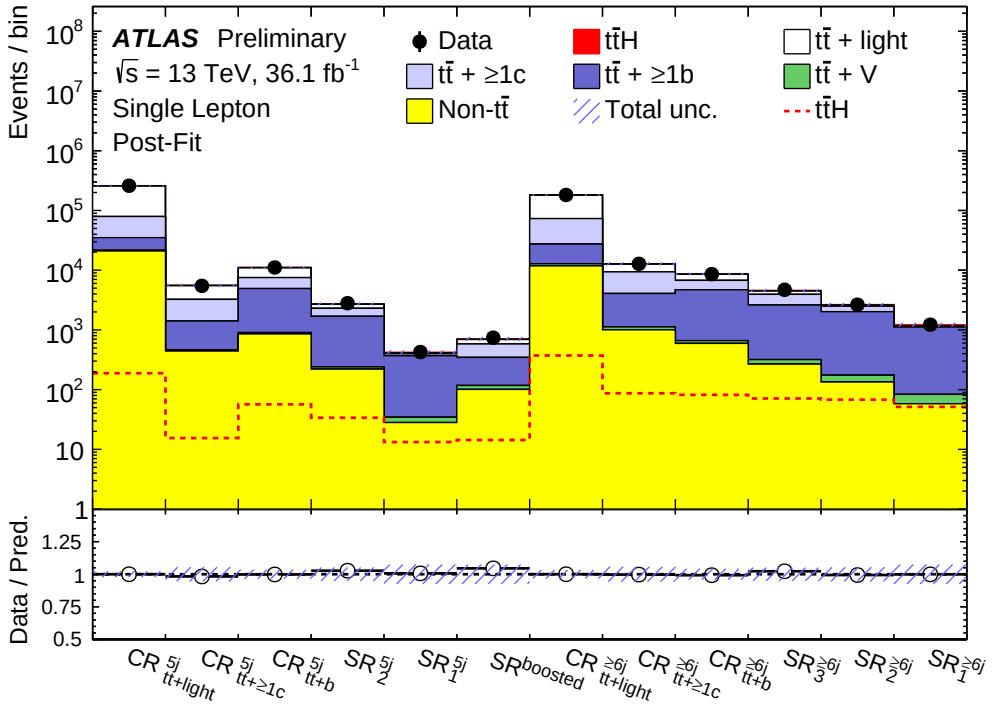


Figure 5.19.: Comparison of the predicted and observed yields in all categories of the single lepton channel after applying corrections from the fit to data. The signal contribution is shown both as a filled red area stacked on top of the backgrounds and as a separate dashed red line. The hashed band represent the statistical and systematic uncertainties. Uncertainties on the $t\bar{t}$+≥1b and $t\bar{t}$+≥1c background normalisations are not included as those are free floating parameters of the fit.

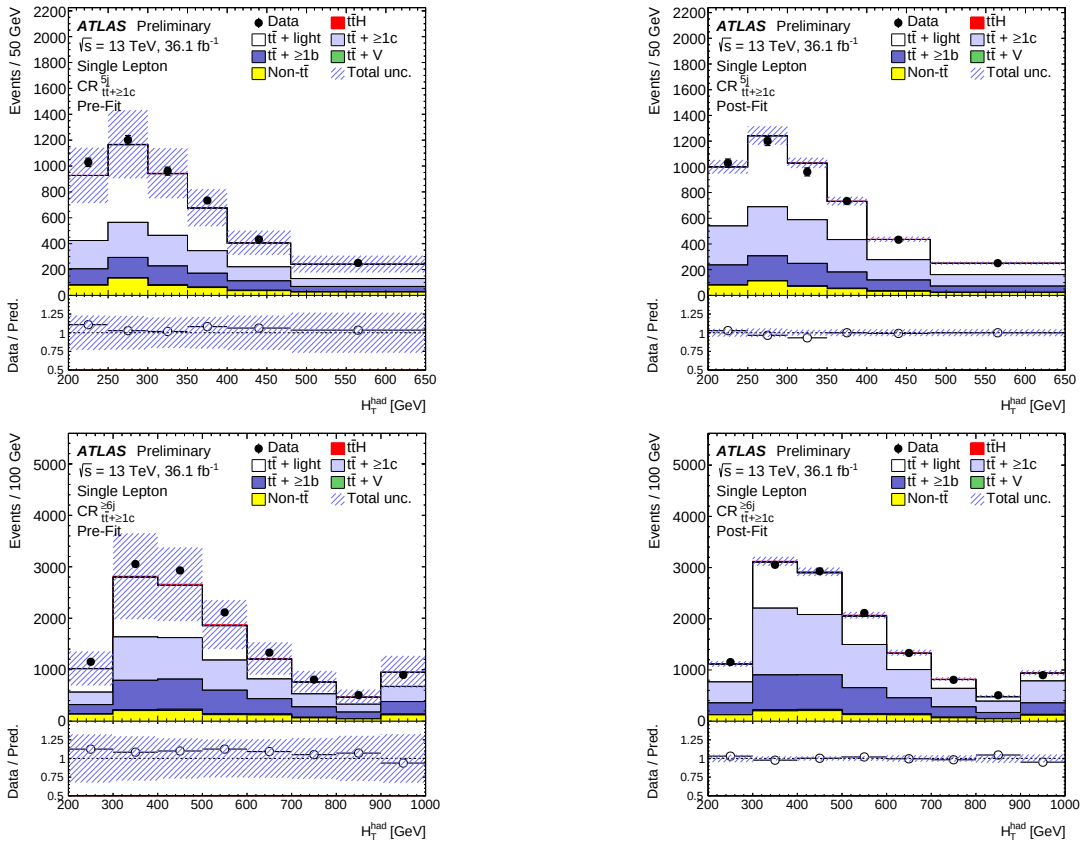


Figure 5.20.: Comparison of the predicted $H_{\mathrm{T}}^{\mathrm{had}}$ distribution to the one observed in data in the $t\bar{t}$+$\geq$1c-enriched categories of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

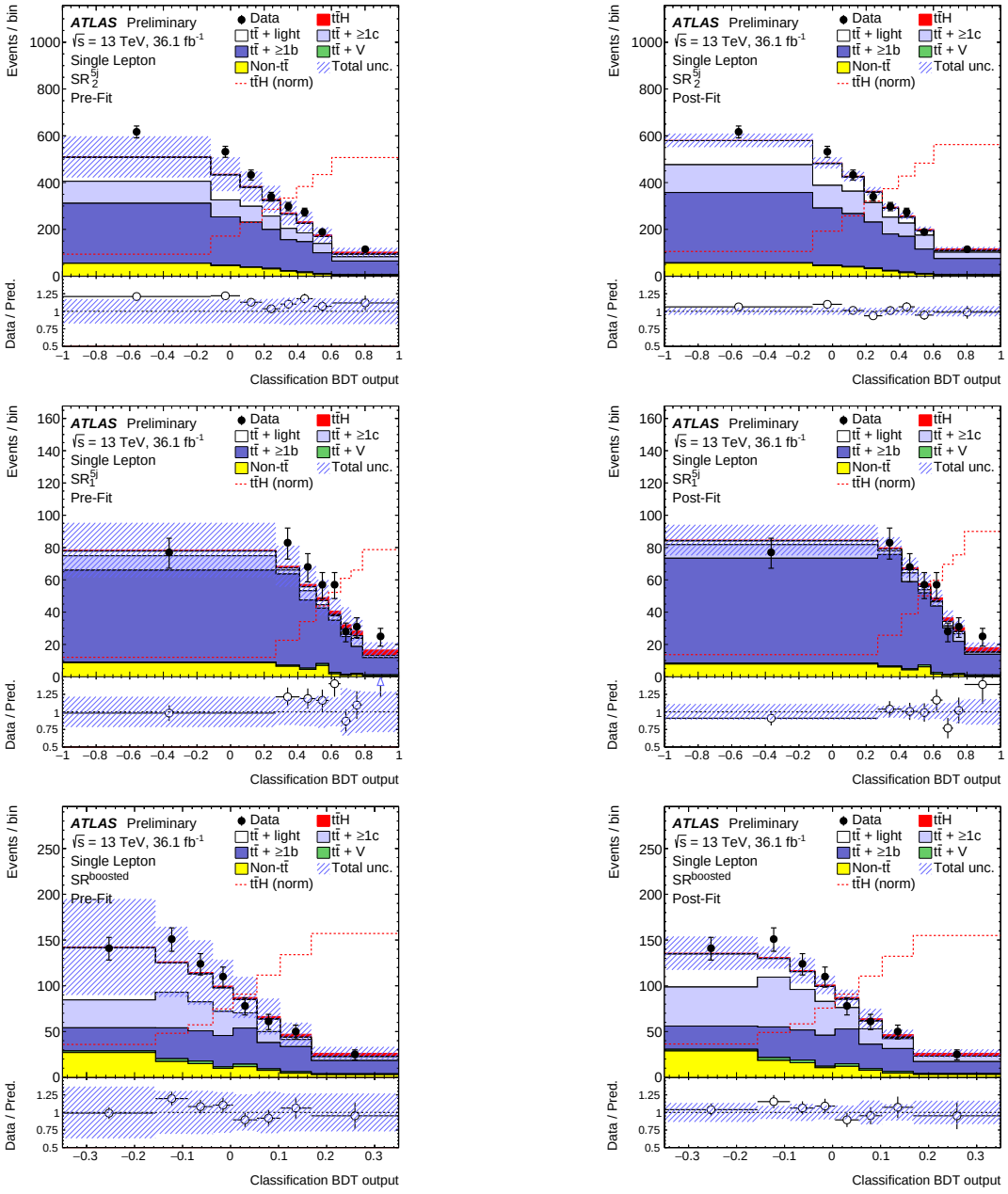


Figure 5.21.: Comparison of the predicted classification BDT distribution to the one observed in data in the $5j$ signal-enriched categories of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+≥1b and $t\bar{t}$+≥1c normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

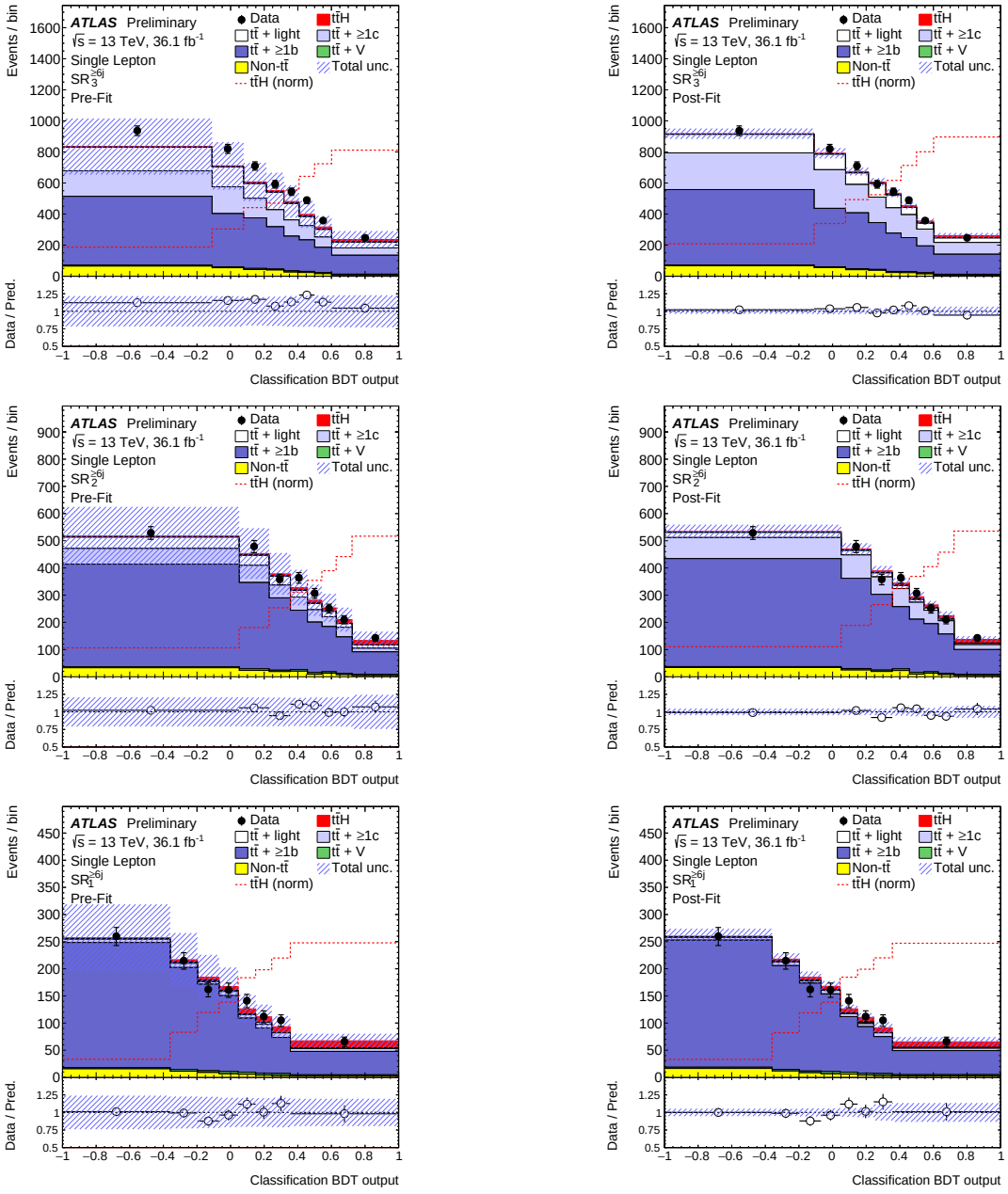


Figure 5.22.: Comparison of the predicted classification BDT distribution to the one observed in data in the $\geq 6j$ signal-enriched categories of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

5.6.2. Post-fit data MC agreement of variables not used in the fit

The classification BDT combines several input variables. A further check of the quality of the fit consists in analyzing the post-fit modelling of the input variables which are not directly used in the fit. A selection of these variables is presented here. In general, the fit to data manages to partially correct all distributions. In particular the normalisation differences between the prediction and observed data are corrected by the fit. The shape of the distribution are not expected to be fully corrected and residual mismodelling can be seen in some variables and categories. However the mismodelling is covered by the uncertainties and no significant deviation from data are observed in the corrected predictions.

The first variable shown in figures 5.23 and 5.24 is the mass of the Higgs candidate from the reconstruction BDT without Higgs. This variable is not well modelled in the $5j$ categories. Especially in the low mass range in the 5j SR1 category. In the $\geq 6j$ signal-enriched categories the reconstructed Higgs boson mass is reasonably well modelled, especially after applying the corrections from the fit.

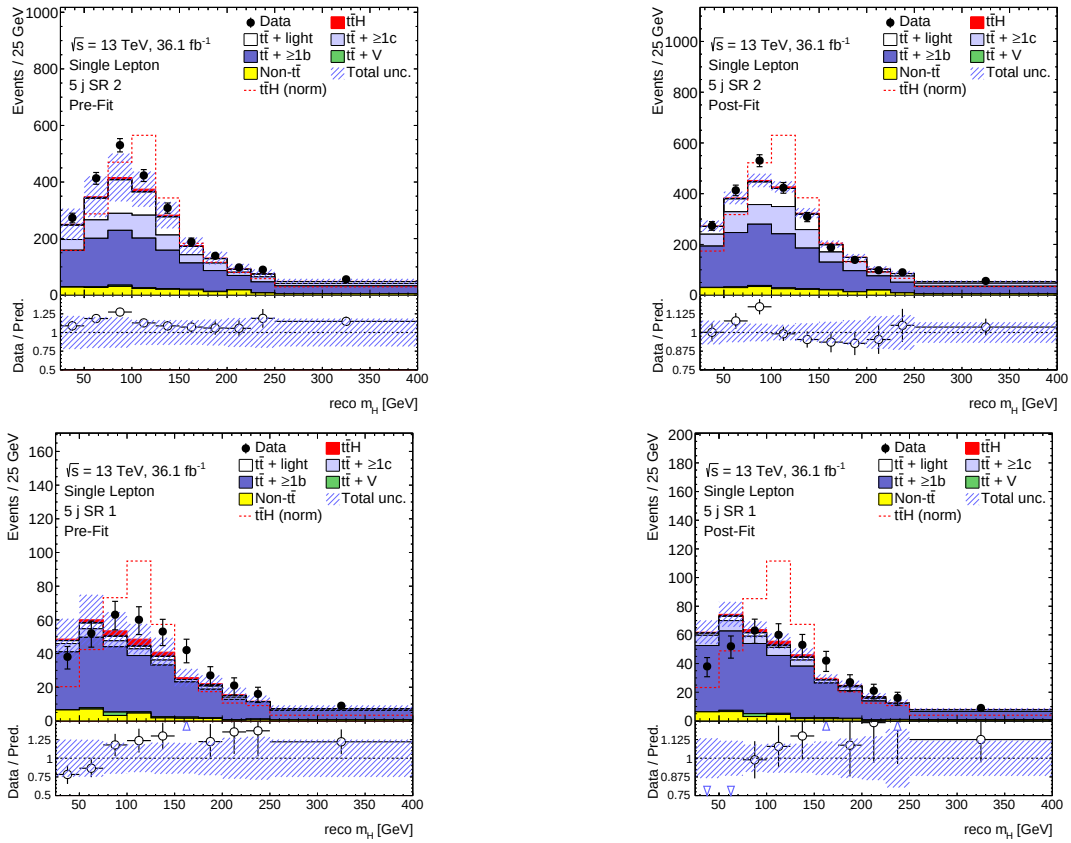


Figure 5.23.: Comparison of the predicted reconstructed Higgs boson mass distribution to the one observed in data in the $5j$ signal-enriched categories of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

Figures 5.25 and 5.26 display the reconstruction BDT with Higgs output distributions in the signal-

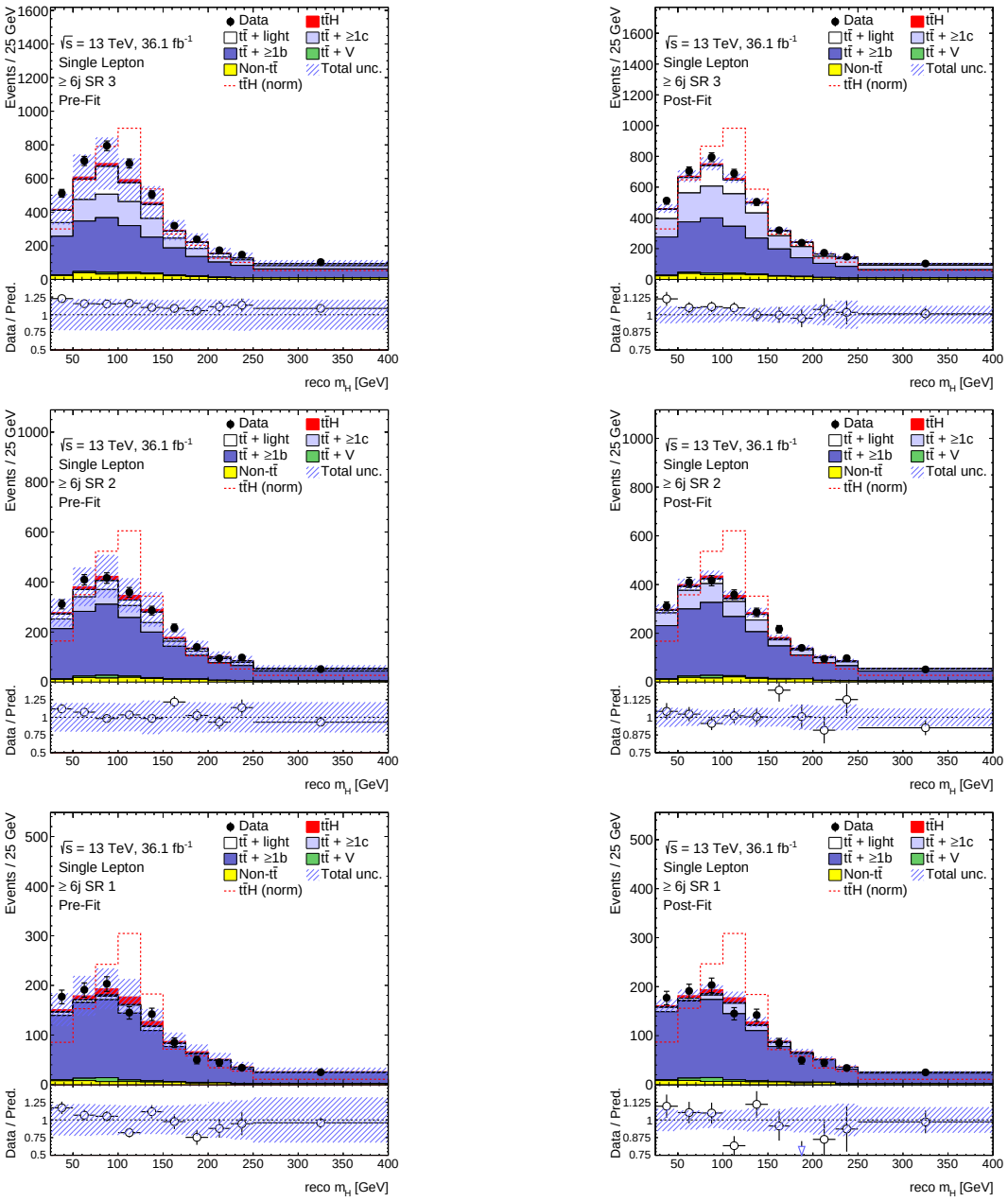


Figure 5.24.: Comparison of the predicted reconstructed Higgs boson mass distribution without to the one observed in data in the $\geq 6j$ signal-enriched categories of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

enriched categories. Some mismodelling is observed in a few categories at low and high BDT output. However in all categories the mismodelling effects are covered by the post-fit uncertainty.

Finally, the output of the Matrix Element method is shown in figure 5.27 and the Likelihood Discriminant is shown in figure 5.28 and 5.29. No significant shape mismodelling is observed in these variables both before and after applying the corrections from the fit. The fit only corrects the normalisation effects in some of the categories.

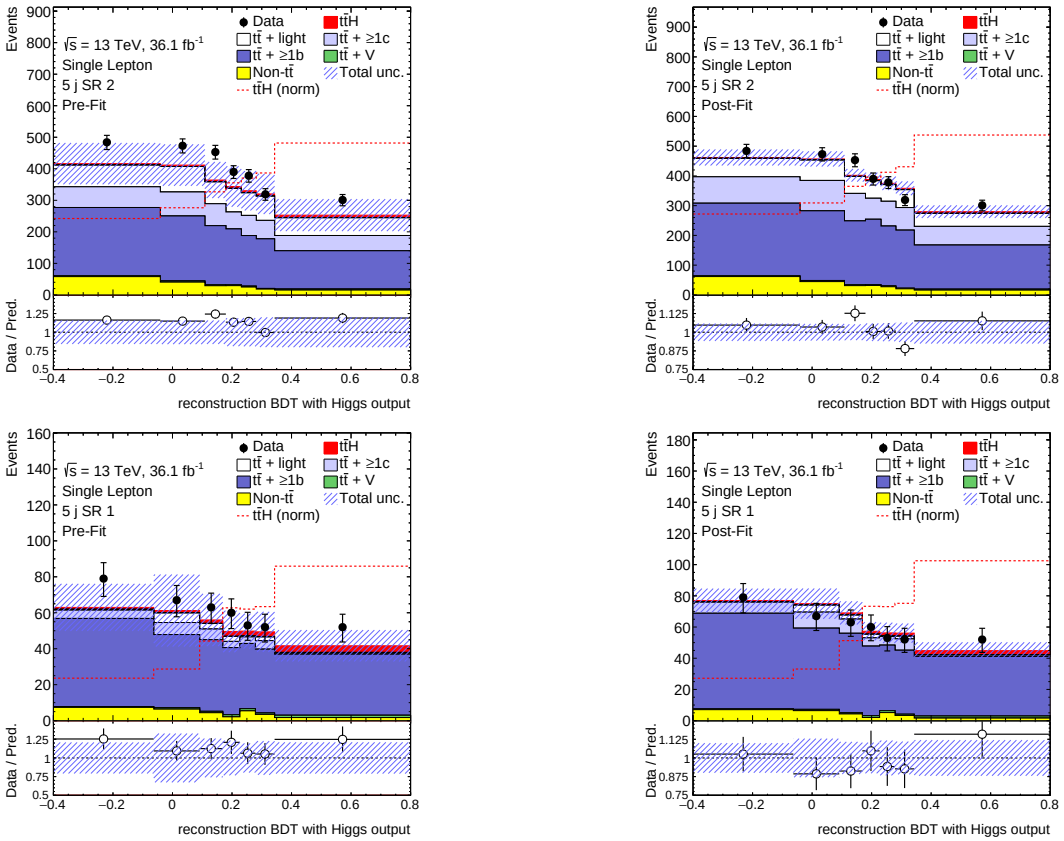


Figure 5.25.: Comparison of the predicted reconstruction BDT with Higgs output distribution to the one observed in data in the $5j$ signal-enriched categories of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

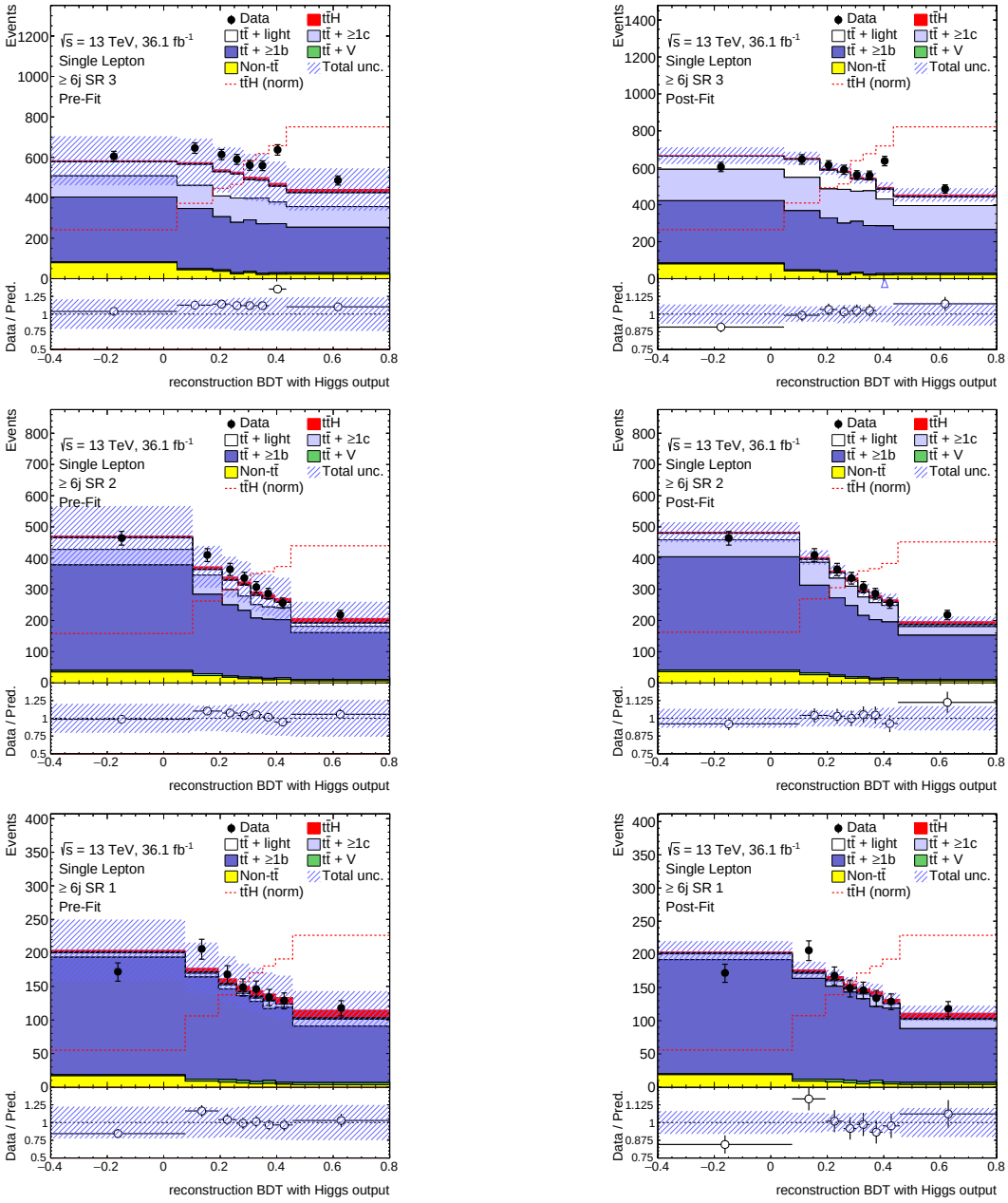


Figure 5.26.: Comparison of the predicted reconstruction BDT with Higgs output distribution to the one observed in data in the $\geq 6j$ signal-enriched categories of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

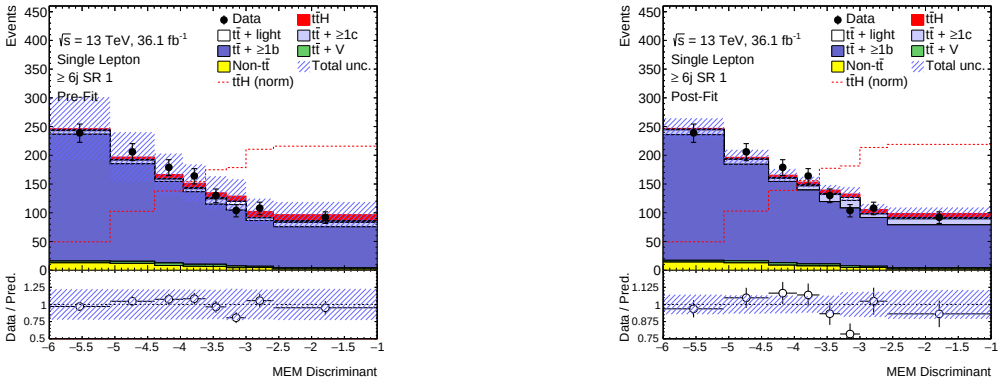


Figure 5.27.: Comparison of the predicted Matrix Element Method output distribution to the one observed in data in the $\geq$ 6j SR1 category of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

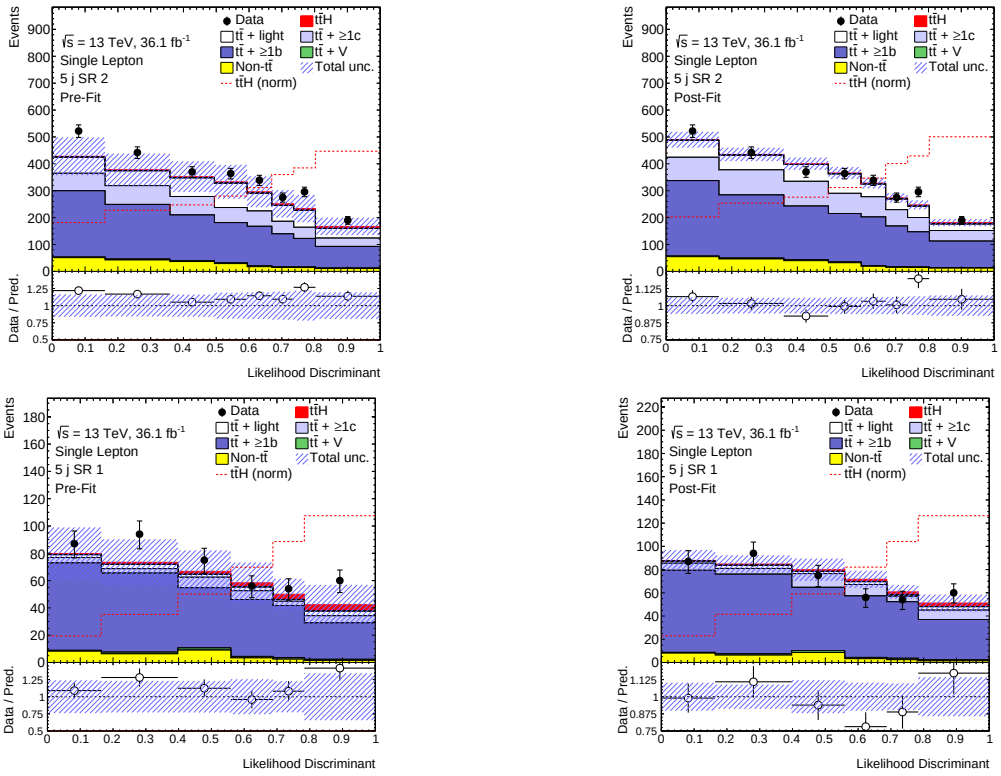


Figure 5.28.: Comparison of the predicted Likelihood Discriminant distribution to the one observed in data in the 5j signal-enriched categories of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

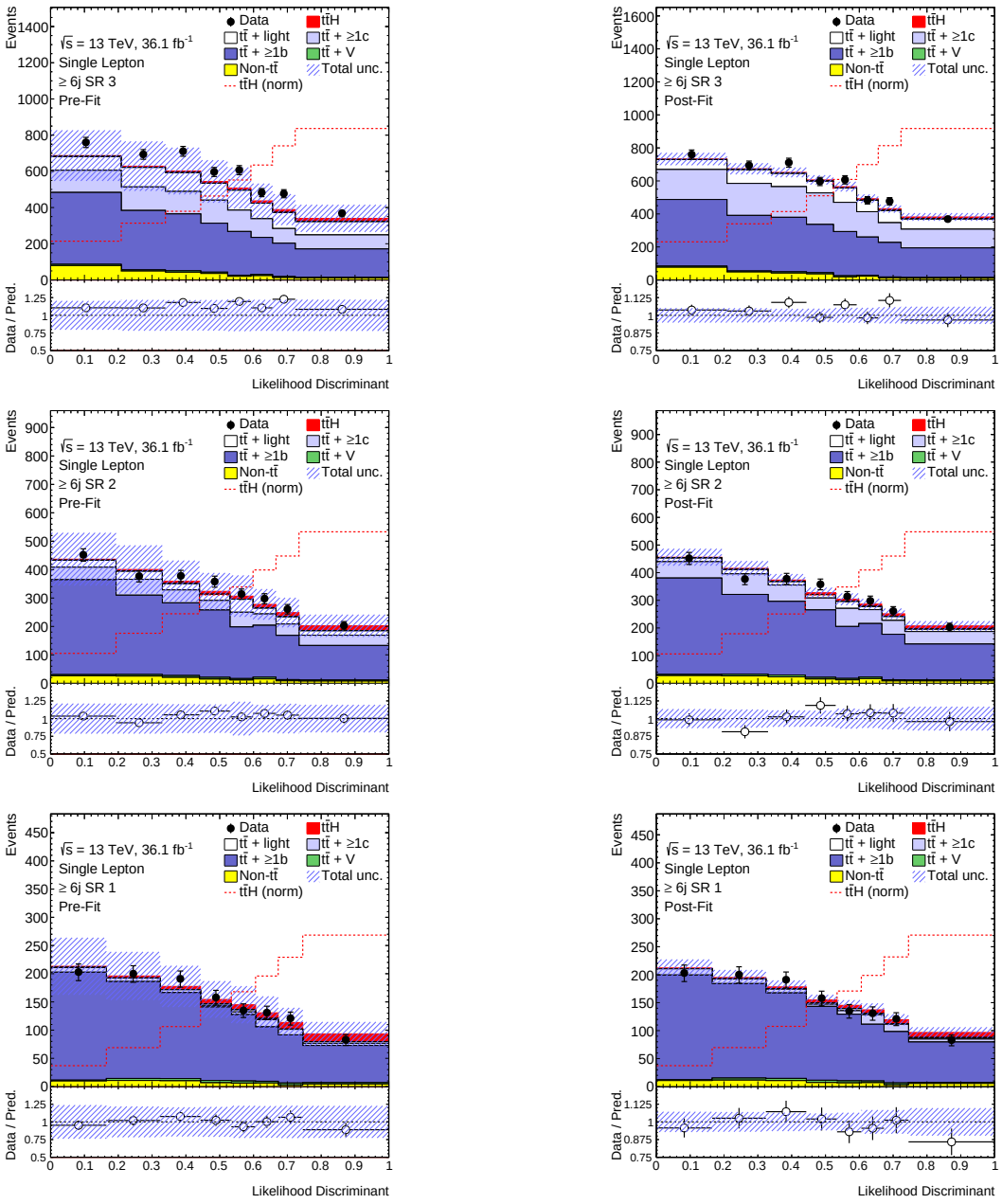


Figure 5.29.: Comparison of the predicted Likelihood Discriminant distribution to the one observed in data in the $\geq 6j$ signal-enriched categories of the single lepton channel. (left) before applying the corrections from the fit, (right) after applying the corrections from the fit. The hashed area represent the statistical and systematic uncertainties. The pre-fit uncertainties do not include the effect of the $t\bar{t}$+$\geq$1b and $t\bar{t}$+$\geq$1c normalisations which are free parameters of the fit. Note that the y-axis range in the lower pad (ratio plot) is twice smaller in the right plots compared to the ones on the right.

5.6.3. Post-fit systematic uncertainties

The modelling of data is a critical subject in the $t\bar{t}H(H \to b\bar{b})$ analysis. The pre-fit mismodelling is expected to pull several nuisance parameters which might bias the signal extraction in data. The use of a second model for the $t\bar{t}+\geq 1b$ background is an important input to estimate the stability of the data description against the most important background.

The best fit values of the theoretical and experimental nuisance parameters are shown in figure 5.30 and 5.31 for the single lepton channel. The normalisations of the $t\bar{t}+$ heavy flavours content in the default model are $k(t\bar{t}+\geq 1b) = 1.18^{+0.14}{}_{-0.13}$ and $k(t\bar{t}+\geq 1c) = 1.29^{+0.41}{}_{-0.34}$.

The two models give almost the same description of data and all common nuisance parameters are compatible. The high compatibility is mainly due to the absence of the shape mismodelling in the $t\bar{t}+b$-enriched categories.

In both models, the data mismodelling is mostly corrected by pulls of the nuisance parameters associated to $t\bar{t}+$ jets systematic uncertainties. Few nuisance parameters from the non-$t\bar{t}$ background uncertainties are also pulled. They mostly provide small shape corrections at low $H_{\mathrm{T}}^{\mathrm{had}}$ in background enriched categories and are not expected to affect the signal.

The pull on the $t\bar{t}+\geq 3b$ normalisation shows that the Sherpa+OpenLoops generator predicts a too high fraction for this category. In fact, the fit reduces the fraction of this component by about 30% when using the Sherpa+OpenLoops prediction while not correcting it when using the POWHEG+PYTHIA8 prediction. This motivated the addition of the 50% prior uncertainty on the normalisation of the $t\bar{t}+\geq 3b$ background in the default model.

The experimental nuisance parameters show very similar pulls for both models. Few pulls in detector uncertainties are also observed, especially related to b-tagging. The second eigenvector of the c-jet efficiency uncertainty and the jet energy resolution in 5j BR($t\bar{t}+\geq 1c$) are the experimental nuisance parameters with the largest pulls. Both these nuisance parameters are used to replace missing degrees of freedom of the $t\bar{t}+$jets background description. In particular, they mostly originate from categories where the $t\bar{t}+$ jets model is not sufficient to correct the $H_{\mathrm{T}}^{\mathrm{had}}$ modelling because of large contributions from the $t\bar{t}+$ light and $t\bar{t}+\geq 1c$ components. More details on these uncertainties are given in section 5.6.4.

Constraints on the data fit are very similar to the ones of the Asimov fit. Moreover they are compatible between the final fit and the blinded fits. This demonstrates that the constraints are mostly originating from the background categories, or the background like bins in the signal categories.

Nuisance parameters are included in the maximum likelihood fit as uncorrelated parameters. However, the fit creates correlations between complementary nuisance parameters. Figure 5.32 displays the linear correlation coefficients of nuisance parameters with at least one correlation above 30% in the default fit. Most of the $t\bar{t}+$ jets modelling uncertainties affect the normalisation and shape of several categories and large correlations are seen between the corresponding nuisance parameters.

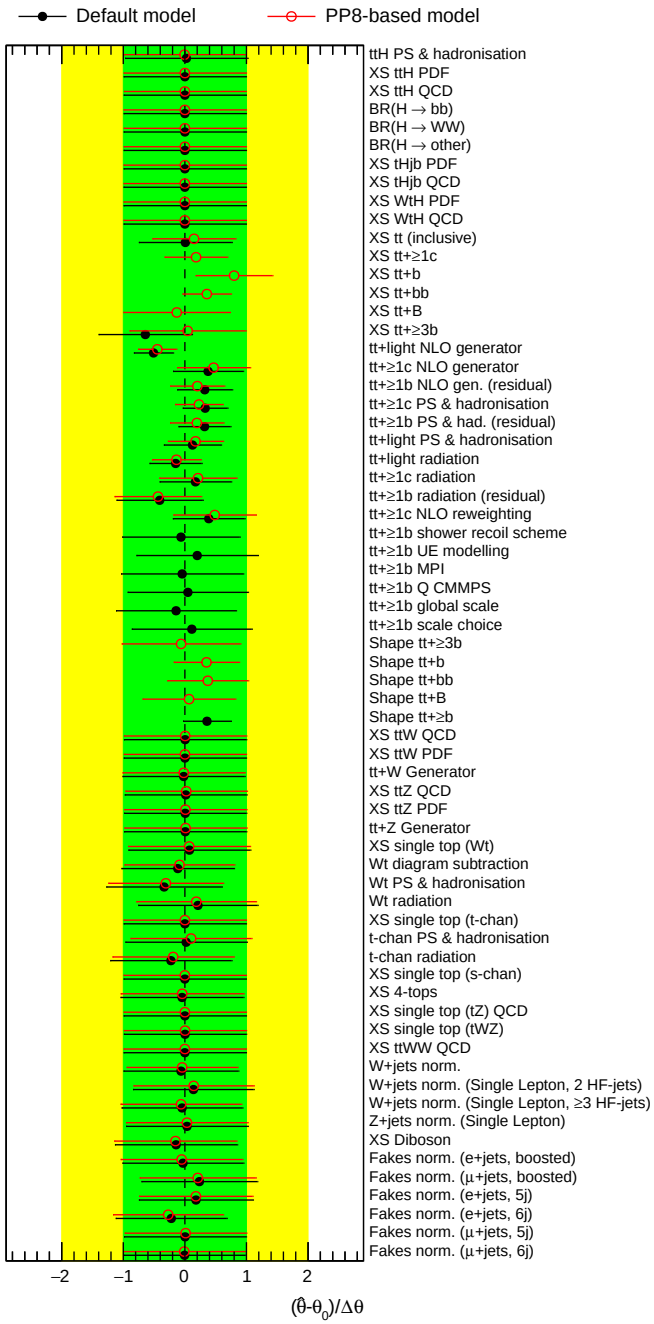


Figure 5.30.: Post-fit theoretical systematic uncertainties from the fit to data. The black points display the values obtained from the default fit. The red points are obtained from the fit to data with the alternative $t\bar{t}+\geq 1b$ model. The green (yellow) area represent the $\pm 1(2)\sigma$ band on the pre-fit systematic uncertainty. The position of the black points and the size of their horizontal bars give the pulls and constraints in units of standard deviation, respectively.

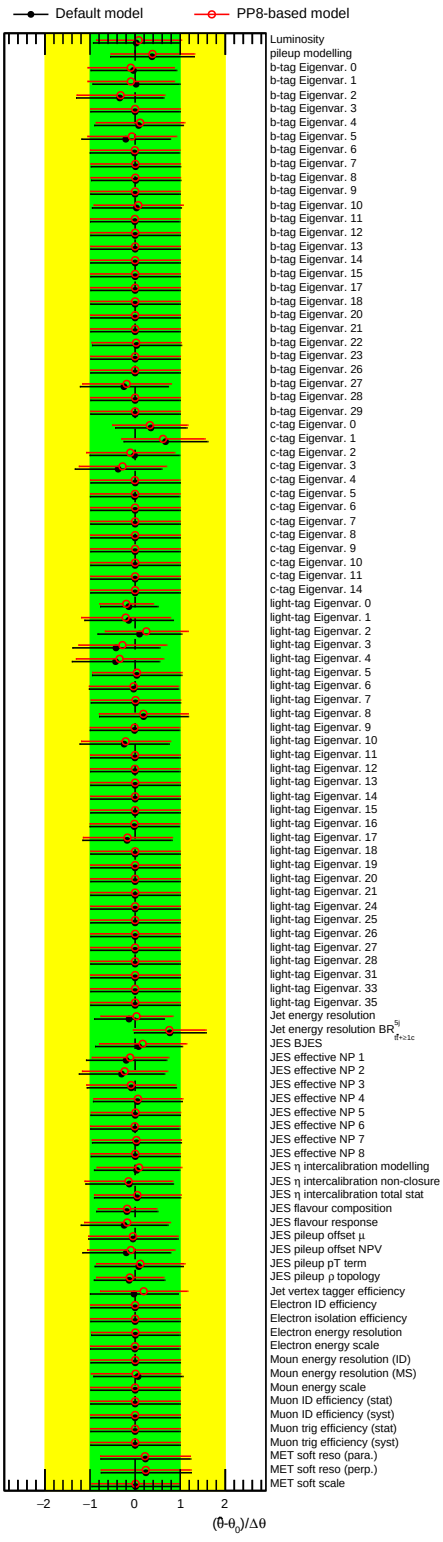


Figure 5.31.: Post-fit experimental systematic uncertainties from the fit to data. The black points display the values obtained from the default fit. The red points are obtained from the fit to data with the alternative $t\bar{t}+\geq 1b$ model. The green (yellow) area represent the $\pm 1(2)\sigma$ band on the pre-fit systematic uncertainty. The position of the black points and the size of their horizontal bars give the pulls and constraints in units of standard deviation, respectively.

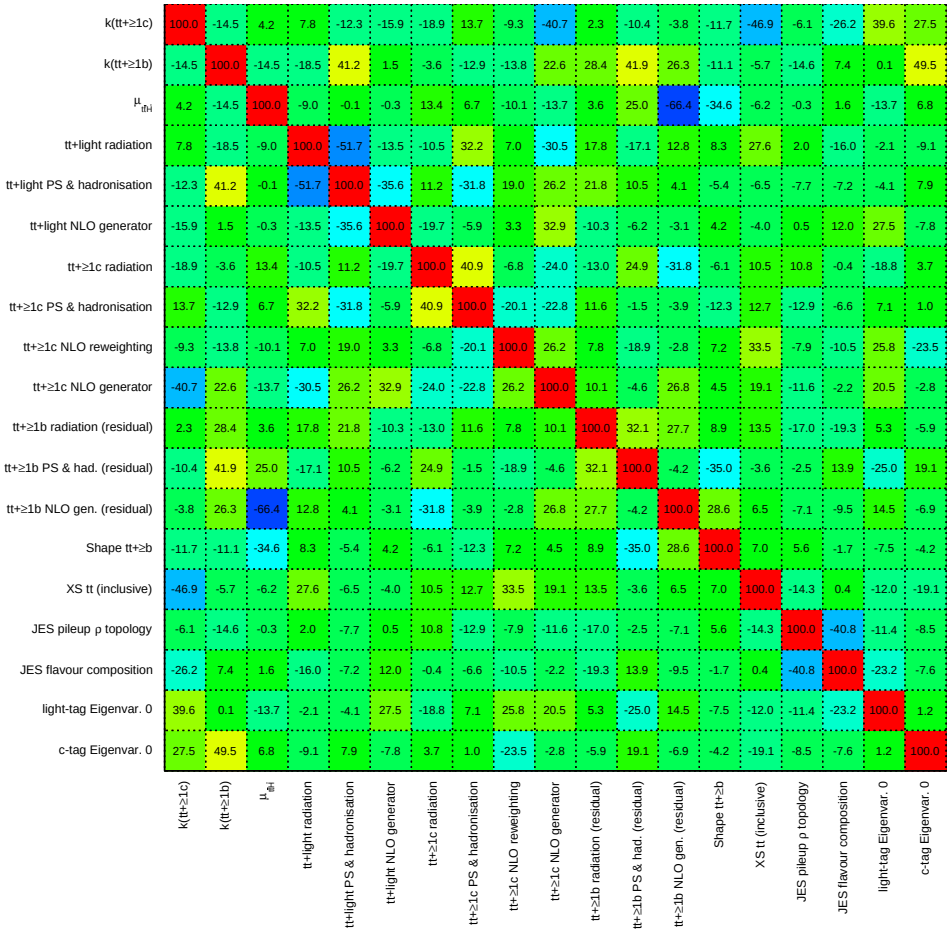


Figure 5.32.: Linear correlation coefficient in the fit to data between the signal strength, k-factors and nuisance parameters. Only nuisance parameters with at least one absolute correlation coefficient above 30% are shown.

5.6.4. Study of the important detector uncertainties

The jet energy resolution presents one of the highest pulls in the fit to data. As mentioned in section 5.2.4, the impact of this uncertainty is small on individual jets but is enhanced by the large number of jets. A selection of the most important components of the jet energy resolution uncertainty is shown in figure 5.33. It can correct the shape of the $t\bar{t}$ + jets background and the signal as well as their normalisations in several categories. In particular, it can correct the low $H_{\mathrm{T}}^{\mathrm{had}}$ slope observed in the 5j BR($t\bar{t}$+≥1c) category. The source of the pull on this uncertainty should be identified to verify that no pulls related to the mismodelling of a specific category are extrapolated towards the signal-enriched categories.

In order to identify the sources of pulls in this nuisance parameter the same study as for the $light$-jet efficiency eigenvector 0 (see section 5.3.3) is performed. The pulls for the different components of the jet energy resolution in the fits using the region, sample and shape/acceptance decorrelation schemes are displayed in figure 5.34(a). The most significant contribution to an upward pull is found in the 5j BR($t\bar{t}$+≥1c) category. The final pull is a combination of several effects. However the same test is

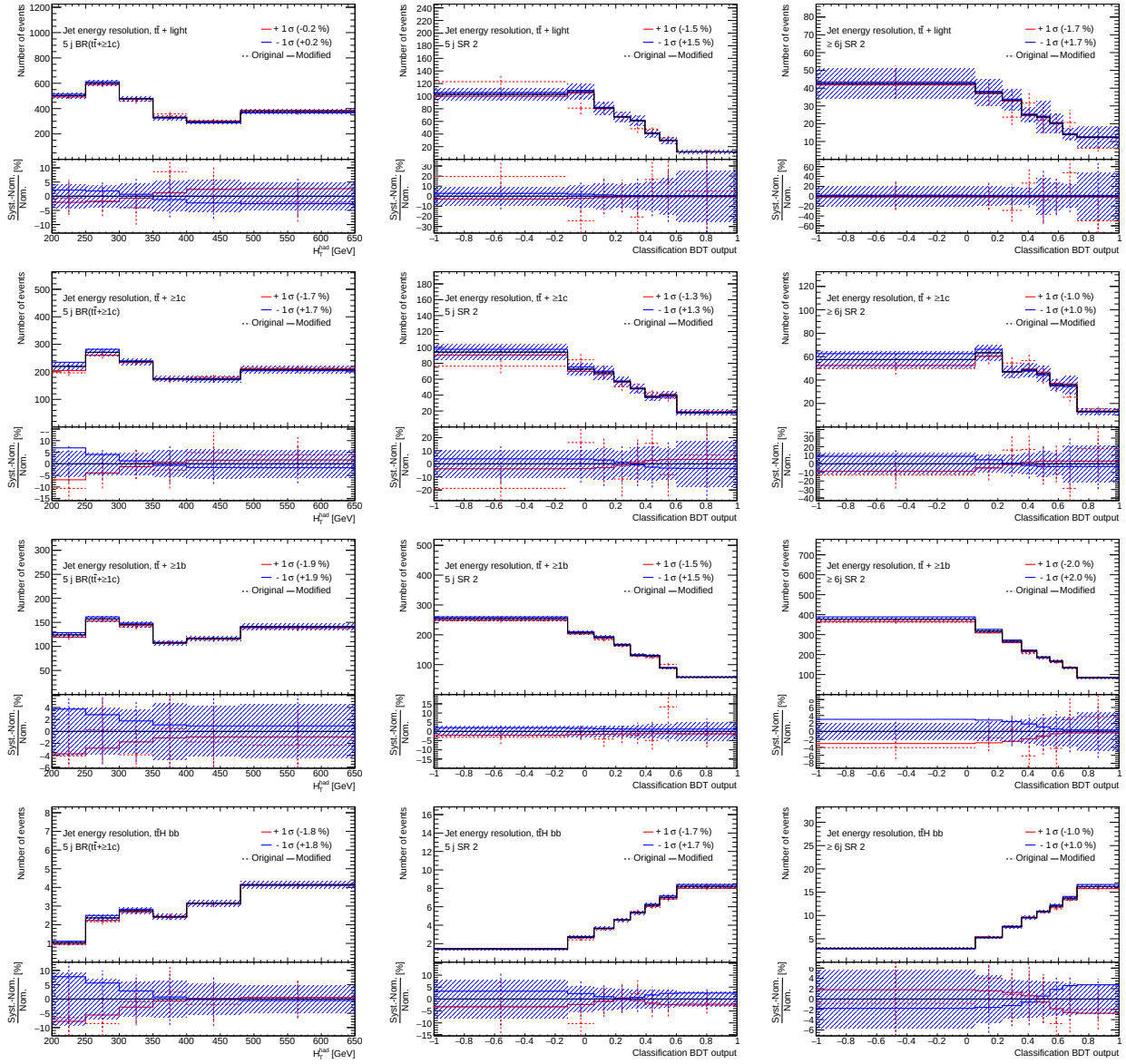


Figure 5.33.: Monte Carlo prediction of the $t\bar{t}+\text{light}$ (top), $t\bar{t}+\geq 1c$ (second row), $t\bar{t}+\geq 1b$ (third row) backgrounds and the $t\bar{t}H(H \to b\bar{b})$ signal (bottom) including the $\pm 1\sigma$ variations induced by the jet energy resolution uncertainty in the 5j BR($t\bar{t}+\geq 1c$) (left), 5j SR2 (middle) and $\geq$ 6j SR2 (right) categories. Colored points (solid lines) display the systematic uncertainty before(after) smoothing. The main-smoothing is used in these plots.

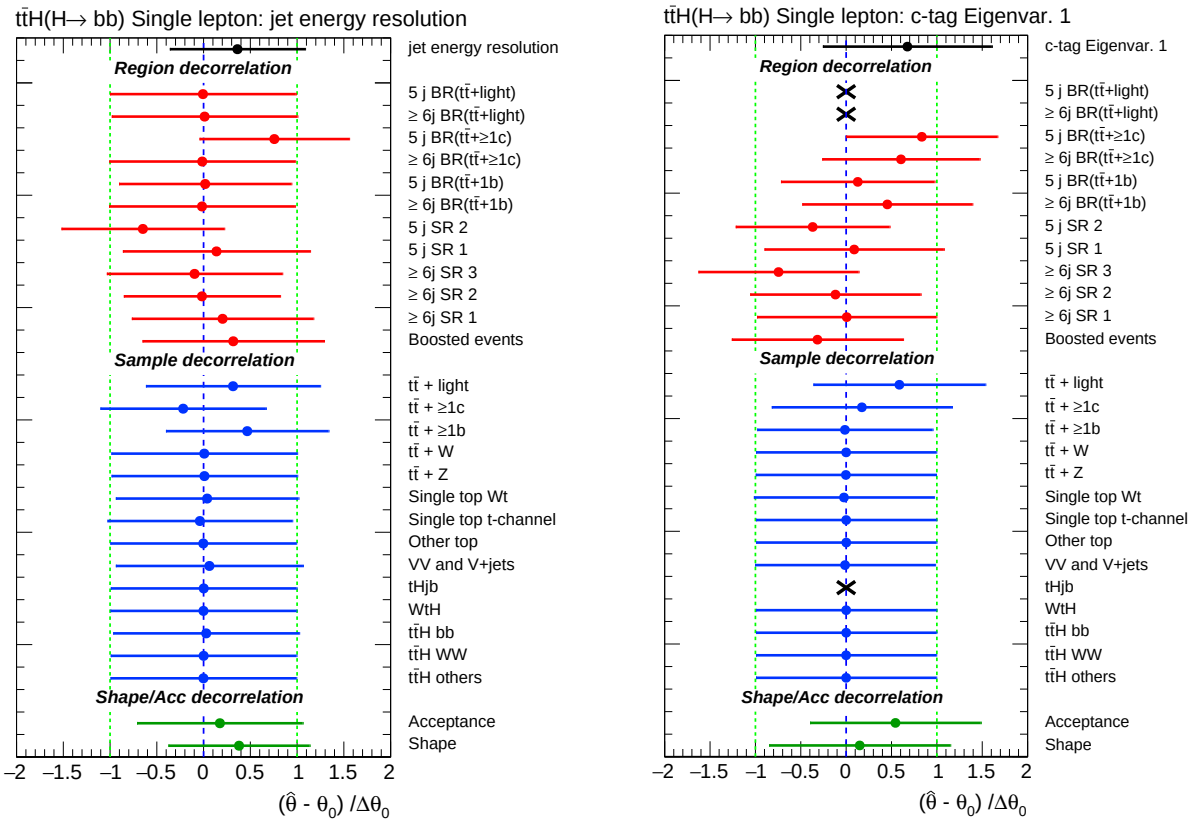
performed in the di-lepton channel and a similar behavior is observed in the category with similar background composition as the 5j BR($t\bar{t}$+$\geq$1c) category. This behavior is not compatible with a jet correction which should affect most of the categories and background components. Moreover the 5j BR($t\bar{t}$+$\geq$1c) category contains a large fraction of $t\bar{t}$+$\geq$1c and $t\bar{t}+$ light backgrounds for which the systematic model has less freedom compared to the $t\bar{t}$+$\geq$1b model. In consequence, the jet energy resolution is considered to be pulled to correct the $t\bar{t}+$ jets modelling in these specific categories. In order to avoid to propagate this correction to signal categories, the jet energy resolution is decomposed in two components, one affecting the 5j BR($t\bar{t}$+$\geq$1c) category (and the corresponding one for the di-lepton channel), and one affecting all others. This decision is based on the blinded analysis of data. The impact on the signal is estimated looking at the difference in μ without looking at the central value and is found to be negligible: $\Delta\mu < 0.02$ which is equivalent to 3% of the uncertainty on the signal strength.

The second eigenvector of the c-jet efficiency uncertainty decomposition (c-tag e.v. 1) is pulled by 0.68σ and is one of the highest pull in the $t\bar{t}H(H \to b\bar{b})$ analysis. However the c-jet efficiencies are measured in $t\bar{t}$ events. A large contribution to the c-jet efficiency uncertainties comes from the $t\bar{t}$ modelling in this dedicated measurement. Thus the c-tag e.v. 1 nuisance parameter can be used to correct the $t\bar{t}$ modelling.

The uncertainty associated to the c-tag e.v. 1 nuisance parameter is shown in figure 5.35. This nuisance parameter provides mainly inter-category normalisation corrections. The decorrelation tests of this nuisance parameter are shown in figure 5.34(b). An upward pull is found to originate mostly from the $t\bar{t}+$ light and $t\bar{t}$+$\geq$1c background in the $t\bar{t}$+$\geq$1c and $t\bar{t}+1b$ enriched categories. These components correspond to the ones that are the most likely to provide mistagged c-jets. In fact, they correspond to categories with several b-tagged-jets at a loose working point and to backgrounds with the final state dominated by $light$-jets[d] and c-jets.

The high upward pull in C1 is not observed in most of the 14 components of the c-jet efficiency uncertainty decomposition. No clear pattern arises from this decomposition. However, all observed pulls are well compatible with the nominal value provided for the c-jet tagging efficiency. Furthermore the c-tag e.v. 1 nuisance parameter has a limited impact on sensitivity and its pull is not expected to bias the signal extraction. The difference in the measured signal strength between the default configuration and the category decomposition is $\Delta\mu = 0.16$. This shift is non-negligible but is small compared to the total uncertainty of the signal strength ($^{+0.71}_{-0.69}$). The other two decomposition tests give a measured signal strength compatible with the default configuration: $\Delta\mu < 0.01$. In the end, no further treatment is required and the default configuration is kept.

[d] About 50% of the $t\bar{t}+$ light events have a c-jet from the hadronic decay of the W-boson.



(a) (b)

Figure 5.34.: Post-fit values on the uncorrelated components of the jet energy resolution (left) and the second eigenvector in the decomposition of the uncertainty on c-jet efficiencies (right). They are obtained from four fits using different decorrelation schemes: (black) default fit with all components correlated, (red) uncorrelated effects per categories, (blue) uncorrelated effects per process, (green) uncorrelating the normalisation and shape effects of the nuisance parameter.

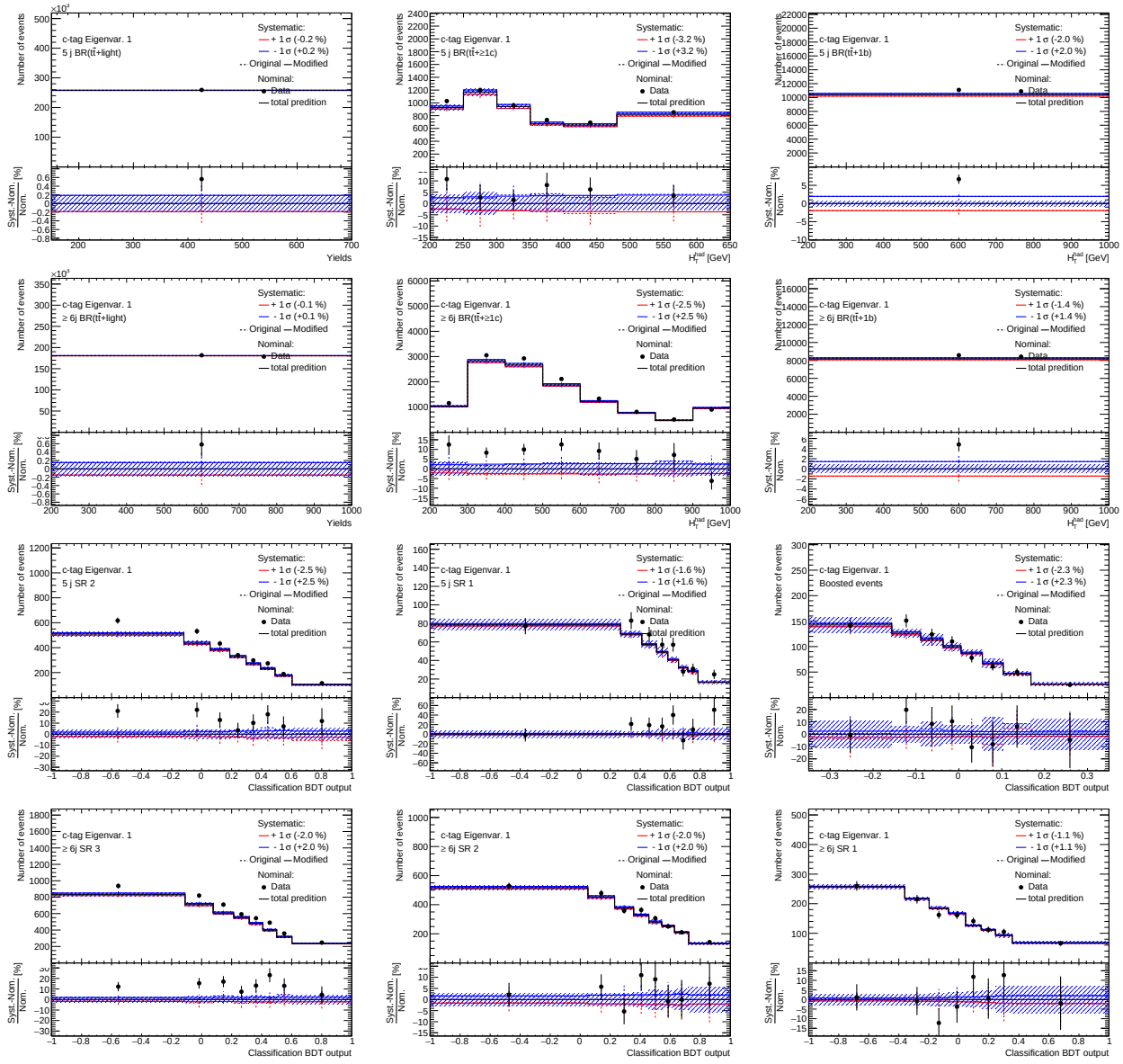


Figure 5.35.: Total prediction of all signal and background samples including the $\pm 1\sigma$ variations induced by the l-tag e.v. 0 uncertainty in all categories compared to data. Colored points (solid lines) displays the systematic uncertainty before(after) smoothing. The main-smoothing is used in these plots. The black points represent data and the black solid line represent the nominal prediction. The lower pad displays the relative systematic uncertainty in percent. This relative uncertainty is compared to the relative difference between the nominal prediction and data (black points).

5.7. Fit results

The fitted signal strength in the single lepton channel of the $t\bar{t}H(H \to b\bar{b})$ analysis is $0.67^{+0.71}_{-0.69}$ for the default model and $1.02^{+0.67}_{-0.65}$ for the alternative $t\bar{t}+\geq 1b$ model (PP8-based model) as shown in figure 5.36. Both are compatible with the standard model expectation within one standard deviation, even though the default model measures a smaller value compared to the alternative model. In both models the uncertainty on the signal strength is largely dominated by the the systematic uncertainties.

The break-down of the uncertainty on the signal strength is shown in figure 5.37 for the 20 most important nuisance parameters after the fit to data. The impact of the individual nuisance parameters is very similar to one observed in the fit to the Asimov data-set (see figure 5.7). The main difference is the increased contribution of the $t\bar{t}+\geq 1c$ nuisance parameters which are enhanced by the increased amount of the $t\bar{t}+\geq 1c$ background ($+29\%$) due to the pull in its normalisation. It is also seen that the systematic uncertainties which have a large impact on the signal are not highly pulled by the fit and do not bias the signal measurement.

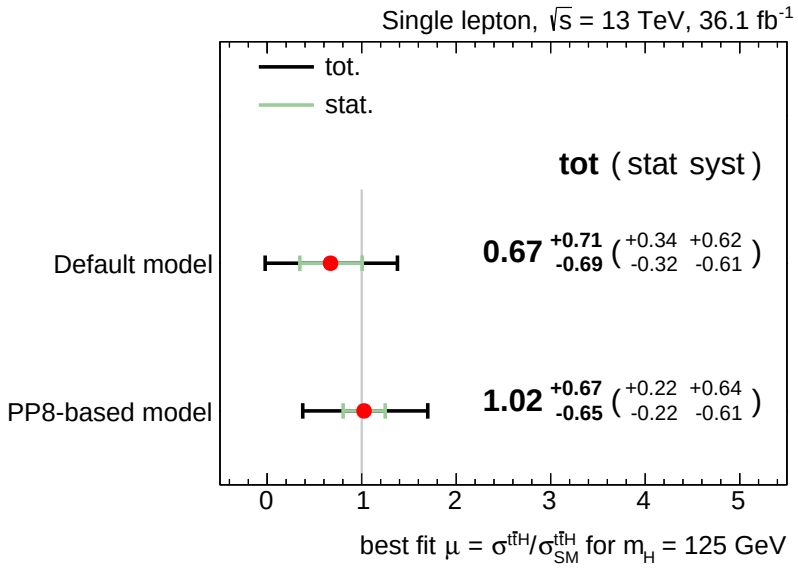


Figure 5.36.: Fitted value of the signal strength and its uncertainty from the fit of the two $t\bar{t}+\geq 1b$ models to data in the single lepton channel.

The di-lepton channel is complementary to the single lepton one since it has different event selections (two leptons and less jets). The systematic models of both channels are almost identical. Only few systematics on non-$t\bar{t}$ backgrounds affects one channel and not the other. In the di-lepton channel a signal strength $\mu = 0.11^{+1.36}_{-1.41}$ is observed which is compatible with the single lepton channel measurement within uncertainties.

The combination of the single lepton channel with the di-lepton channel is expected to improve the signal sensitivity. Common uncertainties between the two channels are treated as fully correlated allowing better constraints, especially for the $t\bar{t}+\geq 1b$ backgrounds. Observed signal strengths after the combined fit are shown in figure 5.38. In this figure, the $t\bar{t}H(H \to b\bar{b})$ production rate in units of SM prediction for each channel is quoted after the so called "two$-\mu$" fit to data. The "two$-\mu$" fit is a combined fit where nuisance parameters and k-factors are correlated between the two channels but considering the signal strengths in the single and di-lepton channels as two separated normalisation factors. After the combined fit the signal strengths of both channels are different to the ones

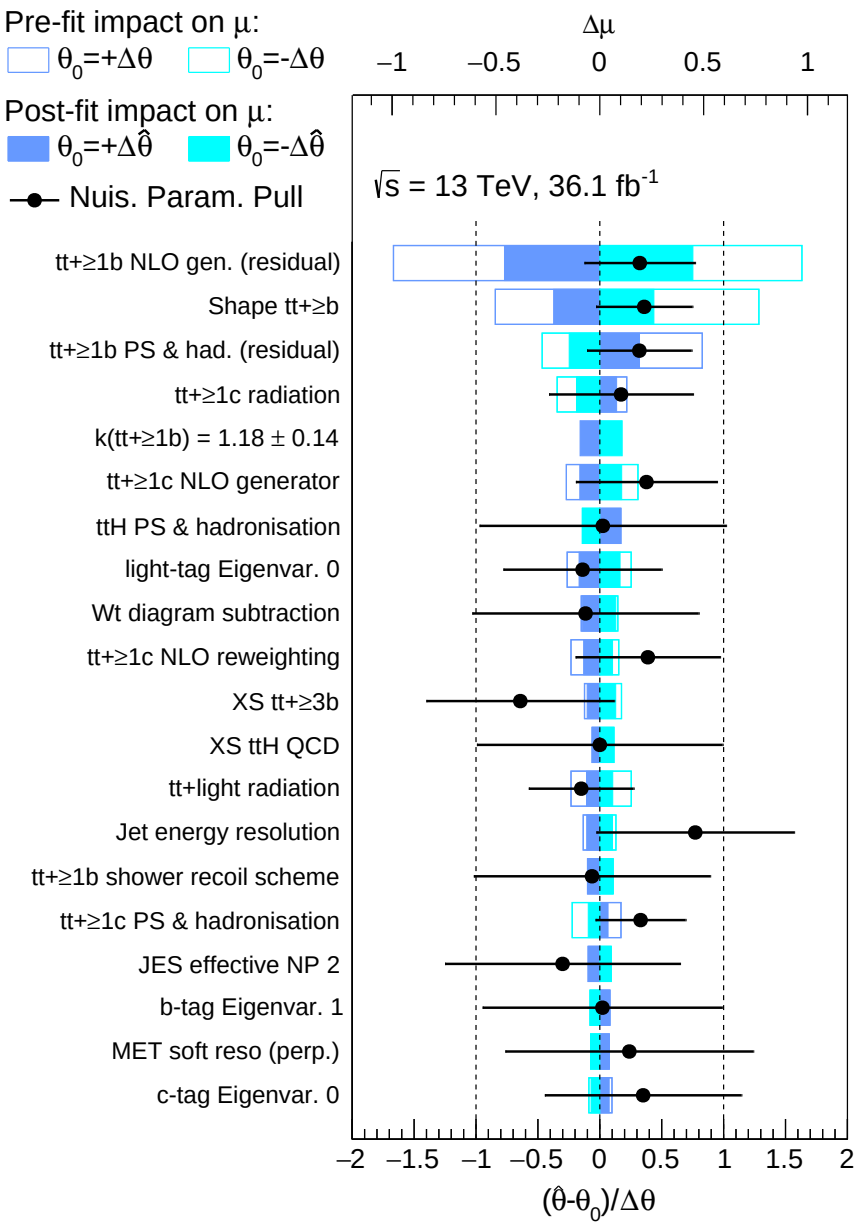


Figure 5.37.: Ranking of the nuisance parameters used in the fit according to their impact on sensitivity. Only the 20 first ones are shown. The filled (open) blue rectangles correspond to the post(pre)-fit contribution of the systematic to the uncertainty on the signal strength and is read on the upper axis. The horizontal bar on the black points show the post-fit uncertainties after applying the constraints from the fit and can be read on the bottom axis. The post(pre)-fit impact on sensitivity is computed performing the fit fixing the nuisance parameter at the post(pre)-fit $\pm 1\sigma$ variation and taking the difference in the fitted μ with the default fit.

observed after the individual fits. These effect can not be explained by the difference in a single nuisance parameter measurement after the combined fit or after the individual fit. It is due to the strong correlations of the systematic uncertainties between the two channels and the different corrections needed to model the $t\bar{t}+$ jets background in the two orthogonal phase spaces. This effect results in an observed combined signal strength for the default model of $\hat{\mu} = 0.84^{+0.64}_{-0.61}$ which is higher than the signal strengths of the individual fits for the two channels but in between the signal strengths of the two-μ fit. The observed combined signal strength is also found to be compatible with the SM expectation within uncertainties.

The two$-\mu$ fit to data also allows to estimate the compatibility between the single lepton and di-lepton channels. The change in the best fit value of the likelihood for the combined and the "two$-\mu$" fits are confronted to a χ^2 distribution with one degree of freedom (in this case the additional μ parameter). This procedure gives the probability to have the same signal strength in the two channels. The observed compatibility between the two channels is 21%.

Figure 5.39 displays the observed and expected 95% confidence level upper limits on the signal strength. The combination with the di-lepton channel provides a 10% improvement of the expected upper limit compared to the single lepton alone. The combined fit finds a 1.4σ excess of $t\bar{t}H(H \to b\bar{b})$ over the background only hypothesis. A signal strength higher than 2.0 is excluded at the 95% confidence level.

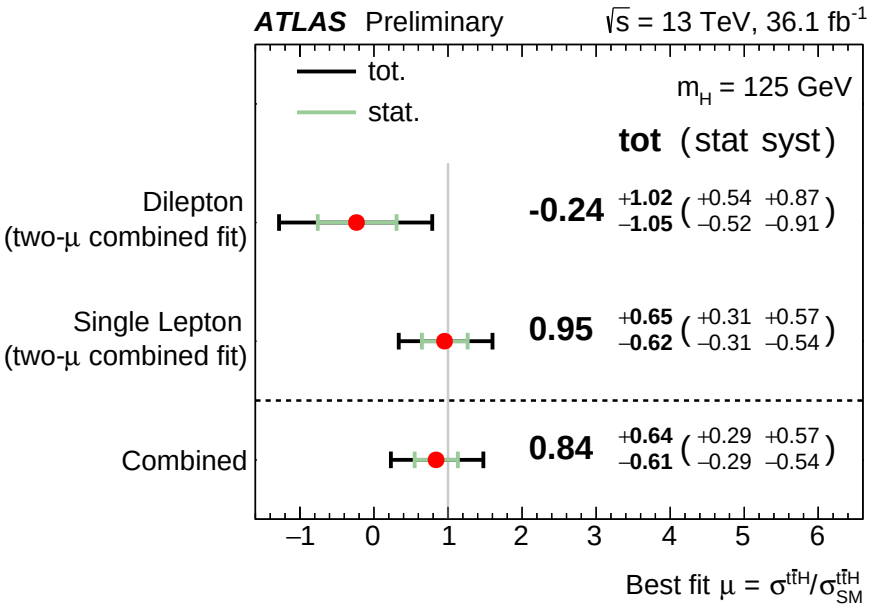


Figure 5.38.: Fitted value of the signal strength and its uncertainty from the fit of the default $t\bar{t}+\geq 1b$ models to data in the single lepton, di-lepton channels and for the combined fit.

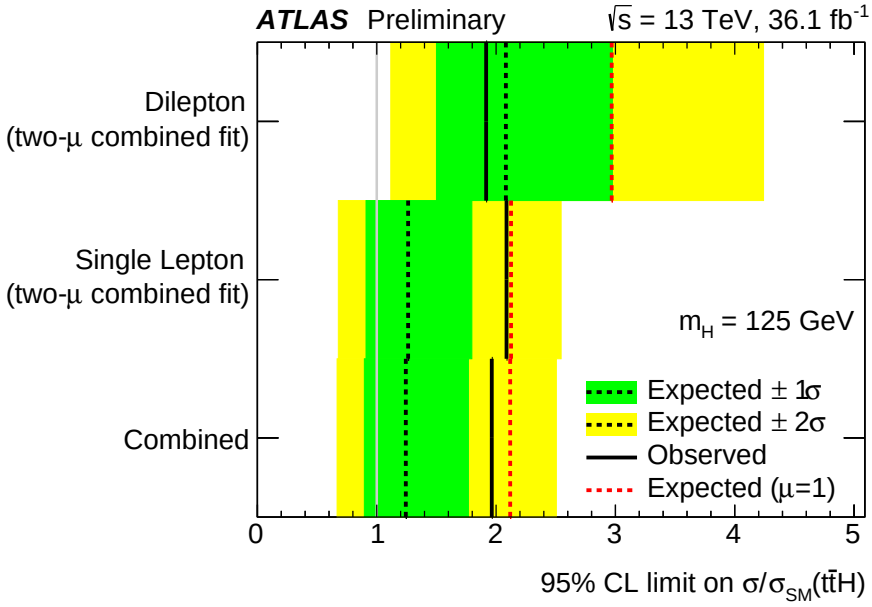


Figure 5.39.: Summary of the upper limits on the $t\bar{t}H(H \to b\bar{b})$ signal strength at the 95% confidence level from both individual channels and for the combination.

5.8. Summary

The search for $t\bar{t}H(H \to b\bar{b})$ production in the 36.1 fb^{-1} of 13 TeV ATLAS data recorded in 2015 and 2016 is presented. The analysis in the single lepton channel is described. The corresponding results and the combination with the di-lepton channel are shown. The measured signal strength in the single lepton channel is $0.67^{+0.71}_{-0.69}$ while the combined value is $0.84^{+0.64}_{-0.61}$. The combined measured (expected) significance is 1.4σ (1.6σ) while $t\bar{t}H(H \to b\bar{b})$ cross-sections 2 times larger than the SM prediction are excluded at the 95% confidence level. The Run 2 analysis presented here provides a 60% improvement with respect to the Run 1 analysis performed with 20.3 fb^{-1}of data at $\sqrt{s} = 8$ TeV.

The $t\bar{t}H(H \to b\bar{b})$ statistical analysis of data relies on a complex fit with many bins, many nuisance parameters and non-trivial correlations. A great attention is dedicated to understand the behavior of the fit and adapt the systematic model accordingly. A selection of the studies I have done in this line, is presented in this chapter. I used decorrelation techniques to detect the origin of the pulls and constraints on the nuisance parameters and to quantify their impact on the measurement. I quantified the impact of the statistical fluctuations of the important systematic uncertainties on the measurement using toys. Whenever this impact is large, dedicated smoothing procedures are adopted for the corresponding systematic uncertainty.

The $t\bar{t}$ + jets background, and in particular the model of the $t\bar{t}$+$\geq$1b component, is the largest source of uncertainty on the signal strength and thus it is scrutinized. The $t\bar{t}$+$\geq$1b default model is the result of a long list of improvements, for which I had a leading role, to the model used in Run 1. It encapsulates the advantages of many models that were proposed to improve the analysis. I also proposed a second $t\bar{t}$+$\geq$1b model, with more flexibility in the systematic uncertainties of the $t\bar{t}$+$\geq$1b sub-components. I validated both using pseudo-data and they yield compatible fit results. In

particular, similar signal sensitivity and measured signal strength are obtained. This greatly increases the confidence in the presented results for such a complex analysis. However, both models suffer from large uncertainties arising from the comparison of several MC samples. In particular, I studied in details the uncertainty on the choice of the MC generator, which is the single nuisance parameter with the largest impact on the sensitivity in this analysis. In depth improvements of the $t\bar{t}$+$\geq$1b model are mandatory for the $t\bar{t}H(H \to b\bar{b})$ analysis to be able to reach an evidence with the full Run 2 data. This will be possible with the incoming $t\bar{t}$+$\geq$1b measurements at $\sqrt{s} = 13$ TeV which will be used to tune the current MC generators and reduce the uncertainties on the $t\bar{t}$+$\geq$1b process.

Conclusion

One of the main goals of the Large Hadron Collider was the discovery of the Higgs boson. After the discovery of a new particle of mass 125 GeV compatible with the Standard Model Higgs boson, a large effort is dedicated to the measurement of its properties. At the beginning of the Run 2 in 2015, many observations support the hypothesis of a Standard Model Higgs boson. However several pieces are still missing, especially the Higgs boson coupling to quarks.

The top-quark is the heaviest known particle and thus has the largest Yukawa coupling to the Higgs boson. The associated production of a Higgs boson with a pair of top-quarks (referred to as $t\bar{t}H$) is the only channel allowing a direct measurement of this coupling at the LHC. No evidence of $t\bar{t}H$ production has been found yet in the 20.3 fb^{-1}of Run 1 data at $\sqrt{s} = 8$ TeV. This thesis presents a search for $t\bar{t}H$ production in the Run 2 data.

The excellent operation of the Large Hadron Collider and the ATLAS detector allowed to record and analyze 36.1 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV in 2015 and 2016. The search for $t\bar{t}H$ events presented in this thesis focuses on the decay of the Higgs boson into a pair of b-quarks, $t\bar{t}H(H \to b\bar{b})$. To reduce the overwhelming presence of the $t\bar{t}$+$\geq$1b background multi-variate techniques are used. This include a complex event categorization, and, in the signal-enriched categories, the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ final state and the use of Boosted Decision Trees to separate the $t\bar{t}$+$\geq$1b background from the $t\bar{t}H(H \to b\bar{b})$ signal. A fit is then performed to statistically match the predictions to the data, simultaneously in all categories. I had a leading role in all these key aspects of this analysis, in particular the responsibility of the statistical analysis, the fit model and the extraction of the result. The analysis in the single lepton channel is described and its results are presented in this thesis, in particular the outcome of the fit and the studies I have done to understand the post-fit systematic uncertainties. The measured signal strength in this channel is $0.67^{+0.71}_{-0.69}$. I also contributed to the combination of the single lepton channel with the di-lepton channel, which leads to a combined signal strength of $0.84^{+0.64}_{-0.61}$ is observed, corresponding to a 1.4σ excess of $t\bar{t}H(H \to b\bar{b})$ over the background hypothesis in data. The analysis of the 36.1 fb^{-1}of Run 2 data allows to exclude $t\bar{t}H(H \to b\bar{b})$ cross-sections 2 times larger than the SM prediction at the 95% confidence level. This represent a 60% improvement with respect to the Run 1 analysis performed with 20.3 fb^{-1}of data at $\sqrt{s} = 8$ TeV.

I have shown that the uncertainty on the reconstructed signal is largely dominated by the systematic uncertainties, including the statistical uncertainty on the MC predictions. Improvements of the $t\bar{t}$+jets background modelling, in particular the description of the $t\bar{t}$+$\geq$1b component, are mandatory for the $t\bar{t}H(H \to b\bar{b})$ channel to reach an evidence with the full Run 2 data. Additional efforts are also needed to increase the amount of Monte Carlo generated events in the small corner of phase space where the signal is present.

Further improvements of the analysis will rely on better separation of the $t\bar{t}H(H \to b\bar{b})$ signal from the $t\bar{t}$+$\geq$1b background. In particular, I developed a new and complementary method for the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ final state to benefit from new sources of information. It is based on a network composed of the final state particles connected by their probability to originate from the same particle. This new technique is not yet used at its full potential and is not included in the main analysis. However, it shows encouraging preliminary results.

High performance and precise understanding of b-tagging algorithms, which identify jets originating from b-quarks (called b-jets), is mandatory for many analyses ranging from Standard Model measurements to Beyond Standard Model searches. In particular the $t\bar{t}H(H \to b\bar{b})$ analysis has four b-quarks in the final state and makes extensive use of b-tagging to separate the signal from the $t\bar{t}+$ light and $t\bar{t}+\geq 1c$ backgrounds.

The definition of the jet flavour (jet labelling) in Monte Carlo simulation can be ambiguous in several cases, for example the fragmentation of a single b-quark into two jets. Moreover, b-tagging performance are evaluated in simulation based on certain Monte Carlo generators and corrected to match efficiencies in data. In order to avoid large extrapolation factors, the definition of b-jets should not be based on Monte Carlo dependent parameters. The jet labelling mostly relies on the particle to jet association. For Run 2, hadrons are preferred over quarks as their definition is less generator dependent. I studied two main algorithms which are presented in this thesis, the "ghost-association" and ΔR matching of jets to hadrons. My studies of the fragmentation of b-quarks in jets and of the track to jet association shows that the ΔR based labelling pairs better with b-tagging algorithm and is therefore chosen as the default labelling scheme for b-tagging Run 2 studies by the ATLAS Collaboration.

Standard b-tagging algorithms are not designed to separate jets originating from a gluon splitting into two b-quarks at low opening angles, bb-jets, from b-jets. However, the identification of bb-jets is important for Run 2. In fact, the large amount of data that is foreseen for Run 2 provides sensitivity to searches for new jet topologies such as boosted $H \to bb$. bb-jet identification can help constraining the background from a gluon splitting into two b-quarks at low opening angles which have large theoretical uncertainties. The algorithms developed for the separation of bb-jets from b-jets based on multi-secondary vertices, the MultiSVbb taggers, are presented in this thesis. In particular, I studied the reconstructed vertices and showed that the usage of only the two leading mass vertices encompass the necessary information for the bb-jet identification. The performance of MultiSVbb taggers after my optimization for Run 2 shows a rejection of b-jets of 20 for a typical working point equivalent to a 35% efficiency to select bb-jets which is seven times higher than what can achieve with standard b-tagging algorithms.

List of Figures

List of Tables

Bibliography

[1] R. M. Wald, *General Relativity*, 1984 (cit. on pp. 23, 25).

[2] C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Mécanique quantique*, Hermann, 1973, URL: http://www.editions-hermann.fr/4438-mecanique-quantique-tome-i.html (cit. on p. 23).

[3] J.-P. Derendinger, *Théorie quantique des champs*, Presses polytechniques et universitaires romandes, 2013 (cit. on p. 23).

[4] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Physics Letters B **716** (2012) p. 1 (cit. on pp. 23, 37, 38).

[5] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Physics Letters B **716** (2012) p. 30 (cit. on pp. 23, 37).

[6] P. A. M. Dirac, *The Quantum Theory of the Emission and Absorption of Radiation*, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **114** (1927) p. 243 (cit. on p. 24).

[7] E. Fermi, *Quantum Theory of Radiation*, Rev. Mod. Phys. **4** (1 1932) p. 87 (cit. on p. 24).

[8] F. J. Dyson, *The Radiation Theories of Tomonaga, Schwinger, and Feynman*, Phys. Rev. **75** (3 1949) p. 486 (cit. on p. 24).

[9] F. J. Dyson, *The S Matrix in Quantum Electrodynamics*, Phys. Rev. **75** (11 1949) p. 1736 (cit. on p. 24).

[10] R. P. Feynman, *Space-Time Approach to Non-Relativistic Quantum Mechanics*, Rev. Mod. Phys. **20** (2 1948) p. 367 (cit. on p. 24).

[11] R. P. Feynman, *Relativistic Cut-Off for Quantum Electrodynamics*, Phys. Rev. **74** (10 1948) p. 1430 (cit. on p. 24).

[12] J. A. Wheeler and R. P. Feynman, *Interaction with the Absorber as the Mechanism of Radiation*, Rev. Mod. Phys. **17** (2-3 1945) p. 157 (cit. on p. 24).

[13] J. Schwinger, *On Quantum-Electrodynamics and the Magnetic Moment of the Electron*, Phys. Rev. **73** (4 1948) p. 416 (cit. on p. 24).

[14] J. Schwinger, *Quantum Electrodynamics. I. A Covariant Formulation*, Phys. Rev. **74** (10 1948) p. 1439 (cit. on p. 24).

[15] S. Tomonaga, *On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields**, Progress of Theoretical Physics **1** (1946) p. 27 (cit. on p. 24).

[16] Z. Koba, T. Tati, and S.-i. Tomonaga, *On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields. IICase of Interacting Electromagnetic and Electron Fields*, Progress of Theoretical Physics **2** (1947) p. 101 (cit. on p. 24).

[17] S. Kanesawa and S.-i. Tomonaga, *On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields. VCase of Interacting Electromagnetic and Meson Fields*, Progress of Theoretical Physics **3** (1948) p. 1 (cit. on p. 24).

[18] S.-I. Tomonaga and J. R. Oppenheimer, *On Infinite Field Reactions in Quantum Field Theory*, Phys. Rev. **74** (2 1948) p. 224 (cit. on p. 24).

[19] R. Feynman, *Quantum mechanics and path integrals*, McGraw-Hill, 1965 (cit. on p. 24).

[20] S. Weinberg, *General Theory of Broken Local Symmetries*, Phys. Rev. D **7** (4 1973) p. 1068 (cit. on p. 25).

[21] B. Odom, D. Hanneke, B. D'Urso, et al., *New Measurement of the Electron Magnetic Moment Using a One-Electron Quantum Cyclotron*, Phys. Rev. Lett. **97** (3 2006) p. 030801 (cit. on p. 26).

[22] F. Minardi, G. Bianchini, P. C. Pastor, et al., *Measurement of the Helium $2^3P_0 - 2^3P_1$ Fine Structure Interval*, Phys. Rev. Lett. **82** (6 1999) p. 1112 (cit. on p. 26).

[23] F. L. Wilson, *Fermi's Theory of Beta Decay*, American Journal of Physics **36** (1968) p. 1150 (cit. on p. 26).

[24] T. D. Lee and C. N. Yang, *Question of Parity Conservation in Weak Interactions*, Phys. Rev. **104** (1 1956) p. 254 (cit. on p. 26).

[25] C. S. Wu, E. Ambler, R. W. Hayward, et al., *Experimental Test of Parity Conservation in Beta Decay*, Phys. Rev. **105** (4 1957) p. 1413 (cit. on p. 26).

[26] R. L. Garwin, L. M. Lederman, and M. Weinrich, *Observations of the Failure of Conservation of Parity and Charge Conjugation in Meson Decays: the Magnetic Moment of the Free Muon*, Phys. Rev. **105** (4 1957) p. 1415 (cit. on p. 26).

[27] P. A. M. Dirac, *The Quantum Theory of the Electron*, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **117** (1928) p. 610 (cit. on p. 26).

[28] R. P. Feynman and M. Gell-Mann, *Theory of the Fermi Interaction*, Phys. Rev. **109** (1 1958) p. 193 (cit. on p. 27).

[29] M. Gell-Mann, *Test of the Nature of the Vector Interaction in β Decay*, Phys. Rev. **111** (1 1958) p. 362 (cit. on p. 27).

[30] E. C. G. Sudarshan and R. E. Marshak, *Chirality Invariance and the Universal Fermi Interaction*, Phys. Rev. **109** (5 1958) p. 1860 (cit. on p. 27).

[31] S. Okubo, R. E. Marshak, E. C. G. Sudarshan, et al., *Interaction Current in Strangeness-Violating Decays*, Phys. Rev. **112** (2 1958) p. 665 (cit. on p. 27).

[32] E. C. G. Sudarshan and R. E. Marshak, *The nature of the four-fermion interaction*, (1994) (cit. on p. 27).

[33] F. Hasert, S. Kabe, W. Krenz, et al., *Observation of neutrino-like interactions without muon or electron in the Gargamelle neutrino experiment*, Nuclear Physics B **73** (1974) p. 1 (cit. on p. 27).

[34] Y. Ne'eman, *Derivation of strong interactions from a gauge invariance*, Nuclear Physics **26** (1961) p. 222 (cit. on p. 28).

[35] M. Gell-Mann, *The Eightfold Way: A Theory of strong interaction symmetry*, (1961) (cit. on p. 28).

[36] G. Zweig, "An SU_3 model for strong interaction symmetry and its breaking; Version 1," tech. rep. CERN-TH-401, CERN, 1964, URL: http://cds.cern.ch/record/352337 (cit. on p. 28).

[37] M. Gell-Mann, "Quarks," *Elementary Particle Physics: Multiparticle Aspects*, ed. by P. Urban, Springer Vienna, 1972 p. 733 (cit. on p. 28).

[38] C. N. Yang and R. L. Mills, *Conservation of Isotopic Spin and Isotopic Gauge Invariance*, Phys. Rev. **96** (1 1954) p. 191 (cit. on p. 28).

[39] H. Fritzsch, M. Gell-Mann, and H. Leutwyler, *Advantages of the color octet gluon picture*, Physics Letters B **47** (1973) p. 365 (cit. on p. 28).

[40] D. J. Gross and F. Wilczek, *Ultraviolet Behavior of Non-Abelian Gauge Theories*, Phys. Rev. Lett. **30** (26 1973) p. 1343 (cit. on p. 28).

[41] H. D. Politzer, *Reliable Perturbative Results for Strong Interactions*, Phys. Rev. Lett. **30** (26 1973) p. 1346 (cit. on p. 28).

[42] S. Bethke, α_s *2002*, Nuclear Physics B - Proceedings Supplements **121** (2003) p. 74, Proceedings of the QCD 02 9th High-Energy Physics International Conference on Quantum ChromoDynamics, ISSN: 0920-5632, URL: http://www.sciencedirect.com/science/article/pii/S0920563203018176 (cit. on p. 28).

[43] P. Zerwas, *W & Z physics at LEP*, The European Physical Journal C - Particles and Fields **34** (1, 2004) p. 41, ISSN: 1434-6052, URL: https://doi.org/10.1140/epjc/s2004-01765-9 (cit. on p. 28).

[44] Y. Nambu, *The Confinement of Quarks*, Sci. Am. **235N5** (1976) p. 48 (cit. on p. 28).

[45] S. L. Glashow, *Partial-symmetries of weak interactions*, Nuclear Physics **22** (1961) p. 579 (cit. on p. 29).

[46] A. Salam and J. Ward, *Electromagnetic and weak interactions*, Physics Letters **13** (1964) p. 168 (cit. on p. 29).

[47] S. Weinberg, *Non-Abelian Gauge Theories of the Strong Interactions*, Phys. Rev. Lett. **31** (7 1973) p. 494 (cit. on p. 29).

[48] Patrignani, C. *et al.* (Particle Data Group), *Review of Particle Physics*, Chin. Phys. **C40** (2016) p. 100001 (cit. on pp. 29, 80).

[49] M. Gell-Mann, *The interpretation of the new particles as displaced charge multiplets*, Il Nuovo Cimento (1955-1965) **4** (1956) p. 848 (cit. on p. 29).

[50] T. Nakano and K. Nishijima, *Charge Independence for V-particles**, Progress of Theoretical Physics **10** (1953) p. 581 (cit. on p. 29).

[51] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (9 1964) p. 321 (cit. on p. 29).

[52] P. Higgs, *Broken symmetries, massless particles and gauge fields*, Physics Letters **12** (1964) p. 132 (cit. on p. 29).

[53] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (16 1964) p. 508 (cit. on p. 29).

[54] N. Cabibbo, *Unitary Symmetry and Leptonic Decays*, Phys. Rev. Lett. **10** (12 1963) p. 531 (cit. on p. 31).

[55] M. Kobayashi and T. Maskawa, *CP-Violation in the Renormalizable Theory of Weak Interaction*, Progress of Theoretical Physics **49** (1973) p. 652 (cit. on p. 31).

[56] CKMfitter group, J. Charles *et al.*, *CP violation and the CKM matrix: assessing the impact of the asymmetric B factories*, The European Physical Journal C - Particles and Fields **41** (2005) p. 1, URL: http://ckmfitter.in2p3.fr (cit. on p. 31).

[57] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, The LEP Working Group for Higgs Boson Searches, *Search for the Standard Model Higgs boson at LEP*, Physics Letters B **565** (2003) p. 61 (cit. on p. 34).

[58] TEVNPH (Tevatron New Phenomina and Higgs Working Group), CDF and D0 Collaborations, *Combined CDF and D0 Search for Standard Model Higgs Boson Production with up to* 10.0 fb^{-1} *of Data*, (2012), arXiv: 1203.3774 [hep-ex] (cit. on p. 34).

[59] LHC Higgs Cross Section Working Group, S. Heinemeyer, C. Mariotti, et al., *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties*, (CERN, Geneva, 2013), arXiv: 1307.1347 [hep-ph] (cit. on pp. 34, 36, 109).

[60] D. LHC Higgs Cross Section Working Group Florian et al., *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, (2016), arXiv: 1610.07922 [hep-ph] (cit. on p. 36).

[61] ATLAS and CMS Collaborations, *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*, Journal of High Energy Physics **2016** (2016) p. 45 (cit. on pp. 37–40).

[62] 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* $\sqrt{s} = 7$ *and* 8 *TeV*, The European Physical Journal C **75** (2015) p. 212 (cit. on p. 39).

[63] ATLAS Collaboration, *Measurement of the Higgs boson mass from the* $H \to \gamma\gamma$ *and* $H \to ZZ^* \to 4\ell$ *channels in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector*, Phys. Rev. D **90** (5 2014) p. 052004 (cit. on p. 39).

[64] ATLAS and CMS Collaborations, *Combined Measurement of the Higgs Boson Mass in pp Collisions at* $\sqrt{s} = 7$ *and 8 TeV with the ATLAS and CMS Experiments*, Phys. Rev. Lett. **114** (2015) p. 191803 (cit. on p. 39).

[65] ATLAS Collaboration, *Evidence for the spin-0 nature of the Higgs boson using ATLAS data*, Physics Letters B **726** (2013) p. 120 (cit. on pp. 40, 41).

[66] S. Bolognesi, Y. Gao, A. V. Gritsan, et al., *Spin and parity of a single-produced resonance at the LHC*, Phys. Rev. D **86** (9 2012) p. 095031 (cit. on p. 40).

[67] A. Denner, S. Dittmaier, M. Roth, et al., *Predictions for all processes $e^+e^- \to$ fermions + γ*, Nuclear Physics B **560** (1999) p. 33, ISSN: 0550-3213 (cit. on p. 40).

[68] A. Denner, S. Dittmaier, M. Roth, et al., *Electroweak corrections to charged-current $e^+e^- \to 4$ fermion processes: Technical details and further results*, Nuclear Physics B **724** (2005) p. 247 (cit. on p. 40).

[69] N. Kauer and G. Passarino, *Inadequacy of zero-width approximation for a light Higgs boson signal*, Journal of High Energy Physics **2012** (2012) p. 116 (cit. on p. 40).

[70] CMS Collaboration, *Constraints on the Higgs boson width from off-shell production and decay to Z-boson pairs*, Physics Letters B **736** (2014) p. 64 (cit. on pp. 41, 42).

[71] ATLAS Collaboration, *Constraints on the off-shell Higgs boson signal strength in the high-mass ZZ and WW final states with the ATLAS detector*, The European Physical Journal C **75** (2015) p. 335 (cit. on pp. 41, 42).

[72] L. R. Evans and P. Bryant, *LHC Machine*, J. Instrum. **3** (2008) S08001. 164 p, This report is an abridged version of the LHC Design Report (CERN-2004-003), URL: https://cds.cern.ch/record/1129806 (cit. on p. 43).

[73] O. S. Brüning, P. Collier, P. Lebrun, et al., *LHC Design Report*, CERN, 2004, URL: https://cds.cern.ch/record/782076 (cit. on p. 43).

[74] J. Stirling, *Parton Luminosity and Cross-section plots*, (), URL: http://www.hep.ph.ic.ac.uk/%5C~%7B%7Dwstirlin/plots/plots.html (cit. on p. 44).

[75] ATLAS Collaboration, *Summary plots from the ATLAS Standard Model physics group*, (), URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CombinedSummaryPlots/SM/ (cit. on p. 44).

[76] CERN, *Taking a closer look at LHC*, (), URL: http://www.lhc-closer.es/taking_a_closer_look_at_lhc/1.home (cit. on p. 45).

[77] CERN, *Computer-generated diagram of an LHC dipole*, (), URL: https://cds.cern.ch/record/39731 (cit. on p. 46).

[78] R. Alemany-Fernandez, E. Bravin, L. Drosdal, et al., *Operation and Configuration of the LHC in Run 1*, (2013), URL: https://cds.cern.ch/record/1631030 (cit. on p. 47).

[79] R. Bruce, G. Arduini, H. Bartosik, et al., "LHC Run 2: Results and Challenges," tech. rep. CERN-ACC-2016-0103, CERN, 2016, URL: https://cds.cern.ch/record/2201447 (cit. on p. 47).

[80] ATLAS Collaboration, *ATLAS: technical proposal for a general-purpose pp experiment at the Large Hadron Collider at CERN*, LHC Tech. Proposal, CERN, 1994, URL: http://cds.cern.ch/record/290968 (cit. on p. 47).

[81] CMS Collaboration, *Technical proposal*, LHC Tech. Proposal,
Cover title : CMS, the Compact Muon Solenoid : technical proposal, CERN, 1994,
URL: http://cds.cern.ch/record/290969 (cit. on p. 47).

[82] LHCb Collaboration, *LHCb : Technical Proposal*, Tech. Proposal, CERN, 1998,
URL: http://cds.cern.ch/record/622031 (cit. on p. 47).

[83] ALICE Collaboration,
ALICE: Technical proposal for a Large Ion collider Experiment at the CERN LHC,
LHC Tech. Proposal, CERN, 1995, URL: http://cds.cern.ch/record/293391
(cit. on p. 47).

[84] TOTEM Collaboration, "TOTEM, Total Cross Section, Elastic Scattering and Diffraction Dissociation at the LHC: Technical Proposal," tech. rep. CERN-LHCC-99-007. LHCC-P-5, CERN, 1999, URL: http://cds.cern.ch/record/385483 (cit. on p. 47).

[85] LHCf Collaboration, *The LHCf detector at the CERN Large Hadron Collider*,
Journal of Instrumentation **3** (2008) S08006,
URL: http://stacks.iop.org/1748-0221/3/i=08/a=S08006 (cit. on p. 47).

[86] The ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*,
Journal of Instrumentation **3** (2008) S08003,
URL: http://stacks.iop.org/1748-0221/3/i=08/a=S08003
(cit. on pp. 48–50, 52–54, 57–59, 61, 62).

[87] ATLAS Collaboration, "Electron efficiency measurements with the ATLAS detector using the 2012 LHC proton-proton collision data," tech. rep. ATLAS-CONF-2014-032, CERN, 2014,
URL: https://cds.cern.ch/record/1706245 (cit. on p. 50).

[88] ATLAS Collaboration, *ATLAS public web page*, (),
URL: http://atlas.cern/resources/multimedia (cit. on pp. 51, 64).

[89] ATLAS Collaboration, *Event Displays from Collision Data public page*, (), URL:
https://twiki.cern.ch/twiki/bin/view/AtlasPublic/EventDisplayRun2Collisions
(cit. on p. 51).

[90] ATLAS Collaboration,
"Studies of the ATLAS Inner Detector material using $\sqrt{s}$ =13 TeV pp collision data,"
tech. rep. ATL-PHYS-PUB-2015-050, CERN, 2015,
URL: https://cds.cern.ch/record/2109010 (cit. on p. 52).

[91] ATLAS Collaboration,
"Particle Identification Performance of the ATLAS Transition Radiation Tracker,"
tech. rep. ATLAS-CONF-2011-128, CERN, 2011,
URL: https://cds.cern.ch/record/1383793 (cit. on pp. 54, 55).

[92] ATLAS Collaboration,
"Alignment of the ATLAS Inner Detector with the initial LHC data at $\sqrt{s} = 13$ TeV,"
tech. rep. ATL-PHYS-PUB-2015-031, CERN, 2015,
URL: https://cds.cern.ch/record/2038139 (cit. on pp. 55, 56).

[93] N. Nikiforou, "Performance of the ATLAS Liquid Argon Calorimeter after three years of LHC operation and plans for a future upgrade," *Proceedings, 3rd International Conference on Advancements in Nuclear Instrumentation Measurement Methods and their Applications (ANIMMA 2013): Marseille, France, June 23-27, 2013*, 2013, arXiv: 1306.6756 [physics.ins-det], URL: https://inspirehep.net/record/1240499/files/arXiv:1306.6756.pdf (cit. on p. 58).

[94] ATLAS Collaboration, *Readiness of the ATLAS Liquid Argon Calorimeter for LHC Collisions*, Eur. Phys. J. **C70** (2010) p. 723, arXiv: 0912.2642 [physics.ins-det] (cit. on pp. 58, 59).

[95] T. Davidek, M. Volpi, and T. Zenis, "Response of the ATLAS Tile Calorimeter to Hadrons in Stand-Alone Testbeam Data," tech. rep. ATL-TILECAL-PUB-2009-004. ATL-COM-TILECAL-2009-002, The note has been updated using comments received by both referees.: CERN, 2009, URL: https://cds.cern.ch/record/1161351 (cit. on p. 60).

[96] G. W. Wilburn, A. Mattillion, G. Facini, et al., "Tile Calorimeter Cell Energy Distribution Using 2015 Collision Data at $\sqrt{s} = 13$ and 0.9 TeV," tech. rep. ATL-COM-TILECAL-2016-003, CERN, 2016, URL: https://cds.cern.ch/record/2131164 (cit. on p. 60).

[97] ATLAS Collaboration, *Commissioning of the ATLAS Muon Spectrometer with Cosmic Rays*, Eur. Phys. J. **C70** (2010) p. 875, arXiv: 1006.4384 [physics.ins-det] (cit. on p. 62).

[98] S. Höche, F. Krauss, M. Schönherr, et al., *QCD matrix elements + parton showers: The NLO case*, JHEP **04** (2013) p. 027, arXiv: 1207.5030 [hep-ph] (cit. on p. 63).

[99] T. Gleisberg, S. Hoeche, F. Krauss, et al., *Event generation with SHERPA 1.1*, JHEP **0902** (2009) p. 007, arXiv: 0811.4622 [hep-ph] (cit. on pp. 63, 91, 111, 138).

[100] S. Frixione and B. R. Webber, *Matching NLO QCD computations and parton shower simulations*, JHEP **0206** (2002) p. 029, arXiv: hep-ph/0204244 [hep-ph] (cit. on p. 63).

[101] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, JHEP **0411** (2004) p. 040, arXiv: hep-ph/0409146 (cit. on p. 63).

[102] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, JHEP **0711** (2007) p. 070, arXiv: 0709.2092 [hep-ph] (cit. on p. 63).

[103] S. Alioli, P. Nason, C. Oleari, et al., *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **1006** (2010) p. 043, arXiv: 1002.2581 [hep-ph] (cit. on p. 63).

[104] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **0605** (2006) p. 026, arXiv: hep-ph/0603175 (cit. on pp. 63, 77).

[105] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) p. 852, arXiv: 0710.3820 [hep-ph] (cit. on pp. 63, 91, 109).

[106] M. Bahr et al., *Herwig++ Physics and Manual*, Eur. Phys. J. **C58** (2008) p. 639, arXiv: 0803.0883 [hep-ph] (cit. on p. 63).

[107] J. Bellm, S. Gieseke, D. Grellscheid, et al., *Herwig 7.0/Herwig++ 3.0 release note*, The European Physical Journal C **76** (2016) p. 196, ISSN: 1434-6052, arXiv: 1512.01178 [hep-ph] (cit. on pp. 63, 138).

[108] T. Cornelissen, M. Elsing, S. Fleischmann, et al., "Concepts, Design and Implementation of the ATLAS New Tracking (NEWT)," tech. rep. ATL-SOFT-PUB-2007-007. ATL-COM-SOFT-2007-002, CERN, 2007, URL: https://cds.cern.ch/record/1020106 (cit. on p. 65).

[109] ATLAS Collaboration, "The Optimization of ATLAS Track Reconstruction in Dense Environments," tech. rep. ATL-PHYS-PUB-2015-006, CERN, 2015, URL: https://cds.cern.ch/record/2002609 (cit. on p. 65).

[110] ATLAS Collaboration, "Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton–proton collisions at the LHC," tech. rep. arXiv:1611.10235. CERN-EP-2016-150, Comments: 52 pages in total, author list starting at page 36, 17 figures, 4 tables, submitted to EPJC. All figures including auxillary figures are available at http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/PERF-2015-01: CERN, 2016, URL: https://cds.cern.ch/record/2235651 (cit. on p. 65).

[111] ATLAS collaboration, *A neural network clustering algorithm for the ATLAS silicon pixel detector*, Journal of Instrumentation **9** (2014) P09009, arXiv: 1406.7690 [hep-ph] (cit. on p. 66).

[112] T. G. Cornelissen, N. Van Eldik, M. Elsing, et al., "Updates of the ATLAS Tracking Event Data Model (Release 13)," tech. rep. ATL-SOFT-PUB-2007-003. ATL-COM-SOFT-2007-008, CERN, 2007, URL: https://cds.cern.ch/record/1038095 (cit. on p. 66).

[113] ATLAS Collaboration, *Comparison of the impact parameter resolution in Run 1 and Run 2*, (), URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/IDTR-2015-007/ (cit. on p. 67).

[114] ATLAS collaboration, *Muon reconstruction efficiency and momentum resolution of the ATLAS experiment in proton–proton collisions at* $\sqrt{s} = 7$ *TeVin* 2010, The European Physical Journal C **74** (2014) p. 3034, ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-014-3034-9 (cit. on pp. 67, 69).

[115] ATLAS collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at* $\sqrt{s} = 13$ *TeV*, The European Physical Journal C **76** (2016) p. 292, ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-016-4120-y (cit. on pp. 67, 69).

[116] ATLAS Collaboration, "Electron and photon reconstruction and identification in ATLAS: expected performance at high energy and results at 900 GeV," tech. rep. ATLAS-CONF-2010-005, CERN, 2010, URL: https://cds.cern.ch/record/1273197 (cit. on p. 69).

[117] 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, CERN, 2008, URL: https://cds.cern.ch/record/1099735 (cit. on p. 69).

[118] "Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data," tech. rep. ATLAS-CONF-2016-024, CERN, 2016, URL: https://cds.cern.ch/record/2157687 (cit. on pp. 70, 108).

[119] ATLAS collaboration, "Expected electron performance in the ATLAS experiment," tech. rep. ATL-PHYS-PUB-2011-006, CERN, 2011, URL: https://cds.cern.ch/record/1345327 (cit. on p. 70).

[120] ATLAS Collaboration, "Photon identification in 2015 ATLAS data," tech. rep. ATL-PHYS-PUB-2016-014, CERN, 2016, URL: https://cds.cern.ch/record/2203125 (cit. on p. 70).

[121] ATLAS Collaboration, *Jet energy measurement with the ATLAS detector in proton-proton collisions at* $\sqrt{s} = 7$ *TeV*, The European Physical Journal C **73** (2013) p. 2304, ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-013-2304-2 (cit. on pp. 70, 71).

[122] M. Cacciari, G. P. Salam, and G. Soyez, *The anti- k t jet clustering algorithm*, Journal of High Energy Physics **2008** (2008) p. 063, URL: http://stacks.iop.org/1126-6708/2008/i=04/a=063 (cit. on pp. 70, 71).

[123] ATLAS Collaboration, "Jet Calibration and Systematic Uncertainties for Jets Reconstructed in the ATLAS Detector at $\sqrt{s} = 13$ TeV," tech. rep. ATL-PHYS-PUB-2015-015, CERN, 2015, URL: https://cds.cern.ch/record/2037613 (cit. on pp. 71, 143).

[124] ATLAS Collaboration, "Tagging and suppression of pileup jets with the ATLAS detector," tech. rep. ATLAS-CONF-2014-018, CERN, 2014, URL: https://cds.cern.ch/record/1700870 (cit. on pp. 71, 72).

[125] ATLAS Collaboration, "Tau Reconstruction and Identification Performance in ATLAS," tech. rep. ATLAS-CONF-2010-086, CERN, 2010, URL: https://cds.cern.ch/record/1298857 (cit. on p. 72).

[126] "Performance of missing transverse momentum reconstruction for the ATLAS detector in the first proton-proton collisions at at $\sqrt{s}$= 13 TeV," tech. rep. ATL-PHYS-PUB-2015-027, CERN, 2015, URL: https://cds.cern.ch/record/2037904 (cit. on p. 73).

[127] R. Zaidan, "A search for a charged Higgs boson in the H+ -> tb channel and b-tagging algorithms with the ATLAS experiment at the LHC.," Theses: Université de la Méditerranée - Aix-Marseille II, 2009, URL: https://tel.archives-ouvertes.fr/tel-00546131 (cit. on p. 75).

[128] ATLAS Collaboration, "Expected performance of the ATLAS *b*-tagging algorithms in Run-2," tech. rep. ATL-PHYS-PUB-2015-022, CERN, 2015, URL: https://cds.cern.ch/record/2037697 (cit. on pp. 76, 87–90).

[129] D. J. Lange, *The EvtGen particle decay simulation package*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **462** (2001) p. 152, BEAUTY2000, Proceedings of the 7th Int. Conf. on B-Physics at Hadron Machines, ISSN: 0168-9002, URL: http://www.sciencedirect.com/science/article/pii/S0168900201000894 (cit. on pp. 77, 91, 109).

[130] M. Cacciari, G. P. Salam, and G. Soyez, *The catchment area of jets*, Journal of High Energy Physics **2008** (2008) p. 005, URL: http://stacks.iop.org/1126-6708/2008/i=04/a=005 (cit. on p. 78).

[131] M. Cacciari and G. P. Salam, *Pileup subtraction using jet areas*, Physics Letters B **659** (2008) p. 119, ISSN: 0370-2693, arXiv: 0707.1378 [hep-ph], URL: http://www.sciencedirect.com/science/article/pii/S0370269307011094 (cit. on p. 78).

[132] ATLAS Collaboration, "Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run," tech. rep. ATL-PHYS-PUB-2016-012, CERN, 2016, URL: http://cds.cern.ch/record/2160731 (cit. on p. 90).

[133] A. Banfi, G. P. Salam, and G. Zanderighi, *Accurate QCD predictions for heavy-quark jets at the Tevatron and LHC*, Journal of High Energy Physics **2007** (2007) p. 026, URL: http://stacks.iop.org/1126-6708/2007/i=07/a=026 (cit. on p. 91).

[134] R. E. Ticse Torres, "Search for the Higgs boson in the $ttH(H \to bb)$ channel and the identification of jets containing two B hadrons with the ATLAS experiment.," Theses: Centre de Physique des Particules de Marseille, 2016, URL: https://tel.archives-ouvertes.fr/tel-01516435 (cit. on pp. 91, 101).

[135] H.-L. Lai, M. Guzzi, J. Huston, et al., *New parton distributions for collider physics*, Phys. Rev. D **82** (7 2010) p. 074024, URL: https://link.aps.org/doi/10.1103/PhysRevD.82.074024 (cit. on p. 91).

[136] ATLAS Collaboration, "ATLAS Run 1 Pythia8 tunes," tech. rep. ATL-PHYS-PUB-2014-021, 2014, URL: https://cds.cern.ch/record/1966419 (cit. on pp. 91, 109).

[137] J. Pumplin, D. R. Stump, J. Huston, et al., *New Generation of Parton Distributions with Uncertainties from Global QCD Analysis*, Journal of High Energy Physics **2002** (2002) p. 012, URL: http://stacks.iop.org/1126-6708/2002/i=07/a=012 (cit. on pp. 91, 109).

[138] A. L. Jeremy G. Siek Lie-Quan Lee, *Boost Graph Library, The: User Guide and Reference Manual*, Addison-Wesley Professional, 2001, URL: http://www.boost.org/doc/libs/1_65_1/libs/graph/doc/index.html (cit. on p. 92).

[139] ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, The European Physical Journal C **75** (29, 2015) p. 349 (cit. on pp. 105, 106, 119).

[140] CMS Collaboration, *Search for the associated production of the Higgs boson with a top-quark pair*, Journal of High Energy Physics **2014** (16, 2014) p. 87 (cit. on p. 105).

[141] ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, To be published soon (2017) (cit. on pp. 107, 119, 139, 142).

[142] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV*, The European Physical Journal C **76** (23, 2016) p. 292, ISSN: 1434-6052, URL: https://doi.org/10.1140/epjc/s10052-016-4120-y (cit. on p. 108).

[143] B. Nachman, P. Nef, A. Schwartzman, et al., *Jets from jets: re-clustering as a tool for large radius jet reconstruction and grooming at the LHC*, Journal of High Energy Physics **2015** (12, 2015) p. 75, ISSN: 1029-8479, URL: https://doi.org/10.1007/JHEP02(2015)075 (cit. on p. 108).

[144] D. Krohn, J. Thaler, and L.-T. Wang, *Jet trimming*, Journal of High Energy Physics **2010** (24, 2010) p. 84, ISSN: 1029-8479, URL: https://doi.org/10.1007/JHEP02(2010)084 (cit. on p. 108).

[145] R. D. Ball, V. Bertone, S. Carrazza, et al., *Parton distributions for the LHC run II*, Journal of High Energy Physics **2015** (8, 2015) p. 40 (cit. on p. 109).

[146] ATLAS Collaboration, "Studies on top-quark Monte Carlo modelling for Top2016," tech. rep. ATL-PHYS-PUB-2016-020, CERN, 2016, URL: https://cds.cern.ch/record/2216168 (cit. on pp. 109, 112, 138).

[147] M. Czakon and A. Mitov, *Top++: A program for the calculation of the top-pair cross-section at hadron colliders*, Computer Physics Communications **185** (2014) p. 2930, ISSN: 0010-4655 (cit. on pp. 109, 138).

[148] M. R. Whalley, D. Bourilkov, and R. C. Group, "The Les Houches accord PDFs (LHAPDF) and LHAGLUE," *HERA and the LHC: A Workshop on the implications of HERA for LHC physics. Proceedings, Part B*, 2005, arXiv: hep-ph/0508110 [hep-ph] (cit. on p. 111).

[149] S. Frixione, E. Laenen, P. Motylinski, et al., *Single-top hadroproduction in association with a W boson*, Journal of High Energy Physics **2008** (2008) p. 029, URL: http://stacks.iop.org/1126-6708/2008/i=07/a=029 (cit. on p. 111).

[150] K. Olive and P. D. Group, *Review of Particle Physics*, Chinese Physics C **38** (2014) p. 090001, URL: http://stacks.iop.org/1674-1137/38/i=9/a=090001 (cit. on p. 120).

[151] A. Barrat, M. Barthlemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, 1st, Cambridge University Press, 2008, ISBN: 0521879507, 9780521879507 (cit. on p. 123).

[152] G. Cowan, K. Cranmer, E. Gross, et al., *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. **C71** (2011) p. 1554, [Erratum: Eur. Phys. J.C73,2501(2013)], arXiv: 1007.1727 [physics.data-an] (cit. on p. 132).

[153] K. Cranmer, G. Lewis, L. Moneta, et al., "HistFactory: A tool for creating statistical models for use with RooFit and RooStats," tech. rep. CERN-OPEN-2012-016, New York U., 2012, URL: https://cds.cern.ch/record/1456844 (cit. on p. 133).

[154] R. M. L. Team, *Minuit2 minimization package*, (), URL: http://project-mathlibs.web.cern.ch/project-mathlibs/sw/Minuit2/html/index.html (cit. on p. 133).

[155] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC*, The European Physical Journal C **76** (28, 2016) p. 653, ISSN: 1434-6052, URL: https://doi.org/10.1140/epjc/s10052-016-4466-1 (cit. on p. 143).

[156] ATLAS Collaboration, *Performance of b-Jet Identification in the ATLAS Experiment*, JINST **11** (2016) P04008, arXiv: 1512.01094 [hep-ex] (cit. on p. 143).

[157] M. Gell-Mann, *A schematic model of baryons and mesons*, Physics Letters **8** (1964) p. 214, ISSN: 0031-9163,
URL: //www.sciencedirect.com/science/article/pii/S0031916364920013.

[158] ATLAS Collaboration, "The Expected Performance of the ATLAS Inner Detector," tech. rep. ATL-PHYS-PUB-2009-002. ATL-COM-PHYS-2008-105, CERN, 2008,
URL: https://cds.cern.ch/record/1118445.

[159] M. Capeans, G. Darbo, K. Einsweiller, et al.,
"ATLAS Insertable B-Layer Technical Design Report,"
tech. rep. CERN-LHCC-2010-013. ATLAS-TDR-19, 2010,
URL: https://cds.cern.ch/record/1291633.

[160] Breiman, Leo and Friedman, Jerome and Olshen, R. A. and Stone, Charles J., "Classification and regression trees," tech. rep., 1984 (cit. on p. 211).

Auxiliary materials

A. Boosted Decision Trees

Boosted Decision Trees are a machine learning technique consisting of a set of binary structured trees combined together with boosting techniques.

A.1. Decision Trees

A Decision Tree (DT) is a multivariate technique developed by L. Breiman *et al.* in 1984 [160]. It generalize the standard cut based analysis in an automated MVA splitting the original sample using a set of input variables and terminating conditions. The final sample subsets are called leaves.

At each node, starting from the original sample, a decision is made. If the node satisfy the terminating conditions it is classified as a background or signal leave based on its purity [e]. In other cases the variable with the highest discrimination power is used to split the node and the procedure is repeated until all nodes are turned in leaves.

In the iteration procedure the same variable can be used several times. This allows the DT to define window cuts (interval of interest) and enhance the sensitivity to the correlations between variables.

Two terminating conditions are used for optimization. The maximum depth of a DT controls the maximum number of layers before classifying the nodes as leaves. The minimum number of events in each leave can also be tuned. Both these parameters are used to balance a maximal use of the available information while avoiding a focus of the DT on topologies that are not statistically relevant (over-training).

A.2. Boosting

A Boosted Decision Tree (BDT) is a weighted averaged sum of DT. This procedure allows to increase the discrimination power and reduces the sensitivity to over-training at the same time.

Suppose one has a sample $\mathbb{S}_1 = \left\{X_1^1, \ldots, X_1^N\right\}$ with N the number of events, $X_1^i = \{x_1^1, \ldots, x_1^m\}$ the values of the m variables for the event i and w_1^i the weight of the i^{th} event. The boosting procedure defines a sequence of N_{tree} samples $\mathbb{S}_k$ to train on (initialized by the input sample) and a combined discriminating variable as follows. Let Y and T_k be the true classification and training classification applications respectively defined as:

$$\begin{array}{rccl} Y: & \mathbb{S}_k & \to & \mathbb{N} \\ & X_k^i & \mapsto & \begin{cases} +1 \text{ if } i \in \text{signal sample} \\ -1 \text{ if } i \in \text{background sample} \end{cases} \end{array} \quad (.5)$$

[e] The purity of a node is defined as $p = \frac{s}{s+b}$ where s(b) are the weighted sum of signal(background) events.

$$
\begin{array}{rccl}
T_k: & \mathbb{S}_k & \to & \mathbb{N} \\
 & X_k^i & \mapsto & \begin{cases} +1 \text{ if } i \in \text{signal leave} \\ -1 \text{ if } i \in \text{background leave} \end{cases}
\end{array} \tag{.6}
$$

Then two boosting algorithms are commonly used for BDTs: the *adaptive boost (AdaBoost)* and the *gradient boost (GradientBoost)*. In the AdaBoost algorithm a weight α_k is assigned to each training:

$$
\alpha_k = \beta \cdot \ln\left(\frac{1-\epsilon_k}{\epsilon_k}\right) \tag{.7}
$$

where β is a free parameter and ϵ_k is the miss-identification rate of the DT k and reads as:

$$
\epsilon_k = \frac{\sum_{i=1}^{N} w_k^i \cdot \text{isMissClassified(i,k)}}{\sum_{i=1}^{N} w_k^i} \tag{.8}
$$

where isMissClassified(i,k) returns 1(0) if $Y(i) \cdot T_k(i) \leq 0$ (≥ 0). A boosting application B is finally defined to go from the sample k to the next one:

$$
\begin{array}{rccl}
B: & \mathbb{S}_k & \to & \mathbb{S}_{k+1} \\
 & w_k^i & \mapsto & w_{k+1}^i = w_k^i \; e^{\,\alpha_k \cdot \text{isMissClassified}(i,k)}
\end{array} \tag{.9}
$$

with α_k replaced by 1 in the GradientBoost algorithm. This down(up)-grading of (in)correctly classified events enforce each DT to focus on a different set of signal and background events and thus improves performance of BDTs compared to DTs and reduces the sensitivity to over-training.

The final output distribution for the AdaBoost $y^{\text{Ada}}(X_1^i)$ and for the GradientBoost $y^{\text{Grad}}(X_1^i)$ are given by a weighted average of DT:

$$
y^{\text{Ada}}(X_1^i) = \frac{1}{\sum_{k=1}^{N_{\text{tree}}} \alpha_k} \cdot \sum_{k=1}^{N_{\text{tree}}} \alpha_k \cdot p_i(X_1^i) \tag{.10}
$$

$$
y^{\text{Grad}}(X_1^i) = \frac{2}{1+\exp\left(-2 \cdot \sum_{k=1}^{N_{\text{tree}}} p_i(X_1^i)\right)} - 1 \tag{.11}
$$

B. The clustering

In MultiSVbb taggers the two b-hadron system kinematic and topology is extracted from the two highest mass vertices only. We have seen in section 3.4.2.1 that the other vertices could potentially provide a valuable information on the original b-hadron but require advanced clustering methods. In this section a novel method for clustering is proposed. The techniques developed here have been exported and largely enhanced in the $t\bar{t}H(H \to bb)$ analysis and will be presented in section 4.7.

The clustering of vertices is done in two step. First a *pairing BDT* is trained to identify pairs of objects originating from the same b-hadron. Then a clustering algorithm runs on all pairs to form clusters.

B.1. The pairing BDT

The pairing BDT aims at the separation of vertex pairs originating from the same b-hadron against pairs originating from different b-hadrons. In other words, the pairing BDT should differentiate a $g \to b$-hadron $\to c$-hadron decay chain from a $g \to bb$ topology. Two differences between these two processes are exploited:

- Kinematic: mainly uses the presence of a c-hadron and a b-hadron in the decay chain which reduces the total mass compared to the $g \to bb$ which has two b-hadrons.
- Topology: as mentioned for the JetFitter algorithm 3.3.1.2 the c-hadron decay vertex is rather aligned with the b-hadron flight axis. On the other hand, a pair of b-hadron or c-hadron vertices would give two separated vertices with similar distance to the PV and a relatively high separation in the plane transverse to the jet axis.

Figure .40 shows rejection of pair of vertices originating from different b-hadrons as a function of the efficiency to select pairs of vertices originating from the same b-hadrons. At a working point of 70% a rejection of 6 is observed.

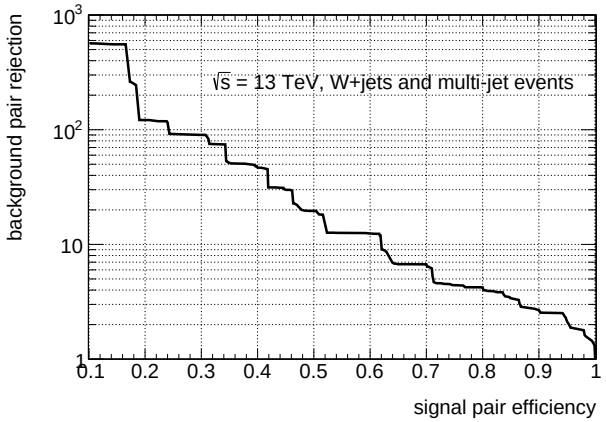


Figure .40.: Pairing BDT performance. The rejection of pair of vertices originating from different b-hadrons is shown as a function of the efficiency to select pairs of vertices originating from the same b-hadrons.

B.2. Clustering

The clustering algorithm developed for bb-tagging is rather simple. The best partner of a vertex vtx_i is define as the vertex which maximize the pairing BDT output: $\max_{\text{vtx}\in\{\text{vertices}\}\setminus\{\text{vtx}_i\}} [\text{BDT}(\text{vtx}_i, \text{vtx})]$. Since the BDT output is a symmetric function, it can be shown very easily that a vertex vtx_i of best partner vtx_j can not be the best partner of a vertex vtx_k. The set of vertices matched to their best partner is thus a set of pairs of vertices. The obtained pairs are used as a set clusters. The 4-momentum and position of clusters is taken as the 4-vector sum and the barycenter of the vertices they include respectively. The procedure is then repeated until only two clusters remains.

Figure .41 shows the resolution of the b-hadron position in the (x, y, z). Overall the quality of the b-hadron reconstruction are very similar. The inclusion of other vertices in the cluster only worsen the mean b-hadron position resolution by ~ 0.1 mm. However clusters provide a higher fraction of b-hadron tracks than the two highest mass vertices.

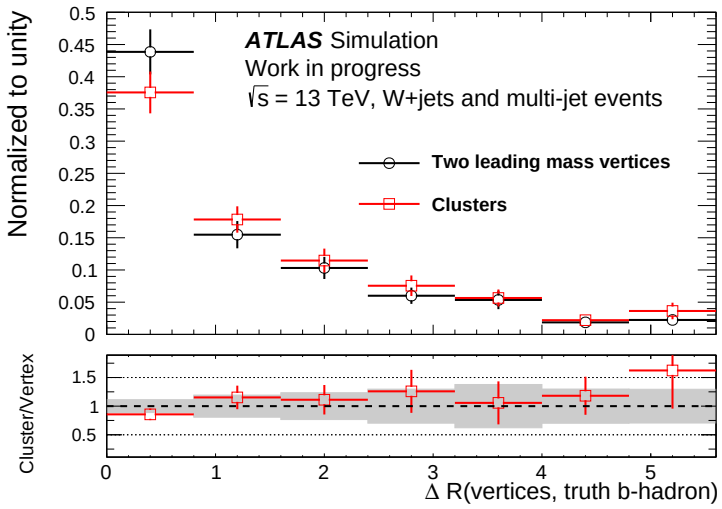


Figure .41.: ΔR distance of the reconstructed clusters and of the two highest mass vertices to their closest b-hadron in (x, y, z).

B.3. Inclusion in MultiSVbb

The pairing BDT and clustering algorithm are used in two ways. MultiSVbb variables can be redefined from the clusters rather than the two highest mass vertices and new variables separating bb-jets from the other jet flavours can be defined. Among many variables considered the minimum BDT output shown in figure .42 is expected to give the best separation between b-jets and bb-jets. Indeed b-jets only contains one b-hadron to c-hadron decay chain and thus only signal pairs.

Several configurations of the MultiSVbb algorithms are tested using the new variables and the re-defined MultiSVbb variables. The small fraction information brought by the additional vertices does not allow to improve the MultiSVbb variables. The clustering technique can however be beneficial to measurements of the $g \to b\bar{b}$ topology inside jets.

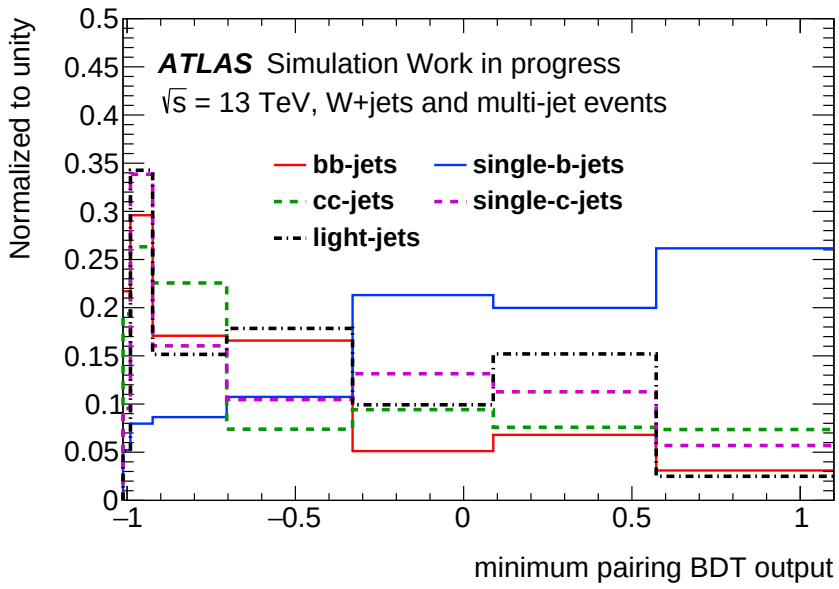


Figure .42.: Minimum BDT output distribution of the pairing algorithm.

C. $t\bar{t}$+$\geq$1b modelling systematic uncertainties

C.1. Systematic uncertainties in the default model

Figure .43 and .44 show the $t\bar{t}+\geq 1b$ systematic uncertainties in the default model (see section 5.2.2). These uncertainties are shown in the signal-enriched categories of the analysis where the $t\bar{t}+\geq 1b$ background dominates.

C.2. Systematic uncertainties in the PP8-based model

Figure .45, .46, and .47 show the $t\bar{t}+\geq 1b$ systematic uncertainties in the PP8-based model (see section 5.2.2). These uncertainties are shown in the signal-enriched categories of the analysis where the $t\bar{t}+\geq 1b$ background dominates.

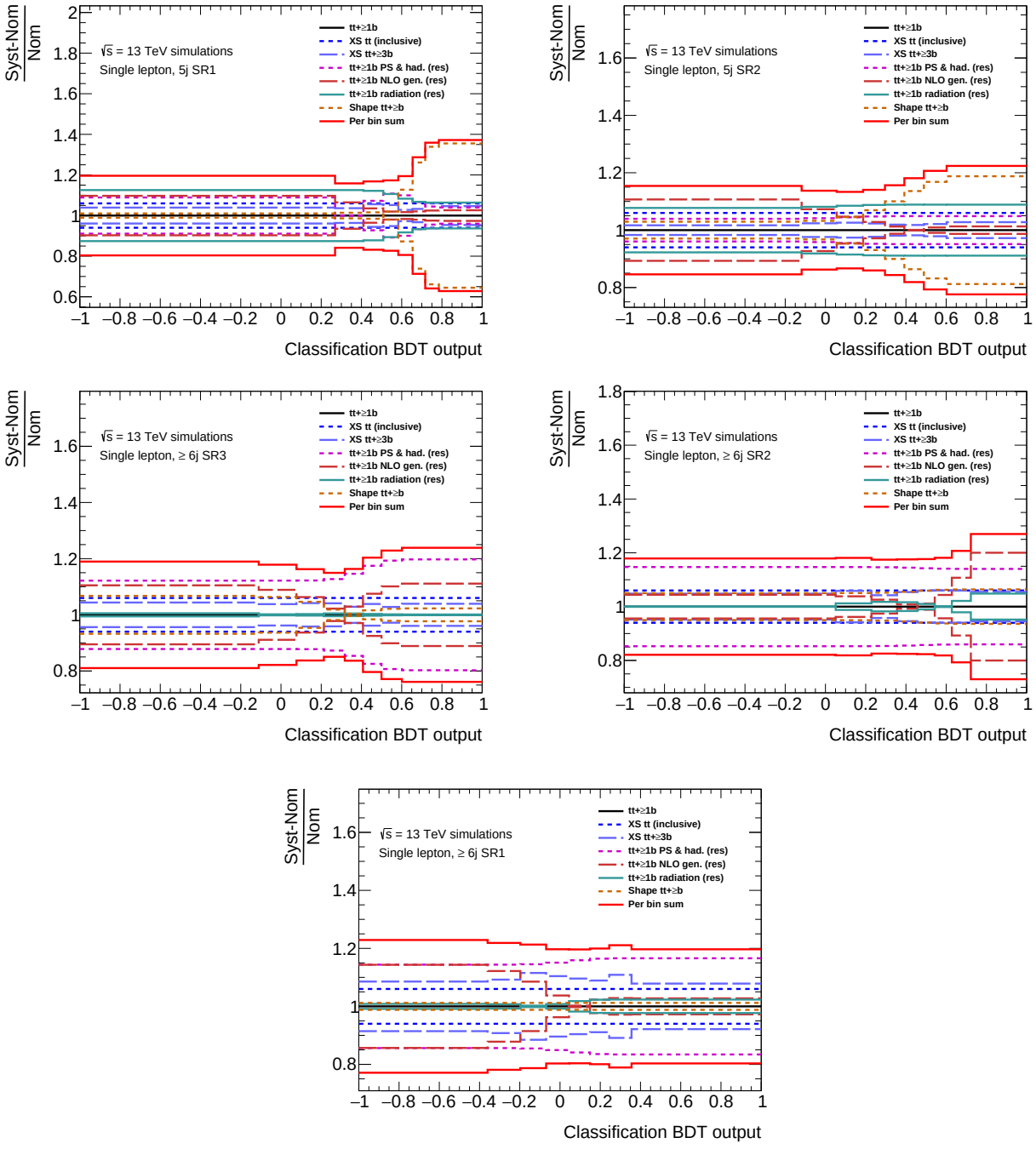


Figure .43.: Relative variations induced by the $t\bar{t}$+≥1b systematic uncertainties on the $t\bar{t}$+≥1b sample at ±1σ for the classification BDT distribution in the signal-enriched categories. The default $t\bar{t}$+≥1b model is used. The uncertainty on the $t\bar{t}$ cross section, the 50% normalisation uncertainty of the $t\bar{t}$+ ≥ 3b sub-component and various MC to MC comparison uncertainties are shown. The red lines show the quadratic sum of the up and down effects systematic uncertainties in each bin.

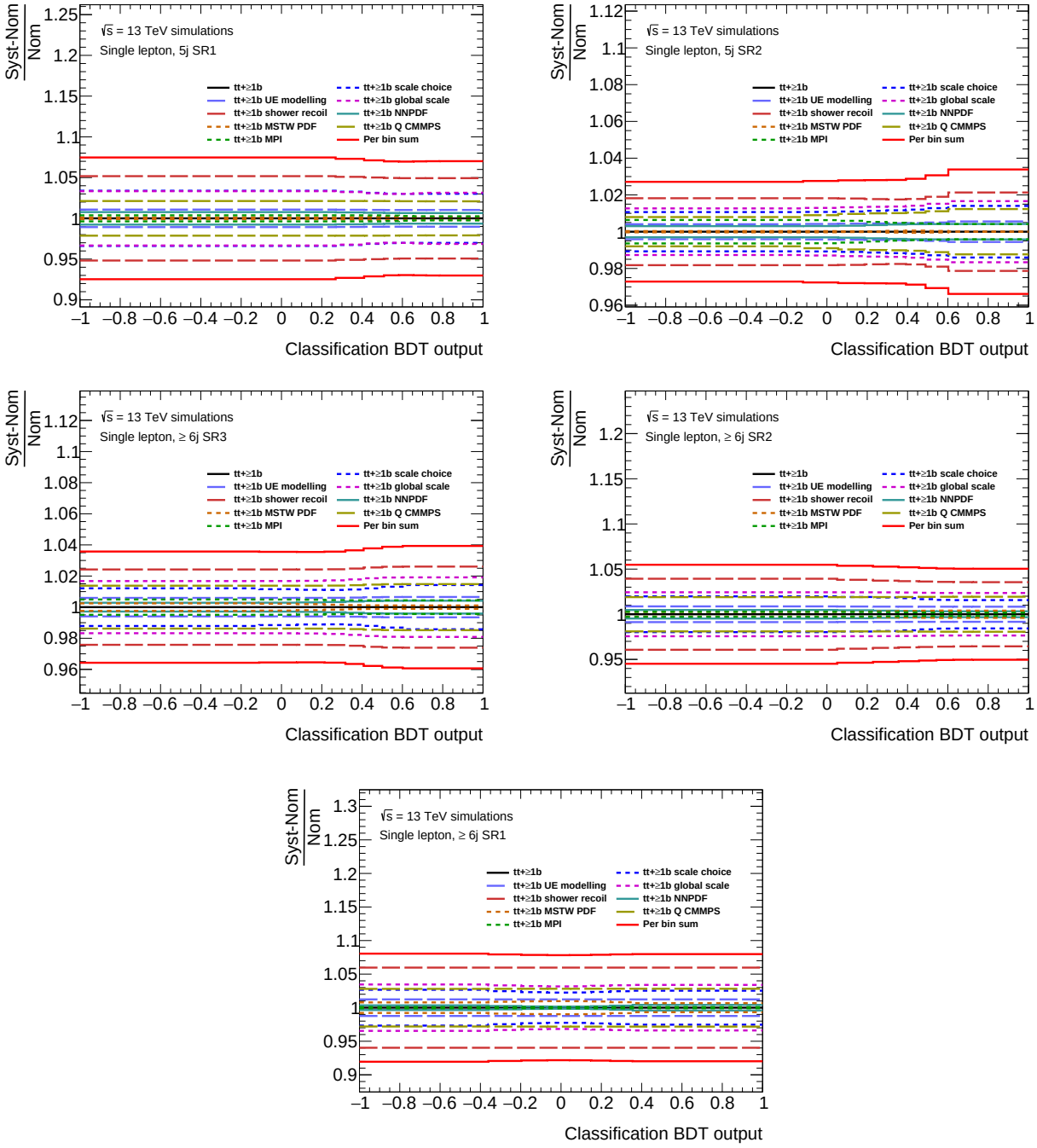


Figure .44.: Relative variations induced by the $t\bar{t}$+≥1b systematic uncertainties on the $t\bar{t}$+≥1b sample at ±1σ for the classification BDT distribution in the signal-enriched categories. The default $t\bar{t}$+≥1b model is used. Systematic uncertainties from SHERPA+OPENLOOPS variations are shown. The red lines show the quadratic sum of the up and down effects systematic uncertainties in each bin.

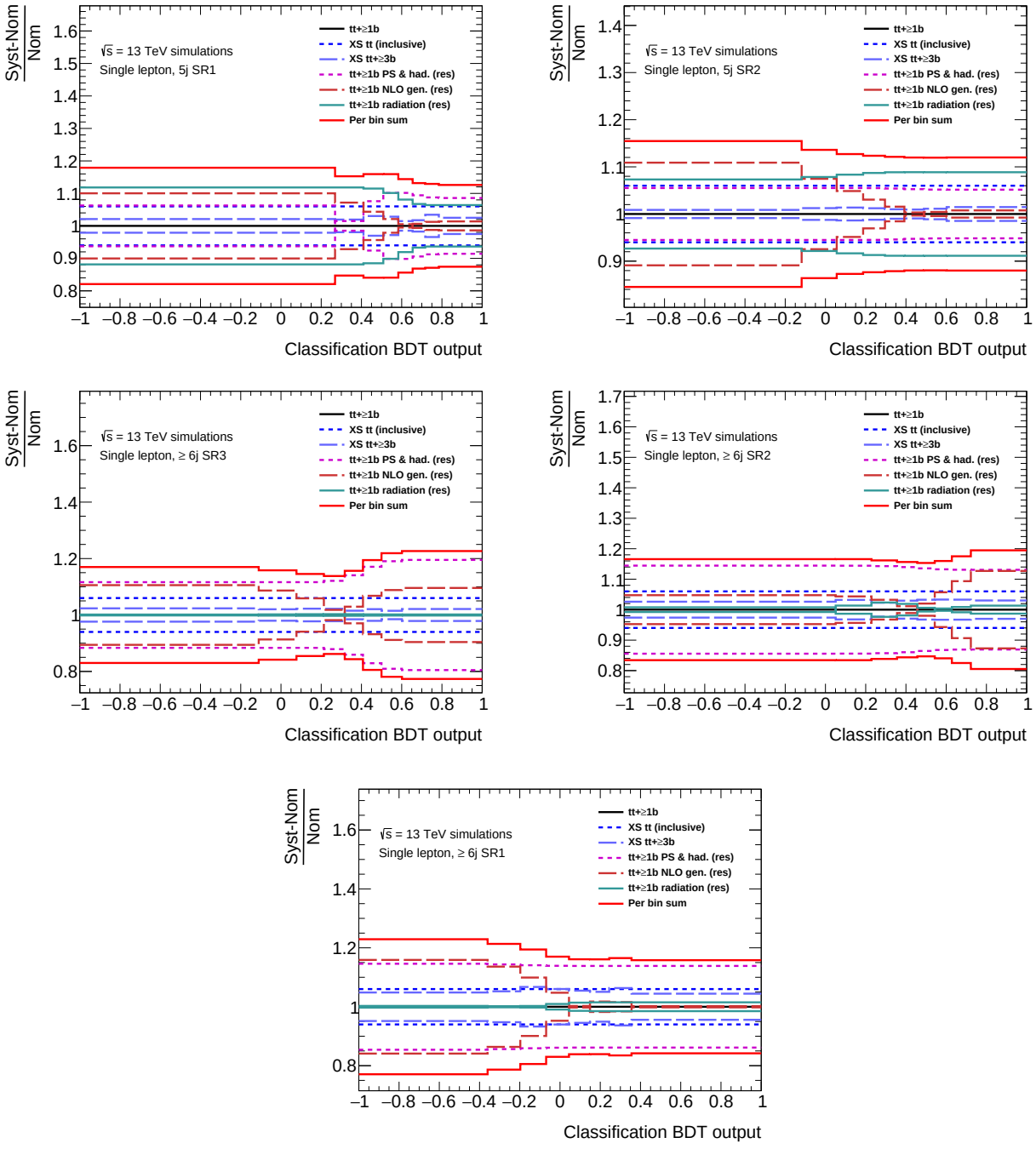


Figure .45.: Relative variations induced by the $t\bar{t}$+≥1b systematic uncertainties on the $t\bar{t}$+≥1b sample at ±1σ for the classification BDT distribution in the signal-enriched categories. The PP8-based $t\bar{t}$+≥1b model is used. Uncertainty on the $t\bar{t}$ cross section, the 50% normalisation uncertainty of the $t\bar{t}$+ ≥ 3b sub-component and various MC to MC comparison uncertainties. The red lines show the quadratic sum of the up and down effects systematic uncertainties in each bin.

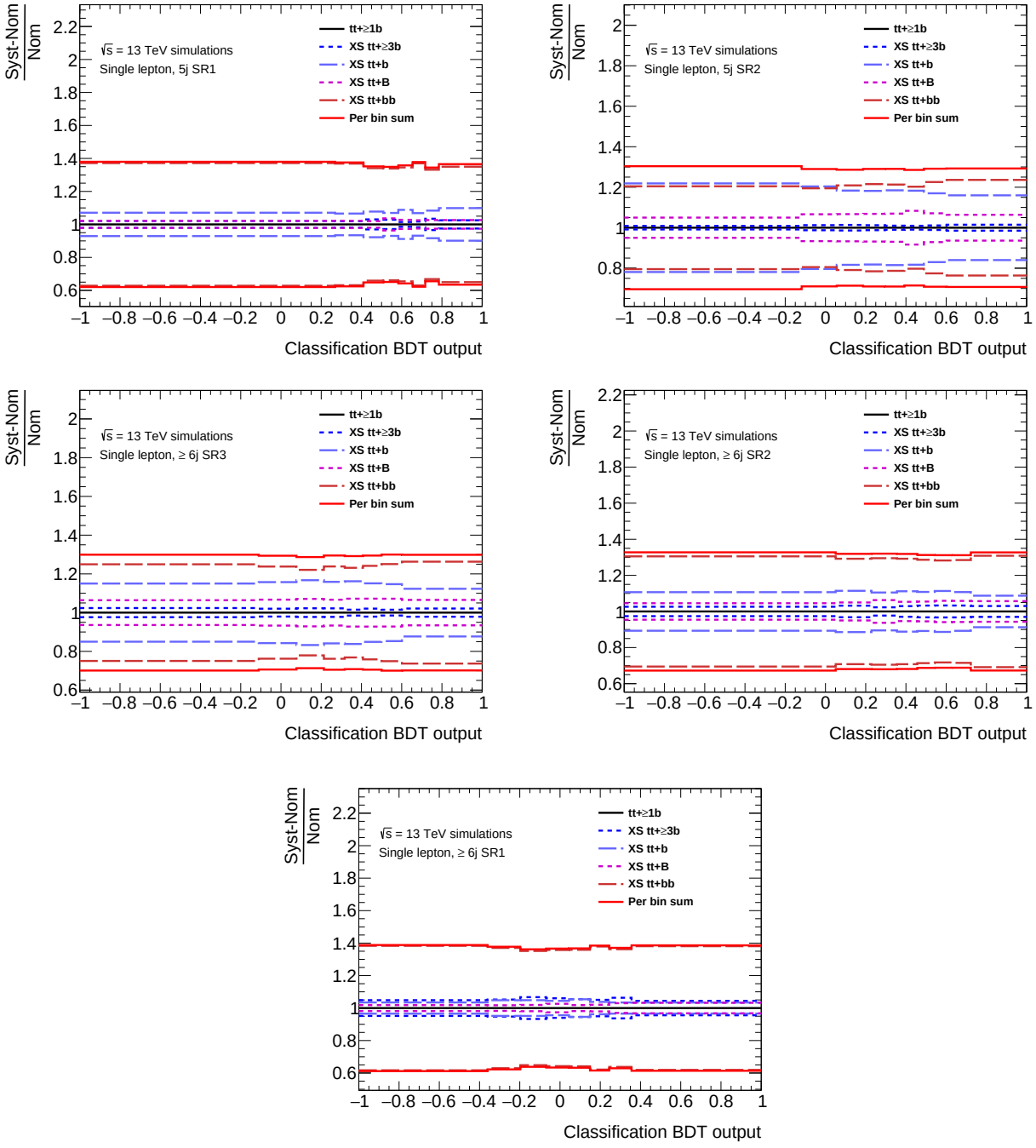


Figure .46.: Relative variations induced by the $t\bar{t}$+≥1b systematic uncertainties on the $t\bar{t}$+≥1b sample at ±1σ for the classification BDT distribution in the signal-enriched categories. The PP8-based $t\bar{t}$+≥1b model is used. The 50% prior uncertainties on the four $t\bar{t}$+≥1b sub-components are shown. The red lines show the quadratic sum of the up and down effects systematic uncertainties in each bin.

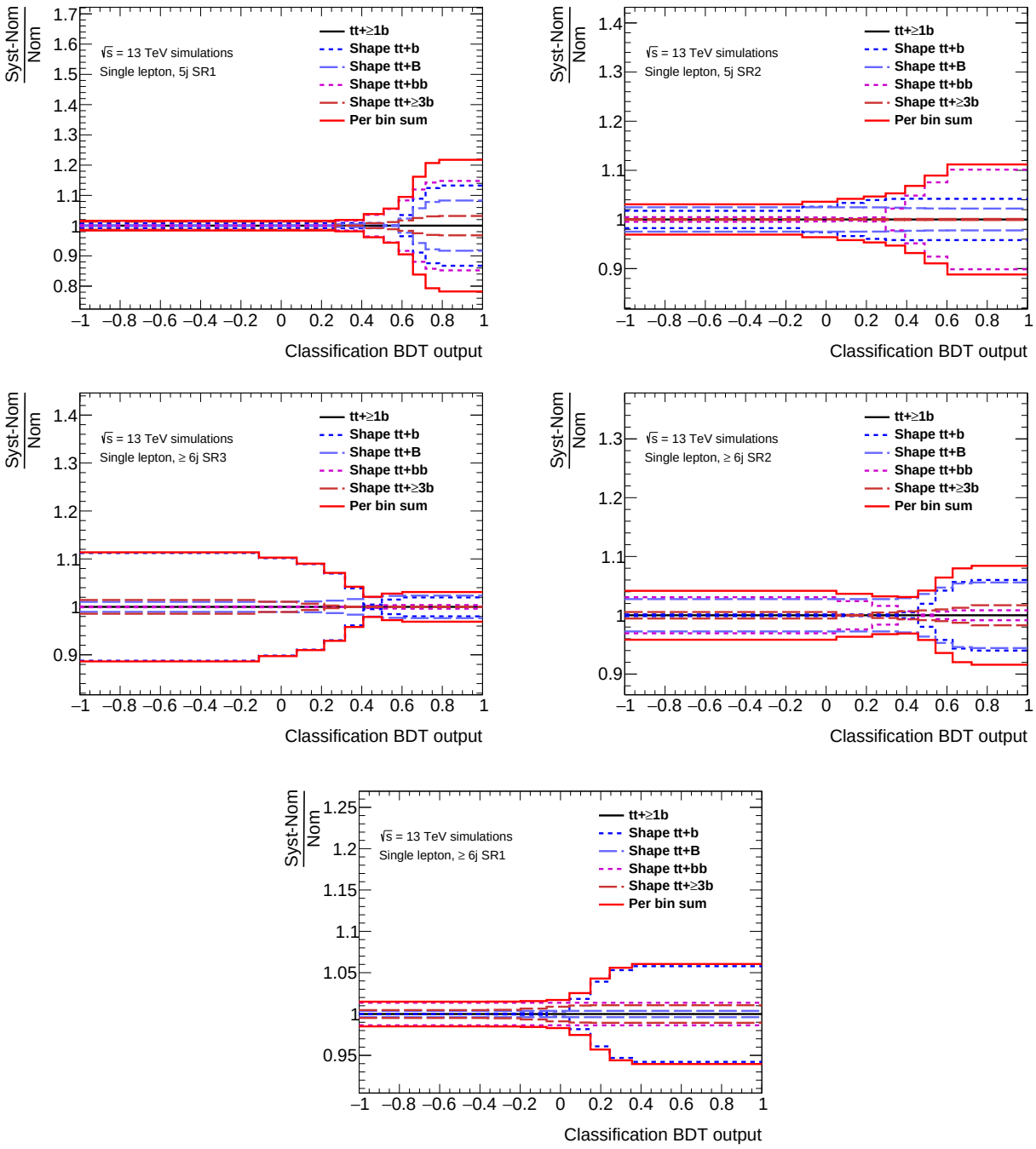


Figure .47.: Relative variations induced by the $t\bar{t}$+≥1b systematic uncertainties on the $t\bar{t}$+≥1b sample at ±1σ for the classification BDT distribution in the signal-enriched categories. The PP8-based $t\bar{t}$+≥1b model is used. The 4FS to 5FS shape comparison uncertainties are shown. The red lines show the quadratic sum of the up and down effects systematic uncertainties in each bin.